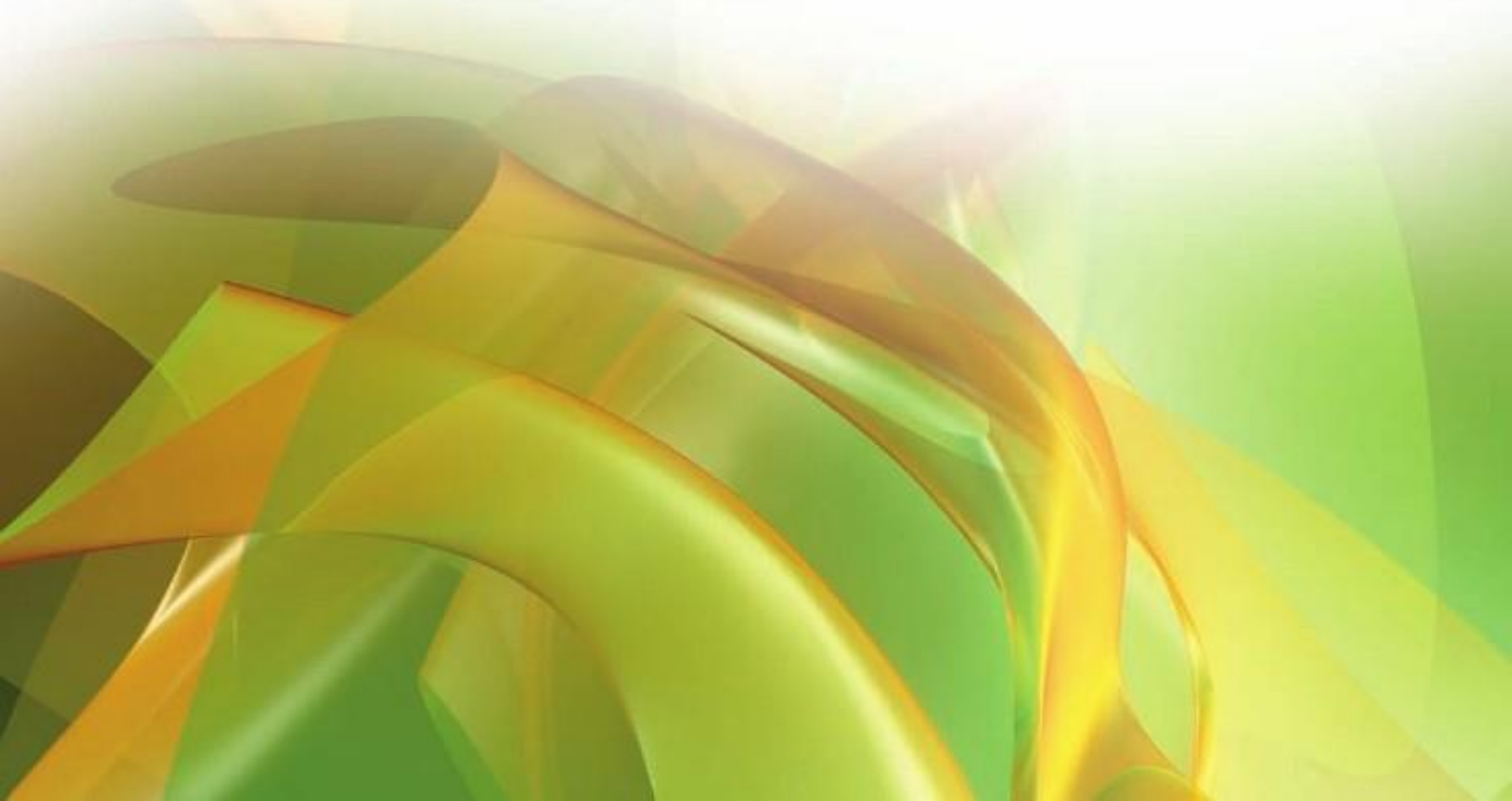
WILEY

Roland Bouman
Jos van Dongen

Pentaho® Soluções

Business Intelligence e Data Armazenamento com Pentaho e MySQL®



Pentaho Solutions

Pentaho® Solutions

Business Intelligence e Data Armazenamento com Pentaho e MySQL®

Roland Bouman
Jos van Dongen

WILEY
Wiley Publishing, Inc.

Pentaho→ Soluções: Business Intelligence e Data Warehousing com Pentaho e MySQL→

Publicado por
Wiley Publishing, Inc.
10475 Boulevard Crosspoint
Indianapolis, IN 46256

www.wiley.com

Publicado simultaneamente no Canadá

ISBN: 978-0-470-48432-6

Fabricados nos Estados Unidos da América

10 9 8 7 6 5 4 3 2 1

Biblioteca do Congresso Número de Controle: 2009930282

Com amor, de Annemarie, David, Roos e Anne.

-Roland

Para Yvonne, Thomas e Lisa, de muitas noites e fins de semana
Passei a trabalhar neste livro. Desculpe ter perdido quase seis meses de
suas vidas, mas prometo fazer isso!

-Jos

Sobre os autores

Roland Bouman tem vindo a trabalhar na indústria de TI desde 1998, principalmente como um desenvolvedor de aplicações web e banco de dados. Ao longo dos anos, ele se concentrou em
tecnologia de código aberto na Web, bases de dados e Business Intelligence. Ele é um membro ativo do MySQL e comunidades Pentaho, e foi premiado com o
MySQL Fórum Líder do Ano em 2006. Roland é orador regular
em conferências. Ele também é co-autor do MySQL 5.1 Cluster Certificação
Guia e revisor técnico de vários títulos relacionados ao MySQL. Você pode
siga seu blog em http://rpbouman.blogspot.com/.

Jos van Dongen é um experiente profissional de Business Intelligence e bem-conhecido autor e apresentador. Ele esteve envolvido no desenvolvimento de software, Business Intelligence e Data Warehousing, desde 1991. Antes de iniciar sua
própria prática de consultoria, Tholis Consulting, em 1998, ele trabalhou por um alto nível integrador de sistemas e uma empresa líder em consultoria de gestão. Ao longo dos últimos anos, tem implementado com sucesso vários armazéns de dados para
uma variedade de organizações, sem fins lucrativos e sem fins lucrativos. Jos abrange novas desenvolvimentos de BI para os holandeses Banco de Dados Revista e fala regularmente em conferências nacionais e internacionais. Além deste livro, ele foi o autor
um outro livro sobre open source de BI. Você pode encontrar mais informações sobre a Jos http://www.tholis.com.

Créditos

Editor Executivo
Robert Elliott

Projeto Editor
Sara Shlaer

Técnico Editores
Tom Barber
Jens Bleuel
Jeroen Kuiper
Thomas Morgner

Editor de Produção Sênior
Debra Bänninger

Copy Editor
Nancy Rapoport

Gerente Editorial
Mary Beth Wakefield

Gerente de Produção
Tim Tate

Vice-Presidente e Diretor Executivo
Grupo Publisher
Richard Swadley

Vice-Presidente e Diretor Executivo
Publisher
Barry Pruett

Editor Associado
Jim Minatel

Coordenador do Projeto, Capa
Lynsey Stanford

Revisor
Josh Chase, uma palavra
Scott Klemp, uma palavra

Indexador
J & J Indexação

Imagem da capa
Ryan Sneed

Cover Designer
Maciej Frolow / Brand X
Fotos / Jupiterimages

Agradecimentos

Este livro é o resultado do trabalho e as ideias de muitas pessoas diferentes. Nós, os autores, acontecerá a ser os únicos que conseguem colocar nossos nomes na capa, mas nós não poderia ter feito isso sem a ajuda dessas pessoas. Portanto, gostaríamos de aproveitar esta oportunidade para prestar nossos respeitos.

Uma coisa que caracteriza saudável projetos de código aberto é a paixão eo nível de envolvimento dos desenvolvedores e engenheiros de software que criam o projeto. Apesar de suas agendas lotadas, descobrimos os desenvolvedores da Pentaho Corporation sempre dispostos a fazer um esforço para explicar um determinado detalhes de seus softwares. Isto faz-lhes não só os desenvolvedores de software grande, mas também valiosos e respeitados membros da comunidade. Em particular, gostaríamos de agradecer Doug Moran, Moran Gretchen, Jens Bleuel, Julian Hyde, Matt Casters, e Morgner Thomas.

Um bom software nunca deixa de criar uma comunidade vibrante e intensa. Esta é ainda mais verdadeiro para o software de fonte aberta. Em uma quantidade relativamente pequena de tempo, o
comunidade Pentaho amadureceu consideravelmente, dando origem a um grupo de Pentaho especialistas que não só escrevem blogs de alta qualidade e ajudar uns aos outros no site oficial
Fóruns Pentaho ea (não oficial) # canal de IRC sobre Pentaho freenode.net, mas também participar ativamente e contribuir para o desenvolvimento do produto Pentaho. Gostaríamos de agradecer a todos que nos ajudaram na nos fóruns e no canal de IRC. Em particular, gostaríamos de agradecer a Daniel Einspanjer, Ward Harris, Goodman Nicholas, Raju Prashant Barbeiro Tom, e Yassine Elassad pelo seu papel na formação desta comunidade. Como é de se esperar com um projeto de código aberto como o Pentaho, alguns membros da comunidade de casal como os desenvolvedores de produtos. Um agradecimento especial para Ingo Klose, e mais do que especial
agradecimentos a Pedro Alves. Juntos, eles criaram o Painel da Comunidade Quadro, e Pedro foi muito útil para nós, explicando a sua arquitetura e design. Outras pessoas que merecem um agradecimento especial nota são Mark Hall, o

principal desenvolvedor do projeto Weka, Kasper Sørensen, o arquiteto da eobjects DataCleaner, e Ronald Damhof, por seus insights valiosos no cofre de Dados modelagem técnica.

Finalmente, gostaríamos de agradecer a Sara Shlaer e Bob Elliott, para gerir essa projeto, e observe o grande trabalho que tanto Sara e Nancy fez Rapoport com os documentos que entregamos. Percebemos que levou um esforço extra para transformar os escritos destes dois caps Holandês Inglês em texto legível. A colaboração com todos os envolvidos em Wiley sempre foi muito eficiente e agradável, talvez seremos capazes de trabalhar juntos novamente em outro projeto.

Roland-Bouman e Jos van Dongen

Primeiro, eu gostaria de agradecer a José, meu co-autor. Sem ele, escrever este livro teria sido tão demorado que eu duvido que ele teria sido praticamente viável. E, claro, eu gostaria de agradecer os leitores da http://rpbouman . Blogspot.com /. A idéia de escrever este livro ao longo do tempo e foi alimentados principalmente pelos comentários que recebi em resposta a uma série de posts que eu dedicado a Pentaho. Além disso, eu recebi muitos comentários encorajadores e e-mails de leitores do meu blog durante o processo de escrita, que não poderia ter sido melhor incentivo para continuar trabalhando para terminar esse livro.

Roland-Bouman

Minha viagem de BI open source começou há quase três anos atrás, quando eu publiquei um dos meus artigos de primeira no Holandês Banco de Dados Revista, intitulada"Pentaho, Prometendo Open Source BI Suite". Eu não poderia imaginar naquela época que isso levar a escrever um livro completo sobre o assunto, mas de alguma forma fiz! Então deixe-me agradecer aos meus co-autor de Roland por seus esforços incansáveis no sentido de obter este projecto fora do chão. Francamente, eu não tenho idéia de como ele conseguiu ficar todo esse trabalho feito com um emprego a tempo inteiro e três filhos pequenos em casa para cuidar. Um agradecimento especial também para Jeroen Kuiper, meu bom amigo e ex-colega, para formatar as secções de armazém de dados deste livro, e para sendo um crítico muito crítica do material.

-Jos van Dongen

Conteúdo at a Glance

Conteúdo

Introdução

Em 1958 a IBM Research Fellow Hans Peter Luhn escreveu um artigo seminal para o Jornal Sistemas IBM "Uma chamada de Business Intelligence System". Neste trabalho o termo inteligência foi definido como"a capacidade de apreender as inter-RELACIONAMENTO dos fatos apresentados de forma a orientar a acção para uma desejada objetivo."Hoje em dia este papel é geralmente considerada como a faísca que iniciou o desenvolvimento de Business Intelligence (BI), sistemas como os conhecemos hoje.

Por um longo tempo o mercado de software para apoiar BI tem sido o domínio de um punhado de fornecedores proprietários que poderia carregar grandes somas de dinheiro para suas soluções especializadas. O final dos anos noventa marcou uma importante viragem ponto para o mercado mundial de software de soluções open source, quando começou para se tornarem alternativas viáveis para o apoio aos sistemas de missão crítica. Primeira os vários sabores de Linux veio da idade, e em 2001 e 2002, vários novos projetos foram iniciados, todos nas áreas de integração de dados, relatórios, análise e mineração de dados, os pilares de uma solução de BI moderna. Em 2004, JasperSoft e Pentaho foi fundada para oferecer suítes completas de BI que alavancou o actual stand-alone soluções de BI. Desde então, essas empresas viram forte crescimento e adoção do mercado, tornando as soluções de BI de fonte aberta uma grave alternativa para os fornecedores estabelecidos proprietários.

Você pode se perguntar agora, o que é Business Intelligence e por que assunto? Na verdade, a definição dada por Luhn ainda é válido, mas em 1993, Analista do Gartner Howard Dresner reintroduziu o termo Business Intelligence e definiu-o como segue:

> Conceitos e métodos para melhorar a tomada de decisões comerciais usando baseadas em fatos
> sistemas de apoio.

Esta é uma definição um pouco menos abstrata em comparação com Luhn, e um que ainda é usado pela maioria das pessoas para explicar o que é BI. A definição é

não essencialmente focado em tecnologia ("conceitos e métodos"), mas é o último
systems" apoio parcial"fact baseado que é o tema deste livro. Este livro
É tudo sobre como criar um sistema de apoio baseado em factos usando as ferramentas entregues
pela Pentaho. Para ser capaz de tomar melhores decisões baseadas em fatos, você vai precisar
obter essa informação factual de um ou mais sistemas de informação, integração
esses dados em uma forma útil, e os usuários apresentam relatórios e análises que
ajudá-los a compreender o passado eo presente do desempenho organizacional.

O valor real dos sistemas de BI está na sua utilização para apoiar as organizações
para tomar decisões bem informadas que levará a uma maior rentabilidade,
custos reduzidos, eficiência, crescimento da quota de mercado, maior colaborador
satisfação, ou o que quer os objectivos da sua organização pode ser. O acrescentado
benefício do uso do open source Pentaho para fazer isso é o grande
valor para o dinheiro ea flexibilidade do software. Isso permite que qualquer organização,
com fins lucrativos ou sem fins lucrativos, grandes ou pequenas, para implementar e utilizar este software para
tomar melhores decisões.

Sobre este livro

Os primórdios da Pentaho Solutions voltar para mais de um ano atrás, quando nós,
os autores, observamos um interesse crescente em código aberto e software livre
soluções, combinado com uma consciência crescente de que ainda software de BI é essencial
em medir e melhorar o desempenho de uma organização.

Durante a década passada, as variantes de código aberto mais e mais tipos de
software tornaram-se alternativas comumente aceita e respeitada a sua
mais caro e menos flexível homólogos proprietários. O facto de software
é open source é muitas vezes confundido por ser livre de custos, e embora isso possa
ser verdade, se você olhar apenas o custo de licença, uma solução de BI não pode (e nunca
vai) ser livre de custos. Existem custos associados a hardware, implementação,
formação, manutenção e migração, e se tudo isso é resumido despeja
que os certificados representam apenas uma pequena parte do custo do ciclo de vida total de qualquer
solução de software. Open source, porém é muito mais do que uma maneira mais barata de
aquisição de software. O fato de que o código fonte está disponível livremente para qualquer um
garante melhor qualidade de código, uma vez que é mais provável que os problemas são encontrados quando mais
as pessoas têm acesso à fonte do que apenas os desenvolvedores do núcleo. O fato de que
software de fonte aberta é construído sobre padrões abertos usando a programação normal
línguas (principalmente Java) torna-o extremamente flexível e extensível. E o
fato de que software de código aberto não é mais vinculado a um determinado sistema operacional
estende essa flexibilidade e liberdade ainda mais.

O que normalmente é insuficiente, porém, é um bom conjunto de documentação e manuais.
A maioria dos projetos de código aberto oferecem software de excelente qualidade, mas os desenvolvedores
geralmente se preocupam mais com sair um grande programa de entrega adequada
documentação. E, embora você pode encontrar muitas boas fontes de informação
sobre cada pedaço de soluções de BI Pentaho, sentimos que existia a necessidade de

uma única fonte de informação para ajudar o usuário iniciante em sua forma descobrir o conjunto de ferramentas Pentaho e implementar a primeira solução. Isso é exatamente o que este livro é para a ajudar a construir sua primeiras soluções de BI Pentaho usando, desde o início (descobrindo Pentaho) até o fim (Painéis de construção para os usuários finais).

Quem deve ler este livro

Este livro é destinado para quem quer saber como oferecer soluções de BI usando Pentaho. Talvez você seja um gerente de TI procura um custo eficiente de BI solução, um profissional de TI que desejam ampliar seu conjunto de habilidades, ou de um BI ou dados consultor de armazém responsável pelo desenvolvimento de soluções de BI em sua organização. Talvez você seja um desenvolvedor de software com um monte de construção de experiência soluções de open source, mas ainda novo no mundo de Business Intelligence. E Talvez você já experimentou um BI ou dados desenvolvedor armazém com profundo conhecimento de uma ou mais das ferramentas existentes de propriedade. Em qualquer caso, Vamos assumir que tem um hands-on mentalidade já que este é um hands-on do livro. Nós esperamos alguma familiaridade com o uso de computadores para fornecer informações, instalação de software, e trabalhar com bases de dados, mas a maioria dos tópicos que serão explicado desde o início. Então, se você não é um especialista em SQL experiente, não se preocupe: nós vamos cobrir o básico dessa linguagem de consulta para que você obtenha em seu caminho.
É claro que os conceitos de BI e data warehouse são explicados como bem, mas o foco principal é sobre como transformar estes conceitos em uma solução de trabalho. Isso é exatamente por isso que o livro é chamado Soluções Pentaho.

O que você vai precisar usar este livro

Para poder utilizar este livro, você só precisa de duas coisas: um computador e uma Ligação à Internet. Todos os softwares e discutimos o uso deste livro é livremente disponíveis na Internet para download e uso. Os requisitos de sistema para o computador que você precisa é bastante moderado, de fato, qualquer computador que é inferior a quatro anos vai fazer o trabalho muito bem, contanto que você tenha pelo menos 1 Gigabyte de memória RAM instalada e 2 GB de espaço livre em disco disponível para baixar e instalar o software.

Os diversos capítulos contêm URLs onde você pode encontrar e baixar o software que está sendo utilizado e as instruções de instalação. Quanto à Pentaho, existem, para além do código fonte real do curso, três versões do software que você pode usar:

- Estes lançamentos GA- são estáveis versões do software, não é geralmente o Os mais recentes, mas certamente o mais confiável.
- Release Candidates-A "versões"quase pronto seguinte do software, possivelmente com alguns pequenos bugs ainda neles.

- Milestone libera-Estes são criadas com mais freqüência e permite que você para trabalhar com versões recentes introduzindo novas características.
- Nightly constrói-A versões mais atualizadas do software, mas também os menos estáveis.

Ao escrever este livro, que trabalhou principalmente com a nightly builds que geralmente precedem os lançamentos GA por três meses ou mais. Isso significa que que quando você ler este livro, o software usado neste livro é, no mínimo, um milestone ou já GA. Isto permite-lhe trabalhar com o material usando um produto estável, livre de bugs e você pode se concentrar em soluções de construção,
Não correção de bugs.

NOTA Como este livro vai para a imprensa, o próximo grande lançamento do Pentaho é esperado no outono de 2009. O número da versão final para esta versão ainda não é do conhecimento público, mas a versão de trabalho é actualmente designada como Citrus''.''

A lista completa com as opções de download está disponível online em http://wiki.pentaho.com/display/COM/Community+Edition+Downloads.

O que você vai aprender com este livro

Este livro vai ensinar-lhe:

- O Business Intelligence é, e porque você precisa dele
- Os componentes e produtos que formam a Pentaho Business Intelligence suite, e como estes produtos e componentes de BI cumprir necessidades específicas
- Como instalar e configurar o Pentaho e como conectá-lo a uma base de dados armazém
- Como projetar um data warehouse utilizando ferramentas de código aberto
- Como criar e carregar um armazém de dados com Pentaho Data Integration (Kettle)
- Como configurar uma camada de metadados para permitir relatórios ad-hoc e de auto-atendimento sem o uso de consultas diretas SQL
- Como criar relatórios utilizando as ferramentas Pentaho Reporting
- Como criar Pentaho Analysis Services (Mondrian), cubos, e visualizar los usando o navegador cubo JPivot
- Como configurar o agendamento, assinatura e distribuição automática de BI conteúdo
- Como começar com o Pentaho Data Mining (Weka)
- Como construir painéis usando a Comunidade Dashboard Framework para Pentaho

Como este livro está organizado

Este livro explica os conceitos de BI, tecnologias e soluções. Nós usamos uma ficção vendas on-line de vídeo e empresas de aluguel (acho Netflix), que aparece em toda o livro. Para cada parte distinta, as implementações de exemplo são criados usando Pentaho. Quando o exemplo se baseia em um banco de dados, nós temos tido o cuidado de assegurar
o código da amostra é compatível com o MySQL popular e ubíqua
banco de dados (versão 5.1).

Estas amostras de fornecer os detalhes técnicos necessários para entender como você pode construir soluções de BI para as situações do mundo real. O âmbito destas BI soluções é principalmente sobre o nível de data mart departamentais, que
acredito ser o caso de negócio mais comuns para o BI / data warehousing.

Parte I: Introdução ao Pentaho

Parte I está focado em obter um entendimento rápido e de alto nível do
Pentaho software, sua arquitetura e suas capacidades. Além disso, essa parte apresenta-lhe uma série de ferramentas de suporte de código aberto que pode ajudar na desenvolvimento armazéns do mundo real de dados e aplicações de BI.

Capítulo 1: Exemplos Pentaho: Quick Start

Business Intelligence é um assunto vasto e Pentaho é uma peça complexa de software. Ao mesmo tempo, é fácil de entender porque você precisa dele, e como poderia se aplicar a você se você seguir junto com alguns exemplos. Este capítulo faz exatamente isso: ele oferece um prático e eficiente na demonstração de que você pode fazer com BI Pentaho e como ajuda a fazê-lo.

Capítulo 2: Pré-requisitos

A fim de desenvolver aplicações de BI e arquitetura de apoio como um
data warehouse, você precisa de diversos produtos de software, tais como modelagem de dados
ferramentas e um servidor de banco de dados. Este capítulo apresenta uma série de produtos que são essenciais para os exemplos desenvolvidos neste livro, bem como um número de ferramentas de apoio para aumentar a produtividade. Como Pentaho, todos os produtos
aqui mencionadas são de código aberto / software livre. Estamos confiantes que você vai encontrar um
número de adições valiosas para o seu conjunto de ferramentas profissionais aqui.

Capítulo 3: Instalação e Configuração do Servidor

Embora este livro não fornece uma referência completa para administração de Pentaho e as tarefas de configuração, o mais importante e instalação são
descrito neste capítulo. Como tal, este capítulo não é tanto sobre a explicação conceitos, mas sim, é saber o que editar arquivos de configuração para ajustar as coisas

até ao seu gosto. Você deve pelo menos ler todo este capítulo uma vez antes construção de aplicações Pentaho. Muitas das informações aqui tem o caráter de uma referência. À medida que abrange os aspectos mais Pentaho todo o livro, você pode querer revisitar este capítulo para procurar determinados elementos da Pentaho configuração.

Capítulo 4: O Pentaho BI Stack

Este capítulo fornece uma visão geral do Pentaho, seus componentes, as suas capacidades, e sua arquitetura. Ele apresenta a você os conceitos importantes Pentaho, como seqüências de ação e solução de repositório. Embora exista uma quantidade razoável da teoria neste capítulo, também fornece explicações práticas, unindo uma grande quantidade de material coberto nos capítulos anteriores.

Parte II: modelagem dimensional e dados Armazém Design

Parte II apresenta-lhe os principais conceitos e técnicas relativas à dimensão modelagem e armazenamento de dados. Estes conceitos são feitos tangíveis usando um estudo de caso único exemplo com base em um aluguel de DVD (fictício) em linha negócios, de Classe Mundial Filmes. Através da teoria constantemente e conscientemente a mistura
e práticos, exemplos práticos, esperamos estabelecer uma fundação sólida para desenvolvimento de aplicações de BI no restante do livro.

Capítulo 5: Caso de Negócio Exemplo: Filmes de Classe Mundial

Neste capítulo, nós introduzimos o World Class aluguer de DVD de filmes online negócios. Nós fornecemos uma explicação detalhada dos seus negócios e subjacentes esquema de banco de dados OLTP.

Capítulo 6: Primer Data Warehouse

Este capítulo apresenta e explica os conceitos fundamentais da dimensão modelagem e armazenamento de dados. Ele explica os benefícios do uso de dados warehouse e como um data warehouse difere de outros tipos de bancos de dados. O capítulo aborda a história, estado atual e perspectivas futuras de dados tecnologia de armazém e arquitetura.

Capítulo 7: Modelagem de Negócios Usando esquemas Star

Este capítulo leva a modelagem dimensional e conceitos de data warehouse do capítulo anterior e aplica-los para a Classe Mundial Filmes business case para desenvolver as várias partes do modelo de data mart. Este modelo serve como base para as aplicações de BI (que são desenvolvidos na próxima parte do livro).

Capítulo 8: os dados do processo de Design Mart

Neste capítulo, o projeto lógico do capítulo anterior é mais desenvol-
desenvolvidos, culminando em uma série de esquemas em estrela, que servem para alcançar o desenvolvimento físico
execução dos Filmes Classe Mundial de data warehouse, que é a base
de praticamente todos os exemplos práticos no restante do livro.

Parte III: Integração de dados e ETL

A Parte III é dedicada ao processo de enchimento do depósito de dados usando o Pentaho ferramentas de integração de dados e funcionalidades.

Capítulo 9: Pentaho Data Integration Primer

Este capítulo fornece uma visão geral de todas as ferramentas que compõem o Pentaho Data Integration (PDI) toolkit. Ele explica a arquitetura e apresenta
a uma série de conceitos que são fundamentais para a concepção de ETL dentro do Pentaho plataforma. Ao mesmo tempo, fornece-lhe com as mãos básicas sobre as habilidades que irá ajudá-lo a usar ferramentas Pentaho Data Integration efetivamente para construir ETL aplicações.

Capítulo 10: Criando Soluções Pentaho Data Integration

Usando os conceitos e competências básicas adquiridas a partir do capítulo anterior, este capítulo se concentra na concepção e construção de um hands-on solução prática
para carregar as Ordens mart de dados do data warehouse de Classe Mundial Filmes.
As transformações exemplo, são acompanhados por uma descrição detalhada de comumente usado etapas de transformação.

Capítulo 11: Pentaho Data Integration Implantar soluções

Este capítulo se concentra na gestão e implantação de Pentaho Data Integration soluções. Além disso, explica como as transformações individuais podem ser combinados para criar empregos. Várias técnicas para a gestão dos recursos estáticos, como conexões de banco de dados e arquivos são discutidas, junto com alguns dos mais recursos avançados de PDI, como a execução remota e clustering.

Parte IV: Aplicações de Inteligência de Negócios

Parte IV explica como usar o data warehouse para criar conteúdo de BI no final usuários se preocupam.

Capítulo 12: A camada de metadados

Este capítulo apresenta Pentaho metadados e do editor de metadados. Além
para explicar conceitos de metadados e da finalidade dos metadados em soluções de BI,

Este capítulo fornece instruções detalhadas para criar um domínio de metadados que podem ser usados para criar relatórios de auto-atendimento.

Capítulo 13: Usando as ferramentas Pentaho Reporting

Este capítulo fornece um tutorial em profundidade sobre o projeto e implantação relatórios usando o Pentaho Report Designer. Você vai aprender como criar consultas usando o designer visual SQL ea ferramenta de consulta de metadados, adicione parâmetros para o relatório para análise interativa, e construir uma grande procura e relatórios perspicazes usando tabelas, gráficos e tabelas.

Capítulo 14: Assinatura, agendamento e de ruptura

Este capítulo é sobre toda a produção automática e entrega de conteúdo de BI. Você vai aprender a usar o Pentaho's built-in scheduler e como ela se liga em recursos tais como inscrição e agendamento.

Capítulo 15: Soluções OLAP Utilizando Pentaho Analysis Services

Este capítulo explica os componentes Pentaho OLAP. Além de explicar OLAP e MDX em geral, este capítulo ajuda você a criar análise Pentaho cubos, pastilhas e pontos de vista. A última parte deste capítulo apresenta o designer agregado
que ajuda a melhorar o desempenho da ferramenta Pentaho Analysis.

Capítulo 16: Mineração de Dados com Weka

Neste capítulo, vamos introduzir os conceitos básicos de mineração de dados e práticas como agrupamento e classificação utilizando Weka, o Pentaho componente de mineração de dados. Terminamos este capítulo com um exemplo de como você pode usar um modelo de mineração de dados
criado com Weka em uma transformação Pentaho Data Integration.

Capítulo 17: Construindo Painéis

Este capítulo explica os conceitos subjacentes ao Dashboard Comunidade Quadro. Utilizando um método passo-a passo, este capítulo explica em detalhes como combinar uma série de itens de solução Pentaho diferentes e levá-los juntos em um painel.

Sobre o Website

Todo o material utilizado no exemplo do livro está disponível para download a partir do Web site do companheiro em Wiley (www.wiley.com / go / pentahosolutions) E na www.worldclassmovies.com. O download inclui os seguintes itens:

- Power * Architect modelos de dados para bases de dados no livro
- Os arquivos de dados de clientes, produtos e funcionários

- MySQL criar scripts para os bancos de dados
- scripts MySQL para gerar transações de vendas
- Todos os trabalhos PDI e transformações
- modelos de metadados para criar relatórios
- Exemplos de relatórios
- esquemas Mondrian
- definição de arquivos Dashboard
- seqüência de exemplos de Acção

Recursos adicionais

Existem vários livros disponíveis sobre os temas específicos abordados neste livro. Muitos capítulos contêm referências para outras leituras e links para sites que contêm informações adicionais. Se você é novo e Business Intelligence armazenamento de dados em geral (ou quiser acompanhar os desenvolvimentos mais recentes),
Aqui estão alguns bons lugares para começar:

- inteligência http://en.wikipedia.org/wiki/Business
- http://www.kimballgroup.com
- http://b-eye-network.com
- http://www.tdwi.org

Nós também encorajamos você a visitar o nosso site, http://rpbouman.blogspot.com e www.tholis.com, Onde você pode encontrar as nossas informações de contato em caso de quero entrar em contato conosco diretamente.

Pentaho Solutions

Parte

I

Começando com Pentaho

Nesta parte

CAPÍTULO

1

Quick Start: Exemplos Pentaho

Pentaho é um poderoso Business Intelligence Suite oferece muitos recursos: relatórios, tabelas dinâmicas de OLAP, dashboards e muito mais. Neste livro você vai encontrar um monte de informações detalhadas sobre os componentes Pentaho, como eles funcionam e interagir, os recursos que oferecem, e como usar o Pentaho BI Suite para criar soluções para os problemas do mundo real. No entanto, é uma boa idéia para tentar compreender o quadro geral antes de mergulhar nos detalhes.

Este capítulo ajuda você a começar por mostrar-lhe onde conseguir o software e como instalar e executá-lo. O Pentaho BI Suite inclui muitos exemplos demonstrando suas características para dar aos usuários uma nova idéia de que tipo de soluções você pode construir com ela. A maioria destes trabalhos exemplos"fora do"caixa e são portanto, ideal para uma introdução ao produto. Ao ler este capítulo, você se familiarizar com a Pentaho, olhando para alguns exemplos.

Começando com Pentaho

Nesta seção, descrevemos como obter o software, instalá-lo e executá-lo. Para executar o software, você precisa de um desktop ou laptop regular execução qualquer sistema operacional popular, como o Ubuntu Linux, Mac OS X ou Microsoft Windows 7, XP ou Vista. Para baixar o software necessário, você vai precisar de um conexão à Internet com banda suficiente para fazer o download de dezenas a centenas de megabytes.

Baixar e instalar o software

O Pentaho BI Suite é um software de fonte aberta, você é livre para usar e distribuir seus programas, e se você quiser, você pode estudar e até mesmo modificar seu código fonte. Você pode fazer tudo isso gratuitamente.

Pentaho é programado em linguagem de programação Java. Antes que você possa executar programas Java, você precisa instalar o Java. Para Pentaho, você precisará de pelo menos
Java versão 1.5. Você também deve ser capaz de usar o Java 1.6. Vamos supor que você já tem uma versão recente do Java instalado em seu sistema. Você pode encontrar mais detalhes sobre como baixar e instalar o Java no Capítulo 2.

Você pode baixar todo o software lançado Pentaho da Fonte Forge site. A maneira mais fácil de encontrar o software para navegar http://sourceforge.net/projects/pentaho/ e clique no link Download. Você verá uma lista de produtos que você pode baixar.

Por enquanto, você não vai precisar de todos os programas-tudo o que você está interessado em nos
momento é o Business Intelligence Server. Clique no link Download no extremo coluna da direita. Isso leva você para uma página contendo uma lista das diferentes versões do software. Aqui você deve tomar cuidado para encontrar a versão mais recente do geralmente liberação (GA) disponíveis, embalados em uma maneira que seja apropriada para sua
plataforma. Por exemplo, usuários do Microsoft Windows deve baixar o Zip. pacote compactado e usuários de sistemas baseados em UNIX deve baixar o . Tar.gz compactado pacote.

NOTA Em páginas de download Pentaho no SourceForge, geralmente você pode encontrar em pelo menos as últimas geralmente disponíveis liberação (GA), bem como um marco chamados lançamento da nova versão, programado. Se você realmente quiser estar a sangrar margem do desenvolvimento, você pode baixar nightly builds do software de http://ci.pentaho.com/. Para este livro, que trabalhou principalmente com a noite compilações da versão Citrus, que ainda estava sendo desenvolvido no momento da escrita, mas que deverá estar disponível como um milestone GA ou pelo tempo de publicação.

É sempre uma boa idéia tentar o marco lançamentos para acompanhar futuras alterações e aditamentos. Mas cuidado que libera marco ainda estão em desenvolvimento, não são destinados à produção, utilização e você pode descobrir bugs ou questões de usabilidade experiência. No entanto, esta é uma das melhores razões pelas quais você deve executar lançamentos marco por reportar qualquer problema que você experimentar, você pode influenciam diretamente na melhoria do software para seu próprio benefício (bem como a de todos os outros usuários).

Depois de baixar o Zip. ou . Tar.gz pacote compactado, você deve extrair o software real do pacote compactado e copiá-lo para alguns lugar que você achar conveniente. Usuários do Windows podem direito do mouse no Zip. arquivo e
escolha Extrair Aqui (na nova pasta) no menu de contexto. Alternativamente, você pode usar um programa de terceiros, tais como PeaZip para extrair os programas do

o pacote compactado. Usuários de sistemas UNIX-like pode abrir um terminal e extrair o pacote na linha de comando.

Extração deve resultar em uma única pasta que contém todos os BI Pentaho Servidor de software. Usuários do Windows podem colocar essa pasta em qualquer lugar que quiser, mas faz mais sentido colocá-lo no diretório Program Files. Para UNIX-like sistemas, o local adequado depende do sabor de UNIX exata, mas para verificar os exemplos, é melhor mover o diretório do servidor Pentaho para seu diretório home. No restante deste capítulo, referimo-nos para o diretório contendo o software do Servidor Pentaho como o diretório home ou Pentaho simplesmente casa Pentaho.

Executando o Software

Agora que você tenha baixado e instalado o software, você pode começar usá-lo.

Iniciando o Servidor Pentaho BI

No diretório home Pentaho, você vai encontrar alguns scripts que podem ser usados para iniciar o servidor. Os usuários do Microsoft Windows pode dar um duplo clique no script chamado start-pentaho.bat.

Para sistemas baseados em UNIX, o script é chamado start-pentaho.sh. Você pode primeiro é necessário para permitir que este script para ser executado. Modern ambiente desktop Linux mentos como o GNOME e KDE vai deixar você fazer isso nas propriedades do arquivo diálogo, que você pode invocar a partir do navegador de arquivos. Por exemplo, no Ubuntu Linux, você pode botão direito do mouse no arquivo e escolha Propriedades no menu de contexto para invocar a caixa de diálogo. Na guia Permissões no diálogo, você pode selecionar um caixa de seleção para permitir que o arquivo a ser executado, conforme ilustrado na Figura 1-1.

Figura 1-1: Tornar o script start-pentaho.sh executável

Alternativamente, você pode abrir um terminal e alterar diretório (usando o cd comando) para o diretório home Pentaho. De lá, você pode usar o acompanhamento comando contribuem para fazer todos os Sh. scripts executáveis:

```
Shell> chmod ug + x *. sh
```

Agora você pode simplesmente iniciar o script clicando duas vezes (você pode precisar confirmar em uma caixa de diálogo) ou digitando-o no terminal:

```
Shell>. / Start-pentaho.sh
```

Depois de iniciar o script, você verá alguns bastante saída constante do console. Você deve deixar aberta a janela de terminal em que você começou o script.

NOTA A start-pentaho script faz duas coisas.

Primeiro, ele inicia um servidor de banco de dados HSQLDB, que é usado pelo servidor para Pentaho armazenar os dados do sistema, bem como um banco de dados da amostra, que é usado pela maioria dos exemplos.
Por padrão, o banco de dados HSQLDB rodando na porta 9001. Você deve ter certeza de que outro servidor está em execução no porto.

Em segundo lugar, ela começa um servidor Tomcat. Por padrão, o servidor Tomcat escuta na porta 8080 para solicitações da Web. Você deve fazer o servidor sem a certeza de outros está sendo executado no
porto, ou o Pentaho BI Server não será iniciado com êxito.

Registro em

Depois de iniciar o servidor, você pode iniciar o seu navegador de Internet para se conectar ao servidor. Você deve ser capaz de usar qualquer um dos principais navegadores (como Mozilla Firefox, Microsoft Internet Explorer, Safari, Opera ou Google Chrome)
para fazer isso. Navegue seu navegador para o seguinte endereço:

```
http://localhost:8080
```

Você será automaticamente redirecionado para o seguinte:

```
http://localhost:8080/pentaho/Login
```

Logo, você deve ver uma página de boas-vindas para o usuário Pentaho console. De lá, você pode fazer logon no servidor pressionando o grande botão laranja Login.
Se você pressionar o botão, uma caixa de login é exibida. De lá, você pode selecionar um nome da lista drop-down. Por agora, faça o login como o usuário Joe, como mostrado na Figura 1-2.
Depois de selecionar o nome de usuário, você pode pressionar o botão Login para realmente Entrar!

Figura 1-2: A tela de boas-vindas e Pentaho de diálogo de login

Manto, o usuário Pentaho Console

Após a confirmação do login, você verá que o usuário Pentaho console, como mostrado na Figura 1-3.

No console do usuário, você encontrará alguns elementos para controlar o BI Pentaho Servidor:

- Uma barra de menu, que está localizado na parte superior da página e se estende da página horizontalmente. Aqui você pode encontrar alguns itens de menu padrão: Arquivo, Exibir, Ferramentas e Ajuda.
- Uma barra de ferramentas que contém vários botões, localizados imediatamente abaixo no menu.
- Um painel lateral, localizado à esquerda da página, podem ser redimensionados dinamicamente usando a barra cinza vertical na extremidade direita do painel. O painel também pode ser oculto / apresentado em sua totalidade utilizando o botão Toggle Browser, que é o botão mais à direita da barra de ferramentas.

- A exibição em árvore que é visível na metade superior do painel lateral é chamado o Browser de Repositório. Na Figura 1-3, este é rotulado Procurar. Você pode usar esse recurso para navegar por todo o conteúdo de BI disponíveis no BI Pentaho Server.

- Uma pasta painel de conteúdo está localizado no painel lateral, logo abaixo do solução navegador repositório. Na Figura 1-3, esta é rotulado arquivos. Ela mostra qualquer conteúdo da pasta selecionada na solução de repositório (como relatórios, dashboards e tabelas dinâmicas de OLAP) como uma lista de itens. Você pode
abrir um item com um duplo clique sobre ele.
- Um espaço de trabalho. Este é o maior painel do lado direito. Quando você clicar duas vezes
um item no painel de conteúdo da pasta, ele será exibido aqui, usando um guia interface.

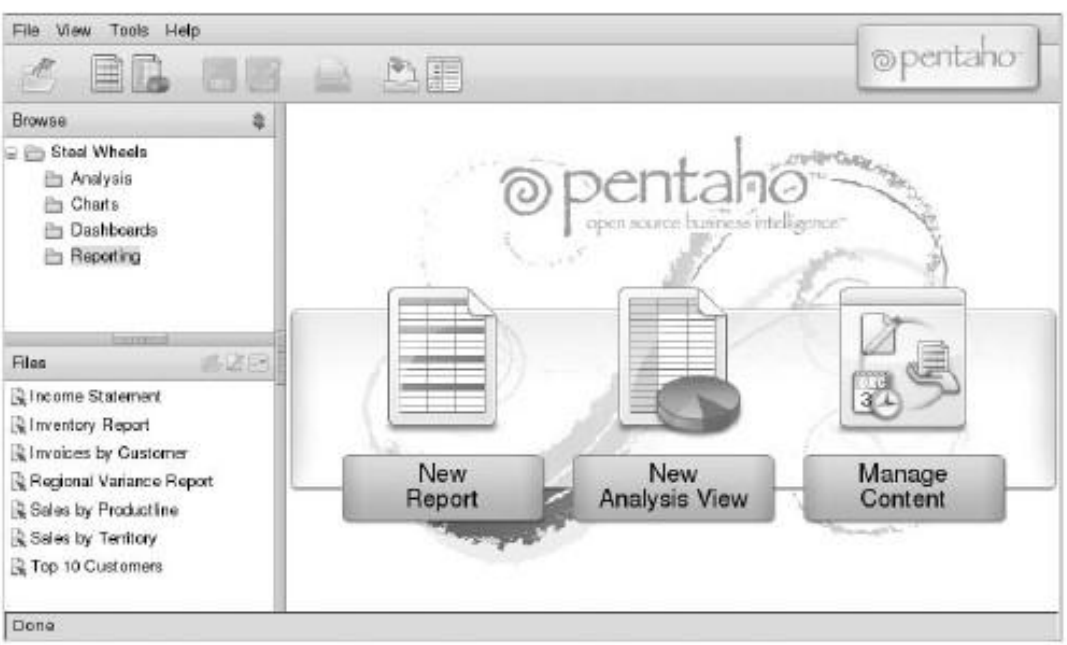


Figura 1-3: O usuário Pentaho console, também conhecido como Manto

Trabalhando com o contexto

A edição da comunidade do Pentaho BI Server vem com dois conjuntos de exemplos:

- Exemplos BI Developer
- Steel Wheels

Cada conjunto de exemplos reside na sua própria solução Pentaho e é visível na browser solução de repositório (veja a Figura 1-4).

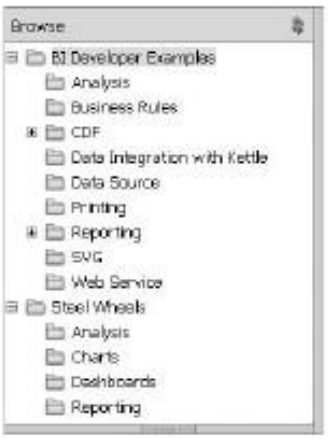


Figura 1-4: Duas soluções de exemplo incluídos no Pentaho BI Server

Ambas as soluções Pentaho contêm bons exemplos para demonstrar a tipos de relatórios que você pode criar com o Pentaho. Ambas as soluções de usar o mesmo dados de amostra. Os exemplos BI Developer concentrar mais nas técnicas aspecto de realizar uma tarefa específica, enquanto que os exemplos Steel Wheels ilustrar como combinar técnicas para construir uma aplicação para suportar um negócios de automóveis clássicos. Os exemplos Steel Wheels também dar mais atenção à personalizar olhar e sentir.

Usando o Browser de Repositório

Você pode acessar todos os exemplos usando o navegador de repositório. (Este é o topo painel da barra lateral esquerda, em que o usuário do console, chamado Browse). O repositório browser oferece uma exibição de árvore que pode ser usado para abrir e fechar as pastas o repositório. Para abrir uma pasta e revelar suas subpastas, basta clicar uma vez sobre o ícone, mais imediatamente no lado esquerdo do ícone da pasta. A pasta é subpastas se tornará visível à direita abaixo da pasta-mãe, eo ícone da esquerda do ícone da pasta para exibir um sinal de menos, indicando que a pasta está expandida. Para fechar uma pasta e esconder as subpastas, clique no ícone de subtração.

Para visualizar o conteúdo de uma pasta, clique no ícone da pasta ou o nome da pasta que aparece diretamente à direita do ícone de pasta. O título da pasta irá mostrar uma cinza destacando e seu conteúdo ficará visível no conteúdo da pasta painel diretamente abaixo do navegador repositório (na figura 1-3, este é rotulado Files).

Para abrir um item que aparece no painel Arquivos, clique duas vezes nele. Isto irá abrir uma nova página no espaço de trabalho, mostrando o resultado gerado pelo item.

Compreender o contexto

Embora você possa aprender muito com os exemplos, basta executá-los, você pode aprender ainda mais se você pode ver como elas foram construídas. Especialmente se você é um

Business Intelligence desenvolvedor, você deve considerar a análise de exemplos mais estreitamente com Pentaho Design Studio.

Você vai aprender os detalhes sobre Pentaho Design Studio no capítulo 4, mas você pode seguir estes passos para começar rapidamente:

1. Download Pentaho Design Studio a partir da página de downloads em Pentaho SourceForge.net.
2. Descompacte o download para algum local que você achar conveniente.
3. Iniciar Pentaho Design Studio. Os usuários do Microsoft Windows pode dar um duplo clique PentahoDesignStudio.exe, Usuários de sistemas baseados em UNIX pode executar o PentahoDesignStudio arquivo binário.
4. Utilize o menu principal (Arquivo Switch Workspace) para alterar o espaço de trabalho para o diretório onde você instalou o Pentaho BI Server. O programa será reiniciado. Na tela de abertura, escolha Workbench.
5. Crie um novo projeto escolhendo novo arquivo de projeto. Na caixa de diálogo, expandir a pasta Geral e escolha do projeto para criar um projeto simples. Clique em Avançar.
6. Na caixa de diálogo seguinte, introduza pentaho soluções para o nome do projeto. Faça certeza de que quer que escreva aqui corresponde exatamente ao nome do pentaho soluções diretório localizado no diretório home do Pentaho BI Server. A caixa de seleção Use Default Local deve ser selecionada, e o local deve apontar automaticamente para a casa Pentaho BI Server diretório.
7. Confirme o diálogo.

Na página da guia Navegador no painel do lado esquerdo em Pentaho Design Studio, você deve ver agora o pentaho soluções pasta do projeto (o que corresponde exatamente com o real pentaho soluções pasta). Você pode expandir esta pasta e navegar através da solução Pentaho repositório.

Clicando duas vezes sobre os itens dentro das pastas normalmente carregar o arquivo em uma nova página no Pentaho Design Studio Workspace. Você pode aprender muito, especialmente desde a abertura do . Xaction arquivos que estão presentes em todo o repositório. Consulte o Capítulo 4 para obter mais detalhes sobre esses arquivos.

Tenha em atenção que os itens que aparecem no browser do repositório do usuário console do Pentaho BI Server, normalmente têm uma etiqueta que é distinto do nome do arquivo real. Isso complica um pouco as coisas no caso de você está procurando o item correspondente na Pentaho Design Studio, como o navegador não só exibe nomes de arquivos. Para descobrir o nome do arquivo correspondente para cada item mostradas no navegador de repositório, clique no item e escolha Propriedades no o menu de contexto. Isto irá mostrar uma janela com algumas abas. O arquivo real nome é indicado na guia Geral.

NOTA A . Xaction extensão indica uma seqüência de ação. Seqüências de ação Pentaho são específicos processos leves de executar ou fornecer conteúdo de BI. Neste caso particular, a seqüência de ação simplesmente chama um relatório Pentaho. Ação seqüências são codificadas em um formato XML específico e normalmente armazenados in.xaction arquivos. seqüências de ação são discutidos em mais detalhes no Capítulo 4.

Executar os exemplos

No restante deste capítulo, discutimos alguns itens a partir desses exemplos para lhe dar uma sensação de que você pode fazer com soluções Pentaho. Para cada item, que incluem referências aos capítulos deste livro que se relacionam com o exemplo. Esperamos que esta irá permitir-lhe obter rapidamente uma visão geral dos recursos Pentaho e ver como este livro pode ajudá-lo a dominá-los.

Exemplos de relatórios

Reportagem é muitas vezes um dos primeiros requisitos de qualquer solução de BI. Reportagem é abordado em detalhes no Capítulo 13. A maioria dos relatos aqui discutidos são invocados a partir de uma seqüência de ação, você pode encontrar mais detalhes sobre seqüências de ação em Capítulo 4.

Os capítulos seguintes analisam alguns dos exemplos de relatórios.

Exemplos BI Developer: Vendas Regional - HTML

A Regional de Vendas - HTML é um exemplo a mais simples exemplos de relatórios, como você diria, ele mostra os números de vendas de um empresa exemplo discriminados por região. Você pode encontrá-lo no Reporting pasta no contexto BI Developer set. O nome do arquivo correspondente é JFree_Quad.xaction.

Quando você executar o exemplo, a saída do relatório é imediatamente mostrado no espaço de trabalho (ver Figura 1-5).

Na saída do relatório que você vê uma organização detalhada por região (Central), departamento (Executivo de Administração, Finanças) e título de posição (SVP Parcerias, CEO, e assim por diante). Para o nível de título da posição, você vê o real de dados. Neste caso, os dados se refere às vendas e mostra as reais e previstos (Orçados) os números de vendas nas duas primeiras colunas ea variância na terceira coluna. Você também pode ver uma linha de totais que resume os dados para o departamento nível, e se você pudesse rolar mais longe você ver, também, os totais para o nível regional, seguido pelos números de outra região. Todo o caminho para baixo em na parte inferior do relatório que ver os totais para o negócio inteiro.

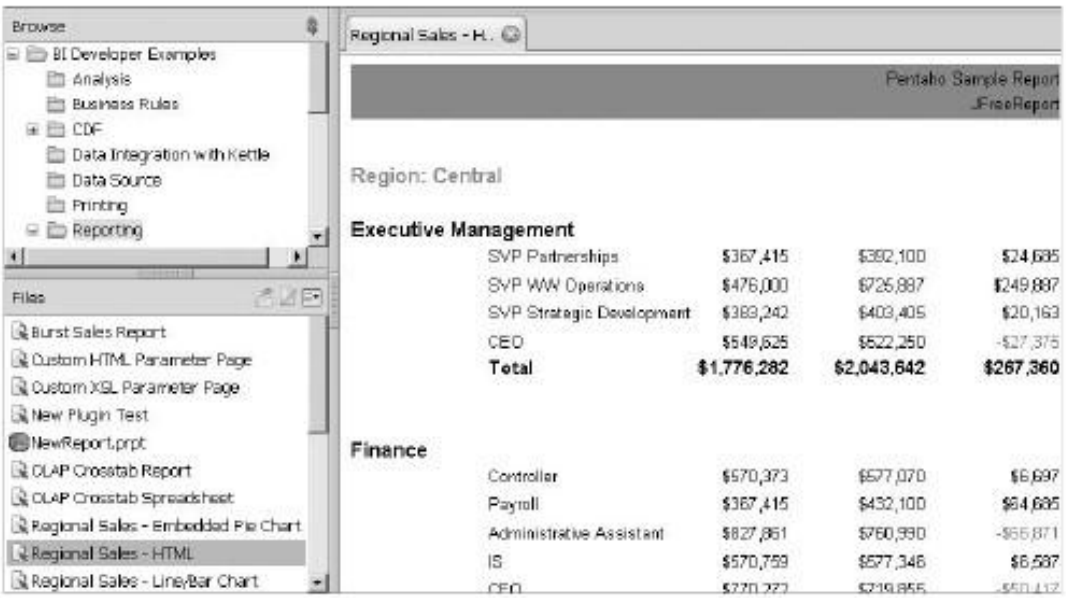


Figura 1-5: A Regional de Vendas - HTML relatório de exemplo

Rodas de Aço: Demonstração de Resultados

O relatório de exemplo Declaração de Renda a partir do conjunto exemplo Steel Wheels outro relatório típico com um nome auto-explicativo. Você pode encontrá-lo no Reportagem pasta abaixo a solução rodas de aço, eo arquivo correspondente nome é Renda Statement.xaction. Figura 1-6 mostra o relatório.

Figura 1-6: As rodas de aço de Renda Declaração relatório

Algumas diferenças a partir do relatório de vendas regional no exemplo anterior são o estilo eo formato de saída. Embora ambos os relatórios foram criados com o Pentaho Report Designer, e ambos são prestados pela Pentaho relatórios do motor (que é o componente responsável por interpretar os relatórios e saída de geração de relatórios), eles parecem bem diferentes. Considerando que o Regional Relatório de vendas gera uma página HTML, este relatório oferece um arquivo PDF como saída.
Além disso, este relatório mostra adereços usando uma imagem de um logotipo e um imagem de fundo da página.

Rodas de Aço: Top 10 clientes

Na seção anterior, mencionamos que o relatório de Declaração de Renda
proporciona uma saída na forma de um arquivo PDF, enquanto o exemplo Regional de Vendas
saídas de uma página web simples.
características importantes do formato de saída do relatório. Você pode encontrar este relatório também
na pasta de informação no conjunto de exemplo Steel Wheels, e seu nome do arquivo é Início Dez Analysis.xaction ProductLine Cliente. Executando esse exemplo não
não mostrar imediatamente a saída do relatório, mas exibe o diálogo mostrado na Figura 1-7 vez.

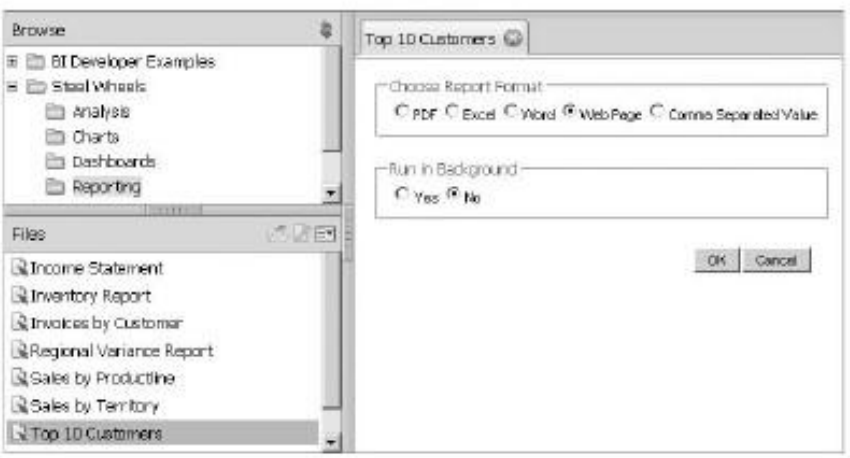


Figura 1-7: O Top 10 clientes relatório

Como indicado pelo diálogo, você pode escolher até cinco diferentes
formatos de saída. Nos exemplos de relatórios anteriores, o formato de saída desejado
foi armazenado como parte do relatório, mas não há nada no relatório que o motor
forças presente. Isso permite aos usuários escolher o formato é mais adequado
para o efeito na mão.

O diálogo mostrado na Figura 1-7 ilustra uma outra característica importante da
Pentaho relatório. O usuário pode optar por aguardar a saída do relatório agora, ou
ter o Pentaho BI Server executar o relatório em background. O último
opção irá executar o relatório, mas não espera para a saída a ser retornado.
Pelo contrário, a saída será armazenado no espaço de armazenamento do usuário pessoal sobre o
servidor. Esse recurso é especialmente útil para os relatórios de execução longa.

Você pode encontrar mais informações sobre a execução de fundo e funções relacionadas, como
programação e inscrição no Capítulo 14.

Exemplos BI Developer: botão único parameter.prpt

Os relatórios de exemplo anterior foram todos chamados a partir de seqüências de ação. Em
a versão Citrus programados, relatórios podem também ser chamado diretamente. Exemplos

usar este recurso estão todos localizados na pasta Relatórios do BI Developer Exemplos set.

Este exemplo tem um olhar mais atento à botão único-parameter.prpt exemplo. Quando você iniciá-lo, carrega o relatório imediatamente no espaço de trabalho. No entanto, a saída real do relatório não será exibido até que você pressione uma da Região botões que aparecem na seção de parâmetros de relatório na parte superior da página. Figura 1-8 ilustra o que você pode ver depois que você pressiona o botão Central.

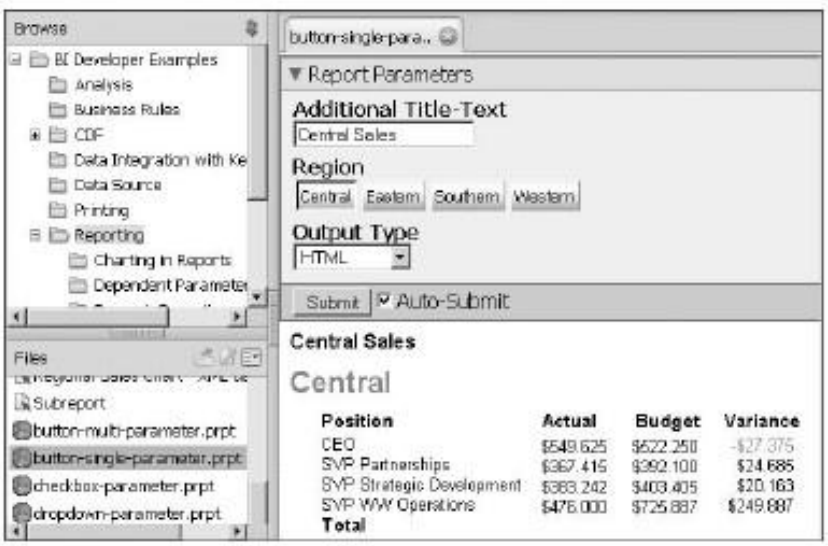


Figura 1-8: O exemplo de botão único parameter.prpt

Este exemplo mostra ainda uma outra característica da Pentaho, nomeadamente o relatório-param tros. Através de parâmetros, o usuário pode interagir com o relatório e especificar valores para influenciar o comportamento do relatório. Geralmente, esse recurso é usado para permitir que o usuário selecionar apenas uma parte de todos os dados do relatório possível.

Neste exemplo, existem dois parâmetros. A param Título Texto Adicional-eter permite ao usuário especificar um título que aparece acima de todas as restantes relatório saída. Há um outro parâmetro para a Região, o que permite que o relatório processar a saída referentes a apenas a região especificada.

Há muitas mais coisas que você pode fazer com os parâmetros do relatório, e estes exemplos, bem como o capítulo 13 deste livro, deve oferecer orientação suficiente para que você possa usar este recurso de forma significativa.

Traçando Exemplos

Considerando que os relatórios são ótimos para comunicar informações detalhadas, estão menos apropriado para obter uma visão geral dos dados como um todo. Para este efeito,

tabelas e gráficos geralmente funcionam melhor. Gráficos também são mais adequadas do que
relatórios para mostrar tendências ao longo do tempo.

O Pentaho BI Server vem com dois tipos diferentes de soluções de gráficos:

- JFreeChart-A 100% biblioteca de gráficos em Java.
- Pentaho Flash gráficos, uma solução de gráficos com base em cartas abertas flash (Que requer Adobe Flash).

relatórios Pentaho oferece integração completa com JFreeChart, e você vai encontrar informações detalhadas sobre a integração das paradas com seus relatórios no capítulo 13. Você pode encontrar mais informações sobre gráficos JFreeChart e como integrar
los com painéis no capítulo 17.

Rodas de Aço: Lista de Escolha Gráfico

O Gráfico exemplo Lista de Escolha está localizado na pasta Gráficos em as rodas de aço exemplo dado. O nome do arquivo correspondente é ChartComponent_ChartTypes . Xaction. Executando as cargas item uma caixa de diálogo na área de trabalho que permite que você
escolher um tipo de gráfico específico. Depois de escolher o tipo de gráfico, você pode pressionar
o botão Executar para realmente mostrar a carta. Figura 1-9 mostra como isso funciona para uma grade de pizza.

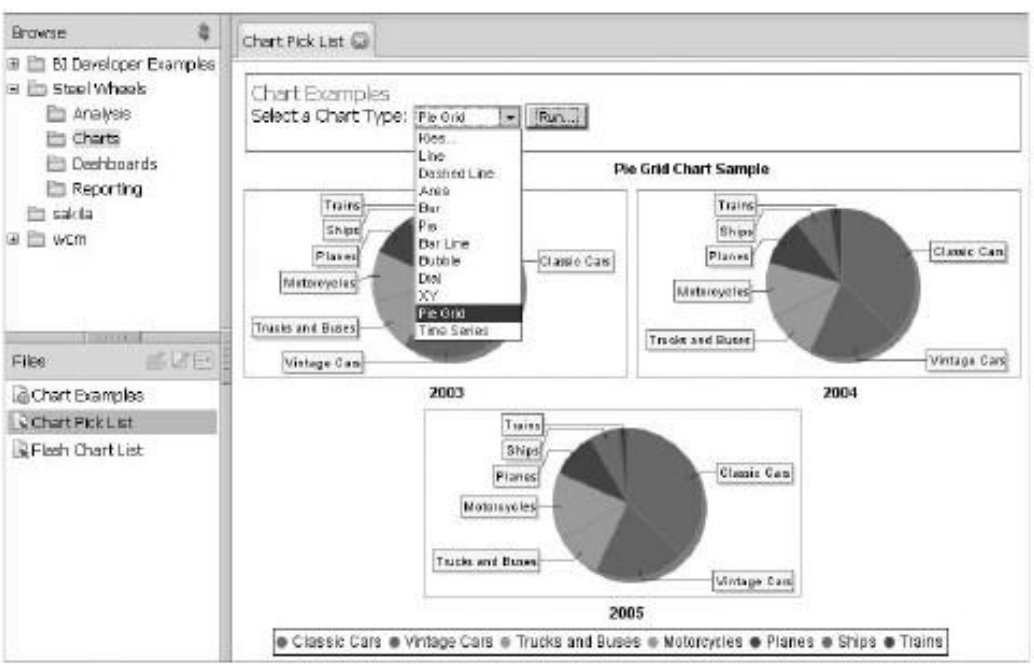


Figura 1-9: Pentaho gráficos usando o JFreeChart Gráfico Lista de Escolha

Rodas de Aço: Lista Gráfico Flash

Funcionalmente, o Flash exemplo Lista de gráfico é semelhante à Lista de Escolha Gráfico exemplo (que é baseado no JFreeChart). A diferença é que o JFreeChart

Pick List exemplo é baseado no projeto Open Flash Chart. Você pode encontrar o Flash Lista gráfico também na pasta de gráficos no exemplo Rodas de Aço definido. O nome do arquivo correspondente é pentahoxml_picker.xaction.

Exemplos BI Developer: Vendas Regional - Linha de gráfico de barras /

A Regional de Vendas - Linha / Bar exemplo Gráfico está localizado no Reporting pasta na solução de BI Developer exemplos. O arquivo correspondente é JFree_ SQLQuery_ComboChart.xaction.

Este relatório apresenta um exemplo gráfico na parte superior da página, e abaixo que, um relatório mais detalhado mostra os números reais. Neste caso, o gráfico é incorporado ao relatório. O relatório de exemplo é mostrado se a Figura 1-10.

Figura 1-10: Vendas Regional - Linha / Bar exemplo Gráfico

Exemplos de Análises

Como informação, a análise é outra característica essencial de todas as soluções de BI. Relatórios
são tipicamente estáticos (excepto para os parâmetros) e usado principalmente para apoiar as decisões
que afetam os negócios a nível operacional. Análise tende a ser muito mais
dinâmica, e é normalmente utilizada pelos gestores para a tomada de decisões na tática
e nível estratégico.

Um dos elementos típicos em soluções analíticas é que eles permitem a
usuário para explorar a dinâmica de dados em um modo ad hoc. Normalmente, os dados
é apresentado pela primeira vez em um nível muito agregado, por exemplo, as vendas totais
por ano, e
então o usuário pode navegar para um nível mais detalhado, por exemplo, vendas por mês
por região. Quaisquer diferenças interessantes entre as regiões e / ou meses, pode
então ser utilizada para perfurar em uma nova direção até que uma nova percepção ou
compreensão
do negócio é obtida, o que poderia então ser usada para afetar os planos para os novos
promoções, catálogo de produtos da próxima temporada, ou o desenvolvimento de novos
produtos.
Esta, em resumo, é o que a análise se destina.

Intimamente relacionado com questões típicas de análise e soluções, é a dimen-
modelo profissional. Em última análise, é isso que permite a visualização de dados de forma
agregada
e recursos como drill up / down. Você encontrará informações detalhadas sobre
o modelo dimensional nos capítulos 6, 7 e 8 deste livro. No capítulo 15,
discute a aplicação prática das aplicações analíticas usando Mon-
drian e JPivot. Todos os exemplos analítico apresentado neste capítulo baseiam-se
em Mondrian / JPivot.

Exemplos Desenvolvedor BI: Slice and Dice

O exemplo de Slice and Dice está localizado na pasta de análise no BI Developer
Exemplos. Seu arquivo correspondente é chamada query1.xaction.

O Slice and Dice exemplo é o exemplo mais básico de análise incluído
Servidor Pentaho BI. Executá-lo produz uma tabela de referência cruzada dinâmica, também
conhecida
como um tabela dinâmica. A tabela dinâmica mostra os valores atuais e orçamentados
vendas, como
bem como a variação real versus orçamento. No contexto do Analytics, figuras como
estes são chamados medidas ou métricas. As medidas podem ser divididas de acordo com
Região, Departamento e Cargo. Estes títulos são mostrados no lado esquerdo
da tabela dinâmica e representam dimensões, que são aspectos que descrevem o
contexto das métricas.

Uma característica típica é que a tabela dinâmica não apenas mostra as próprias figuras
mas também os totais, e que o total pode ser calculada em vários níveis da
dimensões (ver Figura 1-11).

Na Figura 1-11, você pode ver as colunas para a Região, Departamento e
Posições. A primeira linha na tabela dinâmica mostra os resultados de todas as regiões,
Departamentos, e posições, e os valores são agregados ou''enrolada''
ao longo dessas dimensões. Isto representa o maior nível de agregação. Abaixo
que, você verá que os dados são divididos, na primeira coluna, todas as regiões é dividido em
Europa Central, Oriental, Austral e Ocidental, formando o segundo nível mais alto de
agregação para a dimensão Região. Na primeira linha para cada indivíduo
região, você verá os dados acumulados somente através do Departamento e de posições.
Para
na região Central, os dados são novamente dividida, desta vez mostrando todos os indivíduos
departamentos. Finalmente, para o departamento de Gestão Executiva, os dados são
novamente
dividido de acordo com a posição.

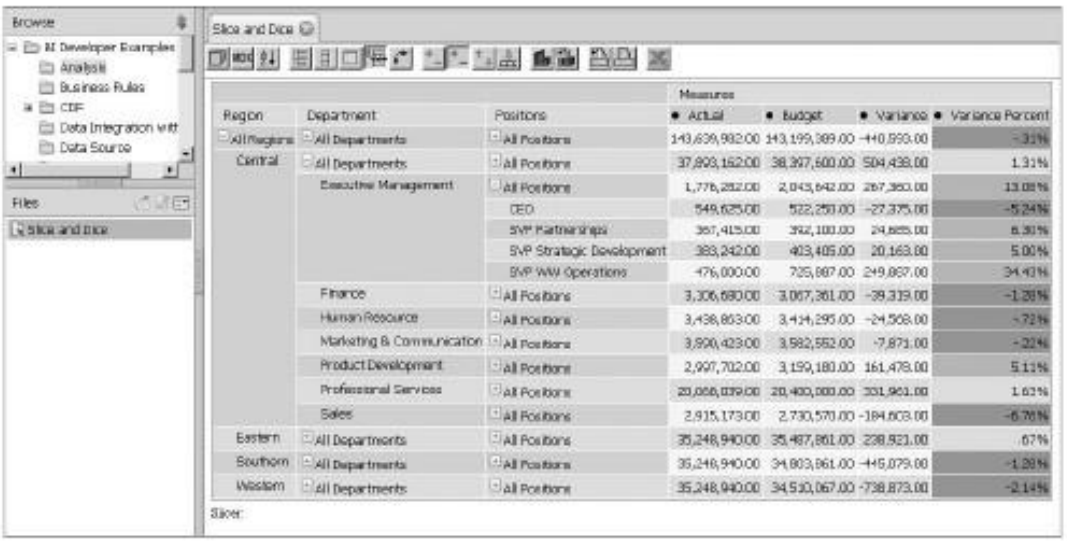


Figura 1-11: O Slice and Dice exemplo de tabela dinâmica

A divisão e rolando "é obtida dinamicamente, clicando no sinal de mais e os ícones menos que aparecem ao lado das etiquetas de identificação Região, Departamento e posições. Por exemplo, clicando no ícone de adição ao lado de qualquer um dos All Departamentos rótulos que na segunda coluna, você pode navegar e ver como o valor enrolado total para qualquer uma das métricas de vendas pode ser dividido. Ao clicar em um ícone menos vai rolar os valores de volta em conjunto para o total de novo, assim drill up.

Rodas de Aço Exemplos de Análises

Além da fatia de base e exemplo dado, você pode encontrar outros interessantes exemplos Analytics na pasta Análise no exemplo Rodas de Aço definido. Lá você vai encontrar dois exemplos:

- Análise de Mercado por Ano
- Análise de Linha de Produtos

Como a fatia de base e exemplo Dice, esses exemplos mostram uma tabela dinâmica, mostrando números agregados de vendas. Nesses exemplos, os números de vendas pode ser cortado ao longo do produto, do mercado (região), e Tempo.

Considerando o exemplo Slice and Dice exibido somente as medidas relativas à eixo horizontal, esses exemplos mostram um pouco mais de variedade, colocando no mercado no eixo horizontal. A Linha de Produtos exemplo de análise também coloca o tempo no o eixo horizontal, sob a Mercados.

Se você gosta, você pode usar caminhos alternativos para configurar os eixos usando o OLAP Navigator. Você pode chamar o Navigator OLAP, pressionando o botão com o ícone do cubo na barra de ferramentas que aparece no topo das páginas mostrando

os exemplos de análise. O Navigator OLAP e uma parte dessa barra de ferramentas são mostrado na Figura 1-12.

Figura 1-12: O Navigator OLAP

O Navigator OLAP mostrado na Figura 1-12 foi tirada da linha de produtos Análise exemplo. Na parte superior do Navegador OLAP, você pode ver a legenda Colunas e abaixo que são duas linhas, Mercado e Tempo. Isso corresponde diretamente com o Mercado e Tempo mostrado ao longo do eixo horizontal do tabela dinâmica. Na seção abaixo, você verá uma legenda linhas, com uma linha abaixo dele, Produto. Isso corresponde com os produtos que estão listados ao longo do eixo vertical da tabela dinâmica. Você pode mover os itens na seção Colunas para a seção de linhas e vice-versa, clicando no pequeno quadrado na frente dele.

Há uma terceira parte no navegador OLAP rotulados Filter. Nesta seção, encontrar clientes, Medidas, e status do pedido. Esses itens não são actualmente aparecem junto a um dos eixos da tabela dinâmica. Você pode mover os itens da Linhas e colunas seções para o filtro, clicando no ícone de filtro. Que se deslocam itens do filtro a qualquer um dos eixos é feito clicando no pequeno quadrado ícone que corresponde ao eixo ao qual você deseja mover o item.

Nós discutimos o Navigator OLAP em detalhes no Capítulo 15.

Exemplos Dashboarding

Os painéis são discutidos em detalhe no Capítulo 17. Se você está interessado no tablier placas, que são fortemente encorajados a verificar o Dashboard Comunidade exemplos Framework (CDF) painéis incluídos no Pentaho BI Server. Você pode encontrá-los na pasta CDF na solução de BI Developer exemplos.

Uma boa maneira de começar com Pentaho Dashboards está navegando para o Amostras subpasta da pasta CDF na solução de BI Developer exemplos. Aqui você vai encontrar exemplos para usar gráficos, relatórios, tabelas analíticas do pivô, e Mapas em um painel, e veja como você pode amarrar estes elementos juntos.

Uma vez que você tem um gosto para o que você pode fazer com painéis, você pode ler Capítulo 17 e siga as etapas detalhadas descritas lá para construir o seu próprio painel de instrumentos. Quando você está no processo de construção de seus painéis próprios, você

vai encontrar a documentação incluída com os exemplos CDF indispensável. Você pode encontrar documentação detalhada na subpasta de Documentação da CDF pasta. A documentação encontrada na pasta de Referência do Componente ser um companheiro em especial inestimável.

Outros exemplos

Muitos outros exemplos estão incluídos no Pentaho BI Server. Estes incluem exemplos para iniciar processos de ETL, para chamar serviços web, para enviar a saída de relatório para
uma impressora ou por e-mail, e muito mais. No entanto, não vamos discutir esses exemplos aqui. Muitos desses exemplos exigir configuração adicional, e outros não são particularmente instrutivo menos que você tenha necessidade de que o recurso especial.
No entanto, os leitores são encorajados a experimentar com os restantes exemplos.

Resumo

Este capítulo apresentou uma introdução ao software Pentaho e caminhou lo através de alguns dos exemplos que vêm com ele. Depois de instalar o software e explorando os exemplos, você deve ter uma boa idéia do que você pode fazer com Pentaho. O restante deste livro vai lhe ensinar como trabalhar com cada parte do Pentaho para criar suas próprias soluções Pentaho.

CAPÍTULO

2

Pré-requisitos

A intenção deste livro é permitir que você, leitor, desenvolver um Intel Negócios
ligence solução do início ao fim (e além porque as soluções de BI
necessidade de manutenção, bem). Para fazer isso, você vai precisar de algum suple-
ferramentas adicionais que não fazem parte da plataforma de BI Pentaho. Trata-se essencialmente
ferramentas de bancos de dados utilizados para conceber e criar um banco de dados, validação da qua-
dade dos dados nos sistemas de origem, e executar tarefas de manutenção, tais como fazer backups, criação de usuários ea configuração de restrições de acesso a dados. Como para utilizar os diferentes instrumentos é descrito nos respectivos capítulos, por exemplo, objecto de dados de perfil é abordada no capítulo 8 e um hands-on do tutorial
a ferramenta que usamos para essa tarefa está incluído ali também. Este capítulo explica onde obter o software e como ele pode ser configurado em um sistema de desenvolvimento.

NOTA Se você já tem um sistema rodando com Java habilitado e MySQL instalado, você pode pular a maior parte deste capítulo. Seu objetivo é fazer com que o novato usuário a funcionar com as ferramentas necessárias para projetar, construir e gerir bases de dados em uma instalação Pentaho / MySQL.

ciclos de lançamento de software, especialmente no mundo do código aberto, são relativamente
curto. Desde o tempo este livro foi finalizado em julho de 2009, novas versões do o software que usamos pode já se tornaram disponíveis. A fim de evitar possível confusão que lista os números de versão do software de pacotes que utilizados no livro:

- Ubuntu 9.04 (64 e 32 bits)
- Windows XP Pro SP3
- Sun Java 6.0.13

- MySQL Server 5.1.34
- Ferramentas GUI MySQL 5.0.12
- Power * Architect 0.9.13
- Squirrel 3.0.1
- eobjects DataCleaner 1.5.1

Configuração Básica do Sistema

Antes que você pode configurar qualquer coisa, há algumas coisas muito básicas para cuidar, como se certificar de que você tenha configurado corretamente no seu Java máquina. Para algumas tarefas, você precisará trabalhar com o terminal. Se você está que não estão familiarizados com isso, não se preocupe, vamos explicar o básico para você ir. Nós supor que a maioria dos nossos leitores estarão familiarizados com o uso de um baseado no Windows computador, mas espero que você aproveite esta oportunidade para explorar a opção de Pentaho usando em uma instalação Linux. Todos os exemplos Linux oferecemos aqui será baseado no Ubuntu Linux com desktop GNOME, mas usando outro Linux distribuições não deve ser muito diferente na maioria dos casos.

Instalar Ubuntu

Se você está entre os leitores de que ler este livro como uma boa oportunidade para iniciar experiências com o Linux, bem como o Pentaho BI Suite, você pode apreciar algumas notas quickstart sobre como ir sobre fazer isto. Existem vários opções para obter um sistema Ubuntu instalado e funcionando, dependendo de qual sistema pretende instalar-lo e se você quer substituir ou aumentar uma corrida Sistema Windows. Esta seção não é um guia completo de instalação passo-a-passo, mas salienta as diferentes opções e por onde começar.

Primeiro, você precisa obter o software. Ele pode ser baixado gratuitamente http://www.ubuntu.com, Onde você vai encontrar diferentes versões para diferentes arquiteturas de hardware. O site exibe tanto um desktop e um servidor edição, mas se você quiser dar um mergulho muito profundo em Linux, não comece com a edição do servidor. Isto só tem uma interface de terminal baseado em caracteres (sem GUI) e não todos os sinos e assobios que você poderia esperar. A próxima escolha é aquela entre 32 e 64 bits. A maioria dos computadores modernos será capaz de executar a versão de 64 bits do Ubuntu, e se você quiser usar o Ubuntu como o seu principal sistema operacional que provavelmente é sua melhor escolha. Todas as outras opções, por exemplo, que língua você quer usar, são parte da instalação em si.

O arquivo baixado é uma imagem ISO e você vai precisar de um gravador de imagem criar um CD a partir do arquivo baixado. A maioria dos programas de gravação de CD são capazes de fazer isso, mas se você usa o Windows e não tem como um programa, você pode baixar e instalar gratuitamente o Active @ ISO Burner da free.htm queimador http://www.ntfs.com/iso.

Usando o Ubuntu no modo nativo

Por nativo queremos dizer que você vai instalar o software em um computador para o máquina pode inicializar e executar o Ubuntu directamente a partir do disco rígido. Se você não está
certeza ainda sobre se ou não o Ubuntu é o caminho a percorrer, basta inserir o disco e aguarde a tela de boas-vindas. A primeira opção é executar o Ubuntu em modo Live, o que significa que você não instalar nada, mas executar o software a partir do CD. Se você gosta, clique no botão Instalar para iniciar o processo de instalação, se você não como ele, apenas ejete o CD e continuar usando o Windows.

O processo de instalação irá perguntar-lhe algumas perguntas. A maioria é bastante fácil de responder, tal como a língua, fuso horário e local, enquanto para outros é pode precisar de uma ajudinha. O mais importante é sobre a configuração do disco. Ubuntu terá todo o prazer que você mantenha seu sistema operacional existente, resultando em
uma configuração dual-boot. Nesse caso, cada vez que você ligar o seu computador você pode escolher para executar o Windows ou Ubuntu.

DICA Se você tiver um computador novo e gostaria de criar um sistema dual-boot, primeiro instalar o Windows e, em seguida o Ubuntu. Se você instala o Ubuntu, o instalador do Windows irá substituir a partição de boot do disco rígido.

Um passo-a-passo conjunto de instruções sobre como instalar o Ubuntu pode ser encontrada em
a ajuda on-line https: / / help.ubuntu.com / comunidade / GraphicalInstall.

Usando uma máquina virtual

A maioria dos computadores modernos têm amplo poder de processamento e memória, por isso
pode ser uma opção para usar uma máquina virtual para rodar o Ubuntu. Nesse caso, Ubuntu funciona como um convidado sistema, enquanto que seu sistema de operação regular atua como o host. Existem muitas soluções disponíveis para o funcionamento do Ubuntu como uma máquina virtual, a maioria deles pode ser encontrada na ajuda on-line em https: / / help.ubuntu.com / comunidade / VirtualMachines. Nós trabalhamos muito com VirtualBox da Sun (http://www.virtualbox.org), Na verdade, metade deste livro foi escrito em uma máquina virtual com Windows XP usando o VirtualBox em um Ubuntu De 64 bits do sistema host. A parte interessante do uso de software de virtualização é que há também uma grande coleção de ready-to-run imagens de máquinas virtuais disponíveis para download, o que significa que não há quase nenhuma instalação necessária para obter um novo sistema instalado e funcionando. Em caso de Pentaho, a melhor opção disponível é criado por Infobright, um fornecedor de banco de dados analíticos. Você pode baixar uma com-
imagem completo a partir de http://www.infobright.org/Downloads/Pentaho ICE VM que contém o seguinte:

- Ubuntu 8.04 do sistema operacional de servidor
- Infobright Community Edition
- Pentaho BI Suite Community Edition

A imagem é uma . Vmdk arquivo, o formato de arquivo nativo VMware, o VirtualBox, mas pode abrir esse tipo de arquivo, bem sem quaisquer problemas.

Trabalhando com o Terminal

Um terminal do Linux é basicamente o equivalente a tela de comando do Windows e pode ser usado como tal. Há um par de pequenas diferenças a ter em consideração, tais como o fato de que no Linux não há letras de unidade e nomes de caminho contém barras. Existem também diferenças maiores, como o fato de que a maioria dos comandos Linux são totalmente diferentes do seu Windows homólogos, e que, em Linux, tudo é case-sensitive. Passando para um Diretório com cd / opt / Pentaho pode retornar uma mensagem de erro informando que não há nenhum arquivo ou diretório, enquanto cd / opt / pentaho vai funcionar muito bem. Lembre-se que no Linux, pentaho e Pentaho são duas totalmente diferentes palavras!

Há duas maneiras para iniciar uma tela do terminal ou para colocá-lo mais precisamente, Existem dois tipos de telas de terminal: o terminal X básicos e do GNOME terminal. O terminal X é muito parecido com uma tela de comando do Windows, que tem um fundo preto com caracteres em branco e não há opções de menu. Você pode iniciar um terminal X pressionando Alt + F2 e digitando o comando xterm. O diálogo Executar Aplicativo aparece, exibindo a linha de comando pronto para aceitar o seu comando. Diferentemente de um terminal Windows, no entanto, o diálogo não pode retornar nada, ele só executa o comando quando você pressiona Enter.

A tela do terminal é o segundo terminal do GNOME, que tem um menu e um fundo branco com caracteres pretos. Isso é o que chamaremos a partir daqui. Você pode iniciar o terminal do GNOME, selecionando Applica-ções Acessórios Terminal ou pressione Alt + F2 e usando o comando gnome-terminal.

Não podemos cobrir o conjunto completo de comandos, mas o que você vai precisar a maioria são fornecidos nas seções seguintes.

NOTA Para uma referência útil sobre a utilização da linha de comando em um sistema Linux, experimente *Ubuntu Linux Toolbox*, por Christopher Negus e Caen Francois, Edições Wiley, 2008.

Lista de Navegação

Você pode usar os seguintes comandos para navegar pelo directório estrutura:

- cd -Mude o diretório, mesmo que no Windows.
- cd .. -Move um nível acima (cd ../.. move-se dois níveis para cima, e assim por diante).

cd ~ -Vá para seu diretório home.

cd / -Vá para o diretório raiz (sistema).

O sistema operacional""sabe os caminhos para que quando você entra no primeiro caracteres de um diretório e pressione a tecla Tab, o nome completo é concluído automaticamente. Por exemplo, digite CD / o, Pressione a tecla Tab, eo caminho é completada automaticamente para cd / opt /. Adicionar pe pressionar Tab novamente para obter o cd
/ Opt / pentaho / comando. Claro que estas listas têm de existir, portanto, se houver não é / Opt / pentaho diretório, é óbvio que o sistema operacional não pode encontrá-lo.

Quando você quiser limpar a tela e ter o prompt exibido na superior da janela novamente, basta digitar o comando claro e pressione Enter.

História de comando

Usando a cima e para baixo, você pode navegar através anteriormente emitido comandos. Todos os comandos são armazenados em um comando arquivo histórico, que pode ser
visto digitando história na linha de comando. Se a lista é longa, você pode uso history | more e na página a lista com a barra (CTRL + C para final). Se você quiser reutilizar um comando específico da história, você pode digitar uma ponto de exclamação seguido do número da linha histórico de arquivo (por exemplo,! 174) e que o comando será executado novamente. Uma opção mais flexível é a utilização do CTRL + R combinação chave, que inicia um reverse-i pesquisa de texto na história arquivo, o que significa que o comando mais recentemente emitidos contendo a seqüência será encontrado em primeiro lugar. Observe que você pode continuar a digitar a seqüência de pesquisa, que
dinamicamente alterar o argumento para a pesquisa. Por exemplo, pressionar CTRL + R e digitando eexibe o seguinte:

```
(Reverse-i-search) `e ': claro
```

Adicionar cexibe o seguinte:

```
(Reverse-i-search) `ec": echo $ JAVA_HOME
```

O comando pode ser simplesmente aceite (e executadas) pressionando a tecla Enter, mas se você quiser modificar o primeiro comando, pressione a seta para a esquerda ou direita, que vai abrir o comando na linha de comando para edição.

Utilizando Links Simbólicos

A maioria dos programas ir diretamente para o diretório home quando você deseja abrir ou salvar um arquivo. Às vezes não é possível alterar o caminho padrão onde o aplicativo deve procurar pastas e arquivos. O Pentaho ferramentas de projeto como

Designer de Relatórios ou do Mondrian esquema Workbench principalmente trabalhar a partir de
o diretório home do usuário que iniciou o programa. Isso nem sempre é
uma localização conveniente para começar, especialmente quando o último local utilizado não é
lembrado pelo programa quer. Abrindo um caminho diferente a cada vez que você
deseja abrir ou salvar um arquivo leva tempo, então a capacidade de abrir a pasta certa
diretamente do seu diretório home seria uma comodidade bem-vinda. Esta é
onde links simbólicos vêm a calhar.

Criar links simbólicos no Ubuntu

Um link simbólico no Linux se parece com um atalho no Windows, mas criá-los
é um pouco diferente.

NOTA atalhos no Windows são arquivos comuns que podem ser resolvidos apenas pelo Windows
eo Windows Explorer. O Windows Vista suporta ligações"verdadeiro"simbólico.

Há duas maneiras de criar links simbólicos (ou symlinks como eles são normalmente
chamado): usando o navegador de arquivos Nautilus do GNOME, ou digitando os comandos
na linha de comando. Ao usar o Nautilus, clique com o botão direito do mouse no arquivo ou pasta
para o qual você quer criar uma ligação, selecione Criar link do drop-down
menu e copie o link resultante para o local desejado, após o que pode ser
renomeado. Usando a linha de comando exige conhecimento da sintaxe, que
é bastante simples. O comando é um simples ln, Seguido pelo de opções,
o local para a ligação, eo nome do link. Para criar um link simbólico em sua
diretório home que aponta para o diretório soluções Pentaho, a seguinte
comandos podem ser inseridos:

```
ln-s / opt / pentaho / biserver-ce / pentaho soluções ~ pentaho /
```

A -S denota opção que você está criando um link para um diretório, não um único
arquivo. Agora, qualquer referência a ~ Pentaho / (Subdiretório pentaho na atual
pasta home do usuário) é traduzido automaticamente para a pasta de destino.

Criando Symlinks no Windows Vista

Criar links simbólicos em Vista funciona de maneira semelhante como em Linux quando o
linha de comando é usado, mas o comando e os parâmetros são diferentes. Para fazer
as coisas mais confusas, a ordem é invertida argumento: no Windows link
é especificado antes do alvo. A ligação simbólica mesmo que no exemplo anterior
podem ser criadas com o seguinte comando:

```
mklink / DC: \ Documents and Settings \ Administrador \ Meus Documentos \ pentaho
 C: \ Program Files \ pentaho \ ce-biserver \ pentaho soluções
```

Java Instalação e Configuração

Todos os programas Pentaho são desenvolvidos em Java e exigem uma Máquina Virtual Java a estar presente no computador que irá executar Pentaho. Instalando o Java foi facilitou muito tanto em Linux e Windows, graças a Sun Microsystems.
Além de instalar o Java, um passo importante é a configuração da esquerda para definir o variável de ambiente JAVA_HOME. Sem isso, os programas Java não sei onde procurar as bibliotecas Java e seu software de Java ainda não será executado. (Você pode ignorar esta seção se você já instalou o Java e configurar seu ambiente variável).

Instalando o Java no Linux Ubuntu

Você pode instalar o Java em uma máquina Linux de duas maneiras. A primeira opção é fazê-lo
manualmente, indo ao www.java.com, Baixar o instalador, e funcionando em seu sistema. A segunda opção ea melhor é usar o Synaptic Package Manager. Os pacotes Java fazem parte dos repositórios do Ubuntu regular, de modo Abra o menu Sistema-Administração e selecione Gerenciador de Pacotes Synaptic. Digite o su (Superusuário) senha para iniciar o Synaptic. Na caixa de pesquisa rápida, tipo java6, Que irá exibir todos os pacotes disponíveis a partir de Java dom. Selecione o pacote sun-java6-jdk (Java Development Kit), que tem um casal de pacotes necessários, que serão selecionados automaticamente pelo Synaptic. Clique em Aplicar
para baixar e instalar o software. Isto irá instalar o Java no subdiretório / Usr / jvm / lib /.

Alternativamente, você pode usar a linha de comando para instalar o Java SDK. Abrir uma tela de terminal e digite os seguintes comandos:

```
sudo shell> apt-get update
sudo shell> apt-get install sun-java6-jdk
```

O primeiro comando garante que todas as informações do repositório é atualizado; o segundo comando irá instalar os pacotes Java depois de ter confirmado a instalação digitando Y. Durante a instalação você terá que concordar com o termos do contrato de licença.

Quando você abrir o diretório de instalação / Usr / lib / jvm você vai notar duas entradas novas: o diretório Java real com um número de versão do postfix, e um link simbólico que aponta para este diretório primeiro. O link simbólico é o que você vai usar para a variável de ambiente. Primeiro verifique se a instalação conseguiu através da abertura de uma tela de terminal e entrar no java-version comando. Isto deveria lhe dar uma saída semelhante ao listados a seguir:

```
java version "1.6.0_13"
Java (TM) SE Runtime Environment (build 1.6.0_13-b03)
Java HotSpot (TM) Client VM (build 11.3 b02, de modo misto, a partilha)
```

A variável de ambiente pode ser definida por adicionar uma linha ao arquivo / Etc / environment, Mas você precisa de privilégios de root para fazê-lo. No mesmo terminal tela, digite o comando sudo gedit / etc / environment. Isto irá iniciar o editor com o arquivo aberto. Basta adicionar a seguinte linha a este arquivo:

```
JAVA_HOME = "usr/lib/jvm/java-6-sun /"
```

Salve o arquivo e feche o editor. Você pode verificar se a variável é definir corretamente emitindo o eco $ JAVA_HOME comando, mas você vai perceber que nada é retornado ainda. Se você quiser ativar a variável no terminal sessão, você pode usar o comando source / etc / environment, Mas para ativar a variável para todas as sessões, só fazer logoff e logon novamente (não há necessidade de reiniciar o sistema).

Instalando o Java no Windows

Para a instalação, basta abrir um navegador, vá para www.java.com, E clique sobre o Download gratuito do Java botão. Siga as instruções no site para instalar o Java. O próximo passo é definir a variável de ambiente. As variáveis de ambiente pode ser adicionado ao abrir Propriedades do Sistema no painel de controle e escolher o Clique na guia Avançado e selecione Configurações do Sistema (ou do Sistema Avançado As configurações diretamente no Vista). Adicionar uma nova variável do sistema chamada JAVA_HOME que aponta para o caminho de instalação do Java, como mostrado na Figura 2-1 (o caminho pode ser diferente em seu próprio sistema).

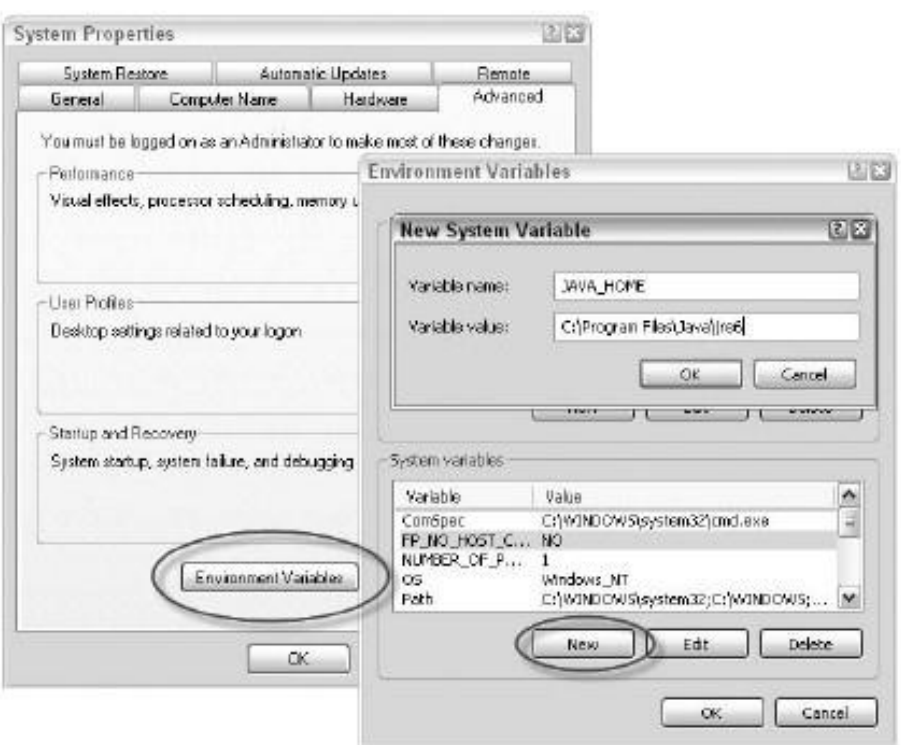


Figura 2-1: Definir a variável JAVA_HOME

Reinicie sua máquina para ativar a configuração. Se você não deseja reiniciar sua máquina de imediato, é possível definir uma variável do sistema temporário abertura de uma tela de comando e usando a SET comando para definir o meio ambiente variável:

```
SET JAVA_HOME = C: \ Arquivos de programas \ Java \ jre6
```

Você pode verificar o ajuste com o ECHO comando, seguido pela variável cujo valor deverá ser devolvido fechado em sinais de porcentagem. Isto irá mostrar o caminho se ele tiver sido definido corretamente, ou o nome da variável, se esta ainda não é especificados:

```
Shell> echo% FOO%
%% FOO

Shell> echo% JAVA_HOME%
C: \ Arquivos de programas \ Java \ jre6
```

Instalação do MySQL

Este livro vai fazer o uso extensivo do MySQL para que a próxima a ter Java habilitado, Este é o segundo pré-requisito importante para seguir junto com todos os exemplos e tutoriais. Se você usa Windows ou Linux, instalar o MySQL é bastante simples e, no caso do Ubuntu, quase automático.

Instalando o servidor e cliente MySQL em Ubuntu

Você pode instalar o MySQL no Linux de duas maneiras: manualmente, baixando os arquivos de instalação e fazer toda a configuração do mesmo, ou usando o Package Manager. Os pacotes de dados MySQL fazem parte do Ubuntu repositórios, e desde o lançamento do Ubuntu versão 9.04, a versão 5.1 do MySQL está disponível no repositório. Você pode optar por fazer um manual instalar, por exemplo, para tentar os mais recentes recursos ou ter total controle sobre o instalação. Se você precisar fazer uma instalação manual, veja o site do MySQL para instruções de instalação.

Para instalar o MySQL usando o Synaptic, abra o Gerenciador de Pacotes Synaptic a partir do menu Sistema-Administração e busca de mysql. Vá para a servidor msql 5.1 pacote e marcá-lo para a instalação. Note que todos os necessários pacotes adicionais são automaticamente selecionados, basta confirmar selecionando Marcos e selecione Aplicar para iniciar o download e instalar os pacotes.

DICA Você pode, simultaneamente, selecione o pacote mysql-admin e ignorar a passo instalação separada GUI.

O configurador do Debian irá pedir uma nova senha para o root do MySQL"" usuário. Esta é uma nova senha para um novo usuário que será criada para o início

e gerenciar o MySQL exemplo. Infelizmente, a definição de uma senha não é obrigatório, mas aconselhamos que você defina um (e não se esqueça disso!). Para o configurações do servidor de email, basta aceitar o padrão determinado. Quando a instalação estiver
terminado, você terá uma Alterações aplicadas""mensagem. Fechar esta eo pacote Manager como bem o servidor MySQL foi instalado.

Você pode verificar a instalação, conectando-se à instância do MySQL linha de comando. Abra uma tela do terminal e digite o mysql fol comando seguida do -P-uroot <sua_senha>. Isto deve abrir o comando MySQL prompt:

```
shell> mysql-uroot-ppassword
Bem-vindo ao MySQL monitor. Comandos com final; ou \ g.
Seu código de conexão com o MySQL é de 43
Server versão: 5.1.31-1ubuntu2 (Ubuntu)


'Help;' ou '\ h' para ajuda. 'C \' Tipo 'para limpar o buffer.

Mysql>
```

Saia do cliente com o parar comando seguido por um ponto e vírgula:

```
Mysql> quit;
Tchau
```

Parabéns, o servidor está funcionando!

Instalando o servidor MySQL eo cliente no Windows

Os instaladores do Windows para o MySQL 5.1 pode ser encontrada em http://dev.mysql.com / Downloads/mysql/5.1.html. Selecione o instalador MSI e baixar este seu computador. Depois de iniciar o instalador, você é apresentado com três instalar opções: Typical, completas ou personalizadas. A primeira opção vai fazer, mas instalar seus arquivos de dados no diretório C: \ Documents and Settings \ All Usuários \ Dados de aplicativos \ servidor MySQL \ MySQL 5.1. Se você preferir ter os arquivos de dados localizados em outro lugar, selecione a instalação personalizada e mudar o
Arquivos de Dados MySQL caminho. Depois de exibir alguns comerciais MySQL Enterprise telas de informação, a configuração do MySQL é iniciado, selecione a Norma Configuração e aceitar todos os padrões na próxima tela. Agora, a raiz passa- tela aparece a palavra. Embora você pode deixar a senha de root em branco, nós recomendo fortemente contra ele. Se você quiser ser capaz de gerir o banco de dados outras máquinas que não localhost, você tem que marcar a opção de acesso como root também. A tela final permite que você execute as definições de configuração e inicia o MySQL serviço em seu computador.

Ferramentas GUI MySQL

O cliente MySQL é meramente um pedaço de software para conectar a um MySQL servidor. Para trabalhar com o banco de dados você vai precisar adicionar mais duas ferramentas, o
MySQL Administrator eo MySQL Query Browser. Ambos estão incluídos
nas ferramentas GUI, o que pode ser encontrada em http://dev.mysql.com/downloads / Gui-tools/5.0.html.

Instalar Ubuntu

A GUI site de download de ferramentas permite que você baixar o binário Linux instaladores, mas as ferramentas também estão disponíveis nos repositórios do Ubuntu. Porque
esta é uma maneira mais conveniente de instalar o software, abra o Synaptic
Package Manager novamente, procure por mysql-admin, e marcá-lo para a instalação. Note que os pacotes mysql-query-browser e mysql-gui-tools-comum
são incluídos automaticamente. Escolha Aplicar para instalar as ferramentas. Seu menu será agora ser enriquecido com uma nova entrada principal, chamada de programação, com dois itens:
o administrador eo Query Browser.

Instalar o Windows

Basta baixar o instalador do Windows http://dev.mysql.com/downloads / Gui-tools/5.0.html. É uma instalação simples do Windows, que não exige
quaisquer configurações especiais. Aceitar os padrões vai fazer tudo certo. O programa atalhos podem ser encontrados no menu Iniciar do Windows sob a entrada MySQL.

Database Tools

Trabalhando em soluções de BI geralmente significa trabalhar com dados e bases de dados. Cada
banco de dados vem com seu próprio banco de dados e gerenciamento de ferramentas de consulta, mas o que
Se você precisar acessar vários bancos de dados ou precisa de uma ferramenta para desenvolver um novo
banco de dados em um modo visual? Esta seção apresenta três das nossas ferramentas favoritas para
projetar, desenvolver e consultar bancos de dados. Todas as ferramentas são escritos em Java
então eles vão rodar em qualquer plataforma, desde que a JVM está instalado.

Power * Architect e outras ferramentas de design

Pentaho não fornece as ferramentas necessárias para desenvolver projeto de um data warehouse
porque há muitos bancos de dados lá fora, e em muitos casos as organizações
já têm uma ou mais ferramentas de projeto disponíveis. Para apoiar o pleno
ciclo de vida do projeto, a partir de lógicas de negócios para a modelagem técnica, incluindo modelo de comparações entre versões e gerenciamento de ciclo de vida, documentação automóvel

recursos e apoio da equipe, há (tanto quanto nós estamos cientes) não open source soluções disponíveis. Mas se houver, por favor, avise-nos! Para design de banco de dados esquemas, você tem duas opções, ambas fonte freeware e aberta. A
Segue-se uma pequena, mas longe da lista completa:

- Power * Architect (http://www.sqlpower.ca/page/architect) - Nosso ferramenta de escolha para este livro. Capítulo 8 contém mais instruções sobre instalação de energia * Arquiteto e como usar a ferramenta para criar banco de dados diagramas e data marts.
- MySQLWorkbench (http://dev.mysql.com/downloads/workbench / 5.1.html) - Padrão ferramenta de design de banco de dados MySQL eo sucessor do DBDesigner popular.
- ERDesigner Mogwai (http://mogwai.sourceforge.net) - Eclipse base, mas há também um esquilo plugin disponível (ver próxima seção).
- ERMaster (http://ermaster.sourceforge.net) - Ainda com alguns Japanese textos de ajuda aqui e ali.
- Azzurri Clay (www.azzurri.jp) - Amplamente utilizado plugin do Eclipse. O núcleo edição é gratuita.

Squirrel SQL Client

Esquilo é uma ferramenta open source consulta SQL que lhe permite abrir e consultar praticamente qualquer banco de dados que já foram desenvolvidos, desde que um driver JDBC
disponíveis. Instalando o esquilo é fácil: vá para www.squirrelsql.org/ e siga as instruções na seção de Download e instalação para baixar o arquivo.

Instalar Ubuntu

Vamos modificar as instruções de instalação a partir do site um pouco, embora o padrão irá funcionar tão bem. Se você seguir os padrões, a ferramenta será instalada no diretório / Usr / local / esquilo SQL Client. Porque nós não gostamos instalar nada no / Usr / local, Muito menos usando um nome de pasta misturada caso com espaços, recomendamos a instalação da seguinte forma:

1. Abra um terminal e navegue até a pasta onde você baixou a arquivo do instalador. Use o seguinte comando para iniciar o instalador:

   ```
   sudo java-jar squirrel-sql-<versão> install.jar
   ```

   No comando anterior, <versão> deve ser substituído pelo número da versão atual.

2. A terceira tela do instalador pede um caminho de instalação. Alterar esta em / Opt / tools / esquilo, Prima Seguinte e clique em OK para aceitar as criação da nova pasta.

3. A tela seguinte mostra uma longa lista de plugins que podem ser instalados. Dependendo do banco de dados e linguagens que você gostaria de usar, você pode fazer a sua selecção. Apenas certifique-se de selecionar a opção MySQL e deixar o checkbox selecionado plugins padrão porque contém uma o recurso de Conclusão de código. Para obter uma descrição de todos os plugins disponíveis, ir para http://www.squirrelsql.org/index.php?page=plugins.

4. A última tela do instalador pede um local para o atalho. Isso é para Windows e não vai ajudar muito em uma máquina Linux, então basta pressionar Avançar para concluir a instalação.

Esquilo agora pode ser iniciado executando o esquilo sql.sh script de um linha de comando, mas para maior comodidade, vamos criar um lançador para adicioná-lo no menu. Abra o editor de menu clicando com o lado esquerdo da norma painel e selecione Editar Menus. Uma nova entrada pode ser adicionado a qualquer padrão principais opções do menu ou você pode criar um novo. Clique em Novo Item para adicionar um item de menu e preencha os campos disponíveis, como mostrado na Figura 2-2.

Figura 2-2: lançador Esquilo

O campo de comentário é opcional, e os ícones podem ser alterados clicando nele e navegando para o / Opt / ferramentas esquilo / ícones pasta.

Instalar o Windows

Instalando o esquilo no Windows requer um esforço ainda menos do que no Linux, basta clique duas vezes no jar de instalação baixado e começa o instalador. A ferramenta será instalado na C: \ Arquivos de programas \ Client SQL SQuirreL. Se você prefere um local diferente que você pode, naturalmente, alterá-lo aqui. Agora, as seleções para atalhos na última tela irá criar a entrada de menu adicionais automaticamente.

SQLeonardo

A última ferramenta que você pode querer considerar é SQLeonardo, uma consulta SQL ferramenta que pode fazer algo que ainda não está disponível no Squirrel: gráfica

design da consulta. O software pode ser baixado http://sourceforge.net / Projects / sqleonardo em um arquivo zip com apenas um Jar. arquivo dentro. Extraia o zip para o local desejado (no nosso caso, / Opt / ferramentas sqleonardo /) E criar um lançador (Linux) ou o atalho do menu (Windows). Não se esqueça de definir a permissão para o arquivo"Permitir execução do arquivo como programa."Caso contrário, ele não será iniciado. Figura 2-3 mostra a interface para a concepção de consultas com o Pentaho banco de dados exemplo aberto.

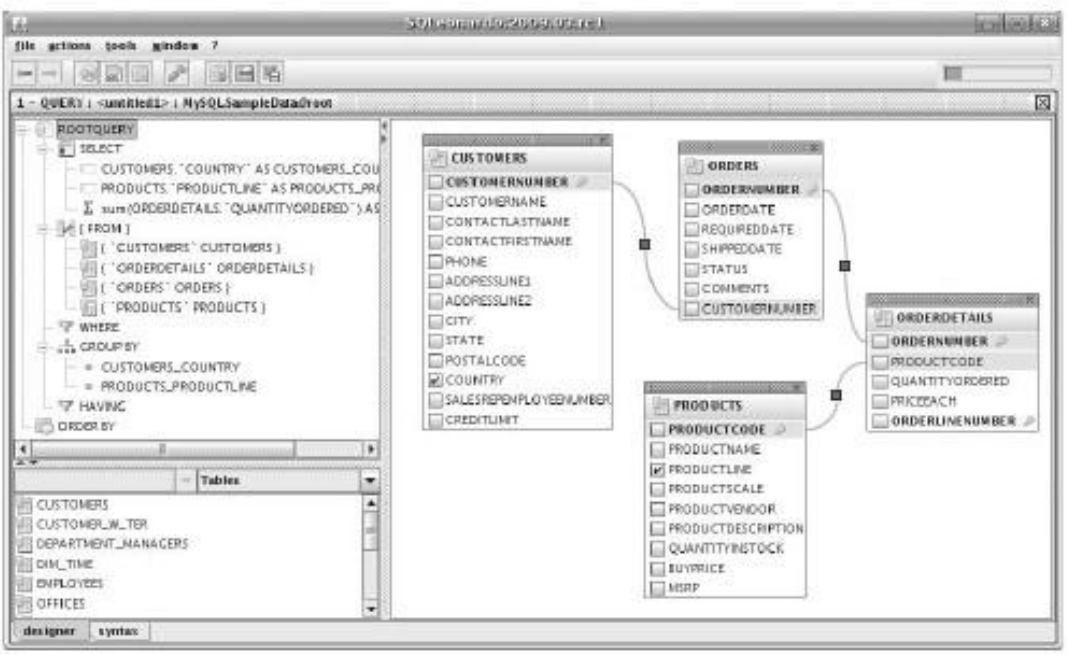


Figura 2-3: SQLeonardo interface de consulta

Cobrimos SQLeonardo com maior profundidade no Capítulo 13, porque é também a ferramenta de consulta gráfico no Report Designer Pentaho.

Resumo

Este capítulo apresenta as ferramentas complementares necessários para desenvolver e gerenciar
uma solução de BI. Na verdade, dificilmente você pode ligar o MySQL uma ferramenta adicional porque
constitui a base da solução iremos desenvolver durante o curso da presente livro. Cobrimos a seguir neste capítulo:

- Trabalhando com a linha de comando no Ubuntu ou outras distribuições Linux
- Configurando Java e JAVA_HOME variável de ambiente
- A instalação do MySQL e as ferramentas GUI MySQL

- Introdução às ferramentas de design de banco de dados
- Instalação de esquilo, uma ferramenta de consulta de banco de dados universal
- Instalação de SQLeonardo, um construtor de consulta gráfica conveniente

Instalação das ferramentas descritas neste capítulo é apenas um pré-requisito para a criação de bancos de dados de exemplo e exemplos, vamos utilizar todo o livro. Todos os scripts de criação de banco de dados, arquivos de dados e instruções de instalação-
ções que utilizamos nos exemplos podem ser encontrados no site do livro na www.wiley.com / go / pentahosolutions.

CAPÍTULO

3

Instalação de Servidor e Configuração

Neste capítulo você irá aprender as tarefas básicas envolvidas na configuração do Pentaho BI Server. Além disso, podemos apresentá-lo a configurar e trabalhar com o Administração Pentaho Console (PAC).

Configuração do Servidor

Como se ilustrou no capítulo 1, você pode executar o Pentaho BI Server imediatamente Depois de baixar e descompactar o software. No entanto, a configuração padrão-ração normalmente não é suficiente para fins de produção para o seguinte razões:

- O BI Server não está configurado para iniciar automaticamente quando o sistema operacional reinicialização do sistema.
- Inicialmente, o BI Server será acessível apenas a partir da porta 8080, o que pode conflito com outro serviço executado em sua máquina.
- Algumas características, tais como mensagens de email e publicação de conteúdo de BI, requerem configuração extra para que eles trabalhem em tudo.
- Por padrão, um em memória de banco de dados HSQLDB é usado para todo o sistema bases de dados, e você pode querer usar um RDBMS que você é mais familiarizado com o lugar.
- Você pode precisar fornecer drivers JDBC extra para se conectar a um RDBMS particular.

Neste capítulo, descrevemos como alterar a configuração padrão do Pentaho BI Server. Apesar de não fornecer um guia detalhado para Pentaho administração e configuração do servidor, este capítulo mostra como alguns dos seus principais componentes são controlados.

Instalação

Já descrevemos como baixar e extrair o Pentaho BI Server.
Se necessário, consulte o Capítulo 1 para obter instruções detalhadas. O diretório descompactado
contém dois subdiretórios:

- administração console -Este é um serviço administrativo para gerir e configurar o real Pentaho BI Server. Isso também é conhecido como o PAC (Para a Administração Pentaho Console).
- biserver-ce -Este é o atual Pentaho BI Server (Community Edition).

Directório de Instalação

No Capítulo 1, que também mencionou que você pode instalar o Pentaho Server em qualquer local que você deseja. No entanto, dependendo do seu sistema operacional, não são alguns locais que (por convenção) fazem mais sentido do que outros. Para o finalidade do restante do livro, vamos assumir os seguintes locais:

- C: \ Program Files \ pentaho para sistemas Windows
- / Opt / pentaho para sistemas baseados em UNIX

Você deve criar esse diretório e mover-se tanto a administração -Console e biserver-ce diretórios (incluindo seu conteúdo) para este diretório. No restante deste livro, vamos nos referir ao biserver-ce diretório como o diretório home Pentaho, e os administração console diretório como o diretório home do PAC.

NOTA Você precisa de privilégios de root para criar um subdiretório / Opt. Por exemplo, no Linux Ubuntu, você pode usar sudo temporariamente para obter estes privilégios:

```
shell> sudo mkdir / opt / pentaho
```

Conta de Usuário

Para considerações de segurança, você deve considerar a criação de um usuário separado conta para executar aplicações de servidor, como Pentaho. Normalmente, a conta de usuário seria também dono do diretório que contém o software. Além disso,
essa conta de usuário será bloqueado para fora do restante do sistema. Esta é uma questão de controle de danos: um bug no software ou um servidor hackeado simplesmente

não pode fazer nenhum mal fora de seus diretórios próprio software, ou seja, desde a conta de usuário reservado para Pentaho ainda não possui permissões fora a árvore Pentaho. Listagem 3-1 mostra uma possível forma de fazer essa configuração em Linux:

Listagem 3-1: Configurando uma conta de usuário, grupo e diretório para Pentaho em baseados em UNIX sistemas

```
# Crie um grupo para o usuário pentaho
sudo shell> addgroup pentaho
Adicionando 'pentaho "grupo (GID 1001) ...


# Criar um usuário do sistema de Pentaho
sudo adduser shell> - sistema - ingroup Pentaho - Pentaho disabled-login
Adicionando 'pentaho "usuário do sistema (UID 115) ...
Adicionando 'pentaho' novo usuário (UID 115) com 'pentaho' grupo ...
Criar diretório home "/ home / pentaho ...


# Criar diretório de software, e descompactar o software não
shell> sudo mkdir / opt / pentaho
sudo shell> cd / opt / pentaho
shell> sudo tar zxvf ~ / downloads/biserver-ce-CITRUS-M2.tar.gz


# ... Muita saída ...


# Conceder a propriedade de diretórios de software para o usuário pentaho
sudo shell> chown-R pentaho: pentaho / opt / pentaho
```

Após configurar a conta de usuário e concessão de propriedade do Pentaho software, você pode começar a Pentaho da linha de comando usando um comando como este:

```
shell JAVA_HOME pentaho> sudo-u = usr/lib/jvm/java-6-sun. / start-pentaho.sh
```

Observe o -U pentaho no comando: isso garante que o comando será executado com as permissões do pentaho usuário. Se você omitir que, o servidor será executado com a permissão de raiz, que é exatamente o que você quer evitar! Na próxima seção, você verá como você pode definir as coisas para o Pentaho BI Server é iniciado automaticamente durante a inicialização do sistema operacional.

Configurando o Tomcat

O Pentaho BI Server está pré-configurado com base no Apache Tomcat Servlet recipiente. O software Tomcat reside no tomcat diretório, que reside dentro do diretório home do servidor Pentaho. Por padrão, o Tomcat atende na porta 8080. Isso significa que Pentaho é, também, acessível através deste porto. Por exemplo,

se o servidor Pentaho é iniciado em sua máquina local, o endereço web do Pentaho home page é http://localhost:8080/pentaho.

NOTA Há produtos de servidores que usam a porta 8080 por padrão, em particular outros recipientes Java Servlet como o JBoss eo GlassFish, mas também o Oracle Application Express (APEX), que é instalado como parte do Oracle Database Server Express. Vários servidores não podem usar o mesmo número de porta em simultâneo. Por esta razão, você pode precisar configurar o software do servidor para garantir que cada um deles é atribuído um porta exclusiva.

Coisas como a porta do servidor, bem como as características básicas de acesso são configurados
ao nível do Tomcat. A maioria das configurações do Tomcat é controlado através de XML arquivos que residem no tomcat / conf diretório. Por exemplo, se você quiser
alterar a porta usada pelo Tomcat, você pode editar o seguinte trecho
a partir de conf / tomcat / server.xml:

```
<Connector port = "8080" maxHttpHeaderSize = "8192"
        MaxThreads = "150" minSpareThreads = "25" maxSpareThreads = "75"
        enableLookups redirectPort = "false" = "8443" acceptCount = "100"
        connectionTimeout = "20000" disableUploadTimeout = "true" />
```

Assim, a alteração do porto atributo faz com que o Tomcat para escutar em outra porta. Se você
decidir alterar o número da porta aqui, você também precisa mudar o núme-porto ber na web.xml arquivo de configuração que está localizado na tomcat / webapps / WEB-INF diretório. Você deve procurar um trecho parecido com este:

```
<context-param>
   <param-name> base-url </ param-name>
   <param-value> http://localhost:8080/pentaho/ </ param-value>
</ Context-param>
```

NOTA Uma discussão completa de configuração do Tomcat está fora do escopo deste livro. No entanto, você pode encontrar bons recursos online e livros sobre o assunto. A ponto de partida óbvio para saber mais sobre o Tomcat é o manual, que pode ser encontrar em http://tomcat.apache.org/tomcat-5.5-doc/.

Arranque automático

O Pentaho BI Server, bem como o console de administração, são servidores aplicações e, normalmente, você gostaria que inicie automaticamente após a inicialização o sistema operacional.

Arranque automático em sistemas Unix / Linux

Usuários de sistemas baseados em UNIX necessidade de criar um script de inicialização (às vezes chamado de
RC), que começa do Servidor Pentaho. Listagem 3-2 mostra um muito básico, mas totalmente

funcional, script de inicialização chamado pentaho-init.sh. Você deve ser capaz de descobrir o que ele faz lendo o código e comentários.

Listagem 3-2: Um script básico pentaho-init.sh

```
#! / Bin / sh
# Vá para a casa pentaho
cd / opt / pentaho / biserver ce-
# Configurar o comando para o usuário Pentaho, o ambiente java conjunto
cmd = "JAVA_HOME pentaho sudo-u = usr/lib/jvm/java-6-sun
JAVA_OPTS =- Djava.awt.headless = true "



case "$ 1" em
início)
# Executar o script de inicialização original pentaho
$ Cmd. / Start-pentaho.sh>> pentaho-demo.log &
;;
stop)
# Executar o script de inicialização original pentaho
$ Cmd. /> Stop-pentaho.sh> pentaho-demo.log &
;;
reiniciar)
$ 0 parar
$ 0 começar
;;




*)
echo "Uso: $ 0 {start | stop | restart}"
saída 1
esac


exit 0
```

NOTA Note que o script da Listagem 3-2 não se destina a ser um grande exemplo de UNIX scripting. É simplesmente uma abordagem mínima que começa o trabalho feito. Se você está interessado em escrever esses scripts você mesmo, você deve referir-se aos inúmeros recursos de scripting em UNIX / Linux e administração do sistema.

A pentaho-init.sh script deve ser colocado no / Etc / init.d diretório. (Note que isso requer privilégios de root, e você deve usar sudo copiar ou mover o script para esse local.) Você deve então testá-lo e verificar se você pode usá-lo para iniciar e parar o servidor Pentaho BI, como mostrado aqui:

```
Shell>    cp pentaho-init.sh / etc / init.d
Shell>    cd / etc / init.d
Shell>    sudo. / pentaho-init.sh
Uso:      . / Pentaho-init.sh {start | stop | restart | status}
```

```
shell> sudo. / start pentaho-init.sh
shell> sudo. / stop pentaho-init.sh
```

Para distribuições baseadas em Debian Linux, incluindo o Ubuntu, você pode então usar o update-rc.d utilitário, que estabelece uma série de links simbólicos, provocando o script para ser usado em tempo de boot para iniciar Pentaho (e para pará-lo no sistema desligamento):

```
shell> sudo update-rc.d padrão pentaho-init.sh
update-rc.d: warning: / etc / init.d / pentaho-init.sh cabeçalho estilo faltando LSB

 Adicionando a inicialização do sistema para o / etc / init.d / init.sh-pentaho ...
   / Etc/rc0.d/K20pentaho-init.sh -> .. / init.d / pentaho-init.sh
   / Etc/rc1.d/K20pentaho-init.sh -> .. / init.d / pentaho-init.sh
   / Etc/rc6.d/K20pentaho-init.sh -> .. / init.d / pentaho-init.sh
   / Etc/rc2.d/S20pentaho-init.sh -> .. / init.d / pentaho-init.sh
   / Etc/rc3.d/S20pentaho-init.sh -> .. / init.d / pentaho-init.sh
   / Etc/rc4.d/S20pentaho-init.sh -> .. / init.d / pentaho-init.sh
   / Etc/rc5.d/S20pentaho-init.sh -> .. / init.d / pentaho-init.sh
```

Você pode usar o mesmo update-rc.d utilitário para remover um serviço já existente usando uma linha como esta:

```
sudo shell> update-rc.d-f pentaho-init.sh remover
```

Outra ferramenta que vem a calhar para gerenciar os scripts de inicialização do Linux é o gerenciador de boot-up gráfica vagabundo. Você pode instalá-lo usando o Synaptic Package gerente, ou usando o seguinte comando:

```
sudo shell> apt-get install bum
```

Após a instalação do Ubuntu Linux, você pode começar vagabundo do Sistema Administração Gerenciador de Boot-up. Isto proporciona-lhe uma interface gráfica interface para realizar tarefas como iniciar, parar, habilitação e desabilitação init scripts.

NOTA Para Red Hat baseado em distribuições Linux, incluindo o Fedora, o chkconfig utilitário pode ser usado para conseguir algo semelhante. O trecho a seguir irá instalar e permitir que o pentaho-init.sh script:

```
Shell> chkconfig init.sh-pentaho - Adicionar
Shell> chkconfig pentaho-init.sh em
```

Depois de configurar o script de inicialização, você deve reiniciar o computador para verificar que o Pentaho BI Server fato inicia-se como parte da seqüência de inicialização.

Arranque automático em Sistemas Windows

Para o Windows, você deve criar um serviço para habilitar a inicialização automática do Pentaho Server. A maneira mais fácil de fazer isso é usar o service.bat script. Este é distribuída junto com o servidor Tomcat, que é usado para enviar o BI Pentaho Servidor, por isso já está incluído no seu download Pentaho. Para usar esse script, abrir um shell de comando e cd para o diretório home Pentaho, e de lá em tomcat \ bin. Depois, basta executar o seguinte:

```
C: \ Program Files \ pentaho \ ce-biserver \ bin do tomcat \ service.bat> instalar o Pentaho
Instalação 'Pentaho' o serviço ...
Usando CATALINA_HOME: C: \ Program Files \ pentaho \ ce-biserver \ tomcat
Usando CATALINA_BASE: C: \ Program Files \ pentaho \ ce-biserver \ tomcat
Usando JAVA_HOME: D: \ Libraries \ Java \ jdk-1_5_0_15
Usando JVM: D: \ Libraries \ Java \ jdk-1_5_0_15 \ bin \ jre \ server \ jvm.dll
"Pentaho" O serviço foi instalado.
```

NOTA No exemplo anterior, usamos Pentaho como o nome do serviço. Você pode omitir o nome, caso em que o nome do padrão Tomcat5 será usado.

Agora você pode navegar pelo seu serviço usando o Service Manager (Start Controle PanelAdministrative ToolsServices) e configurá-lo para iniciar automaticamente. Note-se que o novo serviço está marcado"Apache Tomcat Pentaho" . Ao invés de simplesmente""Pentaho Desinstalando o serviço é igualmente fácil, basta executar o seguinte comando:

```
C: \ Program Files \ pentaho \ ce-biserver \ bin do tomcat \ service.bat> desinstalar Pentaho
"Pentaho" O serviço foi removido
```

NOTA A service.bat script é realmente um invólucro em torno do Tomcat5.exe programa, e você pode exercer maior controle usando isso diretamente. Você também pode usar esse para modificar o serviço do Tomcat já está instalado. Por exemplo, em vez de navegar para o serviço no gerenciador de serviços você também pode executar o seguinte:

```
tomcat5.exe shell> / / EUA / Pentaho - auto inicialização
```

para modificar o serviço para iniciar automaticamente. O exemplo a seguir ilustra como mudar o nome para exibição do serviço a Pentaho BI Server:

```
tomcat5.exe shell> / / EUA / Pentaho - DisplayName = "Pentaho BI Server"
```

Este último exemplo ilustra como configurar uso de memória do Java Virtual Machine para o serviço. Neste caso, a memória heap inicial é definido como segue:

```
tomcat5.exe shell> / / EUA / tomcat5 - JvmMs = 256M - JvMx = 1024 - JvmSs = 64
```

Você pode encontrar mais informações sobre o service.bat roteiro e tomcat5.exe em a documentação do Tomcat aqui na http://tomcat.apache.org/tomcat-5.5 -doc/windows-service-howto.html.

Gerenciando Drivers de Banco de Dados

Todos os pedidos Pentaho, incluindo o Servidor Pentaho, use Java Database Connectivity (JDBC) para comunicação de dados. A fim de conectar-se a um SGBD particular, Pentaho precisa carregar o driver Java apropriada. Por padrão, o Pentaho Server vem com drivers JDBC para o seguinte bases de dados:

- HSQLDB- hsqldb-x.x.x.jar
- MySQL mysql-connector-java-x.x.x.jar
- PostgreSQL postgresql-x.x-xxx.jdbc3.jar

NOTA As ocorrências de x.x.x e x.x-xxxx na Jar. nomes de arquivo não aparecem, mas indicam um número de versão específico.

Então, se você quiser se conectar a qualquer outro RDBMS, você precisará obter um driver apropriado e assegurar que ele pode ser usado pelo servidor de BI Pentaho. O restante desta seção descreve como fazer isso.

Localização Driver para o servidor

Os drivers JDBC estão localizados na tomcat / common / lib sob o diretório diretório do Servidor Pentaho. Se você precisa se conectar a outro tipo de RDBMS, você deve copiar o adequado Jar. arquivos para este local. O servidor precisa ser reiniciado para que os novos controladores a ser carregado.

Localização Driver para o console de administração

Copiar o motorista Jar. arquivos para o tomcat / common / lib diretório somente permite o servidor para se conectar ao sistema de banco de dados correspondente. No entanto, o PAC é normalmente usado para configurar novas conexões de banco de dados chamado. Assim, em
Para configurar corretamente e testar as conexões de banco de dados, o PAC também deve para carregar o driver JDBC novo.
O driver JDBC Jar. arquivos para o PAC são armazenados no jdbc diretório. Este diretório reside imediatamente abaixo do diretório de instalação do Administração do software do console.

Gerenciando drivers JDBC em sistemas baseados em Unix

Em sistemas baseados em UNIX, você pode utilizar links simbólicos para facilitar a gerenciar drivers JDBC através de vários programas. Com esse método, você pode atualizar um driver com uma única ação.

Para fazer isso, você deve manter todos os seus driver JDBC Jar. arquivos em um único diretório (digamos, / Lib / jdbc). Além disso, este diretório deve conter um link simbólico para cada RDBMS distintos, apontando para a versão preferencial de o Jar. arquivo. Este link simbólico serve para criar um nome de arquivo genérico que ser usado para se referir ao driver Jar. arquivo para esse tipo de RDBMS, independentemente da versão. Por exemplo, um link simbólico chamado mysql-connector-java.jar poderia apontar para um mysql-connector-java-5.1.7.jar ou mysql-connector-java -5.0.8.jar, Dependendo da versão que você prefere para a maioria das aplicações.

Em vez de copiar qualquer Jar. arquivos e diretórios do aplicativo (como / / Tomcat common / lib), Você pode colocar um link simbólico lá, apontando para o genérico link simbólico no / Lib / jdbc diretório. Sempre que você quiser atualizar (ou downgrade) um driver, você pode simplesmente colocar o novo Jar. arquivo na / Lib / jdbc diretório e recriar o link genérico simbólico para apontar para o novo Jar. arquivo.

Uma abordagem um pouco mais simples é diretamente renomear o Jar. arquivo para algo mais genéricos, mas isso torna mais difícil saber exatamente qual a versão você está usando.

Sistema de Bases de Dados

A plataforma Pentaho depende de uma série de bases de dados do sistema:

- hibernar -Esse banco de dados é usado para armazenar a autenticação do usuário e dados de autorização, conteúdos BI (solução de repositório) e de dados chamado fontes.
- quartzo -Esse banco de dados atua como o repositório para o agendador de quartzo, que é um componente que faz com que o Servidor Pentaho.
- SampleData -Esta é a amostra de dados que é usado principalmente pela exemplos discutido no Capítulo 1. Estritamente falando, este não é um banco de dados do sistema porque ela não afeta o funcionamento normal do BIServer Pentaho.

Por padrão, os bancos de dados do sistema são todos administrados por um RDBMS HSQLDB.
Nesta seção, descrevemos como migrar estes a um banco de dados MySQL. Nós suponha que você tenha configurado o software de banco de dados MySQL, e assumimos isso usa a porta padrão do MySQL (3306) e reside na mesma máquina host como o Servidor Pentaho. Para esta configuração particular, todos os MySQL JDBC contato seqüências devem ter esta forma:

```
jdbc: mysql: / / localhost: 3306 / <database-name>
```

Aqui, <database-name> representa um esquema específico (banco de dados), gerido pela instância do MySQL. No restante deste capítulo, iremos usar sempre JDBC conectar strings como este. Você está livre para usar outra máquina, ou outro porto, ou ambos, mas você vai precisar para mudar as seqüências de contato que aparecem no

as seguintes seções conformidade. Claro, o mesmo vale se você quiser
implantar isso em outro RDBMS. Para mais detalhes sobre a configuração do MySQL, consulte
no Capítulo 2.

As etapas para a migração do banco de dados HSQLDB são pré-configurados basicamente o mesmo, independentemente do RDBMS específico. Por padrão, o Pentaho BI Server já oferece alguns recursos (como o esquema de criação de scripts) para configurá-lo para o MySQL, Oracle e PostgreSQL. Isto significa que você terá que ajustar os scripts se no caso de pretender configurar o sistema Pentaho bases de dados para qualquer outro RDBMS.

Configurando o MySQL esquemas

Antes que você possa configurar nada sobre as bases de dados no BI Pentaho final Server, você deve primeiro criar um par de coisas sobre o servidor MySQL que irá substituir o banco de dados HSQLDB. Você pode encontrar os scripts SQL para fazê-lo em o data/mysql5 diretório, que reside em baixo da casa Pentaho BI Server diretório. Os scripts a seguir devem ser executados na seguinte ordem:

- create_repository_mysql.sql -Cria o hibernar banco de dados, que é usado para armazenar a solução de repositório, bem como as credenciais do usuário e permissões.
- create_sample_datasource.sql Adiciona-uma fonte de dados para a amostra dados em que todos os exemplos que vêm com o Pentaho se baseiam. Fonte dos dados definições também são armazenadas no hibernar banco de dados. Por agora, este será ainda apontar para o banco de dados HSQLDB, mas vamos modificá-lo mais tarde, quando também a migração dos dados amostra a MySQL.
- create_quartz_mysql.sql Cria-o repositório para o quartzo planejador.

Você pode usar qualquer ferramenta que você gostaria de executar esses scripts, como o MySQL linha de comando do cliente ou do esquilo. Ao usar a ferramenta de linha de comando mysql, Você pode usar o comando source para executar um script a partir do disco:

```
mysql> FONTE / opt/pentaho/biserver-ce/data/mysql5/create_repository.sql
```

Alternativamente, você pode executá-lo directamente a partir do shell:

```
localhost shell> mysql-h - u root-p \
    > </ Opt/pentaho/biserver-ce/data/mysql5/create_repository.sql
```

NOTA Nota assinar o menos (<). Isso faz com que o conteúdo do arquivo de script para ser executada pelo mysql linha de comando da ferramenta.

Configurando quartzo e Hibernate

Nesta seção, você vai aprender a editar os arquivos de configuração Pentaho, a fim para o quartzo e os componentes do Hibernate para se conectar ao banco de dados MySQL.

Quartzo

O quartzo é um agendador de tarefas. Pentaho utiliza para automatizar tarefas e executar assinatura de conteúdo. lojas de quartzo definições de trabalho em um banco de dados relacional.
Para permitir quartzo para se conectar ao MySQL, você precisa abrir / Tomcat / webapps / pentaho / META-INF / context.xml, Que reside na casa do Servidor Pentaho diretório. Procure a seção onde se lê:

```
<Nome do Recurso = "jdbc / Quartzo" auth = "Container"
   type = "javax.sql.DataSource"
   fábrica = "org.apache.commons.dbcp.BasicDataSourceFactory"
   maxActive = "20" maxidle = "5" maxWait = "10000"
   username = "pentaho_user" password = "senha"
   driverClassName = "org.hsqldb.jdbcDriver"
   url = "jdbc: hsqldb: HSQL: / / localhost / quartzo"
   validationQuery = "
      select count (*)
      de INFORMATION_SCHEMA.SYSTEM_SEQUENCES
   "/>
```

Você precisa alterar os valores das seguintes propriedades:

- driverClassName -O valor dessa propriedade deve ser definida para a classe Java nome do driver JDBC do MySQL, que é com.mysql.jdbc.Driver.
- url -Isso deve ser definido para a cadeia JDBC contato. Ele deve ser alterado para o adequado contato string JDBC do MySQL, que é MySQL:: jdbc / / localhost: 3306/quartz.
- ValidationQuery -Isto é usado para verificar se a conexão pode ser criado. Isso deve ser alterado para SELECT 1.

Após a modificação, o trecho deve ficar assim:

```
<Nome do Recurso = "jdbc / Quartzo" auth = "Container"
   type = "javax.sql.DataSource"
   fábrica = "org.apache.commons.dbcp.BasicDataSourceFactory"
   maxActive = "20" maxidle = "5" maxWait = "10000"
   username = "pentaho_user" password = "senha"
   driverClassName = "com.mysql.jdbc.Driver"
   url = "jdbc: mysql: / / localhost: 3306/quartz"
   validationQuery = "SELECT 1 "/>
```

Hibernar

Hibernate é uma camada de mapeamento objeto-relacional que é utilizado por Pentaho para acesso (e cache) o seguinte:

- Objetos da solução de repositório
- fontes de dados com o nome, que são usadas por elementos tais como relatórios de recolher
  dados de bancos de dados JDBC
- A autenticação do usuário e dados de autorização

Você deve modificar os arquivos de configuração Pentaho relevantes do ponto los para o banco de dados MySQL em vez do banco de dados HSQLDB. Primeiro, modrificar a seção correspondente ao Hibernate / Tomcat / webapps pentaho / / META-INF/context.xml. Na subseção anterior, você já mudou essa arquivo para configurar a conexão para o Quartz. Desta vez, você precisará executar o mesma tarefa para o Hibernate, e alterar esse trecho:

```
<Nome do Recurso = "jdbc / Hibernate" auth = "Container"
   type = "javax.sql.DataSource"
   fábrica = "org.apache.commons.dbcp.BasicDataSourceFactory"
   maxActive = "20" maxidle = "5" maxWait = "10000"
   username = "hibuser" password = "senha"
   driverClassName = "org.hsqldb.jdbcDriver"
   url = "jdbc: hsqldb: HSQL: / / localhost / hibernate"
   validationQuery = "
      SELECT COUNT (*)
      DA INFORMATION_SCHEMA.SYSTEM_SEQUENCES "/>
```

para isso:

```
<Nome do Recurso = "jdbc / Hibernate" auth = "Container"
   type = "javax.sql.DataSource"
   fábrica = "org.apache.commons.dbcp.BasicDataSourceFactory"
   maxActive = "20" maxidle = "5" maxWait = "10000"
   username = "hibuser" password = "senha"
   driverClassName = "com.mysql.jdbc.Driver"
   url = "jdbc: mysql: / / localhost / hibernate"
   validationQuery = "SELECT 1 " />
```

Em seguida, cd no pentaho-solutions/system/hibernate diretório no Pentaho diretório Home. Você precisa modificar dois arquivos aqui:

- hibernate-settings.xml
- mysql5.hibernate.cfg.xml

Primeiro, edite hibernate-settings.xml. Neste arquivo, você encontrará uma linha que lê-se:

```
sistema <arquivo-de-configuração> / hibernate / hsql.hibernate.cfg.xml </ config->
```

O valor no <arquivo-de-configuração> elemento é o nome de um arquivo que também reside nesse diretório. Como você pode ver, ele ainda se refere a um HSQLDB específicas arquivo de configuração. Você precisa mudar este mysql5.hibernate.cfg.xml, que contém a configuração específica do MySQL. Isso é tudo que você precisa mudança na hibernate-settings.xml. Há duas coisas que você pode precisar mudança na mysql5.hibernate.cfg.xml:

- Este arquivo contém a string JDBC contato para se conectar ao hibernar banco de dados, por isso, se você estiver usando um outro host, porta, ou ambos você vai precisar
  ajustar a seqüência de conexão JDBC aqui conformidade.

- Você pode querer adicionar uma ligação robusta pooling. Banco de dados de ligações não são criadas à toa. Pelo contrário, o servidor mantém um conjunto de conexões, que permanecem abertas enquanto o servidor está executando.

A string JDBC de conexão é configurada na linha que lê:

```
<property name="connection.url"> jdbc: mysql: / / localhost: 3306/hibernate
</ Property>
```

(Como acabamos de mencionar, você não precisa mudar isso se você não sabe definir o MySQL, diferentemente do que sugerimos).

Para adicionar um pool de conexão, adicione o seguinte trecho, logo abaixo do <session-factory> tag de abertura:

```
propriedade <  name = "hibernate.c3p0.acquire_increment"> 3> </ property
propriedade <  name = "hibernate.c3p0.idle_test_period"> 14400 </> propriedade
propriedade <  name = "hibernate.c3p0.min_size"> 5> </ property
propriedade <  name = "hibernate.c3p0.max_size"> 75> </ property
propriedade <  name = "hibernate.c3p0.max_statements"> 0> </ property
propriedade <  name = "hibernate.c3p0.timeout"> 25200> </ property
propriedade <  name = "hibernate.c3p0.preferredTestQuery"> selecione um </> propriedade
propriedade <  name = "hibernate.c3p0.testConnectionOnCheckout" <>> true / propriedade
```

Este trecho faz com que o pool de conexão C3P0 para ser usado. Você vai precisar garantir que este é carregado pelo servidor com antecedência. Para fazer isso, você precisa colocar o C3P0-x.x.x.x.jar arquivo na tomcat / common / lib diretório. Além disso, também é preciso colocá-lo no lib sob o diretório home do PAC. (Nota que x.x.x.x. representa o número da versão.) Você pode obtê-lo http://sourceforge.net/projects/c3p0. Você só precisa baixar o C3P0-bin pacote. Os usuários do Windows devem obter o Zip. arquivo, e há uma . Tgz Arquivo para sistemas operacionais baseados em UNIX.

NOTA C3P0 é um pool de aplicação livre compatível com JDBC de conexão. Se você gosta, você pode usar um pool de aplicação alternativa de conexão, como o DBCP, que é fornecido pelo Apache (ver http://commons.apache.org/dbcp/). Para mais informações sobre como usar o DBCP para a configuração do Hibernate Pentaho, leia este artigo no blog de Tom Barber: http://pentahomusings.blogspot.com/2008/05/pentaho-server-fails-every-night-we.html.

Estritamente falando, você não precisa configurar um pool de conexão como esta as coisas, continuará a funcionar na maioria das vezes sem ele. No entanto, isso acontece para resolver um problema que é frequentemente encontrados durante a execução Pentaho em cima do MySQL.

O problema é que o Hibernate sempre tenta manter a conexão com o MySQL aberto. No entanto, as conexões MySQL expiram automaticamente após um determinado período de inatividade. Isso pode ser configurado ao final do MySQL, definindo a valor da wait_timeout servidor variável. Por padrão, conexões inativas termina após 8 horas. Se isto acontecer em um sistema de produção, o Hibernate pára

funcionando corretamente, o que essencialmente requer que você reiniciar o servidor Pentaho antes, torna-se útil novamente. Normalmente, você começa a perceber esse problema em um servidor de produção: durante a noite, é possível que o servidor ociosa, e na manhã seguinte você verá as coisas não funcionam mais. Configurando o pool de conexão como descrito resolve esse problema.

NOTA Para saber mais sobre os problemas específicos que podem surgir durante execução Pentaho em cima do MySQL, na ausência de um pool de conexão, por favor leia esta discussão nos fóruns Pentaho: http://forums.pentaho.org/ t showthread.php? = 54939.
Para obter mais informações sobre o MySQL wait_timeout variáveis, consulte http://dev.mysql.com/doc/refman/5.1/en/server-system -Variables.html # sysvar tempo limite de espera e http://dev.mysql.com/doc/ refman/5.1/en/gone-away.html.

Configurando a segurança JDBC

Você também vai precisar se adaptar a autenticação do usuário e configura-autorização ção para o novo hibernar banco de dados. Para fazer isso, editar applicationContext-Acegi -Segurança jdbc.xml. Este arquivo reside na pentaho-solutions/system dentro do Pentaho BI Server diretório home. Você precisa olhar para o seguinte trecho:

```
feijão <id = "dataSource"
    class = "org.springframework.jdbc.datasource.DriverManagerDataSource">
    <property name="driverClassName" "value="org.hsqldb.jdbcDriver" />
    <Nome da propriedade = "url"
    value = "jdbc: hsqldb: HSQL: / / localhost: 9001/hibernate" />
    <property name="username" value="hibuser" />
    <property name="password" value="password" />
</ Bean>
```

Modificar esta para coincidir com o banco de dados MySQL, assim:

```
feijão <id = "dataSource"
    class = "org.springframework.jdbc.datasource.DriverManagerDataSource">
    <property name="driverClassName" value="com.mysql.jdbc.Driver" />
    <property name="url" value="jdbc:mysql//localhost:3306/hibernate" />
    <property name="username" value="hibuser" />
    <property name="password" value="password" />
</ Bean>
```

No mesmo diretório um arquivo chamado applicationContext-segurança Acegi -Hibernate.properties. Seus conteúdos são os seguintes:

```
org.hsqldb.jdbcDriver jdbc.driver =
jdbc.url = jdbc: hsqldb: HSQL: / / localhost: 9001/hibernate
hibuser jdbc.username =
password = jdbc.password
org.hibernate.dialect.HSQLDialect hibernate.dialect =
```

Você precisará editar esta e ajustar as propriedades de banco de dados para coincidir com o MySQL hibernar banco de dados, como segue:

```
com.mysql.jdbc.Driver jdbc.driver =
jdbc.url = jdbc: mysql: / / Localhost: 3306/hibernate
hibuser jdbc.username =
password = jdbc.password
org.hibernate.dialect.MySQLDialect hibernate.dialect =
```

Dados da Amostra

Se você também gostaria de mover a amostra de dados de HSQLDB ao MySQL, você deve primeiro fazer o download de um script para carregar os dados de exemplo para MySQL. Este script é gentilmente cedidas por Prashant Raju, e você pode baixá-lo www. prashantraju.com/pentaho/downloads/sampledatamysql5.sql.

NOTA Prashant Raju também oferece bons guias para configurar Pentaho para o MySQL. Você pode encontrá-los aqui: http://www.prashantraju.com/pentaho/guides/biserver-2.0-final/.

Após a configuração do banco de dados, você ainda precisa de actualizar o SampleData definição da fonte de dados que é armazenado no hibernar banco de dados. Mais adiante neste capítulo, discutiremos como editar fontes de dados usando a Administração Pentaho Console. Por agora, vamos usar um método um pouco mais direto, e atualizar diretamente o registro do banco de dados que armazena a definição da fonte de dados:

```
UPDATE hibernate.DATASOURCE
SETDRIVERCLASS = "com.mysql.jdbc.Driver",
    URL = "jdbc: mysql: / / localhost: 3306/sampledata,
    Query = 'SELECT 1'
WHERE NAME = 'SampleData'
;
```

Modificar o Pentaho Scripts de inicialização

Se você trabalhou com as subseções anteriores, você pode descartar o HSQLDB banco de dados completo. O Pentaho scripts de inicialização e desligamento conter uma linha explicitamente iniciar e parar o banco de dados HSQLDB, respectivamente. Você deve remover essas linhas. Ele poupa alguma memória, e também oferece uma boa teste para ver se você mudou corretamente todos os bancos de dados MySQL.

Aqui está um resumo dos roteiros e as linhas a remover:

- start-pentaho.bat:

```
início start_.bat
```

- parar pentaho.bat:

```
início stop_.bat
```

- start-pentaho.sh:

```
start_.sh & sh
```

- parar pentaho.sh:

```
início stop_.bat
```

NOTA Ao invés de remover as linhas, você pode ativá-los para comentar linhas, o que torna mais fácil de desfazer a alteração mais tarde. Para o . Morcego scripts, você criar uma linha de comentário, definindo a palavra-chave REM seguido por um caractere de espaço logo no início da linha. Para o Sh. scripts, você pode definir o cardinal (#) Logo no início da linha.

E-mail

O Pentaho BI Server tem SMTP (Simple Mail Transfer Protocol) e-mails capacidades. E-mail pode ser usado para distribuir conteúdo de BI (tais como relatórios) para os recipientes apropriados em cenários de ruptura, ou para o envio de monitoramento mensagens. Especificamente, Pentaho utiliza a API JavaMail para atuar como um cliente SMTP enviar e-mails através de um servidor SMTP existente. Observe que você precisa ter um executando o servidor STMP antes que você possa usar este Pentaho, não implementa um servidor de email em si.

E-mail não irá funcionar out-of-the-box na configuração padrão. A fim usar e-mail, você precisa configurar algumas coisas. A configuração de e-mail é controlada através do arquivo email_config.xml, Que reside no smtp-mail diretório dentro do sistema Pentaho solução.

Configuração básica de SMTP

Listagem 3-3 mostra o conteúdo de uma base email_config.xml arquivo.

Listagem 3-3: O conteúdo do email_config.xml

```
<email-smtp>
  <properties>
    <mail.smtp.host> smtp.wcm.com </ mail.smtp.host>
    <mail.smtp.port> 25 </ mail.smtp.port>
    <mail.transport.protocol> smtp </ mail.transport.protocol>
    <mail.smtp.starttls.enable> false </ mail.smtp.starttls.enable>
    <mail.smtp.auth> true </ mail.smtp.auth>
    <mail.smtp.ssl> false </ mail.smtp.ssl>
    <mail.smtp.quitwait> false </ mail.smtp.quitwait>
  </ Properties>
```

```
  <mail.from.default> joe.pentaho @ pentaho.org </ mail.from.default>
  joe.pentaho <mail.userid> @ gmail.com </ mail.userid>
  <mail.password> senha </ mail.password>
</ Smtp-mail>
```

Como você pode ver na Listagem 3-3, há uma propriedades seção que contém a configuração para comunicação com o servidor SMTP. O mais importante propriedades são as seguintes:

- mail.smtp.host -O nome do host ou endereço IP onde o SMTP servidor está executando.
- mail.smtp.port -A porta onde o servidor SMTP está escutando. A porta padrão SMTP é 25.
- mail.transport.protocol -O protocolo usado para se comunicar com o SMTP servidor. Por padrão, esta é smtp.
- mail.smtp.starttls.enable -Por padrão, false. Se verdadeiro, o STARTTLS comando é usado para mudar para uma conexão segura TLS protegido.
- mail.smtp.auth -Por padrão, false. Se for verdade, o comando AUTH será usado para autenticar o usuário. Muitos servidores SMTP exigem autenticação, por isso deve ser normalmente configurado para true.
- mail.smtp.ssl -Por padrão, false. Se for verdade, uma tomada de seguro é utilizado para comunicar com o servidor.
- mail.smtp.quitwait -Por padrão, é verdade, o que significa que o cliente irá esperar para uma resposta à QUIT comando. Se false, a conexão é fechada imediatamente após o QUIT comando.

Fora da propriedades seção, há uma configuração de alguns parâmetros que são usados para autenticar o pedido de SMTP:

- mail.from.default -Padrão O e-mail do remetente. O SMTP protocolo exige que o remetente deve ser especificado, e este endereço de e-mail ser utilizado se nenhum endereço é especificado explicitamente aquando do envio dos e-mail.
- mail.userid e mail.password -As credenciais do remetente. Esta é necessária quando o servidor SMTP requer autenticação (que é o caso quando o mail.smtp.auth propriedade é verdadeira). Normalmente, o remetente do e-mail e-mail e credenciais são associados e os servidores SMTP requerer o endereço do remetente de e-mail para corresponder ao usuário identificado

pelas credenciais. Embora por padrão o protocolo SMTP não
exigir autenticação, na prática, quase todos os servidores SMTP estão configurados para usá-lo.

NOTA Você pode encontrar mais informações sobre JavaMail e sua configuração propriedades na documentação da API Java em http://java.sun.com/products/ javamail / javadoc / com / sun / mail / SMTP / pacote summary.html.

Secure Configuration SMTP

Mais e mais freqüentemente, servidores de email exigem que você use uma comunicação segura
protocolo. Um exemplo bem conhecido é o Google Gmail. Para enviar e-mail usando
como um servidor de email, você vai precisar de uma configuração de e-mail um pouco diferente:

- mail.smtp.port A porta-padrão para SMTP seguro é 465. Às vezes 587 é utilizado. Contacte o seu administrador para obter o número da porta apropriado.
- mail.transport.protocol -Isso deve ser smtps ao invés de smtp.
- mail.smtp.starttls.enable -Você pode precisar definir este como verdadeiro, em vez do que falsa.

Teste de E-mail Configuração

Para testar sua configuração de e-mail, você pode usar o Burst Relatório de Vendas, que reside na seção Reportagem da solução de BI Developer exemplos. Você
pode encontrar indicações sobre como trabalhar com os exemplos pré-definidos em Pentaho Capítulo 1.

Editora Senha

ferramentas de projeto Pentaho são usados para criar definições de conteúdo de BI, como relatórios, cubos OLAP, e metadados. O BI arquivos de conteúdo que são criados por essas ferramentas podem ser implantados manualmente, copiando os arquivos diretamente para o
Diretório solução adequada no sistema de arquivos do host do Servidor Pentaho.
No entanto, a forma típica e preferenciais para implantar conteúdo BI é através de um processo chamado publicação.

Para publicar, as ferramentas de design recorrer a um serviço Web implementados pelo Pentaho Server, que autentica o usuário, bem como verificar a sua
permissões. Quando isso for bem-sucedida, a ferramenta cliente envia os dados de conteúdo para
o servidor, que armazena-la em uma posição desejada em algum local dentro do solução de repositório.

Para habilitar a publicação, você primeiro tem que definir explicitamente a senha do editor. Esta senha deve ser fornecida para o web service, além do usuário
credenciais ao publicar conteúdo de BI. Há uma senha para o editor

todo o servidor, e é configurado no publisher_config.xml arquivo, que reside no pentaho-solutions/system sob o diretório home Pentaho diretório. O conteúdo do arquivo de configuração são mostrados aqui:

```
<publisher-config>
   <publisher-password> publicar </ senha editor->
</ Editor-config>
```

No exemplo anterior, a senha é definida para publicar.

NOTA Por padrão, nenhuma senha é especificada, o que impede a concepção ferramentas de publicar qualquer conteúdo para o servidor.

Tarefas administrativas

Nesta seção, descrevemos como executar tarefas administrativas comuns utilizando o Pentaho Console Administrativo (PAC).

A Administração Pentaho Console

O software PAC é enviado no mesmo pacote como o Pentaho BI Server. Mencionamos antes que ele reside no administração console diretório.
PAC é implementado como um servidor Web leve baseado no Jetty. Tecnicamente, não há razão para que PAC não poderia também funcionar dentro do servidor Tomcat na qual o Pentaho BI Server é baseado, só que agora é possível facilmente separar os recursos administrativos a partir do aplicativo de BI. Por exemplo, você pode facilmente executar PAC em um servidor fisicamente distintos, o que pode torná-lo mais fácil de gerenciar a segurança.

NOTA Jetty é um servidor web e Java servlet container, assim como o Apache Tomcat. A diferença importante é que o Jetty fornece um mínimo muito leve implementação que o torna especialmente adequado para inseri-la. Jetty é também utilizado por Pentaho Data Integration para implementar clustering.You pode encontrar mais sobre o projeto no Jetty http://www.mortbay.org/jetty/.

Configuração básica do PAC

Antes de usar o PAC, você pode precisar configurar algumas coisas. Abra o console.xml arquivo localizado na recurso / config diretório abaixo do PAC home. Seus conteúdos são os seguintes:

```
<? Xml version = "1.0" encoding = "UTF-8"?>
<console>
  <solution-path> </ solução de caminho>
```

```
<war-path> </ guerra caminho>
<platform-username> joe </ username plataforma>
<biserver-status-check-period-millis>
    30000
</ Biserver-status check-período millis>
<homepage-url> http://www.pentaho.com/console_home </ url-homepage>
<homepage-timeout-millis> 15000 </ homepage-timeout millis>
<! - Lista separada por vírgulas de papéis (sem espaços) ->
<default-roles> autenticados </ default-papéis>
</ Console>
```

Você precisa modificar o <solution-path> e <war-path> elementos dentro o <console> elemento a apontar para o local da solução de repositório ea aplicação web Pentaho, respectivamente. Você pode usar caminhos relativos, assumindo assim uma posição padrão como descrito na secção de instalação deste capítulo, esses elementos devem ler-se:

```
<solution-path> .. / biserver-ce / pentaho soluções </ caminho-solução>
<war-path> .. / biserver-ce / tomcat / webapps / pentaho </ caminho-guerra>
```

Iniciando e parando PAC

Iniciar e parar scripts estão localizados diretamente no interior do administração console diretório. Os usuários do Windows podem começar pela execução do PAC startup.bat; Usuários de sistemas baseados em UNIX deve usar start.sh. Da mesma forma, os usuários do Windows podem uso stop.bat PAC para parar, enquanto stop.sh deve ser usado em UNIX sistemas.

O Front End PAC

A extremidade dianteira do PAC é uma página web. Você pode acessá-lo com qualquer ambiente moderno JavaScript habilitado para navegador. Por padrão, o PAC atende a pedidos no porto 8099. Por exemplo, durante a execução do PAC na máquina local, você pode acessar o console, navegue até http://localhost:8099/.

Ao navegar na página, você é primeiro solicitado suas credenciais. O nome de usuário padrão é adminEa senha padrão é senha. Depois o login, você verá uma página como a mostrada na Figura 3-1.

A home page do PAC oferece pouco mais do que alguns aparentemente estático textual informações sobre o Pentaho Enterprise Edition. No entanto, as informações na home page é transferido ao vivo pela internet, para que ele possa ser usado para show up-to-date informações.

Observe o grande botão verde na Administração do lado esquerdo do PAC home page. Clicando que lhe dá acesso ao real console administrativo.

Figura 3-1: O Pentaho Administração home page Console

Configurando a segurança do PAC e Poderes

Jetty apresenta sua própria autenticação plugáveis. Este é separado do Plataforma Pentaho sistema de segurança. PAC da segurança é configurado através do login.conf arquivo de configuração. Este arquivo define o tipo de segurança especificando o chamado módulo de login. Jetty navios com um número de padrão módulos de login, como as propriedades de um arquivo do módulo de login ou um login JDBC módulo.

O conteúdo padrão do login.conf arquivo são os seguintes:

```
PropertiesFileLoginModule {
   org.mortbay.jetty.plus.jaas.spi.PropertyFileLoginModule necessária
   debug = "true"
   file = "recursos / config / login.properties";
};
```

Como você pode ver, o PAC utiliza as propriedades do arquivo de módulo de login padrão (que é implementado pelo org.mortbay.jetty.plus.jaas.spi.PropertyFile LoginModule Java classe). Com este tipo de autenticação, os nomes de usuários e as senhas são armazenadas em um arquivo de propriedades. Neste caso, o arquivo é recurso /

config / login.properties. O caminho é relativo ao repouso e ao PAC
conteúdo desse arquivo é mostrado aqui:

```
admin: 1v2j1uum1xtv1zej1zer1xtn1uvk1v1v, administrador do servidor,: OBF
administrador de conteúdo, admin
```

Se você quiser adicionar novas entradas aqui, ou alterar a senha padrão, você pode usar o org.mortbay.jetty.security.Password classe, que é parte do Jetty. Você pode usar isso a partir da linha de comando, como segue:

```
Shell> cd / opt / pentaho / administração console / lib /
shell jetty.jar java-cp>; org.mortbay.jetty.security.Password molhe-util.jar
java org.mortbay.jetty.security.Password [<user>] <senha> - Uso
Se a senha for?, O usuário será solicitado a senha
```

NOTA Na realidade, o Jar. nomes de arquivos incluem um número de versão, por exemplo, molhe-6.1.2.jar e molhe-util-6.1.9.jar. Por favor, olhe em sua recurso / lib directório para descobrir quais os números de versão se aplicam à sua distribuição.

Então, se você deseja alterar a senha para o usuário administrador do segredo'',"você pode fazer o seguinte:

```
Shell>    java \
   >      Cp-molhe-6.1.2.jar: molhe-util-6.1.9.jar \
   >      org.mortbay.jetty.security.Password \
   >      segredo
```

Este comando gera o seguinte resultado:

```
segredo
OBF: 1yta1t331v8w1v9q1t331ytc
MD5: 5ebe2294ecd0e0f08eab7690d2a6ee69
```

Agora você pode modificar o recurso / config / login.properties arquivo e alterar a ocorrência de OBF: 1v2j1uum1xtv1zej1zer1xtn1uvk1v1v para OBF: 1a ta1t331v8w1v9q1t331ytc.

NOTA Para obter mais informações sobre a APA de autenticação conectável, por favor consulte o Pentaho documentação. Você pode encontrá-lo em http://wiki.pentaho.com/ display/ServerDoc2x/Configuring + Segurança + com + Pentaho + Administração + Console.

Gerenciamento de Usuário

Por padrão, o Pentaho BI Server usa um sistema de autorização simples, consistindo de usuários, papéis e permissões que são armazenados em um banco de dados relacional. PAC
permite a criação de funções e usuários, e faça as associações de utilizadores / papel. A permissão real pode ser controlado a partir da utilização Pentaho Console (isto é, a partir do final do Pentaho BI Server). Aqui, o usuário atual pode conceder ou revogar permissões para itens da solução de repositório para usuários individuais ou a

funções (conferindo, assim, a permissão para que todos os usuários que essa particular papel é atribuído).

O gerenciamento de usuários do console é mostrado na Figura 3-2. Ele pode ser invocado por
clicando na aba Usuários e Roles no topo do console de gerenciamento.

Figura 3-2: O gerenciamento de usuários do console

Se você quiser adicionar um novo usuário, não se esqueça de ativar o modo de usuário, clicando em
o botão da barra de ferramentas do Usuário. No lado esquerdo do console, você verá a lista de
usuários existentes. Clique na +botão na parte superior direita da lista para abrir um diálogo onde você pode digitar os dados que descrevem o novo usuário (veja a Figura 3-2).

Você pode clicar em qualquer entrada na lista de usuário para selecionar o usuário. Ao clicar no
botão x pequeno na parte superior direita da lista de usuários irá remover o usuário selecionado.
Alternativamente, você pode atualizar os dados do usuário selecionado na forma a
lado esquerdo do console de gerenciamento de usuários. No fundo desta forma, há uma lista que mostra os papéis atribuídos ao usuário.

Você pode chamar o usuário atribuições / função de diálogo clicando no +botão
na parte superior direita da lista de atribuições. Este diálogo é mostrado na Figura 3-3. Você pode adicionar tarefas, selecionando um ou mais papéis na lista Disponível em
no lado esquerdo da janela, e clicando no >botão. Da mesma forma, você pode
revogar papéis, selecionando-os na lista e clicando no Assigned <botão.

Use o botão Atualizar na parte inferior do formulário para confirmar quaisquer alterações você fez com os dados do usuário ou as atribuições de função.

Se você quiser criar uma nova função, clique no botão da barra de funções. Isso proporciona uma interface semelhante ao mostrado na Figura 3-2, mas do outro lado: Em
Do lado esquerdo, este ecrã oferece uma lista de funções em vez de uma lista de usuários e, em vez

de uma lista de funções atribuídas, ele fornece uma lista de usuários que foram atribuídos a selecionados papel.

Figura 3-3: Atribuir funções ao usuário selecionado

NOTA A plataforma Pentaho não se limita à autenticação padrão e sistema de autorização armazenadas em um banco de dados relacional. A plataforma Pentaho torneiras no sistema Spring Security (anteriormente conhecido como Acegi). Esta é uma forma flexível sistema de segurança que fornece acesso a uma vasta gama de autenticação e autorização de back-ends, como o LDAP.

Para obter mais informações sobre o Spring Security, veja: www.springsource.com/produtos / springsecurity. Para obter mais informações sobre como aplicar Primavera conceitos de segurança a Pentaho, consulte a documentação do Pentaho em http://wiki.pentaho.com/display/ServerDoc2x/Security.

As fontes de dados

PAC permite criar e editar fontes de dados chamado JDBC (dados JNDI fontes). Você pode acessar a fonte de dados clicando no console de fontes de dados guia no topo do console de gerenciamento. A gestão de Fontes de Dados console é mostrado na Figura 3-4.

A lista de fontes de dados disponíveis no lado esquerdo do console. Se você selecionar um item na lista, você pode alterar suas propriedades no formulário à direita. Dados fontes têm as seguintes propriedades:

- Nome-O nome do JNDI para o fonte de dados. Isso pode ser usado na relatórios e outros conteúdos de BI para se referir a este respeito.
- Driver de classe-nome da classe Java que implementa o JDBC motorista. Para o MySQL, você pode usar com.mysql.jdbc.Driver.
- Nome do utilizador-O nome do usuário do banco.
- Senha de senha do usuário do banco de dados.
- URL A seqüência de conexão. O formato da seqüência de conexão é dependente sobre o condutor, e você deve consultar a documentação do condutor descobrir qual o formato a utilizar. Para o MySQL, o formato é:

jdbc: mysql: / / <hostname> [: <porta>] / <schema_name>

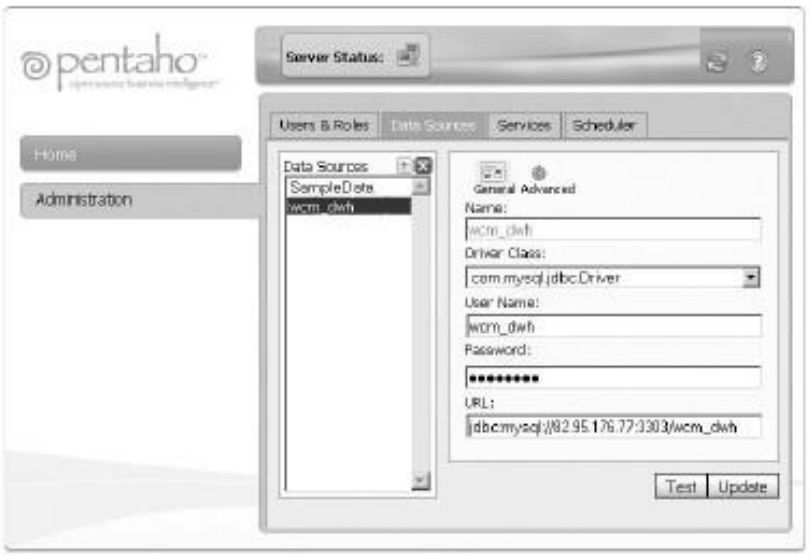


Figura 3-4: A gestão da fonte de dados do console

Ao usar um banco de dados local chamado wcm_dwh que escuta na porta padrão, a URL deve ser:

jdbc: mysql: / / localhost: 3306/wcm_dwh

Quando terminar de modificar as propriedades, você pode usar o botão de teste para confirmar a conexão está funcionando corretamente. Use o botão Update para armazenar as modificações.

Você pode adicionar uma nova fonte de dados clicando no +botão que aparece na superior direito da lista de fontes de dados. Isso abre uma janela onde você pode editar o fonte de dados de propriedades. Afirmar o diálogo irá adicionar a fonte de dados para o lista do lado esquerdo do console.

NOTA Para obter mais informações sobre como criar fontes de dados com o PAC, consulte o Pentaho documentação em http://wiki.pentaho.com/display/ServerDoc2x/.04 Configurando + + + Data Sources.

Outras tarefas administrativas

PAC também permite que você gerenciar agendas e assinaturas. Este tópico é explorados em detalhe no Capítulo 14.

Resumo

Neste capítulo, você aprendeu algumas noções básicas sobre a instalação do Servidor Pentaho,
, configuração e administração, incluindo as seguintes:

- O diretório de instalação preferida
- Configurando uma conta de usuário separada para executar o servidor Pentaho

- Habilitar início automático de e desligamento do Pentaho BI Server
- Permitindo a conectividade de banco de dados de um RDBMS de escolha, acrescentando novos drivers JDBC
- Gerenciando drivers JDBC em sistemas baseados em UNIX usando links simbólicos
- Configurando o sistema de bancos de dados MySQL em Pentaho
- Configurando um pool de conexão C3P0 para o Hibernate
- Habilitando a plataforma Pentaho para enviar e-mail com um existente SMTP servidor.
- Configurando a senha do Publisher
- Configurando e iniciando o Pentaho Administration Console (PAC)
- Gerenciando a autenticação básica PAC
- Gerenciando Pentaho BI Server usuários e os papéis com o PAC
- Gerenciando fontes de dados com o PAC

CAPÍTULO

4

O BI Pentaho Stack

Pentaho é uma suite de business intelligence, em vez de um único produto: ele é feito por um conjunto de programas de computador que trabalham juntos para criar e oferecer soluções de business intelligence. Alguns destes componentes fornecem funcionalidades que são muito básicos, como autenticação de usuário ou banco de dados gerenciamento de conexão. Outros componentes fornecem funcionalidades que opera a um nível superior, como a visualização de dados por meio de tabelas e gráficos.

Frequentemente, mas nem sempre, os componentes que oferecem funcionalidades de alto nível confiar em outros componentes que oferecem funcionalidade de baixo nível. Como tal, a coleção de programas que formam o pacote completo pode literalmente ser visto como um pilha de componentes, cada nível trazendo mais funcionalidade para o usuário final. A Pentaho BI pilha é mostrado na Figura 4-1, onde todos os componentes que fazem até a solução completa são mostrados.

Neste capítulo, descrevemos os diferentes componentes, suas funções e, se for caso disso, as relações que existem entre eles. Na Figura 4-1, o principais camadas da pilha estão claramente identificados, com a camada de apresentação na o topo ea camada de dados e integração de aplicativos na parte inferior. A maioria das finais os usuários irão interagir com a camada de apresentação, que pode assumir muitas formas. Pentaho pode ser acessado por um navegador web simples, mas os componentes também podem ser incorporado em um portal existente, tais como o Liferay ou um gerenciamento de conteúdo sistema, tais como Alfresco. Talvez a forma mais comum de apresentação é Pentaho envio de conteúdo como um arquivo PDF para Caixa de Entrada de um usuário via e-mail.

As principais áreas funcionais da pilha de relatórios de BI, análise, dashboards e gerenciamento de processos, constituem a camada do meio da pilha, enquanto a plataforma de BI em si oferece recursos básicos para a segurança e administração. A integração de dados completa da pilha e é necessário para obter dados de várias sistemas de origem para um ambiente compartilhado de data warehouse.

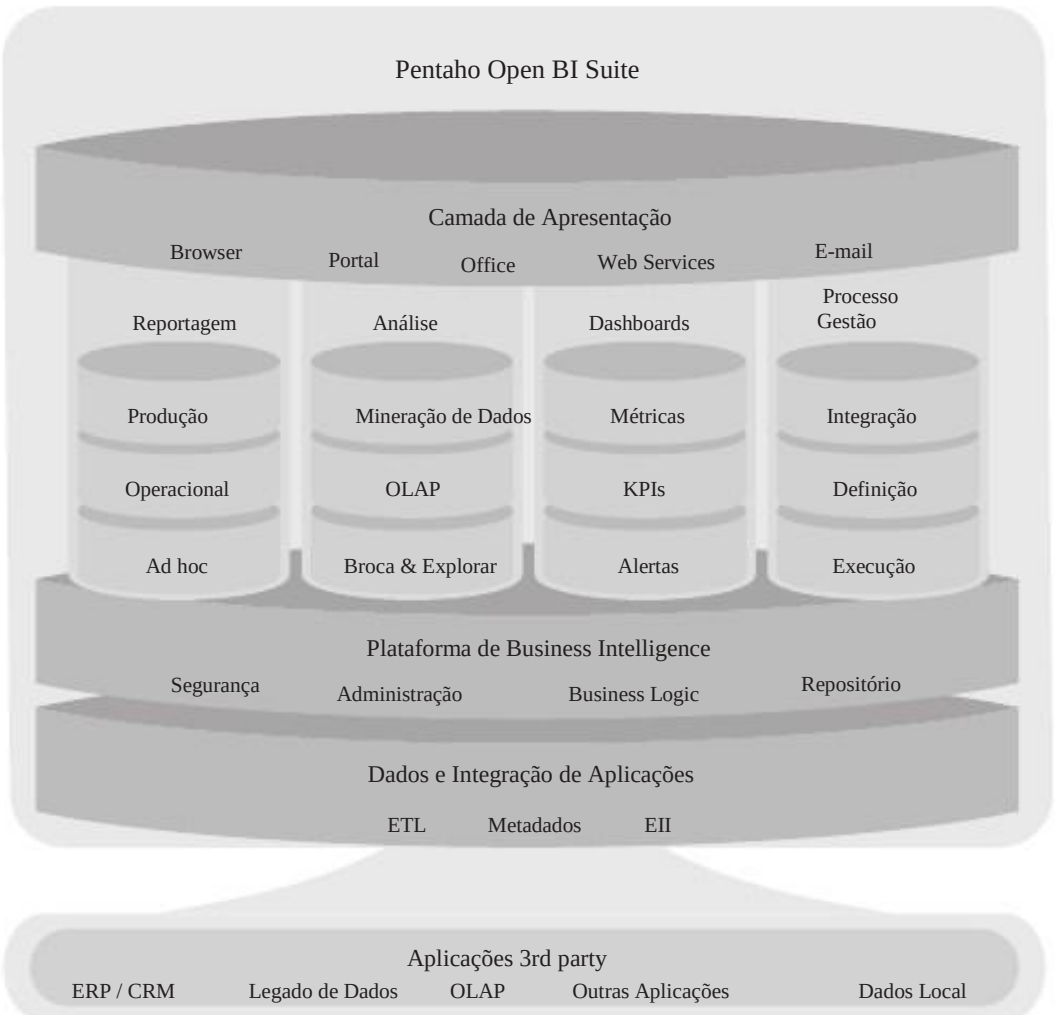


Figura 4-1: Pentaho BI pilha

Em certo sentido, o diagrama não mostra apenas o BI Pentaho pilha a um nível elevado, mas a arquitetura de BI Pentaho também. Uma arquitetura define a estrutura eo esboço de uma solução, mas não exatamente prescrever como a estrutura deve ser construída. No caso da Pentaho, a arquitetura define as camadas e blocos de construção, mas não necessariamente força a usar tudo, desde a pilha, ou para dar um passo além, de Pentaho. Apesar de existirem várias vantagens em usar software Pentaho para construir a pilha, você está livre para misturar em outros componentes também. Por exemplo, Pentaho possui duas formas de criação de relatórios, o Ad Hoc web-based de Consulta e componente de relatório e o Pentaho Report Designer, mas a plataforma pode ser executado tanto Jasper e BIRT relatórios. Mesmo no mundo Pentaho, uma infinidade de alternativas disponíveis, na maior parte iniciadas por projetos comunitários e, posteriormente, aprovada pelo
da empresa. Nós vamos cobrir todas as opções disponíveis mas você deve ser ciente do fato de que novos projetos são iniciados regularmente, às vezes, cobrindo uma

faltando parte da funcionalidade Pentaho, às vezes substituindo ou aumentando uma parte existente da pilha. O Pentaho BI pilha é, portanto, uma evolução entidade, como uma cidade onde os edifícios são criadas novas e as antigas são restauradas expandida, ou substituídos em uma base contínua.

Pentaho BI Stack Perspectivas

Podemos classificar os programas que compõem o Pentaho BI Suite de acordo com uma série de critérios. As seções seguintes oferecem diferentes lentes para a visão os componentes da pilha.

Funcionalidade

Uma maneira de classificar os componentes do Pentaho pilha é de funcionalidade. Por funcionalidade entendemos a tarefa ou as tarefas que um determinado programa foi projetado para executar. Do ponto de vista do usuário, a funcionalidade é o que define o objectivo do programa.

Alguns dos componentes do Pentaho BI oferecem funcionalidades típicas, como ETL, Relatórios e OLAP. Os componentes que fornecem essas funções típicas de BI condicionamento sociocultural são apoiadas por uma série de componentes que oferecem funcionalidade
a um nível consideravelmente mais baixo. Juntos, esses componentes de apoio são conhecida como a plataforma Pentaho.

As funcionalidades oferecidas pela plataforma de BI não são específicos, mas oferecer uma infra-estrutura de software básico. Tarefas como a autenticação de usuário e autori- do o cateterismo, banco de dados de gerenciamento de pool de conexão, ea execução do previsto
tarefas fazem parte da plataforma.

Programas de servidor, o Web Client e Desktop

Outro critério é se o programa pode ser classificado como um cliente, um servidor, ou um programa desktop. Alguns programas são facilmente reconhecidos Pentaho como programas de servidor. Estes são tipicamente executadas em um computador central que
é acessado por clientes baseado na Web através de uma rede (intranet ou Internet). Non-servidor programas da suíte Pentaho pode ser melhor classificado como desktop programas. Estes são tipicamente instalados no computador local do usuário. Estes programas de desktop, na maioria dos casos, ser usado por designers e desenvolvedores pedidos de emissão de um programa servidor (Pentaho), ao qual estão ligados. A bom exemplo da divisão entre o servidor, desktop e cliente é o design, publicação e execução de um relatório. O Report Designer desktop é usado para criar o relatório, que é publicado ao Servidor Pentaho. O Pentaho Server pode executar o relatório sobre a solicitação usando o built-in Report Engine e saída é exibida através do Pentaho Web Portal, que serve como o cliente.

Front-ends e back-ends

Outra forma de distinguir os programas é front-end versus back-end.
programas de front-end são os programas que fornecem uma interface amigável
que permite aos usuários humano para interagir com o programa. programas de back-end
normalmente não são projetados para suportar a interação humana direta. Ao contrário, eles
receber comandos de alguma ferramenta de front-end que sabe como traduzir a
ações do usuário em comandos em que o back-end pode operar a
executar a tarefa real.

O aspecto front-end/back-end está relacionada, mas distinta da diferença
entre servidores, desktops e programas cliente. Apesar de desktop e do cliente
programas são programas também front-end, é possível implementar um
front-end como um programa de servidor. Também é possível implementar o back-end
como um programa desktop.

Há também uma relação entre o aspecto ea funcionalidade
front-end/back-end aspecto: a maioria dos componentes da plataforma não tem
qualquer front-end em tudo. Em contrapartida, os programas que oferecem funcionalidades de BI específica
geralmente têm um claramente distinguíveis front-end.

Subjacente Tecnologia

Praticamente todos os programas no Pentaho pilha são programados em Java
linguagem de programação. Alguns dos componentes do lado do servidor implementar web
aplicações baseadas na tecnologia AJAX, mas isso ainda é alcançado por um aplicativo Java
programa.

Da perspectiva do usuário final, as linguagens de programação e
tecnologia sobre a qual Pentaho é construída são completamente irrelevantes. No entanto, é
extremamente importante do ponto de um administrador de sistemas ou desenvolvedor de
vista.

Uma das propriedades mais notáveis da plataforma Java é que os programas Java
são extremamente portáveis entre arquiteturas de hardware e sistemas operacionais.
Como conseqüência, Pentaho está disponível para muitos sistemas operacionais diferentes.

Do ponto de um administrador de sistemas de visão, a gestão é muito Pentaho
como gerenciar outros aplicativos Java. Há algumas coisas sobre Java
pode exigir algum esforço de um administrador de sistema:

- As aplicações Java são normalmente projetados para serem compatíveis com um determinado
  versão minimalista da plataforma Java, e às vezes um grande específicas
  versão.
- Outro problema pode ser a aplicação particular da plataforma Java.
  As especificações do Java têm sido sempre bastante aberta, e há muitos
  diferentes implementações de ambos a máquina virtual Java, bem como a
  padrão de bibliotecas de classe. programas Java funcionam geralmente bem quando usar
  Sun implementações de referência.

- Os programas Java são executados por uma máquina virtual, que geralmente é um programa nativo. Às vezes é necessário ajustar a máquina virtual parâmetros para efetivamente executar o programa.

O servidor Pentaho Business Intelligence

O servidor Pentaho é uma coleção de programas que trabalham juntos para fornecer uma série de funções essenciais do Pentaho BI Suite. Estes programas são implementado como Java servlets. Servlets não operar de forma autônoma, mas são executados
dentro de uma chamada servlet container, que é em si um servidor HTTP (a web servidor), ou parte dela.

Normalmente, o servlet container é executado em um computador remoto centralizado, onde ele responde a pedidos de programas clientes que estão conectados ao servidor através de uma rede. O Java Servlet Technology é discutido em mais detalhe mais adiante neste capítulo.

Em um nível funcional, o servidor Pentaho pode ser dividida em três camadas:

- A plataforma
- componentes de BI
- A camada de apresentação

A Plataforma

A coleção de componentes conhecidos coletivamente como a plataforma oferece o seguintes serviços:

- Solução motor de repositório e solução
- Banco de dados de gerenciamento do pool de conexão
- A autenticação do usuário e os serviços de autorização
- Exploração madeireira e de serviços de auditoria
- Agendamento de tarefas-
- serviços de e-mail

A funcionalidade oferecida por esses serviços é relativamente de baixo nível e constitui a infra-estrutura básica da plataforma de BI. Um certo número de com-ponentes nesta camada podem aparecer igualmente bem em outros tipos de aplicação servidores.

No restante desta seção, descrevemos brevemente a função de cada componente.

A solução de repositório e do Mecanismo de Solução

A plataforma Pentaho BI organiza o conteúdo em soluções de chamada. A Pentaho solução pode ser pensado como uma pasta de sistema de arquivo com todo o conteúdo de BI para resolver algum problema de negócios. Uma solução Pentaho pode conter pastas e itens chamada seqüências de ação.

As pastas servem apenas para fornecer uma organização geral do conteúdo de BI. As pastas podem conter outras pastas e seqüências de ação (AS). Seqüências de ação são serviços que podem ser invocadas para fornecer alguns conteúdos BI. Eles podem ser chamado diretamente através da interação do usuário, ou tratada como um serviço web a partir de outro aplicativo. A última propriedade permite a integração de Pentaho com outras aplicações

seqüências de ação podem conter várias etapas, algumas vezes chamado def-ação inições. Na forma mais simples, uma sequência de acção contém apenas um passo, para exemplo, para executar um relatório. A seqüência de ação um pouco mais avançada pode consistem de um passo para levar um usuário para a entrada, e uma segunda etapa de execução um relatório, com a entrada do primeiro degrau, como valores de parâmetro. Ao adicionar mais etapas, seqüências de ação avançados podem ser construídas, por exemplo, executar um consulta de banco de dados para localizar todos os armazéns que estão com pouco estoque, loop sobre os depósitos encontrados para executar um relatório de detalhes da fotografia, e distribuir o saída de relatório via e-mail aos gestores de armazém em causa.

seqüências de ação são representados usando XML e são armazenadas em texto simples arquivos com extensão xaction.. seqüências de ação são, portanto, também chamado xactions, após a extensão do arquivo. Em princípio, você pode criá-los com um editor de texto simples. Algumas das ferramentas Pentaho front-end, tais como o Relatório Designer, pode gerar simples, seqüências de ação passo a passo. Mais avançado seqüências de ação são as melhores criadas usando Pentaho Design Studio, ou utilizando Eclipse com a seqüência de ação Pentaho plugin. Estes fornecem uma interface gráfica editor de seqüência de ação bem como o controle sobre a fonte de recurso XML.

seqüências de ação são executadas pelo componente da plataforma conhecida como o solução do motor. Sempre que algum cliente invoca uma seqüência de ação, o motor lê a definição da seqüência de ação e, em seguida, executa a sua etapas.

Logicamente, as soluções Pentaho são armazenados e mantidos na solução repositório. Os aplicativos que se conectar ao servidor Pentaho pode navegar soluções e pastas, e armazenar seqüências de ação novo, um processo chamado publicação.

Fisicamente, a solução de repositório podem ser armazenados como arquivos no sistema de arquivos, ou armazenadas em um banco de dados relacional. Para execução de base seqüência de ação, tanto métodos suficiente. No entanto, a solução baseada em arquivo repositório atualmente não suporte à autorização. Assim, para um controle preciso sobre quais usuários podem acessar quais conteúdo, a solução de repositório precisa ser armazenado em um banco de dados.

Database Management pool de conexão

Na maioria dos casos, os dados apresentados na aplicação de business intelligence
ções são armazenadas em um banco de dados (relacional). Para acessar os dados no banco de dados,
a aplicação precisa estabelecer uma conexão com o banco. A conexão é
então usada para enviar pedidos (queries) ao servidor de banco de dados, que envia de volta os dados como uma resposta.

Estabelecer uma conexão com um banco de dados pode ser uma tarefa relativamente caro. É necessário algum tempo para procurar o host do banco de dados, e algum tempo pode ser gasto em protocolos de negociação, que autentica o usuário, ea criação de uma sessão. Em muitos casos, a conexão é necessária para executar apenas muito poucos consultas. Por exemplo, muitos relatórios são baseados em apenas uma consulta de banco de dados, e
número de consultas vai usar o mesmo banco de dados para recuperar seus dados.

Para evitar a sobrecarga de estabelecer uma nova conexão para cada consulta ou lote de consultas, as conexões de banco de dados pode ser aberto uma vez e armazenados em um
piscina. Sempre que um cliente precisa de uma conexão de banco de dados, uma ligação gratuita pode ser
colhidos a partir da piscina, que serve para fazer algum trabalho, e depois é lançado de volta a piscina novamente.

Database pool de conexão também é uma maneira fácil de limitar o número de conexões simultâneas de banco de dados aberto. Ao insistir que os pedidos sempre escolher uma conexão livre de uma piscina de tamanho fixo em vez de estabelecer uma nova ligação directa, o banco de dados podem ser protegidos para não serem inundadas com solicitações de conexão.

Pool de conexão JDBC é comum na maioria dos servidores de aplicação Java, e muitas implementações diferentes estão disponíveis. Pentaho não oferecer os seus próprios conexão de execução da piscina.

User Authentication and Authorization

A plataforma Pentaho utiliza Spring Security (anteriormente conhecido como Acegi Sistema de segurança para a Primavera) para tratar de autenticação e autorização. Esta é a solução de segurança padrão do framework Spring Java.

Spring Security fornece muitos componentes diferentes para implementar todos os tipos esquemas de autenticação diferentes. Ele fornece a lógica que se mantém informado dos se um usuário precisa ser autenticado, e pode delegar a autenticação
pedidos de um mecanismo de autenticação externa, como um servidor de banco de dados, um diretório LDAP, ou a autenticação NTLM em uma rede Windows.

Agendamento de tarefas

A plataforma Pentaho utiliza quartzo como componente de agendamento de tarefas. Quartzo é criado e mantido pelo projeto e liberado sob OpenSymphony

uma licença Apache 2.0 (ver www.opensymphony.com / quartzo para o projeto detalhado informação).

O agendador de tarefas é usada para uma série de coisas:

- Periódico de execução das tarefas de manutenção
- Contexto de execução de relatórios
- Agendamento de empregos ETL

As capacidades de programação da plataforma são abordados no Capítulo 14.

Serviços de e-mail

A plataforma de BI inclui a capacidade de enviar e-mail usando um padrão SMTP servidor. Um arquivo de configuração para usar uma conta do Gmail também está incluído.

Antes de mensagens podem ser enviadas, o servidor deve ser configurado primeiro. O con- mail figuração deve ser inserido no arquivo email_config.xml, Que está localizado na o diretório <install-path> / pentaho-solutions/system/smtp-email. A arquivos de configuração têm excelentes comentários in-line e deve ser simples para definir este para cima. Reiniciar o servidor depois de mudar o arquivo de configuração não é necessário; o novo inscrições será recolhido automaticamente quando os valores foram inscritos corretamente.

BI Componentes

A plataforma constitui a base para uma série de componentes que oferecem funcionalidade de inteligência empresarial típica. Nessa camada, encontramos o seguinte componentes:

- camada de metadados
- Ad hoc de relatórios de serviço
- motor de ETL
- mecanismo de relatório
- motor OLAP
- Dados do motor de mineração

A camada de metadados

A função do Pentaho Metadata Layer (PML) é proteger os usuários finais da complexidade de SQL e bancos de dados. A PML é baseado na Com- seg Armazém especificação Metamodelo do Object Management Group (www.omg.org / cwm) E é capaz de gerar uma consulta SQL a partir de escritos no Metadados Query Language (MQL). A consulta MQL por sua vez, é criado por um

usuário final, construindo a seleção desejada de um conjunto de objetos expostos em uma modelo de metadados. A camada de metadados consiste em três camadas, conforme especificado pelo
a CWM:

- Esta camada de Físico- é onde a conexão com o banco é armazenado e onde a representação física dos objetos de banco de dados é criado. Ao gerar SQL, esta é a camada de PML, que começa finalmente as informações de atributo do banco de dados.

- A camada de negócios camada intermediária de tradução onde traduções atributos da base de dados técnicos para descrições mais user-friendly são feita. Esta é também a camada onde os relacionamentos entre tabelas e fórmulas e cálculos adicionais são criados.

- visão empresarial Expõe- e re-organiza a camada de negócios para o final usuários e grupos de usuários finais.

Figura 4-2 mostra uma representação gráfica da descrição anterior.

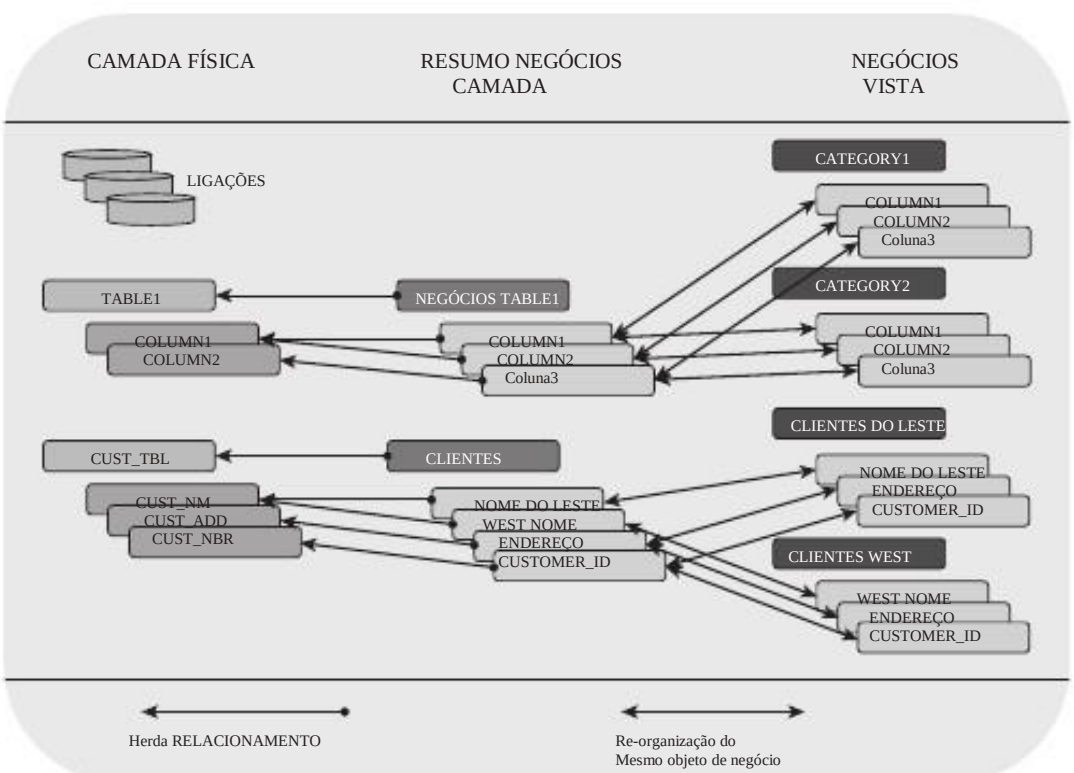


Figura 4-2: modelo de camadas de metadados

O lado direito do diagrama da Figura 4-2 é a única parte dos metadados
camada que é visível aos usuários finais quando eles trabalham com um dos design do relatório
ferramentas. As outras camadas existem para traduzir corretamente o modelo do usuário-expostos

objeto de volta para a consulta correta para o banco no qual a camada de metadados opera.

Ad hoc Reporting Service

A Web Ad Hoc de Consulta e de serviço do Reporting, ou WAQR, oferece aos usuários finais uma maneira fácil de criar relatórios usando a camada de metadados. O WAQR ("Wacker pronunciado") é um serviço separado do relatório de pleno direito motor e é capaz de criar um relatório de lista simples agrupados. WAQR é coberto em mais detalhes no Capítulo 13.

O Mecanismo de ETL

motor Pentaho ETL é o cavalo de batalha para as tarefas de integração de dados e executa os empregos e as transformações criados com as ferramentas Pentaho Data Integration. O motor de ETL é parte da pilha de BI, mas também pode ser executado em um servidor diferente ou mesmo vários servidores em um modo de cluster.

Reportagem Motores

A plataforma Pentaho hospeda vários motores de comunicação. Os motores são nativas o motor já foi mencionado para a ferramenta de consulta ad hoc, e os JFreeReport motor. Além disso, os navios Pentaho com suporte para JasperReports e BIRT já incorporados. Isto significa que a plataforma de BI Pentaho é capaz de manipulação de todos os relatórios criados para os três código aberto mais popular reporte ferramentas.

O mecanismo de OLAP

Mondrian é o motor Pentaho OLAP e traduz as consultas MDX no SQL baseado em um modelo multidimensional. Mondrian faz muito mais do que apenas traduzir de uma língua para outra consulta, que também cuida de cache e buffer de resultados intermédios e os anteriores para otimizar o desempenho. Isto significa que a primeira vez que uma análise é executada em um modelo multidimensional, levará mais tempo do que a posterior análise durante a mesma sessão porque Mondrian tenta manter resultados anteriores, as hierarquias e os cálculos na memória.

Outra característica notável do Mondrian é o seu modelo de segurança, que suporta papéis. As funções podem ser utilizadas para restringir os dados que é acessível por um usuário, assim a limitação do número de pontos de vista diferentes OLAP e relatórios que precisam ser desenvolvidos.

O Mecanismo de Mineração de Dados

motor Pentaho de mineração de dados é indiscutivelmente um dos mais poderosos ainda menor peças usadas da plataforma. É realmente o Weka motor de mineração de dados que tem

foi adotado por Pentaho que controla as tarefas de mineração de dados. É constituída por um coleção completa de algoritmos de mineração de dados, tais como os necessários para cluster, árvores de decisão, regressão e redes neurais. Partes do algoritmos Weka pode ser chamado de uma chaleira transformar a permitir, por exemplo, pontuação direta dos dados de entrada durante uma transformação Chaleira. Capítulo 16 abrange as diferentes ferramentas Weka e mostra um exemplo passo-a-passo de como Weka e Chaleira podem ser usados em conjunto para desenvolver uma transformação de dados que automaticamente notas de novos clientes.

A camada de apresentação

Pentaho vem com um built-in interface web chamado usuário do console. O usuário formulários do console um front-end que permite que um usuário humano de interagir com o servidor.
A camada de apresentação pode ser usado para navegar e abrir o conteúdo existente (Relatórios, painéis, análise), mas até certo ponto também pode ser usado para criar conteúdo de BI novo. Figura 4-3 mostra o usuário Pentaho console onde à esquerda lado uma árvore de pasta é usada para organizar o conteúdo que está listado no painel à inferior esquerdo. Os documentos abertos são exibidos na tela principal e usando guias, o usuário do console pode ter vários painéis, análise e relatórios abertos ao mesmo tempo.

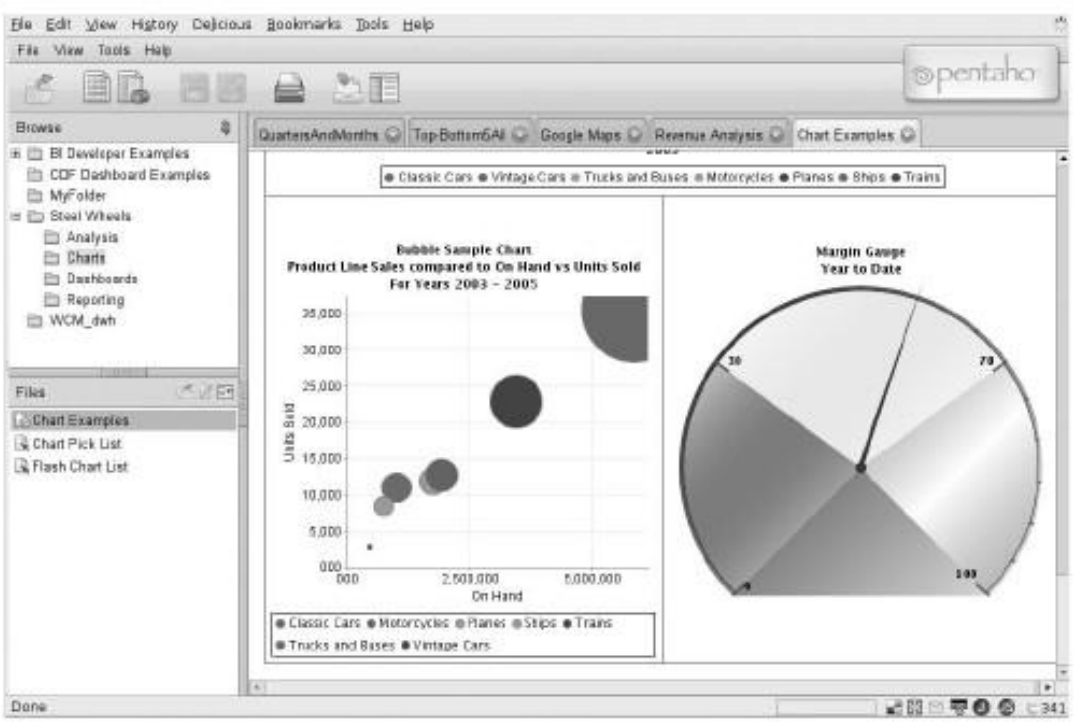


Figura 4-3: Pentaho console de usuário

Novo conteúdo na forma de relatórios e análise de pontos de vista podem ser criados usando a Web Ad Hoc de Consulta e componente Reporting (WAQR) eo JPivot análise de front-end. O WAQR é coberto em profundidade no Capítulo 13,

eo capítulo 15 contém um tratamento detalhado de JPivot e subjacente
tecnologia de Mondrian.

Subjacente a tecnologia Java Servlet

Embora o termo"O Servidor Pentaho"pode sugerir o contrário, há
não um programa que pode legitimamente ser chamado por esse nome. Pelo contrário, Pentaho
fornece uma série de programas chamados servlets que realizam alguma tarefa específica,
um serviço, para qualquer cliente que o solicite. Servlets são programas Java que fazer
não operar de forma autônoma no computador local. Em vez disso, eles são executados dentro de
outro programa, o servlet container.

Normalmente, o servlet container é ela própria um servidor web (ou seja, um servidor HTTP),
ou uma parte dele. O servlet container é responsável por aceitar HTTP
pedidos e encaminhamento deles para um servlet apropriado. O servlet processa então
o pedido, e gera uma resposta adequada, que é transferida para o
recipiente eventualmente percurso de volta para o cliente solicitante.

A organização de um programa Java em um servlet container de servlets e vários
que executam o serviço real é chamado A tecnologia Java Servlet. O Java Servlet
A tecnologia é a padrão de fato para a implementação de aplicações web em Java.
As formas em que o servlet eo seu recipiente pode interagir são precisamente
definida pela API Java Servlet. Esta especificação para esta API foi inicialmente
criado pela Sun Microsystems, e desenvolvido pela comunidade Java
processo.

Pentaho não oferecer o seu próprio servlet container. Pelo contrário, servlets Java pode
executados em qualquer container servlet Java, desde que ambas servlet eo recipiente
apoiar a mesma versão do Java Servlet API, que é a situação usual.
A garantia de que um servlet será executado em qualquer servlet container compatível permite
desenvolvedores servlet para se concentrar no que sabem fazer melhor, que é a adição útil
funcionalidade para servidores web. Por outro lado, os criadores de software de servidor web pode
concentrar completamente na sua tarefa sem se preocupar com qualquer mudança vai quebrar
baseado em servlet extensões de servidor.

Atualmente, o Community Edition do servidor de BI Pentaho é realmente um
Apache Tomcat servlet container com todos os servlets Pentaho pré-instalado.
No entanto, todos os servlets podem ser baixadas separadamente e instruções de instalação-
ções estão disponíveis para outros recipientes de servlet popular também, como o JBoss,
Glassfish, WebSphere, BEA WebLogic, e muitos mais.

Programas Desktop

Como apontado na introdução deste capítulo, a maioria dos servidor Pentaho não
programas podem ser melhor classificadas como programas de desktop. Alguns deles só pode
atuar como um cliente e precisa interagir com um servidor Pentaho, mas pode ser

autônoma usado também. Os programas de desktop servem principalmente como ferramentas de projeto
ou ajudas, porque a maioria da experiência do usuário final é entregue em Pentaho
portal na Internet. As ferramentas de trabalho, portanto, ser usado principalmente pelos desenvolvedores,
embora alguns, como o Report Designer, poderá ser utilizado por usuários avançados.
Todos os programas de computador têm um componente de BI ou componente de servidor para o qual
que se destinam. A Tabela 4-1 mostra as ferramentas disponíveis com o seu homólogo componentes do servidor.

Tabela 4-1: ferramentas de desktop e servidor

Ferramenta Desktop	SERVER / componentes de BI
Design Studio (PDS)	Plataforma de BI
Editor de Metadados (PME)	Metadados camada, Ad Hoc componente de relatório
Esquema (PPS) Workbench	OLAP Engine
Agregado Designer (PAD)	OLAP Engine
Designer de Relatórios (PRD)	mecanismo de relatório
Spoon (PDI)	motor de ETL
Weka	Data Mining Engine

Cada uma dessas ferramentas é coberto em um capítulo posterior, com exceção do Pentaho Design Studio (PDS), que compõe a última parte deste capítulo.
PDS não é uma ferramenta para a criação de novos conteúdos, mas é usado para criar fluxos de trabalho
e ações que trabalham com o conteúdo existente BI. PDS é também uma outra forma outro tipo de ferramenta, é a única parte da suíte de BI que não é autônomo Programa Java, mas um plugin para um ambiente de desenvolvimento já existentes (o Eclipse IDE). A lista a seguir é um breve resumo da área de trabalho diferentes ferramentas e sua posição no conjunto de BI:

- Pentaho Metadata Editor (PME)-Com PME, os designers podem criar meta-camadas de dados que servem como uma camada de abstração entre um banco de dados relacional
  e um usuário final. A camada de metadados pode levar objetos de usuário como Cus-Nome Tomer, País, e Receita e traduzir essa seleção em
  a instrução SQL correta necessário para recuperar essas informações de um banco de dados. Mais sobre PME no Capítulo 12.
- Esquema Pentaho (PPS) Workbench-Este é a ferramenta para a construção esquemas multi-dimensional a ser utilizado pelo Mecanismo de Mondrian.
- Pentaho Designer Agregado (PAD)-A ferramenta separada para (automaticamente) criar tabelas agregadas que são usadas por Mondrian para melhorar a desempenho de cubos OLAP. PSW e PAD são cobertas em profundidade Capítulo 15.

- Pentaho Report Designer (PRD)-O front-end para a construção de relatórios para a plataforma Pentaho e, possivelmente, a única das ferramentas de trabalho disponíveis que pode ser colocada nas mãos de um usuário final conhecedor. PRD é coberto no capítulo 13.

- Pentaho Data Integration (PDI)-O ferramenta de desktop para a construção de postos de trabalho ETL
e transformações é chamado Spoon. PDI contém mais do que apenas Spoon mas esta é a parte mais visível da solução de ETL. O PDI é abordado no Capítulos 10 e 11.
- Weka-A conhecidos open source solução de mineração de dados e as únicas ferramenta não dispor de uma abreviação de três letras Pentaho, pois é não seja mantido pela Pentaho e não estão disponíveis a partir do Pentaho regular sites de download. Weka é um projecto iniciado e mantido pela Universidade de Waikato, na Nova Zelândia, mas foi adotado por Pentaho como sua ferramenta padrão de mineração de dados. Mineração de dados com Weka é objecto de
Capítulo 16.

Todas essas ferramentas têm algumas coisas em comum: eles são escritos em Java e será executado em qualquer plataforma que contém uma máquina virtual Java. Eles são fornecidos com um script de inicialização ou arquivo de lote, mas também pode ser iniciado
diretamente da linha de comando com o Java-jar comando. A segunda uniformização importante é que nenhum deles trabalha com cria ou proprietários formatos de arquivo. Todas as definições criadas com as ferramentas de desktop são diferentes XML
base, assim aberto a qualquer editor e qualquer pessoa. Como conseqüência, você não está obrigada a utilizar uma das ferramentas de design, mas são livres para criar e / ou modificar o Arquivos XML diretamente com um editor de texto simples. Algumas pessoas acham que é ainda mais fácil
trabalhar com os arquivos XML do que com as ferramentas da GUI.

Pentaho Enterprise Edition eo Community Edition

Pentaho oferece duas versões do Pentaho BI Suite. A principal distinção é feita entre o comercial licenciada Enterprise Edition eo completo aberto fonte do Community Edition. Esta distinção tem mais a ver com o tipo do apoio oferecido do que com diferenças de software real, mas a Enterprise Edition (EE) oferece alguns componentes que não estão disponíveis na comunidade versão. Embora não cobrirá componentes EE-específicas neste livro, mencioná-los aqui para ser completo.

- Enterprise Console-A maior parte das adições EE visam prorroga o Community Edition com a funcionalidade necessária em uma empre- ambiente de taxas, como a configuração de segurança, os diagnósticos das aplicações e monitoramento de desempenho, auditoria e registro, gerenciamento de ciclo de vida- mento (conteúdo migrar de desenvolvimento para testar a produção), o conteúdo vencimento, e de backup / restore do Pentaho repositório. A maioria desses

tarefas podem ser executadas com o Enterprise Console. Isso não significa que você não pode fazer essas coisas com o Community Edition, mas exigirá grandes esforços para criar, por exemplo a gestão do ciclo de vida, sem as ferramentas EE.

- Extensões PDI-Pentaho Integração de Dados EE acrescenta um Con-Empresa exclusiva para monitoramento de desempenho, administração remota, e alerta. Há também um plugin extra para mineração de dados, o KnowledgeFlow plugin.

- Single Sign-On com o LDAP e AD-integração Embora o Pentaho Community Edition tem a sua própria autenticação e autorização com-nente, não é integrado com um provedor de autenticação externos, tais como LDAP ou Active Directory. A vantagem de ter essa integração é dupla: os usuários só precisam ser inseridos e mantidos, de vez em central localização, e os usuários não precisam efetuar o logon separadamente e lembre-se outra
senha.

- Dashboard Builder-A componente mais visível da EE é o Dash-placa Builder, que permite aos usuários facilmente preencher um painel de BI com vários tipos de conteúdo, tais como gráficos, relatórios e mapas. Criando-traço placas usando a Comunidade Dashboard Framework (CDF) é coberto em Capítulo 17.

- Serviços e suporte-In Além de maior funcionalidade, Pentaho Enterprise Edition fornece suporte, indenização, manutenção de software manutenção, e recursos técnicos adicionais.

À excepção desta lista, não há diferença entre a Comunidade e Enterprise Edition nos produtos que compõem o BI pilha. Isso significa que praticamente não há limites para o que você pode fazer e construir com o indivíduo ferramentas de BI, porque não há Enterprise Edition construtor do relatório que lhe permite fazer
mais do que você poderia fazer com a norma comunitária Edition. Na verdade, este é o que define Pentaho para além de muitos outros (fonte aberta, mesmo!) fornecedores.

A criação de seqüências de ação com Pentaho Design Studio

Pentaho Design Studio (PDS) é baseado no desenvolvimento integrado Eclipse ambiente (IDE), e pode ser baixado como um completo, pronto para uso solução que inclui o Eclipse. Se você já tiver uma versão do Eclipse em execução, PDS pode ser adicionado a um ambiente existente como um plugin. (Basicamente, o PDS só é
o plugin, mas Pentaho oferece um pacote completo de trabalho por conveniência).
PDS tem um propósito a criação e manutenção seqüências de ação. Como o próprio nome implica, uma seqüência de ação é um conjunto predefinido de ações que podem ser executados
no servidor de BI Pentaho. Execução de uma seqüência de ação pode ser desencadeada por uma

ação do usuário, uma agenda, ou qualquer outro evento, incluindo uma outra seqüência de ação.
complexidade varia de Acção seqüência de muito simples, por exemplo, executar um" relatório"ou"exibição na tela uma mensagem de"a bastante complexo, por exemplo,"encontrar todos os clientes com produtos vencidos e enviar-lhes um lembrete no cliente formato preferido (XLS, PDF, HTML) contendo uma descrição do atraso itens."seqüências de ação são a locomotiva real de uma solução Pentaho e porque eles amarram todos os outros componentes juntos a última parte deste capítulo é orientados para explicar o que seqüências de ação são e como você pode construir e implantá-los na plataforma Pentaho.

A partir da introdução, você deve ter percebido que as seqüências de ação (AS) só podem ser utilizadas para a criação de saída, de uma forma ou de outra. Embora esta seja uma característica como importante, que é apenas uma parte da história. Um AS pode ser usado para muito baixo nível de atividades do sistema, bem como, por exemplo, para definir as variáveis de sessão o momento em que um usuário se autentica, ou para criar listas globais de parâmetros que podem ser utilizado por outro processo ou AS. Suponha, por exemplo, que pretende restringir acesso aos seus dados com base no usuário que está entrando, e suponha que cada usuário é permitido somente a visualização de dados a partir de seu próprio departamento, região, ou qualquer outro critério que você pode pensar. Com um sistema AS, você pode definir o nome do departamento que pertence ao usuário conectado e usar este nome noutros seqüências de ação como um parâmetro para filtrar os dados por diante. Está fora do âmbito de aplicação este livro para explicar Ases do sistema, mas você pode encontrar um instruções detalhadas sobre Pentaho Design Studio (Eclipse) Primer
Wiki Pentaho em http://wiki.pentaho.com/display/ServerDoc2x/Using+ Sistema + Ações + para + Control + + Acesso a Dados.

PDS é, como já explicamos, um plugin para o componente escrito extensamente utilizados Eclipse IDE. Embora não possamos oferecer um Eclipse tutorial completo aqui podemos cobrir o básico para você começar com o PDS para poder criar a sua próprias soluções. Para o restante deste capítulo, usaremos os exemplos PCI eo banco de dados de Aço Rodas, mais tarde, no livro, usaremos o Mundo Classe de banco de dados de filmes para mostrar um pouco do poder de xactions combinado com relatórios e dashboards.

Instalação e configuração do Eclipse e do PDS são abordados no Capítulo 3. Para os exemplos neste capítulo, vamos supor que você tenha um trabalho Pentaho Sistema com os exemplos prontos na mão. Quando você inicia o Eclipse / PDS, o Eclipse tela de boas vindas é exibida ea opção de plataforma de BI aparece em no menu superior. A Plataforma de BI menu tem apenas uma subopção: Nova Ação Seqüência.

Os componentes básicos do Eclipse e uma terminologia que você precisa para começar resumem-se ao seguinte:

- Workspace-Este é o recipiente mais alto nível das soluções que vai criar, um espaço de trabalho é usado para manter uma coleção de projetos Eclipse

logicamente agrupados. Você pode ter apenas um espaço de trabalho aberto a uma tempo, e para a maioria das implementações Pentaho, usando um único espaço de trabalho
vai fazer tudo certo. O espaço de trabalho (o padrão) deve ser definido quando iniciar o Eclipse, após esse primeiro tempo, o espaço de trabalho padrão é aberto automaticamente cada vez que você iniciar o programa.

- Projeto-O coleção de arquivos e pastas que, juntos, compõem uma solução. Os projetos podem ser criados dentro do espaço de trabalho (que é uma pasta em seu sistema), mas que não é obrigatório. A Figura 4-4 mostra um recém-projeto criado chamado My Pentaho. Para a pasta do projeto, já existente pasta Pentaho soluções é selecionado. Agora é fácil de abrir e modificar xactions existentes, como o exemplo de relatório de ruptura, que é aberto na screenshot.

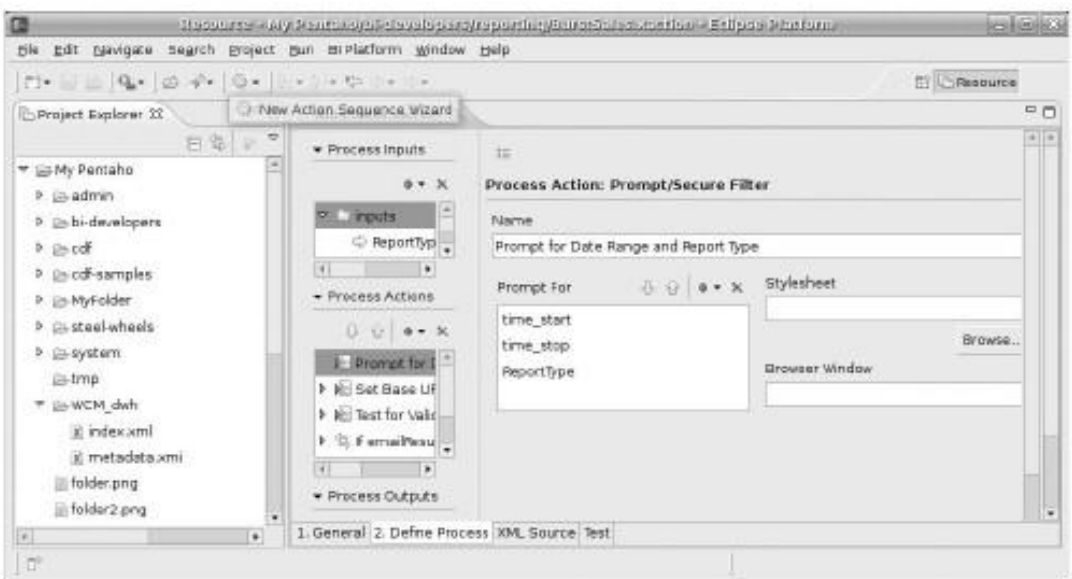


Figura 4-4: Pentaho Design Studio com o editor de seqüência de ação

- Ver-In Eclipse, uma vista é uma janela dentro do IDE, que apresenta algumas conteúdos específicos, tais como a estrutura do projeto, fontes de dados ou um aplicativo Java
pacote explorador. A imagem na Figura 4-4 mostra o padrão de projeto Explorer ponto de vista sobre o lado esquerdo. Eclipse contém uma coleção enorme de outros
vistas, bem, eles podem ser abertos através do menu Window Vista Show ou usando o pequeno ícone de Visualização Rápida no canto inferior esquerdo da tela (O pequeno ícone azul e branco com o sinal de expoente pouco mais).
- Editor-Este é onde você realmente escrever o código, ou, no caso de utilização PDS, definir sequências de ação. O editor de ação Pentaho é mostrado na Figura 4-4. Para iniciar o editor, basta clicar duas vezes sobre uma das existentes xactions amostra ou criar um vazio, selecionando a Ação Novo Seqüência de opção no menu plataforma de BI.

- Perspectiva-Provavelmente o recurso mais poderoso do Eclipse é a capacidade mudar completamente o seu comportamento e as opções disponíveis, alterando a um diferente perspectiva. Uma perspectiva transforma o objectivo geral Eclipse IDE em uma ferramenta concebida especificamente para uma determinada tarefa. Quando você desenvolver código Java, você vai usar a perspectiva Java para ter o built-in editores, depuradores, esboço de um código, e assim por diante a sua disposição. Quando você desenvolver relatórios BIRT, você vai usar o relatório de perspectivas de Design, que de repente se transforma em um ambiente Eclipse comunicação poderoso. Para trabalhar com PDS, a perspectiva de recursos simples fará a multa justa. Outras perspectivas são abertas, selecionando a opção Open Perspective a partir do menu Window.

Para obter mais informações sobre como trabalhar com o Eclipse, visite o site do projeto em www.eclipse.org.

O Editor de Seqüência de Ação

Antes que você possa criar uma seqüência de ação, você precisará definir um projeto para colocar seu novo arquivo dentro Para criar um projeto, selecione File New Project. Eclipse agora inicia o novo Assistente de Projeto, solicitando o tipo de projeto para criar. Para o plataforma de BI, basta selecionar Project na guia Geral. Você precisa dar o projeto um nome (qualquer nome serve; usamos Meu Pentaho, por exemplo) e selecionar um local para seu novo projeto. Você vai perceber que o local padrão do um novo projeto é o espaço de trabalho aberto, mas você pode selecionar qualquer outro local como também. Quando você seleciona uma pasta existente (por exemplo, o Pentaho soluções pasta do servidor de BI), o conteúdo dessa pasta será exibido na Explorador de projeto, logo que você clique em Concluir. Você pode criar uma nova ação Seqüência de várias maneiras:

- Use a nova seqüência de ação do item no menu de plataforma de BI. Isso vai criar um novo arquivo xaction vazio, mas você terá que definir o local para o arquivo (o recipiente) em primeiro lugar.

- Botão direito do mouse em uma pasta existente no Project Explorer e selecione Novo Seqüência de Ação a partir do menu plataforma de BI. Observe que o recipiente é Agora automaticamente preenchidos

- Use a nova seqüência de ação Assistente no menu do ícone de atalho. (A dica para o ícone é exibido na Figura 4-4). Novamente, o local (Recipiente) deve ser selecionado.

Nos três casos, a seqüência de ação Wizard abre para que você pode inserir um nome para a nova seqüência de ação e selecionar um modelo para ajudar você a começar a saltar-construção de uma seqüência de ação nova. Modelos predefinir entradas e ações para a tarefas específicas, tais como uma visão nova análise ou uma ação de ruptura.

O editor xaction é composto por quatro painéis ou guias, que você pode ver ao longo na parte inferior da tela. Uma parte importante da guia Geral, apresentado na

Figura 4-5, é o título, que será exibido no console do usuário. Você pode também encontrar o ícone que vai acompanhar o título de acção no navegador, o Versão, o nível de log, eo Autor na guia Geral. O Visible esconde a caixa de seleção xaction do usuário do console quando selecionado, o que faz possível a criação de""xactions ajudante que não são visíveis aos usuários finais.

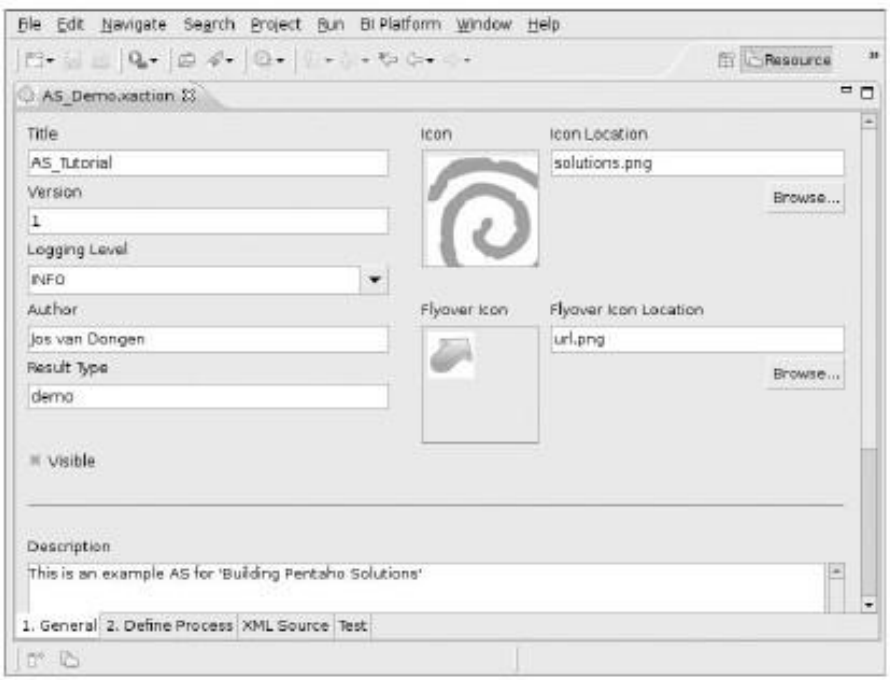


Figura 4-5: Ação editor seqüência, guia Geral

O guia Definir processo mostrado na Figura 4-6 é o editor real onde você pode adicionar entradas, ações e realizações. Para verificar o código XML gerado, você pode abrir a aba Fonte de XML, o que algumas pessoas dizem é a ação real editor. O guia de teste no final permite que você execute a ação diretamente do Design Studio, embora o servidor Pentaho é utilizado para a execução do código.

Antes de começar a construir novas sequências de ação, é uma boa idéia para testar se você pode executar ações na tela de teste. A maneira mais fácil de fazer isso é abrir o HelloWorld.xaction arquivo, que está localizado na pasta bi-developers obtendo-iniciado. Na Definir Processo guia mostrado na Figura 4-6, um processo ação é definida.

Quando você clica sobre isso, a ação do processo Olá Mundo é exibido, que tem apenas dois campos: o nome ea mensagem da acção. Agora você pode alterar a mensagem para algo como Ele está trabalhando! Caso contrário, o padrão string% mensagem será exibida. Após salvar as mudanças que você pode passar para o teste guia de verificar se a plataforma está funcionando.

NOTA Um servidor Pentaho deve ser iniciado antes de executar um teste de PDS, caso contrário, nada vai acontecer.

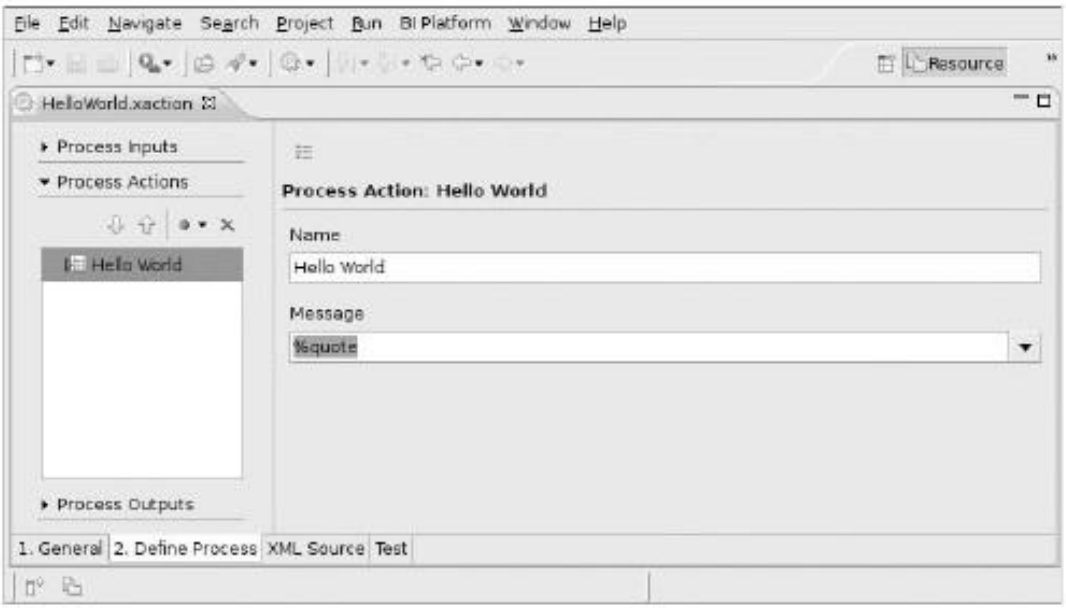


Figura 4-6: O guia Definir Processo

Existem dois campos de texto na guia Test, conforme mostrado na Figura 4-7, uma para digitar a URL do servidor, e um para a URL gerada, o que inclui a xaction chamada. Se você tiver a instalação padrão do Pentaho em execução no seu local computador, o URL do servidor Pentaho é localhost: 8080/pentaho. Quando você entrar neste servidor e prima de teste, a tela de login aparece Pentaho. Primeiro registo em e atualizar o cache de repositório, selecionando Ferramentas Refresh Repositório Cache (caso contrário, o xaction existente com o texto padrão será exibido). Clique em Gerar URL e pressione o botão Executar, à direita do Geração de URL. Você deverá ver o resultado apresentado na Figura 4-7.

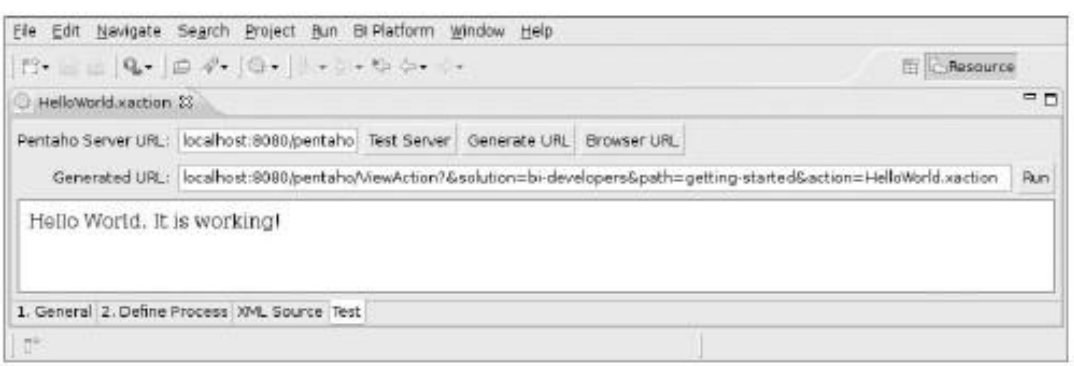


Figura 4-7: Olá Mundo executado

Agora que cobrimos o básico do Eclipse e seqüências de ação Editor, é hora de começar algum trabalho real é feito o uso dessas ferramentas.

Anatomia de uma seqüência de ação

seqüências de ação (Ases), todos têm uma estrutura semelhante, que consiste no seguinte quatro blocos de construção:

- Entradas
- Recursos
- Ações
- Saídas

Quando você abre o Processo de Definição guia do editor como você vai notar que parece que só existem três blocos que precisam ser definidos: Processo Entradas, Ações de Processos, e saídas do processo. Este é o lugar onde há uma ligeira diferença entre o designer e os arquivos XML criados, os arquivos XML fazer uma distinção entre insumos e recursos, enquanto o designer trata tanto como tipos de entrada diferentes. A lógica por trás disso é que você raramente precisará adicionar manualmente os recursos a si mesmo, porque eles vão ser geridos por PDS. Se, por exemplo, você adicionar uma ação de relatório, o local do arquivo é adicionado como um recurso automaticamente.

Entradas

Processo de insumos para as seqüências de ação são os parâmetros que podem ser utilizados em ações processo. A forma mais simples de uma entrada é um parâmetro codificado. Cada entrada do processo deve ter pelo menos um nome e um tipo (string, inteiro, e assim por diante) que pode ser dado um valor padrão, por exemplo, uma seqüência de texto que pode ser passada para a acção Olá Mundo para mostrar. Um AS pode ler os parâmetros a partir de fontes diferentes, tornando possível para passar informações a partir do exterior o como a um processo de produção. As seguintes fontes de entrada estão disponíveis:

- Pedir-Estes são pares nome-valor que pode ser lido diretamente URL. Usando o mesmo exemplo Olá Mundo, você pode adicionar uma entrada nomeado RequestText do tipo string, E adicionar um novo duas palavras com uma origem pedido. O nome padrão dado pelo fonte de entrada é o mesmo que para a entrada do processo em si, mas que pode ser mudado. O processo de nome da entrada é a referência de parâmetro interno; o nome da fonte de entrada é a referência externa. Figura 4-8 mostra um exemplo disso. Usando este exemplo, você pode agora selecionar o Olá Mundo ação do processo e selecione o <RequestText> parâmetro da Mensagem drop-down list. Quando você agora salve o AS, atualizar o cache de repositório e adicione o texto & Req = Este é muito divertido! para a URL, o texto Olá Mundo. Esta é uma grande diversão! será exibido. Note que neste caso, o AS não precisa ser alterada e salvou mais para mostrar uma nova saída. Você pode tentar isso inserindo um texto diferente, após req = e pressionando Execute novamente.

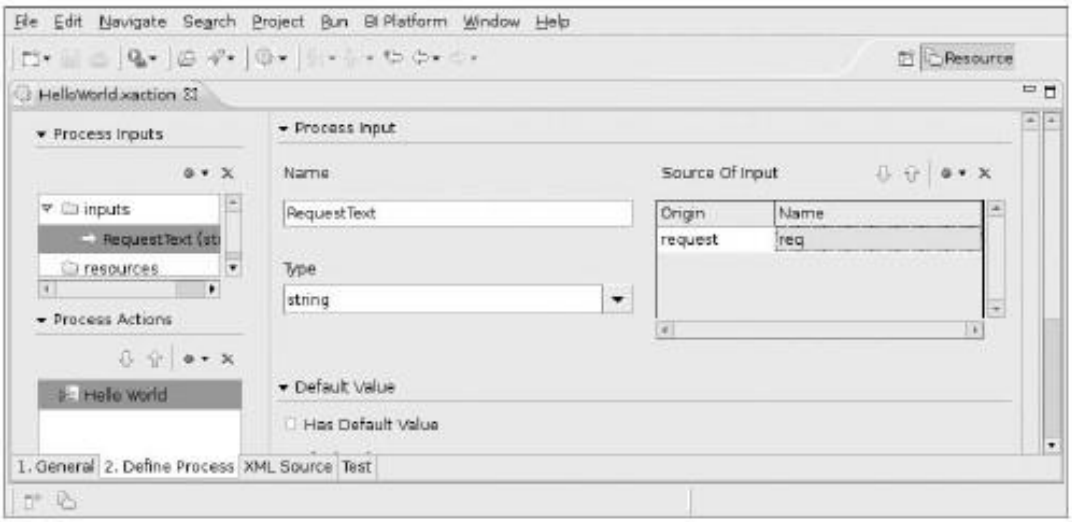


Figura 4-8: Pedido de fonte de entrada

- Estes sessão de são variáveis que vivem para a duração da sessão do usuário. Curiosamente, sessão global, variáveis e tempo de execução pode ser definido através de um
ação do sistema que por sua vez, é criado com PDS. Há também um casal de variáveis de sessão padrão que pode ser acessado, com o nome de usuário do utilizador actual o mais freqüentemente usado um. Isso pode ser testado utilizando os mesmos Olá xaction Mundial pela mudança da fonte de entrada do RequestText entrada do processo para sessão E o nome do de nome. Depois de salvar AS e executá-lo, ele deve agora mostrar Olá Mundo. <username>,
onde <username> é o nome usado para fazer logon no servidor Pentaho.

- Global-Similar das variáveis de sessão, mas estes têm um alcance global, que significa que os valores dessas variáveis são as mesmas para todos os usuários o servidor de BI. A vida útil de uma variável global está vinculada à aplicação, o que significa que enquanto o aplicativo é executado (o servidor de BI é até), O variáveis podem ser acessados.

- Tempo de execução global variáveis, mas sua vida é infinita, significando que quando você desligar o servidor de BI (não a máquina, mas o aplicativo!) a variável tempo de execução permanece na memória. Porque isto de alguma forma limita o
controle você tem sobre essas variáveis, é melhor usar variáveis globais em vez disso.

- Segurança Ativa a recuperação da segurança (sessão) variáveis. O follow-parâmetros podem ser obtidas gratuitamente (note que esses nomes são case-sensitive!):
  - PrincipalName (string) -O nome do momento autenticado usuário. Semelhante à variável de sessão nome.
  - PrincipalRoles (lista de strings) -As funções que actualmente autên-sas sofisticadas usuário é membro.

- PrincipalAuthenticated (string) -verdade se o usuário é autenticado, caso contrário falsa.
- PrincipalAdministrator (string) -verdade se o usuário é autenticado Administrador, caso contrário falsa.
- systemRoleNames (lista de strings) -Todas as funções conhecidas no sistema. Manusear com cuidado, porque essa lista pode se tornar bastante grande.
- systemUserNames (lista de strings) -Todos os utilizadores conhecidos no sistema. Manusear com cuidado, porque essa lista pode se tornar bastante grande.

Saídas

As saídas são o que uma seqüência de ações pode passar para o mundo exterior, que poderiam ser seqüências de ação de outros também. Eles podem ter aproximadamente o mesmo
Destinos como uma entrada pode ter origens, mas há algumas diferenças.
Ao invés de um pedido, uma saída pode passar uma resposta. Saídas também pode salvar em um
arquivo ou ftp-vfs.

Ações

Processo de Ações vêm em todos os tipos e tamanhos. Há ações para recuperar dados, criar ou abrir relatórios e gráficos, programar tarefas, executar fluxos de trabalho ou de dados trabalhos de integração, e para enviar a saída para e-mail ou uma impressora. Embora não possamos
abranger todas as ações disponíveis e combinações de ações, daremos alguns exemplos para ajudá-lo em seu caminho para a construção de suas seqüências de ação própria.
Lembre-se que os capítulos 14 e 17 contêm exemplos adicionais de Ases, mais especificamente para o relatório de ruptura e dashboards.
Um conjunto de ações processo será executado na ordem em que eles estão listados na tela. Ao selecionar uma ação e usando o para cima e para baixo flechas, a ordem de execução pode ser alterado. Este não é o único controle disponível aqui, há duas opções à sua disposição para ramificação (se) E looping (loop) ações. Combinado com o / Secure alerta Ação de filtro com o qual um usuário podem ser feitas para a entrada, estas opções permitem a lógica de execução bastante complexa.
Basicamente, existem dois tipos de xactions: aquelas que os usuários verão no pastas que possam acessar e pode ser executado sob demanda, clicando sobre eles, e os que serão programados e executados em segundo plano. Um bom exemplo do último relatório é de ruptura, que gera conteúdo personalizado para cada usuário ou grupo de usuários. Mas porque um AS pode começar outro (por adição de um Pentaho BI processo na lista de ação do processo), as possibilidades são praticamente ilimitadas.

ATENÇÃO Tenha cuidado ao chamar um xaction de um outro; excluir o "Filho"processo não está impedida de modo que você pode facilmente quebrar o processo principal.

Poderíamos tomar a amostra de modelo de ação de ruptura e explicar como isso funciona, mas apreender o verdadeiro poder de PDS é melhor realizado, iniciando com uma seqüência de ação vazia e estendê-lo passo a passo. Primeiro, vamos explicar o que que deseja realizar:

1. Criar uma lista de gestores com seus nomes, localidades e endereços de e-mail do banco de dados da amostra.

2. Loop através desta lista e envie um e-mail com o orçamento região, receita variância e para os gestores respectiva região.

3. Quando o laço se encontra com o gerente da região central ", enviá-lo uma visão geral adicional da receita total para todas as regiões.

Este exemplo usa muitos dos recursos disponíveis e é um excelente Introdução Se você quer construir seu seqüências de ação própria. Os seguintes etapas orientá-lo através do exemplo.

1. Primeiro, crie um novo vazio AS selecionando a seqüência de ação Wizard. Selecione um contêiner, um modelo em branco e digite um nome de arquivo. Esta é mostrado na Figura 4-9.

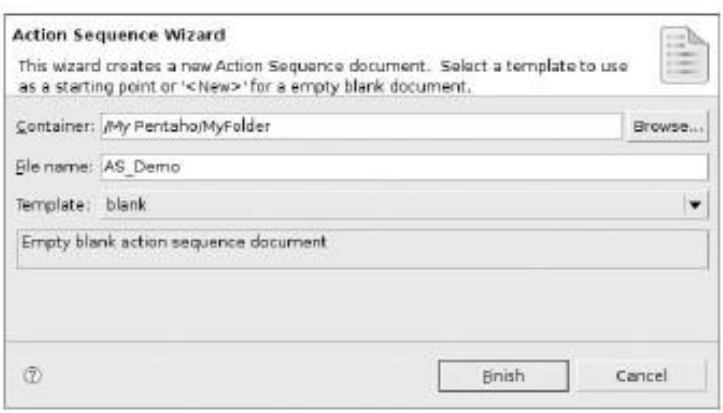


Figura 4-9: Criar uma nova seqüência de ação

2. Na guia Geral, digite o título (obrigatório) e outros campos (opcional).

3. Agora, vá para a segunda guia para definir o processo. Primeiro, você precisa definir onde o e-mail está vindo, então crie uma entrada de um novo processo. Nome do parâmetro de entrada a partir de e fazer certo a origem do pedido está removido da fonte de entrada.

4. Agora você pode adicionar a sua acção primeiro processo para recuperar a lista de dados. Adicionar
um novo processo de ação e escolha Obter dados FromRelational. Esta é

provavelmente a ação do processo mais utilizado para que explicar isso em um pouco mais detalhadamente:

a. O nome da ação é exibido na lista Ação. Processo e deve refletir a função da ação. Neste caso, um nome como GetManagerInfo vai fazer bem.

b. Para ser capaz de recuperar dados de um banco de dados, uma conexão é necessária.
Aqui você pode usar o tipo de JNDI simples com o nome SampleData (Você precisa digitar isto, não está disponível na lista drop-down).
Além disso, o conjunto de resultados tipo deve ser definido como na memória.

c. A própria consulta recupera os dados. Para este exemplo, usaremos a consulta Região distinta seleção, MANAGER_NAME, e-mail de DEPARTMENT_MANAGERS. Esta consulta deverá ser executado imediatamente.

d. Finalmente, o resultado Name Set deve ser inserido, este é o nome pelo qual a lista será referenciado mais tarde, por isso use um nome significativo, novamente,
como o LoopList nome que usamos. Como as colunas de consulta são não são automaticamente traduzidas para variáveis referenciável, o resultado Conjunto Colunas precisa ser inserido explicitamente. Neste caso, você adiciona o colunas REGIÃO,MANAGER_NAMEE EMAIL, Todos do tipo string.

5. Observe que o GetManagerInfo ação agora contém quatro saídas de acção: os três nomes de coluna mais o nome do conjunto de resultados. Agora que você adicione
um loop para ser capaz de lidar com todas as linhas do conjunto de resultados em seqüência.
The Loop Na lista drop-down contém apenas uma opção, o LoopList.
Após escolher esta, você pode continuar adicionando as ações internas para do loop.
6. Com o Loop Ação selecionada, adicione uma outra ação Relacional. Um pop-up tela vai perguntar se a ação deve ser criado dentro ou após o ação selecionada, você precisa adicioná-lo dentro do loop. Nome essa ação GetResults, Use a conexão JNDI mesmo, e digite o seguinte consulta:

```
SELECT   SUM (real) SUM, REAIS (ORÇAMENTO) ORÇAMENTO, SUM VARIÂNCIA (variação)
DA       QUADRANT_ACTUALS
ONDE     REGIÃO = '{região}'
```

Note que usamos {Região}, Que é uma referência ao nome da região de a iteração atual do loop. Existem quatro regiões para essa consulta será executada quatro vezes.

ATENÇÃO Todos os apelidos e nomes de coluna do conjunto de resultados (parâmetros) em um Relacional ação deve ser no mesmo processo ou Pentaho irá gerar uma mensagem de erro.

7. Antes de criar o próprio e-mail, você precisa dar alguns passos.
Primeiro, adicione o resultado as colunas com a segunda ação relacional se não tenham feito isso ainda.

8. Agora note que, mesmo que este conjunto de resultados recupera apenas uma única linha, você ainda
necessidade de adicionar outro circuito PDS, porque não pode saber que você recuperou
apenas uma única linha. Este segundo ciclo, é claro, execute apenas uma vez para cada iteração do laço externo. Dentro deste loop interno, é necessário para preparar o texto do assunto e uma mensagem utilizando um Modelo de Mensagem. Dentro dessa ação, uma mistura de texto fixo e os parâmetros podem ser utilizados, resultando em um texto gerado dinamicamente.
Acrescentar uma mensagem modelo eo nome deste FormatSubject. Use o texto como fonte do template e insira Resultados para a região {região} como texto. A nome de saída é o nome pelo qual este texto pode ser reconhecido na seguintes ações, por isso vamos usar MailSubject aqui.

9. Agora adicione outra mensagem modelo. Chame esse um FormatTextE uso MailText como o nome de saída. O texto que você digitar aqui o e-mail completo corpo do texto, incluindo os parâmetros dos resultados recuperados, o que pode ser observado na Figura 4-10.

10. Finalmente, você pode adicionar uma ação de e-mail, que pode ser encontrado sob a Enviar
Para ações. Nome este Enviar e-mail Região e usar o <from> string
parâmetro de entrada que você criou em primeiro lugar no campo. Em um cenário da vida real,
você usaria o <EMAIL> parâmetro no campo, mas porque estes
são endereços falsos, neste caso, use seu endereço de e-mail próprio ou selecione o <from> parâmetro se você digitou o endereço de e-mail próprio lá. Em o campo Assunto, selecione <MailSubject> e no texto da mensagem, selecione <MailText>. Agora, a ação pode ser salvo e executado. Se tudo foi digitado corretamente, você deve receber quatro e-mails com resultados diferentes.
11. Para completar o cenário, você precisará adicionar algumas etapas extra. Em primeiro lugar,
adicionar um Se declaração para verificar se a região Central. A condição para add (REGIÃO == 'Central') É mostrado na Figura 4-10. Depois, você pode adicionar o GetTotals Relacional ação para recuperar os resultados gerais. E, assim como adicionado um loop de acção para a região de resultados, você adiciona um outro ciclo aqui também com um modelo de mensagem separada e-mail mensagens. A fluxo de conclusão é exibida na Figura 4-10. A imagem também mostra que o Se instrução utiliza os operadores de comparação Java estilo == para igualdade e ! = para a igualdade não.

DICA Você pode criar e adicionar suas seqüências de ação própria como modelos para o projeto Studio salvando o arquivo *. xaction para o diretório do PDS modelo. Ele está localizado na <eclipse instalar o diretório / plugins / org.pentaho.designstudio . Number> <version editors.actionsequence / templates.

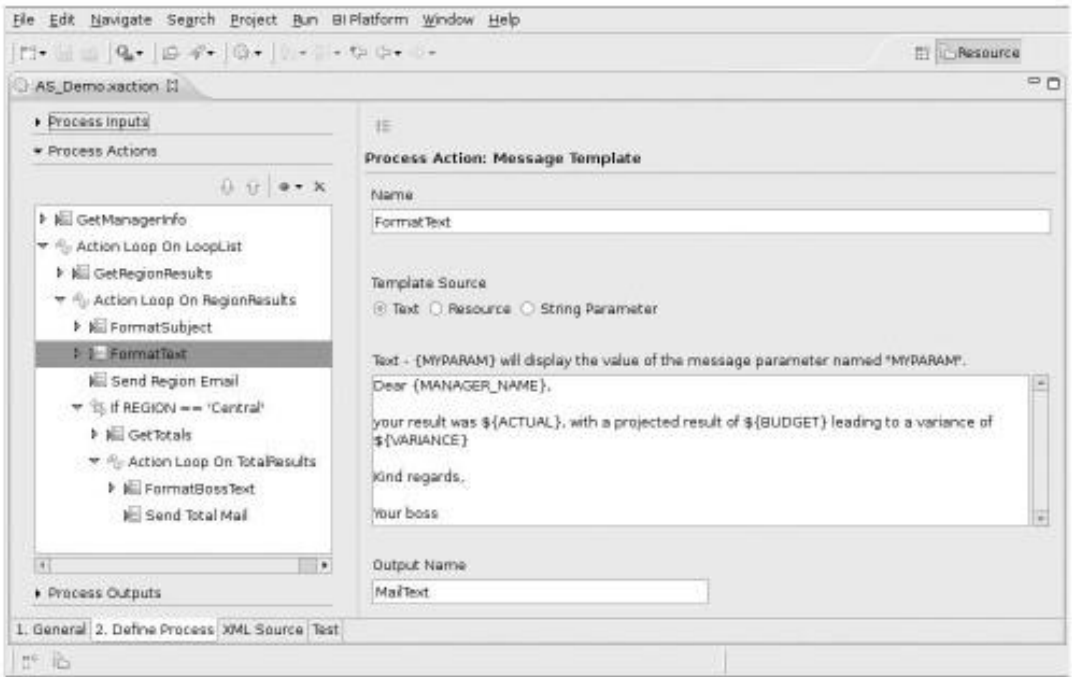


Figura 4-10: Terminado o fluxo de trabalho

Resumo

Este capítulo oferece uma visão panorâmica completa do Pentaho BI pilha e todos os componentes constituintes. Os tópicos abordados neste capítulo, incluiu a seguinte:

- A natureza aberta da plataforma, permitindo que as ferramentas não Pentaho ser integrada
- programas de servidor, cliente e desktop para ser utilizado por designers, administradores e usuários finais
- A arquitetura de servidor Java servlet baseado em plataforma
- O Community Edition da Pentaho e os recursos extras do Enterprise Edition
- Uma visão geral da e introdução ao Eclipse e do Projeto Pentaho Studio

A parte final do capítulo consistiu em um passo-a-passo para a criação de um relatório de ruptura seqüência de ação.

Parte

II

Dimensional e Modelagem Data Warehouse Design

Nesta parte

CAPÍTULO

5

Business Case Exemplo: Filmes Classe Mundial

Os exemplos no restante deste livro são baseadas em uma empresa fictícia chamado World Class Filmes (WCM). WCM é uma oferta firme de varejo em linha tanto as vendas e aluguéis de filme; esta combinação define a empresa além de varejistas online como a Amazon, que apenas vender itens, e empresas como Netflix, onde os filmes podem ser alugados ou vistos on-line.

Por que nós escolhemos uma loja de filmes online? mais Primeiro, é um exemplo que as pessoas podem facilmente relacionar-se: nós amamos filmes, lemos sobre os novos filmes, que se-
baixo''notícia''sobre os atores que a estrela no cinema, assistir filme e revisão programas na televisão. E, claro, todos nós estamos familiarizados com todos os pedidos os tipos de coisas fora da web. Em segundo lugar, a vantagem técnica de usar um retalho online
Exemplo disso é que todas as transações podem ser ligados a um cliente chamado ou identificado,
para que possamos evitar os desafios apresentados por ponto de venda comum vendendo itens
a clientes anônimos. Finalmente, uma quantidade razoável de dados é necessária para ilus-tram alguns dos conceitos relacionados ao armazenamento de dados, business intelligence e análises. Isso significa que precisamos de muitos clientes, muitos individuais pro-dutos, e um lote de transações, o que coincide maravilhosamente com os escolhidos exemplo.

NOTA Um exemplo como este nunca pode cobrir todos os meandros de uma empresa real ou organização, para estar preparado para correr em diversos outros tipos de serviços e processos de negócios no mundo real. Departamentos como finanças, produção, controle de qualidade, TI, RH e todos têm os seus próprios processos de negócios e de apoio sistemas e interagir uns com os outros e com o mundo exterior em seus próprios maneiras particulares. Também esteja preparado para encontrar processos específicos da indústria, tais como
processamento de sinistros (seguros), ambiental, saúde e práticas de segurança (Indústria química), ou gerenciamento de risco (bancário), cada um apresentando seus próprios

desafios quando se trata de processo e modelagem de dados e informações de gestão.

Filmes Classe Mundial: O Básico

World Class Filmes começou a vender e alugar DVDs on-line em abril de 2000 e tem mostrado um crescimento constante desde então.

O modelo de negócio WCM inclui dois processos totalmente integrados: cus- atendimento de pedidos tomer e reposição de estoque, como evidente a partir do seguinte descrição do negócio do WCM.

Um cliente pode encomendar um DVD a partir de um catálogo na web WCM e vê-lo sempre que ele gosta. Se o DVD for devolvido dentro de um determinado período de tempo, é considerado um""transação de aluguel, que, se o DVD é mantido ou não voltou no tempo, é considerado um''comprar''transação. A chave para o modelo reside no fato de que o DVD é inicialmente pago como se fosse comprado, e ao retornar o filme no tempo, a diferença entre as vendas e preço do aluguer é adicionado ao saldo da conta do cliente para encomendas posteriores. Os clientes são obrigados a tornarem-se membros e pagar uma taxa de entrada antes de estão autorizados a pedir um filme. Se um cliente devolve um DVD depois do aluguel período, o filme já está marcado como uma compra, mas o item adquirido está novamente em um armazém WCM e precisa ser enviado de volta para o cliente. Neste caso, a taxa de entrada é usado para cobrir o transporte extra e manipulação despesas.

Para estimular os clientes a comprar / alugar mais produtos, WCM usos diversos tipos de promoções.

WCM opera diferentes sites dirigidos a diferentes grupos de clientes de modo que uma variedade mais refinada pode ser oferecido para cada segmento de mercado, embora os clientes podem adquirir produtos através de múltiplos canais:

- Filmes Classe Mundial Portal-Este é o site principal da empresa, com um ampla oferta de filmes de sucesso e favoritos de todos os tempos, excluindo as últimas lançamentos dos grandes estúdios.
- WCM Premium Premium é um site de alto nível onde os clientes são cobrar, mas são a garantia de receber um novo shrinkwrapped DVD. O site contém apenas os filmes mais recentes e os maiores blocos busters.
- WCM-O Outlet Outlet é o local de negócio, onde os clientes podem obter ''Filmes''utilizado que poderia ter sido arrendado várias vezes, mas já estão disponíveis com um desconto.

- WCM Cool-Este site é destinado a um público mais jovem e mais na moda.
- WCM Exclusive-A site exclusivo oferece edições especiais e importados itens.

Esta divisão em múltiplos canais, cada um com seu público alvo e planos de preços, permite WCM para manter seu estoque movendo rapidamente.

O back office do negócio consiste em uma série de armazéns espalhados todo o país. Quando uma ordem do cliente é colocado, os itens são ordenados enviados a partir do próximo depósito para minimizar o custo de distribuição e transporte tempos. WCM começou com um único armazém, mas, porque o negócio tem crescido ao longo dos anos, foi considerado mais econômico para adicionar vários pontos de distribuição.

A sede da empresa ainda estão localizados no mesmo local que o Armazém da Califórnia, onde começou WCM. Todas as ordens de compra colocadas nas distribuidoras diferentes oriundos desta sede, onde cada ordem especifica o depósito das mercadorias devem ser entregues.

Os dois principais processos de negócios para a empresa podem ser resumidas como seguinte forma:

- Cliente de atendimento de pedidos lida com pedidos de clientes individuais e navios / recebe DVDs de e para armazéns diferentes.
- Reposição de Estoque abrange centralizada e descentralizada ordenação recebimento de mercadorias em armazéns diferentes.

E, claro, para tornar a empresa realmente fazer alguma coisa, as pessoas estão necessários, bem como, para que os funcionários e descrições concluir a ronda de a descrição do negócio de alto nível.

O fluxo de encomendas e produtos entre os distribuidores, WCM, e os clientes é ilustrada na Figura 5-1.

Os dados WCM

Sempre que você embarcar em um projeto de business intelligence, é imperativo que você compreender a origem ea natureza dos dados que serão utilizados na data warehouse. Sem esse conhecimento, é quase impossível de se conceber e construir um sistema que irá suportar o negócio em análise e elaboração de relatórios sobre os dados para melhorar o desempenho empresarial.

World Class Filmes usa dois bancos de dados para suporte às suas atividades, uma para o operação de back-office (gestão de armazém, compras, RH) e um para os diversos sites (cadastro de clientes, vendas). gestão do produto é vinculadas aos processos e WCM faz uma abordagem interessante para isso. Em vez

de ter os empregados inserir manualmente as informações sobre cada produto, o
a empresa oferece, WCM usa uma alimentação de dados externos para o produto da empresa
catálogo. A única coisa que WCM adiciona a esses dados é seu próprio produto interno
IDs para vincular as informações nos sistemas internos para os dados do exterior
fonte. Além do catálogo do filme, que contém detalhes sobre cada peça
de inventário, WCM usa o ISO 639 e 3.166 mesas para o código eo nome do
idioma, país e estado (região).

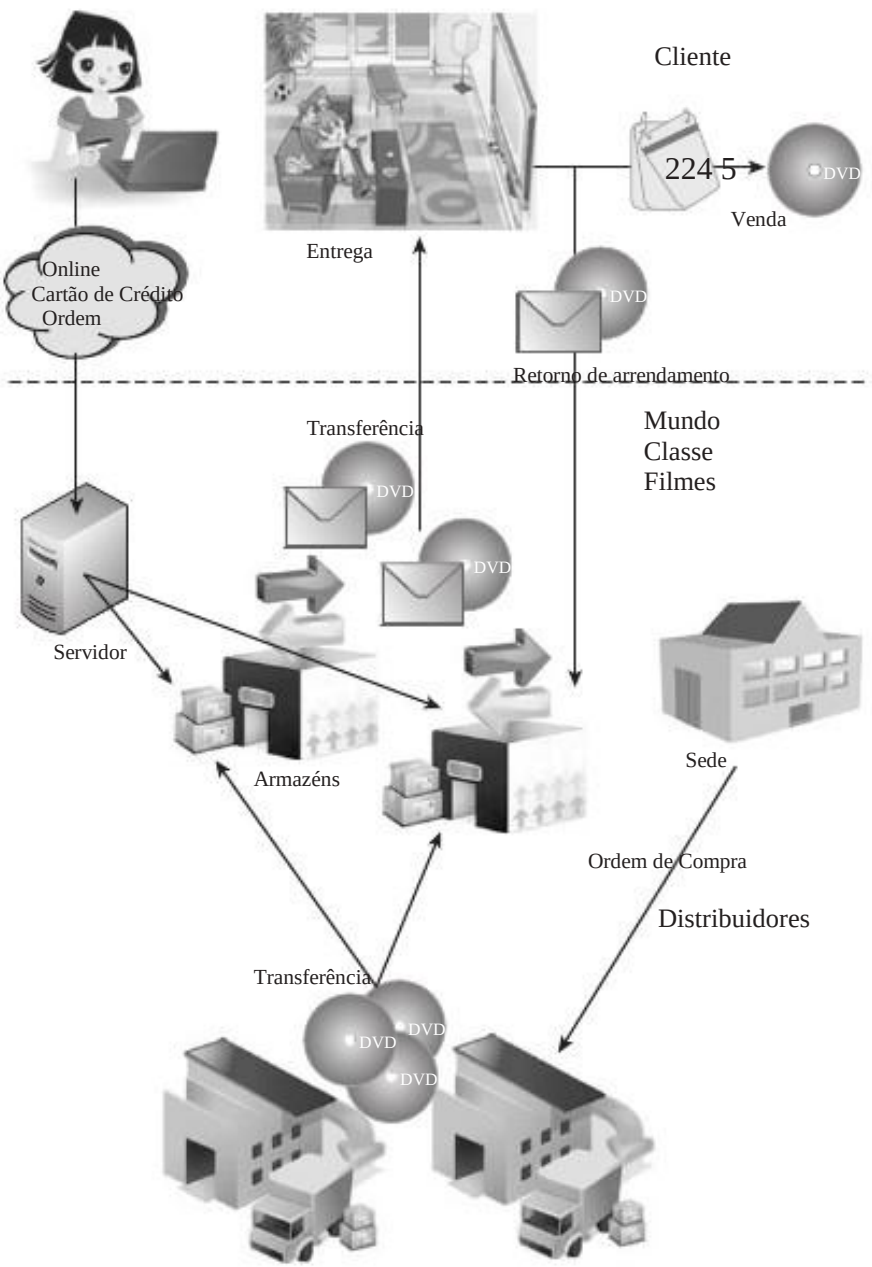


Figura 5-1: fluxos do processo principal no World Class Filmes

ISO: International Organization for Standardization

ISO define e gerencia as normas em uma grande variedade de temas, a maioria nomeadamente a família ISO 9000 de normas de gestão da qualidade. Mais de 17 mil padrões internacionais têm sido desenvolvidos até agora e mais de 1.100 são adicionados a cada ano. Entre os padrões mais conhecidos são: ISO 9660, que define o formato de arquivo ISO de imagem que é usado para gravar CDs e DVDs e ISO 12232:2006, que define a velocidade do filme em digital câmeras. A maior vantagem de usar um padrão para dados tais valores como datas ou nomes de países é a conformidade das menções entre os vários sistemas, o que garante a compatibilidade e fácil tradução de um sistema para outro. A melhor prática é utilizar os dados que se conforma a um padrão ISO em seus sistemas de origem, sempre que possível. Quanto mais o sistema de origem é e faz uso padronizado das definições de dados uniformes, mais fácil a tarefa de construção de um data warehouse será.

Obter e gerar dados

Não há muitas boa amostra bases de dados disponíveis, ea maioria deles não pode ser utilizado sem restrições ou fazem parte de uma oferta de banco de dados comerciais. Pessoas familiarizadas com o MySQL pode conhecer o banco de dados exemplo Sakila, que serviu como ponto de partida para o banco de dados WCM, mas está um pouco mais simples e
contém dados muito pouco para ilustrar os conceitos deste livro. Por esta razão, decidimos desenvolver um livremente disponíveis, dados de exemplo LGPL-licenciados com complexidade suficiente e uma quantidade real de dados em que seja útil como uma fonte para um projeto de data warehouse.

Ao desenvolver um modelo de dados instrutivo é um desafio em si, esta desafio é minimizado através da criação de dados significativos para ele. Felizmente, existem algumas fontes de dados acessível ao público, tais como o Censo dos EUA, o nome falso Gerador, eo Home Theater Info catálogo de DVD que podemos usar para clientes, funcionários, produtos e informações externas. Todos os outros dados no banco de dados seja criado manualmente ou gerados usando scripts. A base de dados conjuntos e os scripts para criar o esquema e criar os dados da transação pode ser baixado do site companheiro deste livro.

WCM Database: The Big Picture

Antes de explicar cada uma das peças do modelo de dados WCM, apresentamos aqui uma visão global do banco de dados que será usado para os exemplos o restante do livro. A forma mais rápida e fácil de se familiarizar com o modelo de dados é a soma das diferentes entidades, relações e papéis desempenhados por cada entidade, que é a finalidade da lista a seguir. O diagrama na Figura 5-2 pode ser usado como uma referência.

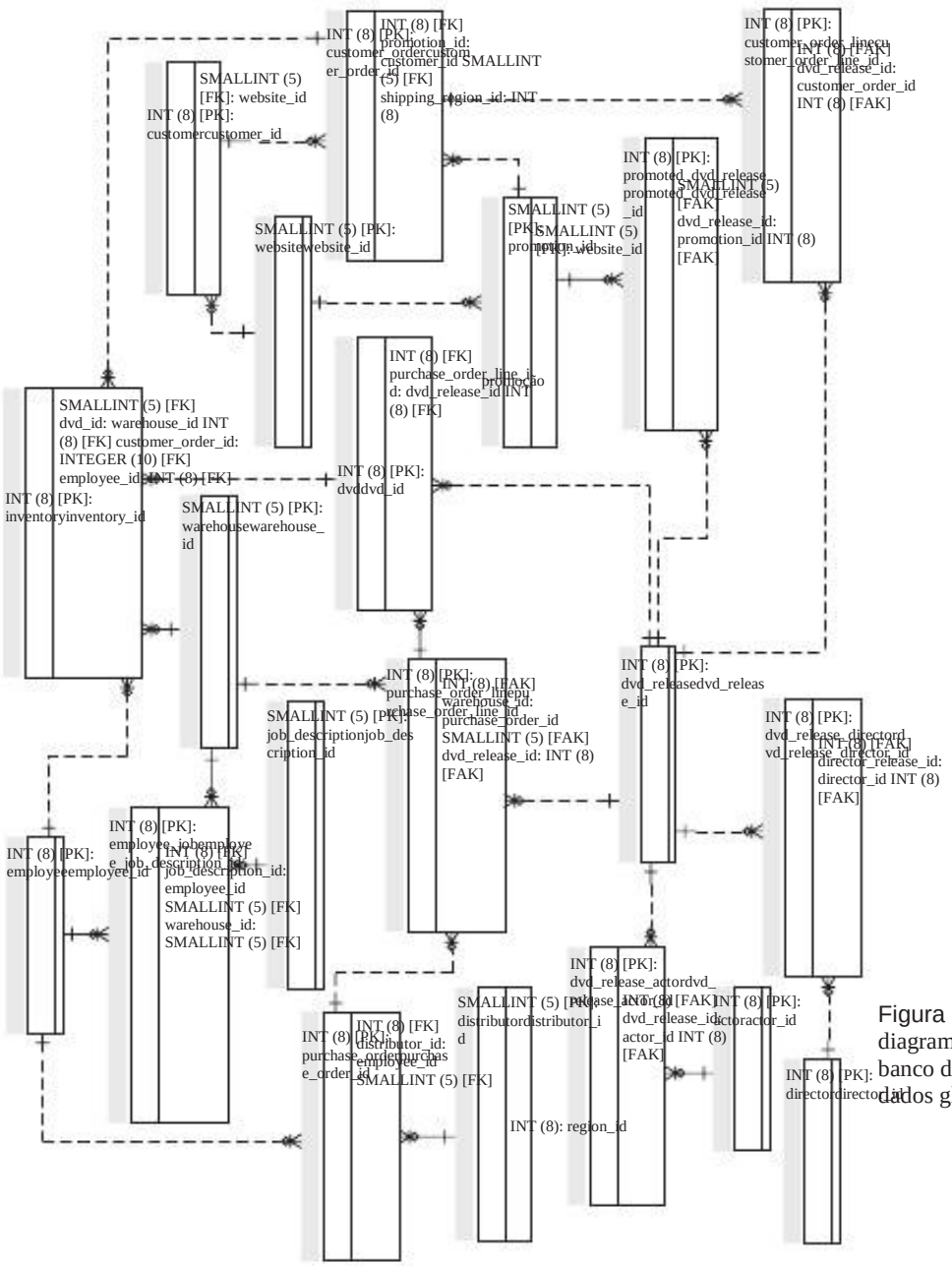


Figura 5-2: diagrama de banco de dados global

- A pedra angular do banco de dados é composto de clientes, produtos e ordens.
- Existem dois tipos de ordens: ordens de compra e ordens de clientes.
- Os produtos são adquiridos de distribuidores e recebidas em um armazém onde são verificados pelos funcionários.
- Cada ordem de compra é feita por um funcionário de uma distribuidora determinados pode ter uma ou mais linhas de pedidos de compra.
- Uma linha de ordem de compra é composto do item, quantidade e preço dos adquirida do produto e também denota o armazém de destino da produtos.
- Os funcionários têm uma descrição do trabalho e trabalhar em um armazém específico.
- Armazéns, funcionários, clientes e distribuidores têm um endereço.
- Cada endereço é localizado em uma determinada região, e uma região é parte de um país.
- ordens do cliente são colocados em um site e pode ter uma promoção que lhes são inerentes.
- Promoções pertencem a certos produtos (lançamentos em DVD) e pode ter um menor preço de venda, a um preço mais baixo de aluguel, um longo período de aluguer, ou uma combinação destes.
- A ordem do cliente consiste em uma ou mais linhas de ordem quando uma linha de pedido é criado para cada produto.

Nem todas as entidades estão visíveis no diagrama, todas as tabelas que contêm o endereço informação estão ligadas ao região e país. Ambos os quadros são deixados de fora para clareza, mas pode ser encontrado no diagrama detalhada do cliente. Duas outras tabelas não estão visíveis no diagrama são lookup_type e valor_procurado, Que contêm diferentes combinações de chave / valor par de informações como código de status e tipo de transação. Construções como estas (várias listas de não-relacionados em um único tabela de referência) são comuns em um sistema de Enterprise Resource Planning (ERP).

As próximas seções fornecem informações adicionais sobre cada parte do banco de dados esquema, incluindo o modelo de dados e conteúdo.

Catálogo de DVD

Existem várias opções para a obtenção de informações cinematográficas a partir da Internet, com o Internet Movie Database (www.imdb.com), Sendo provavelmente a mais bem conhecida e largamente utilizada fonte de informação do filme. WCM investigou a utilização do IMDB como fonte para o catálogo de DVD, mas encontrou a informação contidos na lista Laserdisc IMDB estar longe de ser útil. O melhor e mais fonte completa de informações acabou por ser o Home Theater site Info (www.hometheaterinfo.com), Onde a informação para todos os DVDs disponíveis a partir de

os estúdios podem ser encontradas várias, incluindo informações ator e diretor. WCM decidiu licença deste banco de dados para uso interno, bem como para o web de catálogo.

Cada título será considerado uma Lançamento de DVD, e embora haja uma distinção entre o conceito de um filme e um DVD, apenas este último está disponível na WCM banco de dados. A filme é o artefato que descreve a produção de Hollywood, e quando este filme é lançado em DVD, WCM cria um lançamento em DVD, que podem ser requisitados pelos clientes através da loja virtual. Os clientes são, em seguida, enviou uma
DVD físico, que são colhidos a partir do inventário. Então, teoricamente falando, existe um modelo de três camadas de dados (lançamento do DVD do filme em DVD), que denota uma relação mestre-detalhe detalhe entre as três entidades. Neste caso teórico atributos, tais como título, atores e diretor estaria ligado ao filme e atributos, tais como data de lançamento eo preço do aluguel seria relacionado a um lançamento em DVD. No entanto, o catálogo não tem WCM filme entidade e pois armazena todas as informações do filme disponível no nível de lançamento do DVD.

Adicionado a informação do DVD é de dados sobre os atores e diretores, que também é obtido a partir do banco de dados Info Home Theater. Isso permite que WCM os clientes a procurar filmes com um ator específico ou filmes que são dirigido por um diretor específico. O esquema completo catálogo de filmes é exibida na Figura 5-3.

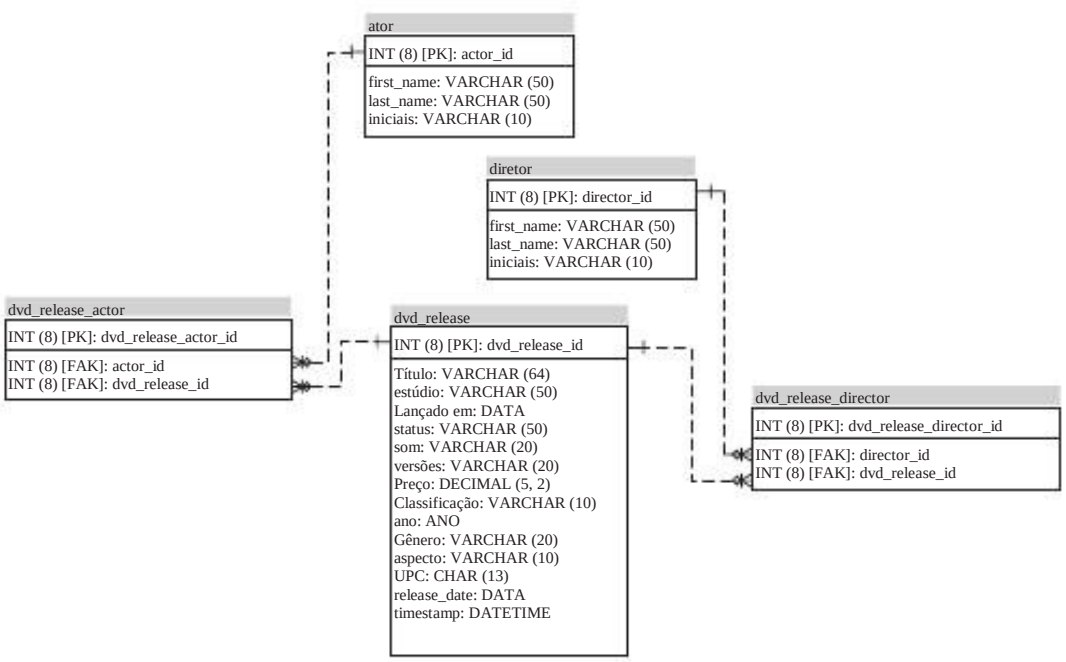


Figura 5-3: Catálogo de Filme modelo de dados

NOTA Como esses dados são carregados de uma fonte externa, o modelo de dados é aprovada, resultando assim em um modelo menos elegante. Uma solução melhor seria a

mesclar atores e diretores em um Pessoa tabela e adicionar uma tabela de ligação entre DVD liberação e Pessoa. Esta tabela ligação poderia ser estendida aos campos do pavilhão como is_actor e is_director (E no futuro talvez is_producer, is_music_composer, etc)

Clientes

Como a maioria das organizações comerciais, WCM tem clientes. E porque os produtos devem ser enviados para um endereço de e pago por cartão de crédito, WCM tem um monte de informações sobre seus clientes. Esta informação inclui o endereço, código postal, data de nascimento e sexo, o que torna a dados de clientes muito adequado para todos os tipos de análise. O banco de dados WCM para data contém informações sobre mais de 145.000 clientes que subscreveram o webstores diferentes ao longo dos últimos oito anos. Claro, isso não é exatamente verdade: a verdadeira história é que essa coleta de dados do cliente é gerado aleatoriamente pela linha nome falso Generator (www.fakenamegenerator.com), Onde dados do cliente pode ser gerado por uma série de países em lotes livres com um tamanho máximo de 40.000, ou podem ser comprados em lotes de um milhão de nomes. A aspecto interessante dos nomes gerados é que eles não são realmente aleatórios, mas representante de determinados padrões demográficos. Por exemplo, você verá mais pessoas a viver na cidade de Nova York do que em Zwolle, Louisiana, o que faz os dados perfeitamente adequado para o banco de dados demo WCM. A cliente tabela também referências a país e Estado as tabelas, pois (novo) os clientes são apenas permissão para selecionar de uma lista fixa de valores para evitar erros de entrada de dados. A última referência diz respeito ao site onde o cliente originalmente aplicada para uma conta no WCM. Figura 5-4 mostra o modelo de dados completos dos clientes.

Empregados

O empregado modelo de dados é simples, mas permite a mudança de trabalho dentro a empresa e até mesmo para realocação dos empregados para armazéns diferentes. Figura 5-5 mostra o diagrama de empregado.

WCM tem um sistema de RH separado que inclui todos os outros agentes relacionados com infor- informações, tais como salários, tipos de contrato, a ausência, os planos de educação, e assim por diante. sistemas de RH são notoriamente complexa para recuperar dados e não estão cobertos neste livro. A fonte das informações do funcionário utilizado é o mesmo que para clientes e é composto por um subconjunto do conjunto gerado nome falso.

As ordens de compra

O processo de compra é bastante simples, WCM: uma ordem de compra é colocada no um distribuidor por um determinado funcionário e contém um ou mais de ordem de compra

linhas. Cada linha de ordem de compra contém uma série de lançamentos de DVD ordenados de um armazém específico. As Relações Diagrama Entidade (ERD) é mostrado na Figura 5-6.

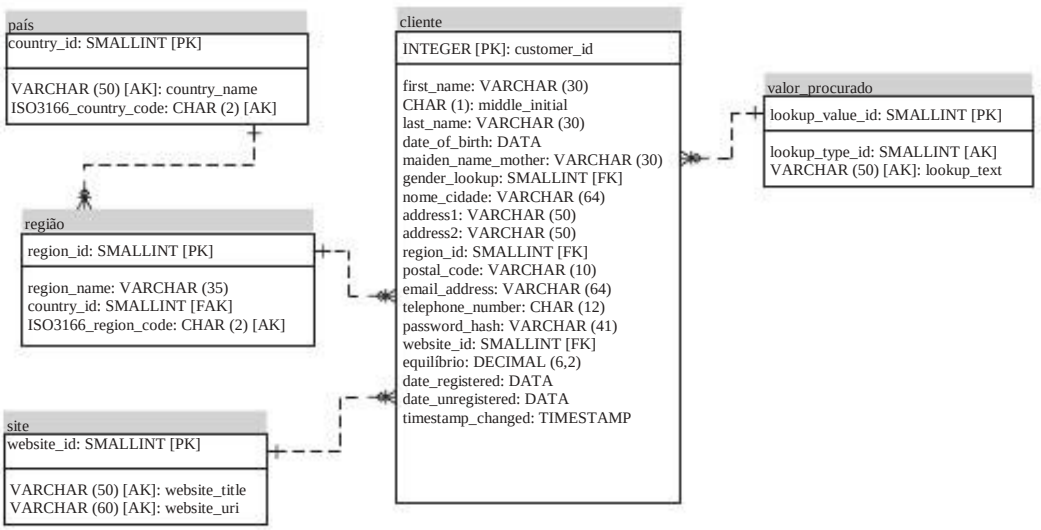


Figura 5-4: Cliente modelo de dados

A linha de ordem de compra também contém o preço de compra obtido a partir de o lançamento do DVD tabela no momento da encomenda, o custo do transporte, e os datas para o transporte, espera, entrega e entrega real. O preço não é história mantida no lançamento do DVD de dados, mas o preço só atualizou lá. A Histórico de preços, no entanto, sempre podem ser obtidos a partir das linhas de encomenda. Note que qualquer alteração de preço intermediário são perdidas dessa forma, por exemplo, quando nenhuma ordem existe em um ponto determinado preço. Além disso, note que este é um modelo simplificado que ignora o fato de que no mundo real, vários distribuidores podem oferecer o mesmo produto a preços diferentes.

Pedidos de clientes e Promoções

Clientes encomenda online e DVDs a aplicação web torna-se que estes pedidos são inseridos no a ordem do cliente e linha do pedido do cliente tabelas. Promoções são usados por WCM para estimular as vendas adicionais ou para limpar ações redundantes. Promoções ativas são convertidas automaticamente para banners e anúncios em diversos sites. Quando uma promoção é selecionado diretamente ou DVDs estão ordenados que pertencem a uma promoção, a ordem correta dos clientes linhas com os lançamentos com desconto de DVD são adicionados ao banco de dados na web pedido. Em outros casos, os clientes poderão encomendar um DVD único que pertence com uma promoção em curso e neste momento a opção é oferecida para selecionar o promoção completo (que pode consistir de vários DVDs).

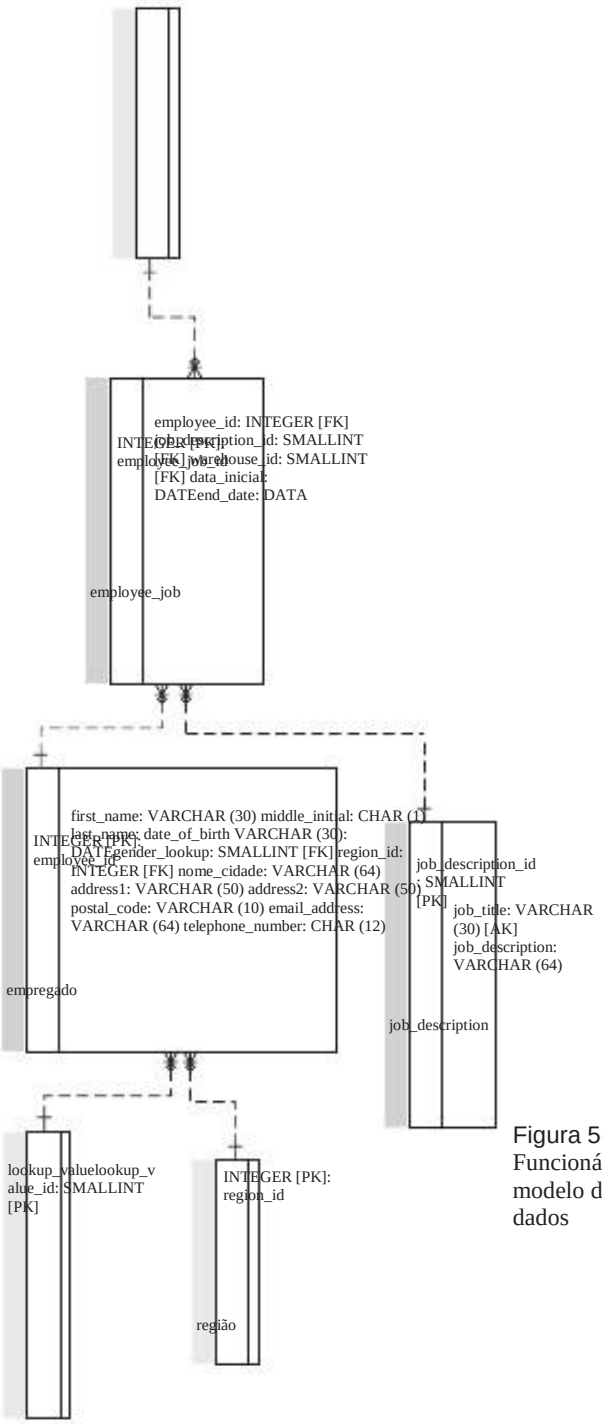


Figura 5-5: Funcionário modelo de dados

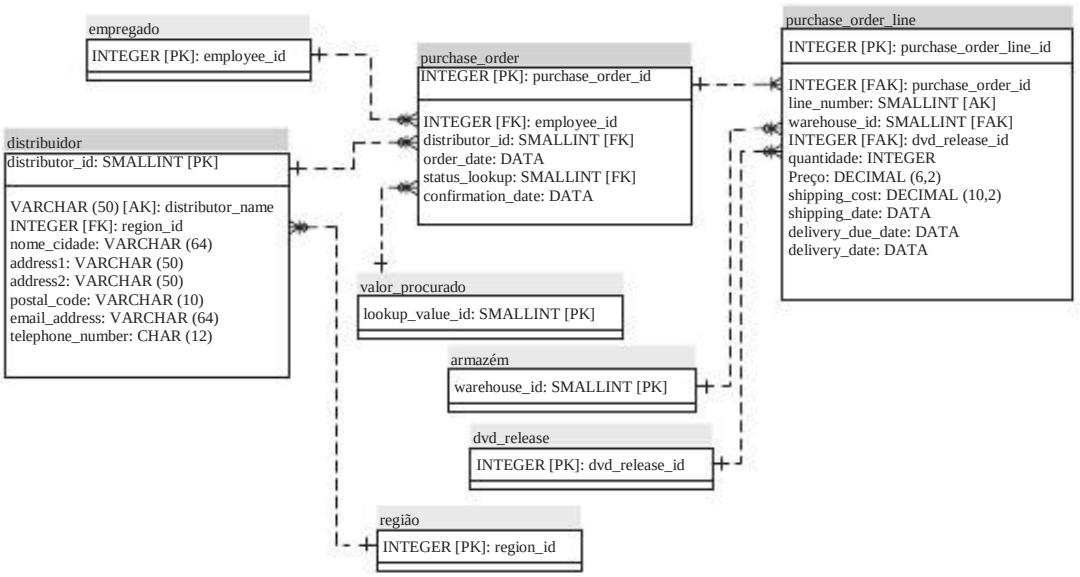


Figura 5-6: ordem de compra do modelo de dados

O que pode parecer estranho inicialmente, é a duplicação do nome do cliente e dados de endereço. Isso garante que WCM sempre tem uma história totalmente rastreável de dados do cliente. O endereço do cliente podem ser atualizados na cliente mesa, mas as ordens sempre refletirá o endereço de um item foi enviado para.
A ordem do cliente e parte de promoção do esquema do banco de dados é exibido na Figura 5-7.

Note que neste modelo um campo quantidade da ordem não está disponível, levando a a limitação de que apenas um item de lançamento de um DVD específico pode ser encomendado na
ao mesmo tempo.

Gestão de Stocks

WCM precisa controlar seu estoque e quer gravar toda a história do produto também. Essa informação é combinada no armazém,DVDE inventário tabelas. Cada DVD físico que é recebida de um distribuidor é registrado e adicionado à DVD e inventário tabelas. A entrada de DVD obtém o status novo, o que significa que ele está disponível para venda ou aluguel. A entrada de estoque fica o mesmo
novo status, o que significa que o novo DVD é adicionado ao estoque de WCM. Após este evento, o inventário tabela reflete a história de cada transacção.
Um DVD pode ser enviado,devolvidoOu vendidos. Quando um item retornado despeja ser danificado, ele obtém o status lixeira. Embora o status atual de um item pode ser recuperada a partir do inventário tabela, a escolha é feita de duplicar o situação atual no DVD mesa para fácil referência. Desta forma, WCM é capaz de relatório sobre os níveis de estoque atuais e históricos. Figura 5-8 mostra a parte da diagrama entidade relacionamento com o submodelo de gestão de inventário.

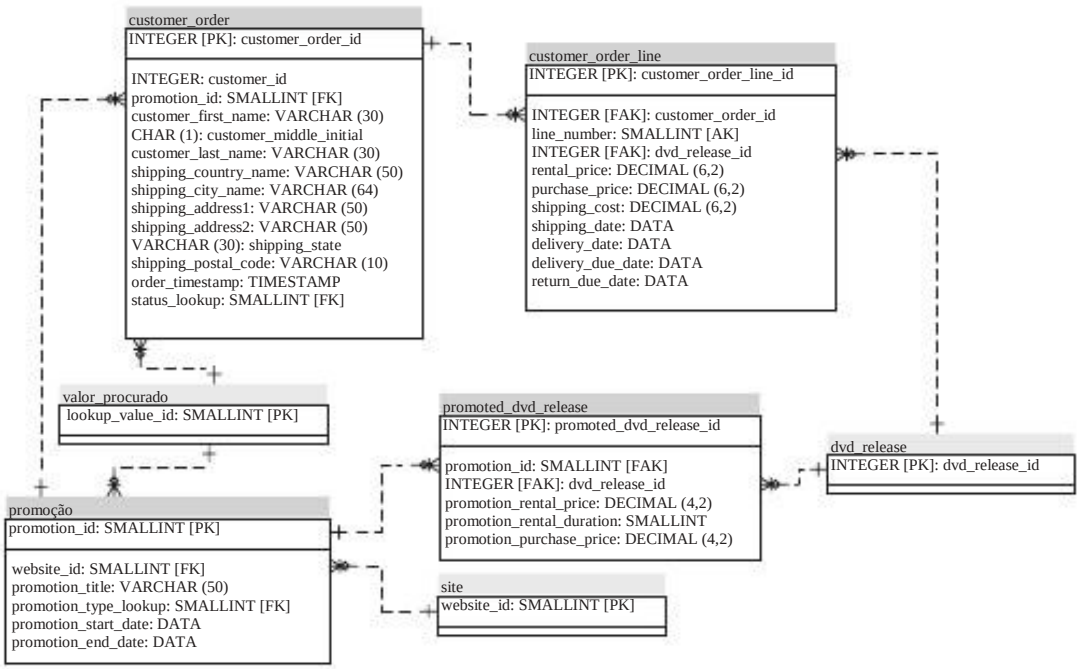


Figura 5-7: ordem do cliente modelo de dados

Cada DVD tem um local físico em cada momento, indicada pelo ID do armazém onde o item é enviado ou devolvido. Normalmente um produto é enviado a partir do mesmo armazém, uma vez que é devolvido ao presente, mas não precisa ser necessariamente o caso. Outras informações que podem ser obtidos da inventário tabela é:

- Cliente (Customer_order_id) Mostra-cliente, o item é enviados para, vendidas ou retornado.
- Empregado (employee_id) Mostra-o agente do warehouse enviados ou recebidos do DVD.
- Timestamp -A data ea hora exatas as informações são inseridas no do sistema.

Gestão do Negócio: A finalidade do negócio Inteligência

Gerir um negócio é realmente muito simples quando você toma uma missão orientada vista. WCM foi criada porque houve uma evidente necessidade de uma conveniente caminho para comprar e alugar DVDs sem ter que sair de uma loja. A missão é se tornar a maior distribuidora de DVD on-line no país, e vários passos intermediários foram definidos para finalmente alcançar esse objetivo (criar o negócios em um estado, expandir a base de clientes, adicionar diferentes canais de vendas).

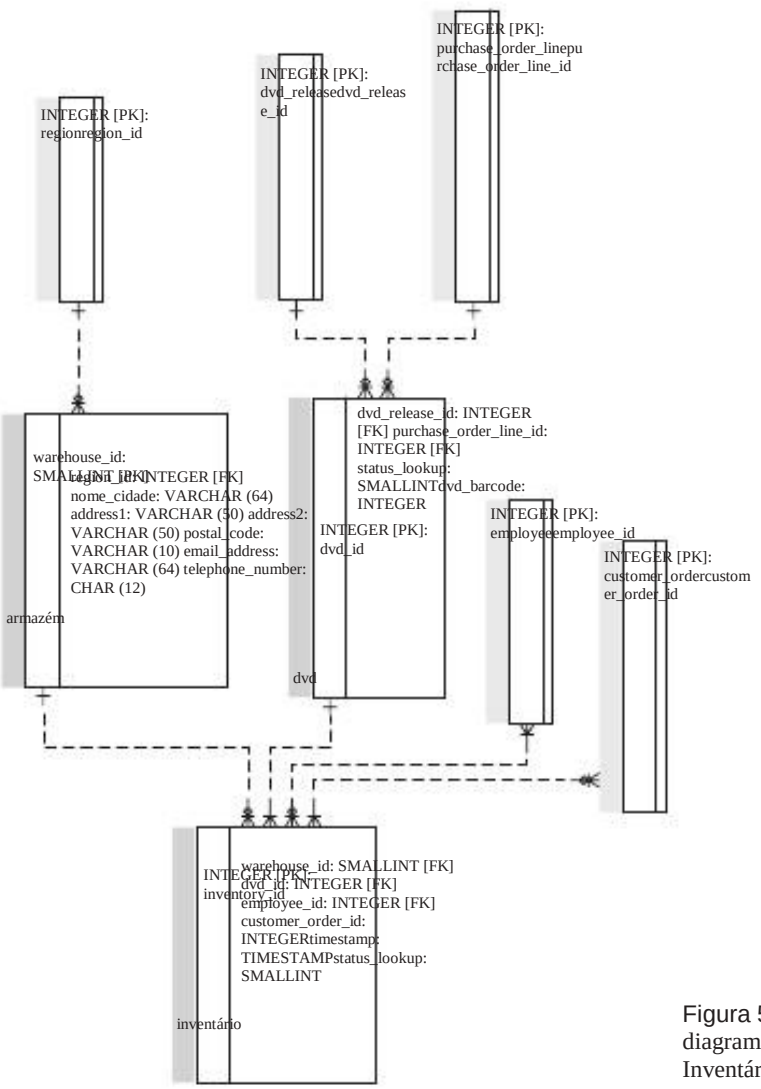


Figura 5-8: diagrama de Inventário

Ao longo do caminho, WCM precisa acompanhar e analisar o seu desempenho para ver Que fatores estão contribuindo para o sucesso e que são desvirtuando, e que é onde a inteligência de negócios chutes pol

BUSINESS INTELLIGENCE

Howard Dresner, analista sênior de longa data do Gartner Group, é muitas vezes creditado com a definição do termo business intelligence (BI). Embora Dresner popularizou o termo, ele não foi o inventor. Na verdade, o termo de negócios inteligência foi usada pela primeira vez em 1958 por Hans Peter Luhn na IBM Jornal. No entanto, sua definição Dresner de BI que é usado na maioria das vezes hoje:"Conceitos e métodos para melhorar a tomada de decisões comerciais usando sistemas de suporte baseado em fatos."Esta definição descreve de uma forma muito maneira elegante que BI é muito mais que tecnologia por si só (fato que é frequentemente negligenciado por muitos dos praticantes BI) e que o BI é um meio para um fim, não o objetivo em si.

Perguntas Business Intelligence típica para WCM

Por uma questão de argumento, vamos supor que de Classe Mundial Filmes corre muito gestão eficiente organização corporativa, mas não tem o discernimento necessário para levar a empresa para o próximo nível. organização WCM contém departamentos tais como Finanças e Vendas, Logística e Compras, Atendimento a Clientes, e Marketing e Gestão de Produto e, claro, tem algum tipo de interesse geral gestão e um CEO. Estas pessoas e todos os departamentos têm diferentes precisa para executar sua parte do negócio melhor e tenho algumas perguntas em comum também. As perguntas típicas podem incluir:

FINANÇAS E VENDAS

- Como a receita que vamos gerar na região, por mês, e do filme categoria?
- Quais as categorias de filme gerar a maior receita e é essa constante ao longo do tempo?
- Como o nosso desempenho para comparar o mercado de entretenimento total?
- Nós estamos crescendo mais rápido ou mais lento do que nossos principais concorrentes?

Logística e AQUISIÇÃO

- Como nossos distribuidores pontuação em termos de variedade de produtos, preço e desempenho da entrega?
- Como podemos otimizar ainda mais nossos custos de distribuição?

ATENDIMENTO AO CLIENTE

- Quantas queixas lidamos?
- Que tipo de reclamações que os clientes costumam ter?
- Qual é a média de processos por representante de serviço?

NOTA Serviço ao cliente é adicionado como um exemplo extra e não é coberto pelo modelos de dados de amostra

MARKETING E GESTÃO DE PRODUTO

- Como o valor da vida dos 100 principais clientes se comparam aos fundo 100?
- Como podemos segmento de nossos clientes com base na RFM (recência, freqüência análise) monetária?
- Não temos os dados do cliente que pode ser usado para indicar rentabilidade futura ou churn?
- Podemos prever receitas futuras para um novo cliente com base na actual perfis de clientes e características, tais como código postal, idade ou sexo?
- Como podemos manter o controle do ciclo de vida de um produto e que as vendas canais devem ser usados (combinações de mercado do produto)?
- Lançamentos em DVD que são mais susceptíveis de gerar receita com base em alta características do produto, como ator, diretor ou gênero de filme?

A partir desses exemplos, é óbvio que algumas perguntas (e suas respostas) referem-se múltiplas áreas de negócio. E, novamente, estes exemplos são típicos de um organização comercial venda de produtos aos consumidores, mas provavelmente não são aplicável a outras indústrias, por exemplo, uma organização de saúde, como um hospital.

Dados é fundamental

Nenhum sistema de business intelligence pode existir sem os dados, e as ques-amostra ções desde já poderia ter mostrado que os dados internos por si só não suficiente para obter uma boa compreensão do desempenho de uma organização. A maioria dos sistemas de BI começam com um foco interno, elaboração de relatórios sobre dados de vendas passadas. Este é ótimo para começar seus pés molhados, como você verá nos próximos capítulos, mas é muito pouco em um ambiente competitivo. Tomemos, por exemplo, um organização que comemorou um crescimento de 10 por cento na receita líquida de vendas mais No ano passado, mas ignorou o fato de que o crescimento global do mercado foi 20 por cento. Eles realmente fizeram um trabalho muito ruim, mas ficou sem saber porque os dados externos que poderiam ter divulgado essa informação não foi usada. Para

pleno conhecimento sobre o desempenho real de uma empresa, os dados externos é mandatório. World Class Filmes, portanto, usa duas fontes adicionais de informações. A primeira é o Censo 2000 CEP conjunto de dados, que permite à empresa relacionar dados de clientes internos global de dados demográficos e trends.1 A segunda fonte de informação é a receita de dados históricos e-commerce obtidas no site E-Stats do Censo dos EUA, que pode ser encontrada em www.census.gov/eos/www/ebusiness614.htm.

Resumo

Este capítulo apresenta a empresa fictícia de Classe Mundial Filmes que nós criado para este livro. Com base neste exemplo, descrevemos a seguir:

- Os negócios principais processos da empresa Classe Mundial Filmes
- Um modelo global de dados do banco de dados de apoio aos processos
- O modelo de dados detalhados para cada entidade do negócio principal (clientes, produtos, pedidos, estoque, funcionários)

A última parte do capítulo ilustrou o propósito eo valor do uso soluções de Business Intelligence em geral e as respostas que uma solução de BI pode fornecer para uma empresa como de Classe Mundial Filmes.

[1] A Censo 2000 CEP conjunto de dados é reutilizado mediante permissão de A análise dos dados e SQL Excel, por Gordon S. Linoff, Wiley Publishing, Inc., 2008

CAPÍTULO

6

Data Warehouse Primer

Um data warehouse não é nada novo, na verdade, armazenamento de dados estava sendo praticado por anos, mesmo antes do termo foi cunhado por Bill Inmon em sua trabalho seminal Construindo o Data Warehouse.[1] Embora Inmon é muitas vezes referida como o pai"de data warehousing,"isso não é inteiramente o caso. Para dar crédito onde crédito é devido, foi o irlandês IBM arquitetos Barry Devlin e Paul Murphy, que, em 1988, lançou as bases para o que hoje chamamos de data warehouse. É interessante ver que o conceito de Dados Corporativos Armazém (BDW) Devlin e Murphy descreveu em seu artigo original não tem mudado muito nas últimas décadas. Eles definem o BDW como"o único armazém lógica de todas as informações utilizadas para informar sobre o negócio,"que ainda é exatamente o que é.

NOTA Sinta-se livre para apenas pular este capítulo, ou mesmo ignorá-lo se você já está familiarizado com os conceitos básicos de armazenamento de dados.

Para obter uma melhor compreensão da natureza de um armazém de dados que irá adicionar
descrição original Inmon também. Ele definiu um armazém de dados como sendo:

- Assunto orientadas para Todos entidades e eventos relacionados a um assunto específico
(Por exemplo,""de vendas) estão ligados entre si.
- Tempo variante Todos alterações nos dados são controlados para permitir os relatórios que mostra a evolução ao longo do tempo.
- Não volátil-Quando os dados são inseridos no data warehouse, nunca é sobrescritos ou excluídos.

[1] Cf. Construindo o Data Warehouse, 4 ª Edição, por WH Inmon, Wiley Publishing, Inc., 2005.

- Integrada-A armazém de dados contém dados de múltiplas fontes sistemas depois de ser limpo e conformado.

Ao longo dos anos, essas descrições (especialmente o aspecto""não-volátil) foram desafiadas e adaptada por outros autores e praticantes, levando para diferentes arquiteturas e formas de modelagem do data warehouse. Trata-se, No entanto, é bom ter em mente que todos os autores ainda concordam com o raciocínio atrás de um armazenamento de dados separados para análise de negócios e prestação de contas foi originalmente definido por Devlin e Murphy:

- Garantir que o desempenho dos sistemas de produção não é interrompida por consultas ad hoc ou de análises
- Exigir que as informações necessárias aos usuários finais não muda enquanto que o utilizam, ou seja, os dados point-in-time

Por que Você Precisa de um Data Warehouse?

Pessoas que nunca foram expostas ao conceito de um data warehouse são muitas vezes confusos sobre a necessidade e finalidade de um banco de dados especializados para fins de apoio a decisão. Mesmo depois que os dados óbvios benefícios integrados a partir de diferentes sistemas, o desempenho da consulta, aliviando os sistemas de código de consultas de longa execução eo acompanhamento da história, foram explicados, é ainda não é claro por que sempre a construção de um data warehouse é uma boa idéia. Bastante muitas vezes, esses usuários se acostumaram a recuperar ou a obtenção de dados de várias fontes, incluindo os dados enviados a eles por e-mail, que depois de importação em um aplicativo de planilha eletrônica que eles usam para posterior análise e relatórios. Não armazém de dados necessários, certo? Não é tão certo, na verdade. Vamos tentar explicar por um armazém de dados é útil, do ponto de vista de um usuário:

- Toda a informação está em um lugar no mais caçar várias dis- parate fontes de informação ou tentar encontrar arquivos com mais de um confuso E-mail do sistema ou estrutura da pasta. Não é preciso tanto para combinar tudo isso dados a si mesmo: ele já está integrado e pronto para uso.
- Up-to-date informações de dados no armazém de dados é automaticamente carregadas e atualizadas em uma base regular, o que significa que você nunca está fora de data ou procurar informações antigas.
- Acesso Rápido O armazém de dados é otimizada para recuperação rápida de da informação. O data warehouse respostas as suas dúvidas muito mais rápido do lojas locais de arquivo ou arquivos de e-mail.
- Não há limite de tamanho Planilhas pode armazenar somente uma quantidade limitada de dados e muitas vezes precisam ser divididos em pedaços para acomodar todas as informações necessárias.
  Um armazém de dados pode armazenar uma quantidade quase ilimitada de dados para que não

mais descarregamento de dados para um banco de dados local ou ainda uma outra planilha é
necessário.

- Toda a história disponíveis O armazém de dados não contém apenas corrente informação, mas também os dados da última semana, no mês passado, no ano passado, e vários
anos atrás também. Isto significa que a análise de qualquer tendência ou comparação ao longo do tempo é suportado pelo armazém de dados. Na verdade, se você nunca apagar
dados do data warehouse, muitas vezes contêm muito mais histórica informações do que os sistemas de origem. A história não é disponível somente "Dados mais antigos,"mas oferece valor adicional quando as alterações são controladas como
também. Isto permite-lhe olhar para os dados como ela realmente era durante a momento em que foi originalmente processado. Quando alguém vive em Boston em 2008
mas se muda para New York em 2009, você ainda vai ver os resultados de 2008 para este
cliente atribuído a Boston, não para Nova York.

- Fácil de entender-A Data Warehouse é modelada em termos de negócios e reflete o modo como você olha para sua organização. Você não precisa decifrar siglas de três letras que ninguém entende, mas pode ter
nomes claros para todos os elementos de dados.

- Definições claras e uniformes, no mais discussões sobre os dados que significa ou o que a definição de receitas""é. Todos na organização usa as mesmas definições, o que simplifica a comunicação.

- Todos os dados padronizados, dados em conformidade com as normas, o que significa que
é apenas uma definição e um conjunto de valores para cada peça de informação. Um bom exemplo disso é a codificação do gênero. Alguns sistemas usam 0
e um, alguns usam masculino / feminino e outro uso M / F / U (para desconhecidos). Todos
traduções em uma única definição padronizada foram atendidos.

Essa lista destaca as vantagens de um armazém de dados, mas é baseado em uma pressuposto importante: que o data warehouse é projetado e construído adequadamente. Mesmo quando a execução é de alto nível a partir de uma perspectiva técnica (Primeiro cinco pontos), ele ainda pode ser considerado um projeto fracassado de um perspectiva do usuário (os últimos três pontos de bala). O"fácil de entender,""clara e definições de dados uniforme, padronizada"e""vantagens são muitas vezes prédio em frente, especialmente quando o data warehouse é executado por uma TI departamento sem o envolvimento dos utilizadores suficiente.

Metadados: denominar e descrever SEUS DADOS

Se os últimos três argumentos para a construção de um data warehouse são invertidos, é também é possível olhar para o fim do armazém de dados de uma maneira diferente: como veículo para chegar a dados padronizados e definições claras e uniformes.
Isto não é totalmente um assunto de armazém de dados relacionados, mas tem um muito mais amplo

(Continuação)

Metadados: denominar e descrever SEUS DADOS (Continuação)

aplicabilidade, que é muitas vezes referida como metadados. Um simples e amplamente utilizado definição de metadados é dados sobre dados. Lotes de equívocos sobre a existir metadados que é ou deveria ser, mas sem querer simplificar demais, ele todos os se resume a uma questão: como o nome e descrever as informações em uma organização de tal forma que toda a gente compreende imediatamente o significado disso e este significado é o mesmo para todos os envolvidos? Se você pode superar este obstáculo no início de seu projeto de data warehouse, você estará pago de volta três vezes. Se você tomar um passo mais adiante, você verá que os metadados abrange muito mais do que apenas descrições dos dados (por exemplo, que se entende por 'Receitas'?). Os seguintes itens também são considerados metadados e são de especial importância em um ambiente de data warehouse:

- Linhagem de Dados-A informação sobre a origem e destino dos dados em cada etapa do processo de transformação de dados. Dados informações de linhagem fornece uma pista de auditoria completa dos dados em um data warehouse, que é essencial para cumprir os regulamentos compliancy, como a Lei Sarbanes-Oxley.

- Dados Oportunidade-A informação sobre quando os dados foram alterados e como "velhos" dados são um usuário está olhando. Muitas vezes, vários carimbos são utilizadas quando as informações são apresentadas: o tempo que um relatório é executado ou impresso, o tempo que os dados foram carregados ou alterada no data warehouse, eo tempo dos dados foi modificada pela última vez no sistema de origem.

- Modelo de Dados, os modelos utilizados no livro são também uma forma de metadados, geralmente chamado metadados estruturais uma vez que não fornecem uma descrição (Como nos itens anteriores), mas apenas a estrutura dos dados. O texto explicar o modelo de dados é o metadados descritivos neste caso.

O grande debate: Inmon Versus Kimball

Existe um consenso generalizado sobre a idéia básica de usar um armazenamento de dados especiais para apoiar as análises e relatórios. É a forma como este armazenamento de dados deve ser estruturado e organizado que tem sido objecto de muitos debates acalorados ao longo dos anos.

No início, havia basicamente duas abordagens para a modelagem dos dados armazém. Tudo começou quando os dois gigantes da indústria, Ralph Kimball e o referido Bill Inmon, começou a publicar seus dados e evangelizadora entreposto idéias. Cada grupo tem um (e às vezes até mesmo fanática) leais seguidores, o que contribuiu ainda mais para a discussão em torno das duas escolas de pensamento. Enquanto Inmon popularizou o termo armazém de dados e é um forte

proponente de uma abordagem centralizada e normalizada, Kimball tomou um rumo diferente perspectiva com a sua data marts e dimensões conformadas.

As principais diferenças entre o centro da abordagem Inmon e Kimball sobre três pontos. A compreensão destes pontos de diferença vai ajudar você a ganhar uma compreensão maior de armazenamento de dados em geral. (Você vai começar a abundância do exemplos de data mart neste livro, portanto, não se preocupe se essas descrições não afunda imediatamente.)

- Data warehouse versus data marts com dimensões adequadas, Nós já apresentaram definição de Inmon o Business Data Warecasa:"o armazém lógica única de todas as informações utilizadas para relatório nos negócios."Em contraste, um mart de dados contém informações relativas para uma função específica do negócio, como vendas ou quadro de pessoal. Este informações podem ser visualizadas a partir de perspectivas diferentes, chamados dimensões.
  Cada dimensão contém todas as informações relativas a um determinado negócio objeto, tais como uma agenda, clientes ou produtos, e pode ser ligado a um ou mais tabelas de fatos contendo itens mensuráveis (receitas, custos, número de funcionários, e assim por diante). O efeito final é que os usuários podem recuperar informações sobre as vendas por departamento, por cliente durante um período específico a partir de um data mart, mas a partir da mesma data mart não é possível recuperar informações tangencialmente relacionados, tais como número de funcionários empregados.
  Isso requer um data mart independente, que pode reutilizar alguns dos mesmos informações de dimensão (neste caso: empregado e de calendário) já utilizados na data mart de vendas. Devido as dimensões empregado e calendário ter uma aplicação que se estende além de um data mart de dados único, eles são chamados dimensões conformadas. O conceito de modelagem dimensional é coberto em profundidade no Capítulo 7.
- Abordagem centralizada versus iterativo / descentralizada abordagem como mencionadas, um data mart contém dados apenas para uma finalidade específica, enquanto um data warehouse contém todas as informações de forma integrada. A principal diferença entre um data warehouse Kimball e Inmon estilo é o facto de Kimball organiza seu armazém de dados como uma combinação de data marts integrados, enquanto Inmon considera um data warehouse como um modelo de dados integrado, normalizado que contém todos os dados necessários para relatórios e análises e usa data marts apenas para acesso ao usuário final.
  Isso soa como uma camada extra de complexidade, mas tenha em mente que, em Inmon um estilo de data warehouse todos os problemas relacionados com a conformação de dados e garantir correção histórica de dados são resolvidos na região central armazém, enquanto que em uma arquitetura de estilo Kimball essas questões precisam ser resolvidos dentro da data marts.
- Normalizada modelo de dados em função dos dados dimensionais modelo Se a commarts combinada de dados em relação ao armazém central eram a única fonte para o debate, a discórdia teria sido resolvido há muito tempo, mas há

outro, e talvez ainda mais contraste marcante entre os dois: o
questão da normalização versus desnormalização. Kimball introduziu o
técnica de desnormalização para as tabelas de dimensão. Agora um produto
pode ter um grupo de produtos relacionados a ele, que em um banco de dados
normalizado
seria armazenado em uma tabela separada e ligados através de chaves estrangeiras.
Para um
exemplo disso, dê uma olhada no WCM cliente tabela, que contém
um link para a região, que por sua vez, contém um link para país (ver Figura 6-1).
Em um esquema de banco de dados normalizado, cliente, região, país e são
armazenadas em três tabelas diferentes para garantir a integridade da região e
nomes de países (que são armazenados apenas uma vez). Um esquema de-
normalizados,
Por outro lado, as lojas de todas as informações em uma única tabela, assim creat-
ção de informações redundantes, que é uma maldição para o estritamente normalizado
acampamento. A Figura 6-1 mostra a diferença entre uma normalizada e um
desnormalizados esquema de banco de dados.

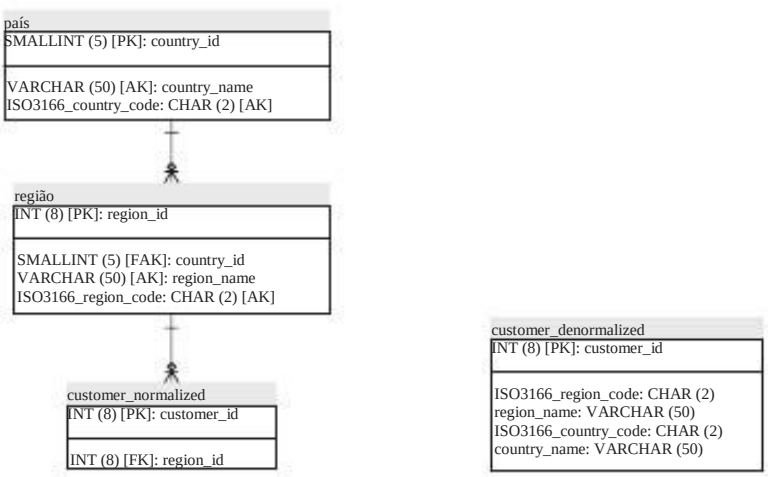


Figura 6-1: A normalização versus desnormalização

Ao contrário da crença popular, de-normalizar os dados em uma dimensão não é propenso
para os problemas usuais de não normalizar um banco de dados de transacção para
pela simples razão de que não é usado para transações. As tabelas de dimensão apenas
é atualizado através da utilização de uma extração, transformação e carregamento (ETL)
processo
que elimina os riscos de alguma forma envolvido com a atualização não-normalizada
de dados. Nós cobrimos o processo de carregamento dimensão no capítulo 10 e cobrirá
técnicas de modelagem dimensional no Capítulo 7. As partes subseqüentes deste
foco capítulo sobre arquitetura e tecnologias utilizadas para o armazenamento de dados.

Arquitetura de Dados do Armazém

Um arquitetura é um conjunto de regras para aderir ao construir algo, e
porque um armazém de dados pode se tornar muito grandes e complexos, utilizando um
arquitetura é essencial para o sucesso. Várias arquiteturas de data warehouse

existem, mas antes de explorar o assunto em profundidade vamos introduzir um general quadro e explicar alguns dos termos que você vai encontrar mais à frente.

O quadro é ilustrada na Figura 6-2. No diagrama, você pode ver:

1. Um ou mais sistemas de origem (arquivos, SGBD, ERP);
2. Um processo para extração, transformação e carregamento de dados (ETL). Muitas vezes, esse processo contém uma área de teste utilizado como local de desembarque para extraíram os dados e para fazer a transformação dos dados iniciais e de limpeza. Para a preparação de dados tanto um banco de dados e arquivos de plano pode ser usado. Em muitos casos usando arquivos simples permite um processamento mais rápido.
3. O armazém de dados, que consiste no banco de dados do armazém central e zero ou mais data marts.
4. A camada de usuário final (EUL), com as várias ferramentas para trabalhar com o dados (relatórios, dashboards, planilhas e documentos publicados).

Geralmente, a combinação do armazém central e os data marts é considerados o data warehouse, eo termo armazenamento de dados é usado para designar todo o processo de construção, carregamento e gestão dos dados armazém (DWH).

NOTA O diagrama na Figura 6-2 é uma estrutura lógica, mas não um físico exigido estrutura. Alguns processos ETL transferir os dados diretamente através""a partir da fonte sistemas para as tabelas do data warehouse, e alguns armazéns de dados não contêm data marts, ou apenas conter data marts e nenhum armazém central.

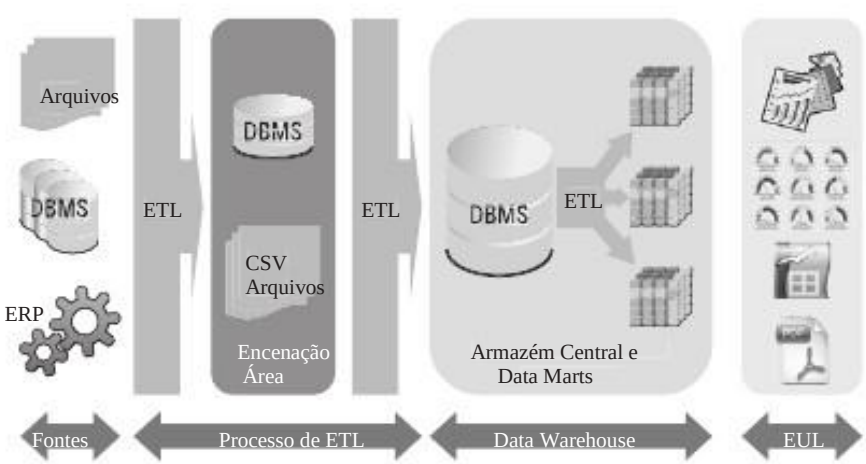


Figura 6-2: Genéricos arquitetura de data warehouse

Alguns outros termos usados geralmente quando se fala de armazenamento de dados são back office e front office. Na definição de Kimball, o back office engloba toda a tecnologia e os processos utilizados para construir e manter o data warehouse, Considerando que o front office é a área onde os usuários finais utilizam os dados da DWH. Em

Figura 6-2, o front office é a combinação de todos os relatórios e análise ferramentas disponíveis para ter acesso aos dados. O back office é composto de ETL processos e data warehouse.

mais claro, este diagrama é uma versão muito abstrata e simplificada de implementações DWH. A parte mais importante do processo é o auto-acasaladas carga periódica de novos dados para o data warehouse. Este é o lugar onde a ETL (para Extract, Transform and Load) ferramenta é utilizada. ""Periódica pode significar "Diário", mas mais frequentes ou às vezes até cargas mais raros são com-seg. Periódico carregamento de dados é também referida como por lotes ETL, ao contrário em tempo real ETL. No primeiro caso, vários registros são transformados em um único lote; na operação de um segundo no sistema de origem é imediatamente capturado e carregados no data warehouse. Atualmente, a maioria dos armazéns de dados são carregados utilizando ETL por lotes, mas também com o lote de carga uma freqüência quase em tempo real pode ser necessário. Considere, por exemplo, a carga de dados do orçamento, quando este alterações de dados apenas uma vez por mês ou uma vez por trimestre, é completamente inútil para carregar o orçamento de cada dia. Por outro lado, quando o data warehouse é usado de relatórios operacionais, bem como, os dados precisam ser o mais atualizado possível.

Ao utilizar ETL por lotes, isso poderia significar a execução do processo de cinco em cinco minutos. ETL é usada não somente para extrair dados dos sistemas de origem e carregamento do armazém de dados, mas também para a manutenção de dados relacionais ou data marts mover dados de uma fonte para outra.

A área de preparo

Cada solução de data warehouse deve usar um área de teste onde os dados extraídos é armazenada e, eventualmente transformado antes de carregar os dados na central armazém. A implementação desta parte da arquitetura como uma parte separada do de arquitetura de dados do armazém é uma prática comum para que siga essa orientação aqui também. A World Class Filmes data warehouse usa uma plataforma separada catálogo (catálogo é o prazo para banco de dados MySQL) chamado wcm_staging mas CSV arquivos são usados também.

Qual é o propósito de uma área de teste, quando poderia simplesmente usar a nossa ferramenta de ETL para carregar os dados diretamente no data warehouse? Existem várias razões:

- Fonte vezes a carga do sistema deve ser mantido a um mínimo absoluto, de modo quando os dados são extraídos de tabelas do sistema fonte, é a melhor prática para copiar os dados"como"para as tabelas de paragem o mais rapidamente possível.
- Usando uma área de teste em separado permite que você trabalhe em um subconjunto específico dos dados, ou para ser mais específico, apenas os dados que é necessário para o execução atual.
- Um esquema dedicado permite triagem específico ou a indexação ainda mais otimizar e apoiar o processo de ETL.

- A área de preparo é uma rede de segurança: um processo pode falhar antes de completar.
  Como uma área de teste contém todos os dados a serem processados, um processo pode
  ser reiniciado a partir do ponto de início ou a meio sem que o
  dados a serem extraídos novamente. Além disso, o conjunto de dados na área de teste não muda durante uma única execução; carregar novamente a partir do sistema de origem
  impõe o risco de os dados serem alterados, pois a carga anterior.

Lembre-se que a área de teste contém apenas os dados atuais extraídos,
o que significa que após uma carga de sucesso ou antes de executar o processo de ETL todos
tabelas estão sendo truncados novamente. Às vezes, um arquivo histórico é adicionado à
a arquitetura contendo todos os dados extraídos com uma hora de carga, mas acrescentou
este é um banco de dados separado, e não uma área de preparo.

Modelando uma área de teste é um processo muito simples, especialmente quando
tabelas são truncadas de cada vez. Basta duplicar a definição da fonte
tabelas sem todas as teclas e os índices e está feito. Remoção de estrangeiros
restrições de chave impõe alguns riscos de inconsistência de dados, mas isso pode ser
combatido com um trabalho de ETL cuidadosamente. Se os índices são usados em tudo no estadiamento
tabelas, sua finalidade deveria ser apenas para ajudar a acelerar a transformação
nada do processo, outra coisa.

O Armazém de Dados Central

regras estritas sobre como arquiteto de um data warehouse não existem, mas nos últimos
15 anos algumas arquiteturas comuns têm surgido. Para ajudar você a decidir qual
é o melhor para os seus fins, é sempre uma boa idéia de olhar para benchmarks
e estudos de caso de implementações DWH, como The Data Warehousing Institute
(TDWI) fez em 2006. A pesquisa realizada TDWI distinguidos cinco possíveis
formas de arquitetar um data warehouse e teve-as com base no sucesso da
arquiteturas diferentes. O diagrama na Figura 6-3 mostra os cinco alternativas.

Descrevemos aqui brevemente essas arquiteturas e explicar algumas das vanta-
vantagens e desvantagens de cada um antes de explicar as escolhas feitas para a
casos de exemplo neste livro.

- marts independentes de dados cada data mart é construído e carregado indiví-
  aliado, não há metadados comum ou partilhado. Esta é também chamado de funil
  solução.
- Dados de ônibus A-mart Kimball solução com dimensões adequadas.
- Hub and spoke (fábrica de informações corporativas)-O solução Inmon
  com um armazém de dados centralizado e marts dependentes de dados.
- armazém de dados centralizado-Similar ao hub and spoke '', mas sem a
  raios, ou seja, todo o acesso do usuário final está diretamente orientada para o data warehouse.
- Federados-An arquitetura, onde vários data marts ou data warehouse,
  casas já existem e estão integrados mais tarde. Uma abordagem comum

para isso é construir um data warehouse virtual onde todos os dados ainda residem nos sistemas fonte original e é logicamente integrados utilizando especiais soluções de software.

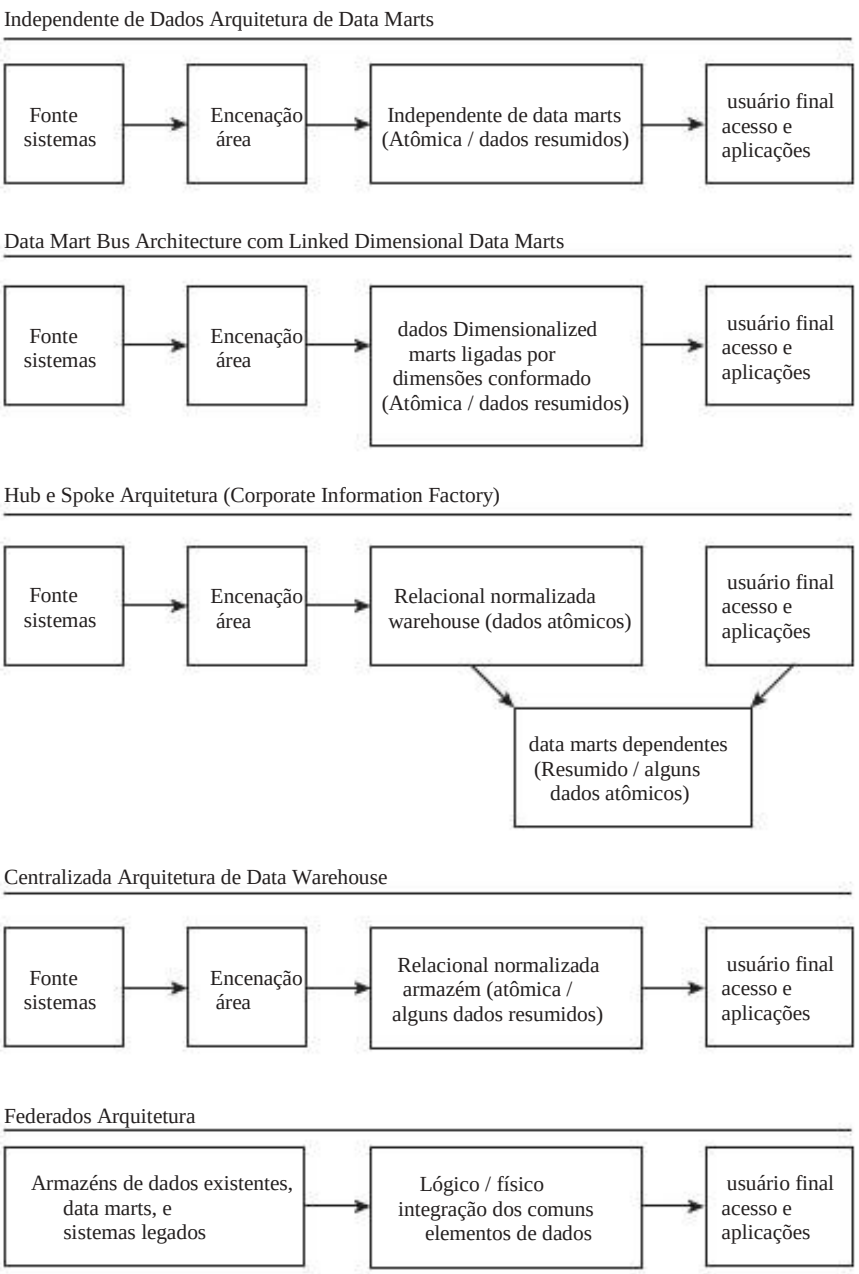


Diagrama de Thilini Ariyachandra e Hugh J. Watson da Data Warehousing Institute, usada com permissão.

Figura 6-3: arquitetura de dados alternativas armazém

Dos cinco abordagens tanto independentes de data marts e federados arquiras receberam uma pontuação significativamente menor em TDWI que os outros três, mas

o mais interessante é o fato de que o hub de ônibus, e falou, e centralizado
arquiteturas de pontuação sobre o igualmente bem.

A diferença entre o hub e falou arquiteturas centralizada e não é
que grande; data marts dependentes são adicionados como raios apenas para o desempenho
razões. Há ainda um grande hub poucos chamados e falou de data warehouse
implementações que usam apenas vistas a definir os data marts, o que resulta
em um data mart lógico, sem qualquer ganho de desempenho. A grande diferença,
No entanto, entre centralizada ou hub and spoke por um lado, eo ônibus
arquitetura sobre o outro é o custo eo tempo de construção do primeiro incremento,
ou parte utilizável do data warehouse. Como seria de esperar, este custo é
consideravelmente maiores em um hub e falou solução. Além disso, as informações
no armazém de dados é, de preferência expostas na forma de dados dimensional
marts de qualquer maneira uma vez que este modelo é fácil de entender para os
desenvolvedores que criam
soluções para usuários finais ou mesmo usuários avançados que podem aceder à data marts
diretamente.
Esta é a razão pela qual nós vamos basear a nossa soluções de dados exemplo armazém no
ônibus
arquitetura.

Nós recomendo vivamente a leitura do relatório completo, que ainda está disponível online
em http://www.tdwi.org/Publications/BIJournal/display.aspx?ID=7890.

Data Marts

A partir da discussão na seção anterior, pode parecer uma data mart
contém apenas um conjunto limitado de dados. A fim de atender às necessidades específicas
de um
organização, um data mart pode realmente cobrir apenas um processo específico e será
restrito aos limites do processo. Você não vai encontrar ausência do empregado
informações em um data mart de vendas, por exemplo, porque um analista de vendas não
preciso dessa informação.

Do ponto de vista técnico ou banco de dados, no entanto, não há nenhuma limitação para
a quantidade ou o tipo de dados que podem ser incluídos em um data mart. Por exemplo,
considere uma empresa de telecomunicações que quer analisar média das chamadas
duração por tipo de subscrição, grupo de clientes, e período. Tal detalhada
análise exigirão que todos os registros de detalhes da chamada ser incluído na data mart,
o que poderia facilmente adicionar até milhares de milhões de registros na tabela de fatos. Na
verdade,
empresas até comprar hardware especializado para este tipo de análises que podem
crunch coleções tamanho terabyte de dados em segundos.

Cubos OLAP

Como você já viu, um data mart pode ser definida como um conjunto de pontos de vista sobre
uma central
depósito (data mart virtual), mas no caso de utilização da arquitetura de barramento, o
armazém de dados consiste em um conjunto integrado de data marts que não precisam de um
camada extra em cima deles. Este não é o fim da história, entretanto. Para completar
a imagem, temos de introduzir um outro tipo de armazenamento, que é freqüentemente usado

para data marts, e isso é o motor OLAP. OLAP é um acrônimo para OnLine Analytical Processing e tem sido há décadas, embora o termo OLAP só foi introduzido em 1993 por Codd2 E.F. Provavelmente o mais conhecido Banco de dados OLAP no mundo é o Analysis Services, desenvolvido originalmente pela Israel empresa de software Panorama, mas mais tarde adquirida pela Microsoft, que agora pacotes do produto com o seu banco de dados. A idéia de um banco de dados OLAP é usar um formato de armazenamento otimizado para análise de dados em um multi-dimensional formato para oferecer a flexibilidade do usuário e acesso muito rápido. A velocidade oferecida pela bancos de dados OLAP é causada pelo fato de que a maioria dos totais e subtotais (aka agregações) são pré-calculadas e armazenadas no OLAP cubo. Embora um cubo OLAP pode ter muito mais do que três dimensões, um banco de dados OLAP é freqüentemente visualizado como um cubo de Rubik, daí o nome cubo.

Formatos de armazenamento e MDX

Três variações de bancos de dados OLAP existir, cada um nomeado com base no armazenamento formato que é usado:

- MOLAP (OLAP Multidimensional)-O formato original em OLAP que os dados são armazenados em um formato proprietário multidimensional. Todos dados detalhados e agregados são armazenadas no arquivo de cubo. Um bom exemplo de uma fonte aberta de banco de dados MOLAP é PALO, desenvolvido pelo alemão Jedox empresa.

- ROLAP (OLAP Relacional)-Em Neste caso, os dados e todos os agregados são armazenadas em um banco de dados relacional padrão. O motor traduz ROLAP consultas multidimensionais em SQL otimizado e, geralmente, acrescenta o cache capacidades, bem como para acelerar as consultas subsequentes analítica. Pentaho Mondrian é um exemplo perfeito de um motor de ROLAP.

- HOLAP (OLAP Híbrido)-In HOLAP, agregados e dados de navegação são armazenados em uma estrutura MOLAP, mas dados detalhados é mantido no relacional banco de dados. Até o momento, não há solução open source HOLAP disponíveis, mas algumas das vantagens foram incorporadas no Mondrian com a Além de quadros gerados automaticamente agregadas para agilizar consultas.

Todas essas soluções têm Multidimensional Expressions ou consulta MDX língua em comum. MDX fornece uma sintaxe especializada para consultar dados armazenados em cubos OLAP, como o SQL faz para tabelas relacionais. A linguagem foi introduzido pela primeira vez pela Microsoft em 1997 como parte do OLE DB para OLAP especificação, mas foi rapidamente adotado pela maioria dos vendedores de OLAP. A razão para o desenvolvimento de uma linguagem especializada para análise de dados OLAP armazena dados é que o SQL não foi muito bem adaptado para fazer isso. Embora ao longo dos

[2] Para um panorama da história dos motores OLAP ver http://olapreport.com / Origins.htm.

anos, a linguagem SQL foi estendido com recursos analíticos, MDX é ainda
o padrão de fato no mundo OLAP. Para aqueles ansiosos para começar bem
distância: Capítulo 15 aborda os conceitos básicos de MDX para utilizá-lo em conjunto com
Mondrian. No ano de 2000, a especificação do XML for Analysis (XML / A) foi
introduzido, que é agora um padrão de fato para consultar bancos de dados OLAP.
XML / A estende a linguagem MDX com tags XML padronizado para permitir
execução de instruções de consulta sobre HTTP utilizando métodos SOAP.

O servidor OLAP Mondrian suporta MDX e XML / A, que faz
é uma solução muito versátil para todos os tipos de cenários, mas tenha em mente que
Mondrian não é nem de armazenamento de dados (dados reside no relacional subjacente
banco de dados), nem uma ferramenta de análise (você ainda vai precisar de um front-end
para análise
de dados). Então, quando você estiver usando Mondrian como a solução OLAP (R) em sua
ambiente de Data Warehouse, os componentes será parecido com o esquema de
Figura 6-4.

Figura 6-4: Data Warehouse com Mondrian

Isto conclui a descrição de data marts a partir de uma base tecnológica mais e
perspectiva conceitual; capítulos 7 e 8 cobrirá os desafios de design
envolvidos com o desenvolvimento de data marts.

Desafios do Armazém de Dados

Descrevendo os diferentes componentes da arquitetura de um data warehouse meramente
serve como ponto de partida e explicar a terminologia comum. Os próximos
algumas seções cobrem os principais desafios envolvidos com o armazenamento de dados:
qualidade dos dados, o volume de dados e desempenho, capturando os dados alterados e
evolução das necessidades. É tentador olhar para o armazenamento de dados como um
desafio técnico, mas como já descrito em"A finalidade da
Data Warehouse"anteriormente neste capítulo, os desafios organizacionais
provavelmente ainda mais importante e mais difícil de enfrentar. Isto é especialmente
verdadeiro
em relação aos dados de gestão da qualidade, o tema da próxima seção, que
assume a liderança em locais de difícil endereço problemas que muitas vezes têm apenas um
número limitado
relação à tecnologia.

Qualidade dos dados

Um dos maiores desafios em qualquer projeto de data warehouse é garantir que os dados qualidade (DQ). Segundo o principal analista do Gartner Group, não há uma organização no mundo que não tenha um problema de qualidade de dados, para ser preparado para lutar algumas batalhas internas antes de ir viver com qualquer relatório ou Análise da solução. DQ problemas vêm em uma variedade de formas que são impossíveis para cobrir completamente, mas o mais importante geralmente caem em uma ou mais das seguintes categorias:

- -Os dados duplicados mesma entidade está inscrita várias vezes em um único sistema, ou a mesma entidade existe em vários sistemas, mas não pode ser ligado para a falta de falta de teclas ou referências.

- Dados incompletos, A entidades estão lá, mas algumas das informações é faltando, por exemplo, um número de casa em um endereço, um número de telefone, ou qualquer outra propriedade de uma entidade empresarial (uma entidade de negócios pode ser qualquer coisa
aqui, que vão desde uma apólice de seguro de um avião). Uma corretamente sistema projetado deve impedir que os usuários insiram dados incompletos, mas ainda existem muitos sistemas mal concebidos lá fora. E
mesmo quando uma empresa usa um sistema ERP como SAP ou Oracle, é provável que os dados a partir de sistemas mais antigos, foi migrado para estes sistemas,
que mais uma vez é motivo de dados incompletos ou inconsistentes.

- Todos os dados incorretos dados disponíveis e completo, mas tem erros devido às entradas com erros ortográficos ou erros de digitação. A fonte mais comum de incorreto
os dados são agentes de call center que terá de introduzir manualmente os nomes, endereços,
e outras informações que lhe são transmitidas pelo telefone. Digitação e ouvir velocidades diferentes, e lá você vai.

- Os dados conflitantes, mesmos dados são armazenados em tabelas diferentes (em o mesmo sistema), ou nos sistemas de origem diferente e contradiz a cada outras.

- Imprecisão metadados-Definições dos dados em um sistema não são claras, levando a ambigüidade dos dados em si. Como exemplo, pense em um utilitário empresa definir os seus clientes como pontos de conexão com um endereço, resultando no envio de contas para postes de luz.

- Falta de dados-Este é uma versão extrema de dados incompletos, onde registros completos que deve ser no sistema ou estavam lá mais cedo têm desaparecido. Este é obviamente o mais difícil categoria de problemas para resolver porque na maioria dos casos não é sequer evidente que existe um problema.

- Campos de valores em NULL em um banco de dados que não têm nenhum valor. O problema
é que isso pode significar coisas diferentes: não se aplica, é desconhecido ou ausente.

Muitas ferramentas estão disponíveis, tanto de origem proprietários e abertos, que podem ajudar
solucionar um ou mais desses problemas, mas não há uma ferramenta no mundo
que resolve o verdadeiro problema, que é a falta de normas, procedimentos e adequada
os processos. Resolvendo problemas de qualidade de dados não é um problema técnico, é uma
problema de organização. E ainda que irá mostrar-lhe algumas soluções
dados através de perfis e ferramentas de qualidade de dados, incluindo as etapas de validação
disponível em Chaleira nos capítulos 9 e 10, você terá que ter muito cuidado na
aplicação destas.

Em muitos projetos de data warehouse a melhoria da qualidade de dados é incluído como um
dos objetivos do projeto, pois o armazém de dados precisa conter válido e
dados corretos. Ou não? Nós gostamos de pensar de forma diferente. A via adequada para exame,
na nossa opinião, é a primeira pôr todos os procedimentos corretos em torno da qualidade dos dados
no lugar como parte do processo de negócio principal, em seguida, limpar a fonte
sistemas, e só depois de todo esse trabalho tem sido feito iniciar o carregamento dos dados.
Porque a partir deste ponto o processo de DQ está guardando a fonte de dados, o
a equipe do armazém de dados pode se concentrar em fornecer informações úteis ao fim
usuários e analistas. Isso faz parte de um tópico mais amplo chamado governança de dados, que,
como finanças ou produção, trata seu objeto gerenciado (neste caso, os dados) como um
ativo corporativo. E, assim como qualquer outro processo empresarial envolvendo ativos como
dinheiro, recursos físicos, ou empregados, um quadro de controlo adequado às necessidades
de estar no lugar.

Infelizmente, isso ainda é utopia para a maioria das organizações. Data Warehouse
projetos geralmente pegar um atalho por etapas, incluindo limpeza de dados como parte de
o processo de carregamento sem alimentar as exceções de volta para o negócio ou outra fonte
sistemas. E porque este é um livro sobre a construção de soluções de BI com Pentaho,
vamos mostrar a você como configurar isso. Basta ter em mente que esta é uma solução para
apenas parte do problema.

Dados Vault e Qualidade de Dados

Há uma nova escola de pensamento com base em um conceito chamado dados vault. Dados
vault (DV) é uma técnica de modelagem de banco de dados que é radicalmente diferente da
modelo tridimensional conformada que estamos usando neste livro. modelos DV se baseiam
no conceito de que todos os dados pertence a um dos três tipos de entidades: hubs, links,
e satélites. Em suma, os centros de conter os principais atributos de entidades empresariais
(Como ordens, produtos e clientes), links definir as relações entre
os hubs (por exemplo, encomendas de clientes ou categorias de produtos), e os satélites
conter todos os outros atributos relacionados com os hubs ou links, incluindo todos os atributos
mudar a história. Figura 6-5 mostra um exemplo parcial de como o mundo
Classe data warehouse Filmes olharia quando traduzido em uma abóbada de dados modelo.

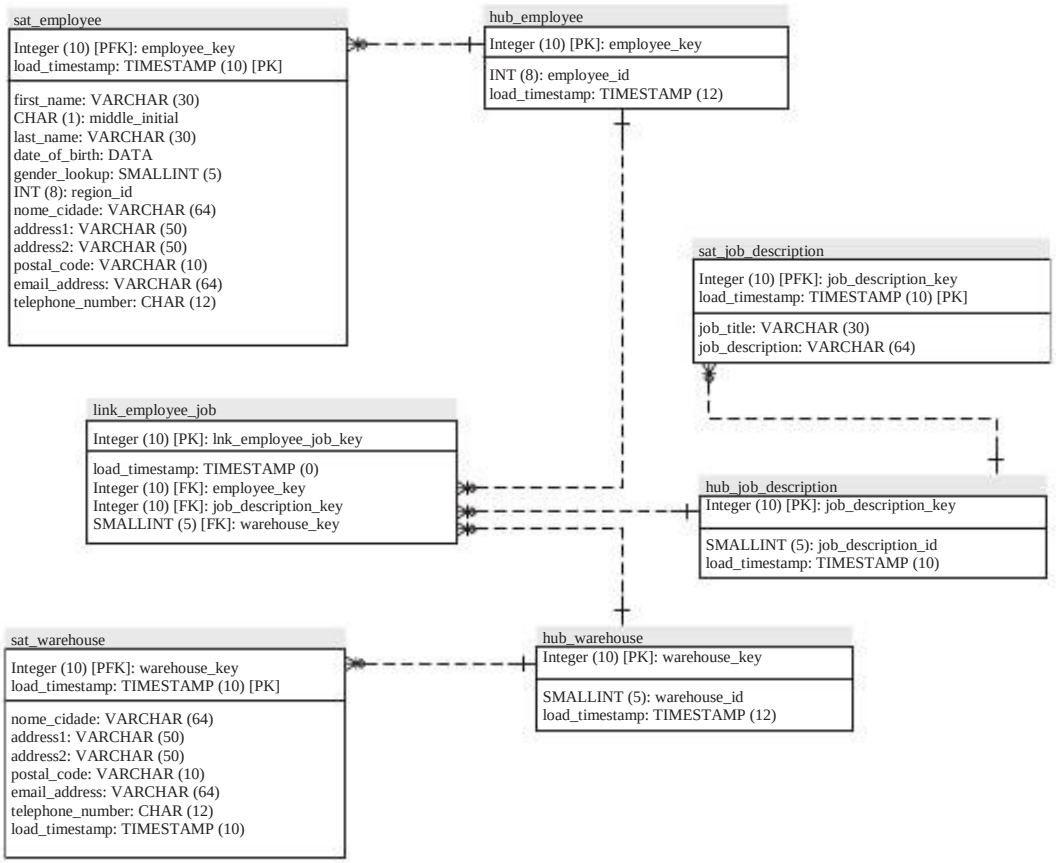


Figura 6-5: Dados exemplo abóbada

A grande vantagem da técnica é a sua flexibilidade. Adicionar atributos ou tipos de transação nova é relativamente simples, não há necessidade de reconstruir a qualquer uma das partes já existentes do banco de dados. Por outro lado, um abóbada de dados não é acessível a qualquer utilizador final, nem mesmo um analista experiente.

Somente após a construção de data marts da abóbada de dados, que inclui processos de limpeza e, possivelmente, conforme dados sujos, as informações podem ser disponibilizar aos utilizadores finais. Outra desvantagem possível é o fato de que o número de tabelas no data warehouse será grande em comparação com um tradicional normalizado ou arquitetura de barramento de dados, tornando o DVD mais difícil de gerir.

Do ponto de vista da qualidade dos dados, utilizando-se uma abóbada de dados tem um grande vantagem: os dados do DV é carregado a partir de sistemas de origem"como", assim se os dados são sujos no sistema de origem, é suja no cofre de dados também. Se uma correção é feita no sistema de origem, os dados corrigidos é adicionada ao abóbada de dados como uma nova versão. Se existirem sistemas de múltiplas fontes, o sistema do registo é incluído na chave dos dados armazenados. O resultado de tudo isso é que uma abóbada de dados sempre pode reproduzir a informação que estava armazenada em um sistema de origem em qualquer ponto do tempo, o que torna a solução ideal em casos onde auditabilidade e rastreabilidade de dados e transações são de preenchimento obrigatório,

como nas empresas bancárias ou de seguros. Todos limpeza, fusão e conversão de dados que geralmente ocorre entre a área de teste e data marts em uma arquitetura de barramento ocorre agora entre a abóbada de dados e data mart. E porque toda a história está disponível no cofre, data marts podem ser facilmente eliminados ou transformados em uma outra maneira para responder a questões de negócio diferentes.

Na comunidade de BI, um dos maiores chavões ao longo dos últimos dos anos tem sido Versão única da verdade, o que significa que os dados em uma base de dados armazém (não importa como ele é modelado) significa a mesma coisa para todos os diferentes pessoas que trabalham com os dados. Dados profissionais vault olhar para esta ligeiramente diferente: sua opinião é que a única verdade é a fonte de dados (Single Version dos fatos), que não deve ser transformada em qualquer forma, quando entrou em o data warehouse. Daquele ponto em diante, cada usuário ou departamento pode ter seu ''própria versão da verdade''criado na data mart: retrocesso ao real ""A verdade é sempre possível, então.

Se você estiver interessado em aprender mais sobre como usar o cofre de dados, por favor visite site do inventor em www.danlinstedt.com/.

Usando dados de referência e Master

O termo gerenciamento de dados mestre (MDM) é usado para descrever todos os processos e tecnologias necessários para cumprir as definições de dados uniforme e conteúdo de dados em uma organização. Mestre de gerenciamento de dados é uma área adjacente aos dados armazenamento com a sua própria gama de conceitos, livros, médicos, e conferir-diferenças a respeito do tema. No contexto de um data warehouse, é necessário para cobrir os princípios básicos do MDM, porque a ausência de uma adequada MDM processo pode fortemente afetar o sucesso de um projeto de business intelligence.

MDM tem a ver com o gerenciamento de informações referentes às entidades empresariais tais como clientes, produtos, fornecedores e locais. Um dos objetivos iniciais de qualquer iniciativa do MDM é identificar os sistema de registro de uma entidade. Em muitos organizações existem vários sistemas que contenham um dos mencionados itens, com os clientes""entidade com a reputação de ser a mais notoriamente complexo de lidar. Um sistema de registro do cliente identifica a ""Ouro cópia dos dados do cliente e não deve ser confundido com o sistema de entrada onde os dados são inicialmente inscritas ou mantida depois. Idealmente, estes dois sistemas são os mesmos, mas isso não é necessariamente o caso, especialmente quando há muitos sistemas de entrada e apenas um sistema de registro.

Um exemplo desta última situação é uma organização que usa um produto como o SugarCRM para manter todas as interações com clientes via telefone e e-mail, e Compiere como um sistema ERP que controla pedidos, transporte, e faturamento de e para os mesmos clientes. Além disso, os clientes também podem entrar e manter seus próprios dados em um banco de dados separado por meio de um web personalizada aplicação como front-end, então você tem três sistemas com os dados do cliente, cada com a capacidade de atualizar esta informação. O problema fica ainda pior na

bancos, onde a maioria dos sistemas são orientados para os produtos em vez de orientada para o cliente,
o que significa que cada produto tem seu sistema de dados próprio cliente.

Uma iniciativa partes características MDM muitos com um data warehouse projeto: os dados precisam ser analisados, integrados e limpos. E, assim como dados armazenamento, o MDM é um processo, não um projeto. Há, no entanto, também alguns diferenças notáveis. A diferença mais importante é que ao contrário de um data mart ou armazém de dados, o MDM não pode nunca ser uma iniciativa do departamento, executado somente dentro de
um departamento de atendimento ao cliente, por exemplo. E, claro, bom de dados mestre é inicialmente destinada a apoiar os sistemas transacionais para assegurar que as ordens estão sendo enviados para o endereço correto e que os inquéritos do cliente pode ser direcionado para o número de telefone correto e pessoa de contato.

Os dados mestre não precisa se originam dentro de uma organização: às vezes é melhor usar os mesmos dados externos que todo mundo está usando. Boa exemplos destes são a região ISO e tabelas país que usamos no Mundo
Classe de banco de dados de filmes, ou o norte-americano Industry Classification System (NAICS) para identificar em qual linha de negócios de uma empresa opera. Neste caso, um organismo de normalização cuida da gestão dos dados.
Às vezes, esses dados mestre externo é chamado dados de referência para distingui-lo a partir de dados mestre interno, que é simplesmente chamado dados mestre.

A partir desta breve introdução, espero que ela imediatamente claro por que o MDM é tão importante com relação a armazenamento de dados e qualidade dos dados: se houver um sistema de identificação que contém a cópia correta de uma entidade empresarial, que é provavelmente a melhor fonte para alimentar as suas tabelas de dimensão a partir.

Volume de dados e desempenho

Como você aprendeu anteriormente, duas razões para a utilização de um ambiente físico separado
para um data warehouse é o ganho de desempenho, em comparação com a pesquisa numa regular
sistema de transação e da incapacidade dos sistemas de apoio tanto transacional e consultas analíticas. Outra coisa que é facilmente esquecido na primeira é a facto de armazéns de dados tendem a ficar muito grande, como resultado de armazenamento de dados
de vários sistemas em um longo período de tempo. Normalmente, os dados de transações podem ser arquivados após um determinado período de tempo, por exemplo, depois de tudo necessário
relatórios externos às autoridades fiscais tenha sido finalizado. Isso libera espaço e acelera o sistema de transação, mas também faz com que o banco de dados inúteis para análise de tendências durante um período prolongado de tempo. Por este motivo, os dados
armazém de retém todos os dados históricos, de preferência no nível mais detalhado, resultando em enormes conjuntos de dados, por vezes, várias centenas de Terabytes em tamanho. Os armazéns de dados maior, no momento da redação deste texto conter mais de um Petabyte (1024 Terabytes =1 petabyte) de dados do usuário.

O maior desafio quando se tenta analisar essas quantidades de dados é conseguir um desempenho aceitável de consulta para usuários finais. Diversas técnicas

foram desenvolvidos ao longo dos anos para atingir esse desempenho, mas infeliz-
Infelizmente algumas destas técnicas estão disponíveis apenas em proprietários caros
sistemas de gerenciamento de banco de dados como Oracle, SQL Server ou DB2. O se-
mugido lista ajuda a determinar qual técnicas pode ser útil em um data
warehouse para melhorar o desempenho da consulta:

- Com a indexação nenhuma das medidas especiais tomadas, qualquer tabela em um banco de dados é apenas
  uma lista não ordenada, com linhas de dados. Qualquer consulta precisa ler todos os dados
  a mesa (a.k.a. um varredura completa da tabela) para encontrar a resposta correta. Este processo pode
  ser acelerada por armazenar ponteiros para os dados corretos nos arquivos de índice especial.
  Índices em um banco de dados são semelhantes aos índices em um livro onde você pode olhar
  uma palavra-chave e encontre todas as páginas que fazem referência a esta palavra-chave. Um especial
  tipo de índice é o chave primária índice, que é o identificador exclusivo de um
  linha. Se você estiver usando InnoDB como o mecanismo de armazenamento do MySQL, o principal
  chave também denota a ordem de classificação física dos registros em uma tabela. Este pode vir a calhar, porque uma grande quantidade de consultas terão uma referência à data. Ou ainda a data chave é a primeira coluna de uma tabela de fatos e de parte da chave primária, a coluna pode ser colocada em ordem decrescente de classificação, resultando
  nas linhas mais recentes foram as primeiras linhas físicas também. Recuperando receita de semana passada ao vendas não será uma tarefa muito demorada.
- Bitmap Indexação-Uma forma especial de índice é o bitmap ou bitmap índice. Índices de bitmap são uma opção ideal para indexação chamado baixo cardinalidade colunas (colunas com um número relativamente pequeno de exclusiva valores, como gêneros ou gênero de filme). A ideia básica é que, para a dis-tinct número de valores na coluna, uma projeção bit é criado permite um acesso muito rápido para os valores indexados. Figura 6-6 contém um exemplo simples disso, uma coluna com três valores possíveis com o acompanham os valores de índice de bitmap. Infelizmente, o suporte para bitmap indexação é planejada para uma futura versão do MySQL, mas não está disponível ainda.
- Particionando-One dos novos recursos mais interessantes no MySQL 5.1 é particionamento, que é a capacidade de cortar uma tabela em várias física peças verticais. A vantagem desta técnica é que quando uma consulta podem ser satisfeitas através dos dados em um ou um número limitado de partições, o outras partições que não precisa ser olhado. Nesse sentido, é uma partição uma espécie de índice de super. Suponha, por exemplo, que uma tabela é muito grande particionada por mês do ano e você tem 36 meses de dados armazenados. A referência de consulta ao último trimestre é automaticamente limitada ao últimos três partições, o que lhe poupa o esforço de olhar para o outro 33. Outra vantagem do uso de partições é a possibilidade de queda e (re) criar partições, por exemplo, ao projetar um round robin""esquema onde os mais recentes 36 meses de dados estão sempre disponíveis. Neste caso, o partição mais antiga pode ser simplesmente descartado quando um novo para a corrente meses podem ser facilmente adicionados para o carregamento de dados atual.

Tabela de dados		Bitmap Index		
ID	COR	LX1	LX2	LX3
1	Vermelho	1	0	0
2	Verde	0	1	0
3	Amarelo	0	0	1
4	Verde	0	1	0
5	Amarelo	0	0	1
6	Amarelo	0	0	1
7	Verde	0	1	0

Figura 6-6: Bitmap exemplo de índice

- Agregação Todos técnicas de aprimoramento de desempenho por objectivo limitar a quantidade de dados a ser pesquisado para responder a uma consulta específica. Isso pode
também ser conseguido através de tabelas de agregados em que os resultados pré-calculados
estão sendo disponibilizadas. Para responder à pergunta:"Qual foi a nossa total receita de vendas por mês por site?", a consulta não precisa de consulta
Os dados pormenorizados e resumi-lo, mas pode obter os dados diretamente de um tabela de agregados em que estes resultados estão prontamente disponíveis. É claro que o
dados do processo de carga do armazém precisa reconstruir ou atualizar o agregado tabelas de dados de cada novo tempo é adicionado, o que pode tornar este um time-consuming
e uma tarefa desafiadora, especialmente quando várias tabelas de agregados foram criado. Neste ponto, o Mondrian Designer vem agregada ao
de emergência (ver Capítulo 15) para criar essas tabelas automaticamente para você.
- As visões materializadas-A visão materializada "É uma visão com dados "infeliz-Infelizmente, este recurso não está disponível em qualquer banco de dados open source. Uma visão do banco de dados regular é apenas uma definição que imita uma mesa, mas não
contendo dados reais. Assim, quando uma tela com a receita total por mês por gênero de filme é criado (com soma""e""pelo grupo), o banco de dados precisa
para calcular estes resultados das tabelas de pontos de vista de base cada vez que o ponto de vista é
consultado. Uma visão materializada não só armazena esses valores calculados, mas também atualiza automaticamente quando essas tabelas de base estão sendo carregados
ou atualizados.
Agora, se isso fosse a única vantagem, ainda seria possível criar
uma solução semelhante em qualquer atualização de banco de dados usando triggers e procedimentos.
A parte que ainda estariam desaparecidos, nesse caso, é a verdadeira beleza
de visões materializadas: redirecionamento de consulta. Esse recurso analisa a consulta orientada para as tabelas de detalhes e redireciona a consulta materializada
vista quando os dados necessários está lá ou não recupera os dados
das tabelas de detalhe. As visões materializadas podem aumentar drasticamente a desempenho de um armazém de dados, mas, infelizmente, são suficientes apenas disponível em
as edições da empresa de bases de dados principais proprietários.

Mesmo que não há banco de dados de código aberto com esse apoio disponíveis, isso não significa que não há solução. Novamente, Mondrian precisa ser mencionado aqui, uma vez que funciona de forma semelhante quando tabelas agregadas foram definidos. Mondrian pode então calcular o custo de resolver o consulta e, quando se utiliza uma tabela de contagens agregadas melhor, ele é usado pelo motor.

- funções Janela-Para fins analíticos, o SQL 2003 Standard é alargado com os chamados funções de janela. Isto permite uma consulta para executar cálculos em cima (parte de) um conjunto de resultados. A vantagem é que uma tabela única precisa ser digitalizado uma vez para gerar múltiplas saídas, tais como médias ou subtotais. Duas importantes adições SQL cuidar dessa funcionalidade: a Mais e Partição cláusulas. As funções de janela de trabalho em conjunto regular com funções agregadas como Soma,MédiaE Contagem mas também permite funções especiais, tais como Ranking () e Row_number ().

Janela de exemplo FUNÇÕES

A declaração a seguir é um exemplo simples de uma tabela com três colunas: OrderID, ProductID, e RECEITAS. A tabela contém cinco linhas, com duas distintas identificações de produto:

```
Selecione CódigoDoPedido, ProductID, receita
 SUM (receita) OVER (PARTITION BY ProductID) AS
 PRODUCTTOTAL, SUM (receita) OVER () AS GRANDTOTAL
DA SQLWINDOW
ORDER BY 1
```

Os resultados da consulta são apresentados na Tabela 6-1.

Tabela 6-1: Janela de resultado da função Set

OrderID	ProductID	RECEITAS	PRODUCTTOTAL	GRANDTOTAL
1	A	3	9	19
2	A	3	9	19
3	B	5	10	19
4	B	5	10	19
5	A	3	9	19

(Continuação)

Janela de exemplo FUNÇÕES (Continuação)

Open Source Apoio janela banco de dados

PostgreSQL suporta funções de janela a partir da versão 8.4, mas o MySQL não
têm essa capacidade, no entanto, nem é parte do roteiro MySQL. O MySQL tem apenas
uma função Rollup para permitir a adição de totais e subtotais dentro do resultado do SQL
definido. A versão do MySQL da consulta anterior parecido com este:

```
SELECT IFNULL (ProductID, "ALL") AS ProductID,
SUM (receita) AS RECEITAS
 DA SQLWINDOW
GROUP BY ProductID WITH ROLLUP
```

o que resulta em dados apresentados na Tabela 6-2:

Tabela 6-2: Rollup Resultados MySQL

ProductID	RECEITAS
A	9
B	10
ALL	19

Por favor, note que os resultados acumulados são rotulados NULL se não mea especiais
Sures foram tomadas. No exemplo anterior, a declaração do IFNULL toma
cuidado de traduzir o NULL valor para o texto ALL.

- Lembre-Arquivamento quando disse que todos os dados históricos devem ser dis-
  poder no armazém de dados no menor nível de detalhe? Bem, nós mentimos (apenas
  um
  pouco). Às vezes é suficiente para manter os dados históricos em um agregado
  nível, que ainda pode ser perfeitamente aceitável para apoiar a análise de tendências
  (Por exemplo, para analisar as tendências de vendas diárias por grupo de produtos basta
  para armazenar dados no grupo de produto / dia). Um cenário possível seria
  seja para manter os dados detalhados on-line em um prazo que se deslocam de 24 ou
  36 meses;
  Após isso, apenas os agregados diária, semanal ou mensal estará disponível.
  É claro que em casos especiais, os dados detalhados podem ser trazidos de volta a
  partir de
  o arquivo para sustentar a análise detalhada.

A lista anterior mostra uma série de maneiras para aumentar o desempenho em uma base de dados
depósito, mas dependendo do banco de dados real sendo usado algumas dessas
recursos podem não estar disponíveis. O Pentaho BI Suite é banco de dados independente,
e enquanto estamos usando o MySQL como banco de dados de exemplo neste livro, haverá
haver casos em que outros bancos de dados pode ser uma alternativa melhor. Seleção

um banco de dados para armazenamento de dados é sempre uma questão de funcionalidade necessária
versus orçamento disponível.

Captura de dados alterados

O primeiro passo para um processo de ETL é a extração de dados de diversas fontes sistemas e armazenamento de dados em tabelas de teste. Este parece ser um trivial tarefa e, no caso de inicialmente o carregamento de um armazém de dados geralmente é, além
dos desafios decorrentes dos volumes de dados e conexões de rede lenta.
Mas, depois da carga inicial, você não deseja repetir o processo de completamente extrair todos os dados novamente, o que não seria de muita utilidade, dado que você Já tenho um quase completo conjunto de dados que só precisa ser atualizada para refletem a situação atual. Tudo o que nos interessa é o que mudou desde a última carga de dados, então você precisa para identificar quais registros foram inseridos, modificados
ou mesmo excluída. O processo de identificação dessas mudanças e só recuperar registros que são diferentes daquilo que já está carregado no data warehouse é chamado Captura de dados alterados ou CDC.

Basicamente, existem duas categorias principais de processos CDC, intrusivo e não-intrusiva. Por intrusiva, queremos dizer que uma operação de CDC tem um possível impacto no desempenho do sistema os dados são recuperados. É justo dizer que qualquer operação que exija a execução de instruções SQL de uma forma ou outra é uma técnica invasiva. A má notícia é que três dos quatro caminhos para capturar os dados alterados são intrusivos, deixando apenas uma opção não-intrusiva. As seguintes seções oferecem descrições de cada solução e identificar as suas prós e contras.

Fonte de Dados Baseado em CDC

Fonte CDC baseadas em dados é baseada no fato de que existem atributos disponíveis no sistema de origem que permitem que o processo de ETL para fazer uma seleção dos alterado
registros. Há duas alternativas aqui:

- Leitura direta com base em timestamps (valores de data e hora)-At pelo menos um timestamp é necessária uma atualização aqui, mas de preferência dois são criados: um timestamp de inserção (quando o registro foi criada) e um timestamp atualização (Quando o registro foi modificada pela última vez).

- Usando seqüências de banco de dados mais- bases de dados têm algum tipo de auto-incremento opção para valores numéricos em uma tabela. Quando tal número de seqüência está a ser utilizado, também é fácil identificar quais registros foram inseridos desde a última vez que você olhou para a mesa.

Ambas as opções requerem mesas extra com o armazém de dados para armazenar as dados sobre a última vez que os dados são carregados ou a última seqüência recuperados número. Uma prática comum é criar essas tabelas de parâmetros ou em um

esquema separado ou na área de teste, mas nunca no armazém de dados central e certamente não em um dos data marts. Uma hora ou uma seqüência baseada em solução é sem dúvida o mais simples de aplicar e também por esta razão uma dos métodos mais comuns para a captura de dados alterados. A penalidade para a essa simplicidade é a ausência de algumas capacidades essenciais que podem ser encontrados
em opções mais avançadas:

- Distinção entre inserções e atualizações somente quando o sistema de origem contém inserir um carimbo e uma atualização pode ser a diferença detectada.
- Excluídos detecção registar-Este não é possível, a menos que o sistema de origem apenas logicamente exclui um registro, ou seja, tem um fim ou excluído data, mas não é fisicamente excluído da tabela.
- Várias detecção update-Quando um registro é atualizado várias vezes durante o período compreendido entre a anterior ea data corrente de carga, estas atualizações intermediárias se perder no processo.
- Em tempo real capacidades, Timestamp ou extração de dados baseada na seqüência é sempre uma operação em lote e, portanto, inadequado para dados em tempo real cargas.

Trigger Baseado CDC

Database triggers podem ser usados para disparar ações mediante o uso de qualquer manipulação de dados
declaração como INSERT,UPDATEOu DELETE. Isto significa que os gatilhos podem também ser usado para capturar essas mudanças e colocar esses registros alterados em quadros intermédios mudança nos sistemas de origem para extrair dados de mais tarde, ou para colocar os dados diretamente nas tabelas de preparo do armazém de dados ambiente. Porque a maioria dos gatilhos para adicionar um banco de dados será proibido no casos (que requer modificações para o banco de dados de origem, que muitas vezes não é abrangidos por acordos de serviço ou não permitido pelos administradores de banco de dados)
e pode severamente abrandar um sistema de transações, esta solução, embora funcionalmente atraente à primeira vista, não foi implementado com muita freqüência.

Uma alternativa para usar os gatilhos diretamente no sistema de origem seria para configurar uma solução de replicação onde todas as alterações às tabelas selecionadas serão
replicado para as tabelas que recebem na parte lateral do armazém de dados. Estes replicado tabelas podem ser estendidos com o exigido gatilhos para apoiar o CDC
processo. Embora esta solução parece envolver uma grande quantidade de overhead de processamento
e requer mais espaço de armazenamento, é realmente muito eficiente e não-intrusiva uma vez que a replicação é baseada na leitura mudanças a partir dos arquivos de log do banco.
A replicação é também uma funcionalidade padrão de gerenciamento de banco de dados mais sistemas, incluindo MySQL, PostgreSQL e Ingres.

CDC Trigger baseado é provavelmente a alternativa mais intrusiva descrito aqui mas tem a vantagem de detectar todas as alterações de dados e permite quase em tempo real

carregamento de dados. As desvantagens são a necessidade de um DBA (o sistema de origem é
modificada) ea natureza de banco de dados específicos das demonstrações gatilho.

Instantâneo baseado CDC

Quando não estão disponíveis timestamps e dispara ou replicação não são uma opção, o último recurso é usar tabelas instantâneo, que pode ser comparada a alterações. Um instantâneo é simplesmente um extrato completo de uma tabela de origem que é colocado nos dados
armazém área de preparo. Os dados da próxima vez precisa ser carregado, uma segunda versão (instantâneo) da mesma tabela é colocado ao lado do original e da
duas versões em relação a mudanças. Tomemos, por exemplo, um exemplo simples de um tabela com duas colunas, ID e Cor. A Figura 6-7 mostra duas versões deste
mesa, um instantâneo instantâneo e 2.

Snapshot_1

ID	COR
1	Black
2	Verde
3	Vermelho
4	Blue

Snapshot_2

ID	COR
1	Grey
2	Verde
4	Blue
5	Amarelo

Figura 6-7: versões Snapshot

Existem várias maneiras de extrair a diferença entre essas duas versões. O primeiro é usar uma junção externa completa na coluna de chave de identificação e marca as linhas do resultado
de acordo com seu status (I para Inserir, U para atualização, D para Excluir e, para N Nenhum) onde as linhas são filtradas inalterada na consulta externa:

```
SELECT * FROM
(Caso selecione

            t2.id quando for nulo, 'D'
            t1.id quando for nulo, 'I'
            quando t1.color <t2.color> depois em 'U'
            else 'N'
        final como bandeira
, Caso
            t2.id quando for nulo, t1.id
            mais t2.id
        final como id
, T2.color
fromsnapshot_1 t1
junção externa completa snapshot_2 t2
ont1.id = t2.id
) Um
onde 'N' flag <>
```

Isto é, naturalmente, quando o banco de dados oferece suporte completo as junções externas, o que não é
o caso do MySQL. Se você precisa de construir uma construção similar com o MySQL Existem algumas opções, como as seguintes:

```
'U' Escolha como bandeira, como t2.id id, como a cor t2.color
de snapshot_1 t1 inner join snapshot_2 t2 em t1.id = t2.id
onde t1.color! = t2.color
União todos os
selecione 'D' como bandeira, como t1.id id, como a cor t1.color
de snapshot_1 t1 left join snapshot_2 t2 em t1.id = t2.id
onde t2.id é nulo
União todos os
selecionar 'I' como bandeira, como t2.id id, como a cor t2.color
de snapshot_2 t2 t1 left join snapshot_1 em t2.id = t1.id
onde t1.id é nulo
```

Em ambos os casos o resultado seja o mesmo, conforme apresentado na Figura 6-8.

BANDEIRA	ID	COR
U	1	Grey
D	3	NULL
I	5	Amarelo

Figura 6-8: Snapshot comparar resultados

A maioria das ferramentas ETL hoje contêm funcionalidade padrão para comparar dois tabelas e as linhas como bandeira I,UE Dnesse sentido, assim você terá mais provável usar essas funções padrão em vez de escrever SQL. Pentaho Kettle, para exemplo, contém o Mesclar linhas etapa. Esta etapa demora dois classificado conjuntos de entrada
e compara-los nas teclas especificado. As colunas a serem comparados podem ser selecionados, também, e uma saída de nomes de domínio da bandeira deve ser especificado.

CDC instantâneo baseado pode detectar inserções, atualizações e exclusões, que é um vantagem sobre o uso de carimbos, com o custo de armazenamento extra para os diferentes instantâneos. Também pode haver um problema de desempenho grave quando as tabelas a serem
comparação são extremamente grandes. Por esta razão, acrescentou a ilustração SQL porque para este tipo de trabalho pesado, o motor de banco de dados é muitas vezes mais adequados
de uma ferramenta de ETL motor baseado.

Log-base CDC

A forma mais avançada e menos invasiva de captura de dados alterados é
usar uma solução baseada em log. Cada inserir, atualizar e excluir executar a operação em um
banco de dados podem ser registrados. Em casos usando um banco de dados MySQL, o log binário deve
ser habilitado explicitamente na ferramenta Administrador (variáveis de inicialização Logfiles). Daquele momento em diante, todas as alterações podem ser lidos em tempo quase real a partir do

log do banco de dados e utilizados para atualizar os dados no data warehouse. As capturas aqui é que isso soa mais simples do que realmente é. Um arquivo de log binário deve ser transformado primeiro em uma forma compreensível, antes das entradas pode ser lido em um processo posterior.

A instalação do MySQL contém uma ferramenta especial para esse fim, mysqlbinlog. Esta ferramenta pode ler o formato binário e converte-lo em um pouco formato legível e de saída podem ler os resultados a um arquivo de texto ou diretamente em um cliente de banco de dados (no caso de uma operação de restauração). Mysqlbinlog tem várias outras opções, com o mais importante para nossos propósitos é o fato que possa aceitar um começo e / ou carimbo do tempo final para ler apenas uma parte do arquivo de log.

Cada entrada tem também um número de seqüência que pode ser usado como um deslocamento, por isso não duas maneiras de evitar duplicados ou valores perdidos durante a leitura de esses arquivos.

Após a saída mysqlbinlog é gravada em um arquivo de texto, este arquivo pode ser analisado e ler, por exemplo, um passo de entrada Chaleira que lê os dados e executa as declarações sobre as tabelas de teste correspondente. Para outros bancos de dados não São soluções similares, e alguns oferecem um quadro completo do CDC, como parte de sua solução de data warehouse.

A desvantagem de usar um conjunto de banco de dados específico de ferramentas é óbvia: ele só trabalha com um único banco de dados. Sempre que houver a necessidade de utilizar um registro baseado em solução em um ambiente heterogêneo, várias ofertas comerciais disponíveis no mercado.

Qual alternativa CDC deve você escolher?

Como temos mostrado nas seções anteriores, cada uma das opções descritas para identificação e seleção de dados alterados têm seus pontos fortes e fracos. Algumas alternativas exigem adaptações ao banco de dados de um banco de dados administrador (DBA), alguns podem apoiar em tempo real o carregamento de dados, e outros suportam apenas uma descoberta parcial de mudanças. Tabela 6-3 resume esses pontos para ajudar você a decidir qual opção é mais aplicável à sua situação.

Requisitos Variáveis de usuário

Isso pode parecer um lugar estranho para falar sobre a mudança de requisitos dos utilizadores desde geralmente isso envolve a parte de análise e projeto de um data warehouse, que é abordada no capítulo seguinte. Não é tão estranho que possa parecer, no entanto. Um novo exigência de usuário geralmente significa aumentar o armazenamento de dados e da carga processo de alguma maneira ou de outra. Novos ou requisitos de usuário alterada pode levar a mudanças em seu data warehouse que vão desde a simples adição de um novo coluna ou um cálculo extra para adicionar um sistema de código inteiramente novo. Usuário mudanças de requisitos não pode ser apenas baseada na procura (a partir da perspectiva do armazém de dados), mas também pode ser impulsionado por mudanças no funcionamento

sistemas. Um dos projetos mais desafiadores que você pode embarcar em se ter um substituição completa de um ou mais sistemas de origem, por exemplo, a migração da aplicação financeira de A para B aplicação financeira de um fornecedor diferente. Agora imagine que os requisitos do usuário não muda em nada, o que significa que você tem que mudar todos os processos ETL sem quebrar a comunicação de informações existentes
ambiente, incluindo o data warehouse.

Tabela 6-3: CDC Opções

	TIMESTAMP	SNAPSHOT	TRIGGERS	LOG
Inserir / atualizar distinção?	N	Y	Y	Y
Várias atualizações detectado?	N	N	Y	Y
Exclui identificados?	N	Y	Y	Y
Não-intrusivo?	N	N	N	Y
suporte em tempo real?	N	N	Y	Y
Independente do SGBD?	Y	Y	N	N
Não DBA necessária?	Y	Y	N	N

É impossível cobrir todas as mudanças que podem ocorrer durante a vigência do um armazém de dados, mas há um princípio que deve ser clara e acordado desde o primeiro dia do início de uma possível solução:

Um armazém de dados é um processo, não um projeto.

Portanto, agora que temos que sair do caminho, podemos começar a fazer planos para acomodar esse processo com algumas diretrizes gerais. Apenas estar ciente das fato de que tudo vai (finalmente) a mudança, seja ela de relatórios, sistemas de origem, pessoas e departamentos que utilizam o sistema, basicamente, tudo relacionado ao o data warehouse irá mudar em algum ponto no tempo. Certifique-se que estes mudanças podem ser acomodados, seguindo as seguintes regras:

- Projeto de data warehouse o como um conjunto de independentes (mas interligadas) blocos de construção que pode ser facilmente substituído por soluções alternativas. Para exemplo: suponha que o armazém de dados preciosos é substituído por um aparelho de armazenamento de dados de outro fornecedor. Quanto não-padrão SQL que você usa em processos de ETL ou consultas utilizadas para a comunicação?

- Use as ferramentas e usá-los direito. Todas as organizações muitas vezes passam baldes de dinheiro com fantasia de painéis de BI e ferramentas de relatórios e contratar consultores caros para implementá-las, mas recusar um pedido de

$ 5.000 ferramenta para modelagem e gestão do armazém de dados. Em vez disso, eles planilhas usar para armazenar as definições manualmente e criar bancos de dados e tabelas usando as ferramentas padrão que veio com o produto de banco de dados. Este é bom para criar rapidamente um protótipo, mas que é tão longe como deveria ir.

- Padronizar. Escolha uma solução e ficar com ela (a menos que hajam Como isso ajuda a acomodar as mudanças? Simples: Quase todo BI ou dados ferramenta de depósito que você pode comprar (ou baixar) vai fazer o trabalho. Alguns pode ter uma interface mais extravagantes ou ter suporte para Flash velocímetros enquanto outros não, mas basicamente o valor do negócio é a informação, não na apresentação. Um padrão que permite ficar com uma organização para desenvolver habilidades de profundidade em torno de um produto. Então, quando uma nova exigência aparece, pode ser rapidamente implementada porque todas as habilidades já estão disponíveis. Este livro ajuda você a atingir esse objetivo para o Pentaho BI Suite mas a mensagem mesmo se aplica a todas as outras soluções também.

Tendências do Armazém de Dados

Concluímos este capítulo, destacando algumas das atuais e recentes desenvolvimentos no armazenamento de dados: armazenamento de dados virtual, em tempo real armazenamento de dados, bancos de dados analíticos, aparelhos de data warehouse, e on-demand armazéns de dados.

Data Warehousing Virtual

Este livro cobre uma abordagem clássica para o armazenamento de dados, o que significa projetar e construir uma nova arquitetura armazenamento de dados e mover fisicamente os dados para este armazenamento de dados, fazendo uso de ferramentas de ETL. Esta abordagem é também conhecido como física armazenamento de dados. E quando há um físico solução, há provavelmente uma solução virtual. Como você já pode ter adivinhado, um data warehouse virtual não armazena uma réplica dos dados extraídos dos sistemas de origem, mas os dados permanecem nos dados operacionais lojas. Do ponto de vista do usuário, no entanto, há uma camada especial criado que traduz os dados transacionais em uma exibição de data warehouse. Dados Virtual soluções de armazenamento também pode integrar os dados dos sistemas de origem diferente, oferecendo uma exibição ao vivo para os dados, pois é no momento da execução de uma consulta.

Esta é provavelmente a mais notável vantagem da utilização deste tipo de solução: é em tempo real, os dados atuais que você estará olhando, e você não precisa de muito grande de ferro para armazenamento de dados antes de ser capaz de analisar e informar sobre ela. Naturalmente, há são muitas desvantagens também: os dados não são limpos, conformada e validado. Pode drenar o desempenho dos sistemas de origem, e não há nenhuma noção de

a história de todos os dados são apresentados na sua forma actual, o que torna difícil de usar de tendências e comparações históricas.

Em muitos casos, soluções virtuais e físicos são usados lado a lado, onde uma consulta dirigida ao armazém de dados recupera dados históricos de o armazém de dados físicos e complementa-as com os dados atuais do os sistemas de origem. Isso requer uma camada muito inteligente de metadados, que é ainda não está disponível no quadro Pentaho. Uma opção a considerar na neste caso, é o banco de dados LucidDB analítica, que pode ser usado para embrulhar"" vistas em torno de qualquer sistema de origem (incluindo, por exemplo, arquivos de texto) e, em seguida, comporta-se como uma solução federados (veja a seção que segue na"Analítica Bancos de dados"). As consultas são então orientados para o banco de dados LucidDB, que por sua vez redireciona as consultas aos sistemas de origem respectiva. Note-se que esta é apenas uma alternativa leve para várias soluções proprietárias no mercado que também conter caching inteligente e mecanismos de indexação, que LucidDB falta. A de pleno direito open source solução de federação de dados não está disponível por muito tempo, mas no início de 2009 a Red Hat fez sua solução adquirida MetaMatrix disponível um projeto open source chamado Teiid (http://www.jboss.org/teiid).

Referência Cruzada Para obter mais informações sobre os dados virtual de armazenagem, consulte as seguintes fontes:

- www.tdwi.org / Publicações / WhatWorks display.aspx /? id = 7305
- www.b-eye-network.com/channels/5087/view/9752/
- www.b-eye-network.com/channels/1138/view/663
- www.ebizq.net/topics/eii/features/8153.html

Real-Time Data Warehousing

Nos primeiros dias de data warehousing, semanais e até mesmo cargas mensais eram uma prática comum. Atualmente, as cargas diárias são consideradas padrão e há uma tendência para se mudar para intradia ou até mesmo dados quase em tempo real cargas. Essa tendência coincide com uma outra, que visa trazer Business Intelligence (BI) para o nível operacional da organização, também chamado BI Operacional. Um bom exemplo disso é os representantes de atendimento ao cliente necessidade de up-to-date informações sobre suas contas, incluindo todos os históricos informações relacionadas aos seus clientes. BI Operacional é sobre um monte de pequenas decisões a serem tomadas por muitas pessoas como parte de seu trabalho diário, em oposição ao BI tático ou estratégico, o que deixa mais tempo para revisão completa. A diagrama na Figura 6-9 mostra a relação entre o tempo necessário para fazer uma decisão e agir sobre ela, eo valor perdido quando a latência entre o original evento de negócios e da decisão tomada é muito alto.

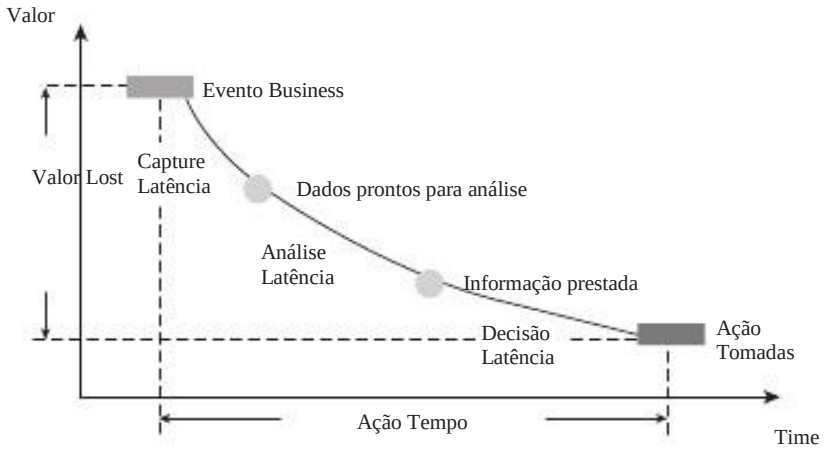


Figura 6-9: Ação valor diagrama

Deste diagrama, você pode ver que obter os dados para o data warehouse em tempo real é apenas uma parte da história. Dados ainda precisam ser analisados e atuada. Existem várias maneiras de implementar uma solução que carrega os dados quase em tempo real, conforme abordado anteriormente no""Mudou Data Capture seção, e diversas soluções comerciais no mercado que pode simplificar a configuração de uma tal alternativa. Uma solução que não foi mencionado lá porque é não faz parte do CDC é o uso de filas de mensagens. Nesse caso, cada evento de negócios e os dados resultantes é imediatamente repassado como uma mensagem que pode ser pegou e transformado pelo ambiente de data warehouse. Em todos os casos, No entanto, essa atualização em tempo real, precisa ser firmemente ligado para as próximas etapas no processo, que também pode ser automatizado. Por exemplo, a análise pode ser feito usando regras de negócio que dispara automaticamente quando um determinado limiar é alcançado, que então inicia ações correspondentes também. Há um Casal de muito boa fonte aberta regras de negócio motores disponíveis, com Drools (www.jboss.org / baba) E OpenRules (http://openrules.com), Provavelmente sendo a mais madura e bem conhecida. Pentaho seqüências de ação Os capítulos 4, 13 e 17) também pode ser usado para essa finalidade.

Em tempo real, soluções baseadas em mensagens que podem ser aplicadas regras de negócio e fazer cálculos em tempo real são normalmente tratados como processamento de eventos complexos (CEP).

Uma solução open source CEP é Esper (www.espertech.com), Que é um solução baseada em mensagens que podem lidar com grandes volumes de dados de streaming.

Estas ferramentas podem complementar ambientes de data warehouse, mas não são uma substituto para o processo ETL conforme descrito anteriormente.

Referência Cruzada Para obter mais informações sobre os dados em tempo real de armazenagem, consulte as seguintes fontes:

- www.tdwi.org / Publicações / WhatWorks display.aspx /? id = 8913

- www.ebizq.net / temas / operacionais bi / recursos / 10604.html
- www.information-management.com/issues/20040901 / 1009281-1.html

Bancos de dados analíticos

Neste capítulo, nós já explicou por que um armazém de dados precisa ser arquitetada de uma maneira diferente do que um sistema operacional. No entanto, ainda mais em casos, o mesmo RDBMS que é usado para o sistema de transação também é selecionada como banco de dados do data warehouse. Nos últimos anos, muitos novos produtos entrou no mercado, que contestou esta abordagem através da oferta de recursos que são especificamente adaptadas para lidar com cargas de trabalho analítico. A utilização de OLAP bases de dados de data marts já foi mencionado, mas uma das mais abordagens radicalmente novo é o surgimento dos chamados bases de dados em colunas ou lojas de colunas. bases de dados Colunar não armazenam dados em linhas seqüencialmente, mas horizontalmente em colunas. Isso resulta em uma série de vantagens:

- Poda-Mais consultas analíticas pertencem somente a algumas colunas. Consultas em bancos de dados baseados em linha sempre ler linhas completas a partir do disco, mesmo quando apenas duas colunas são selecionadas. Em uma loja de coluna, apenas os selecionados colunas precisam ser lidos. Esta característica influencia fortemente o disco I / O, que é um dos limitadores de velocidade maior.
- Não há necessidade de indexação Porque uma loja de coluna já armazena os dados coluna por coluna na ordem correta, cada coluna é automaticamente sua próprio índice.
- Porque compressão uma coluna sempre contém dados de um determinado tipo de dados, os dados podem ser compactados de forma muito eficiente, reduzindo ainda mais disco I / O.

Somado a estas vantagens, as lojas de coluna mais comerciais, como Vertica, ParAccel e XASOL pode ser executado em um processamento paralelo massivo (MPP) cluster, o que reforça ainda mais o desempenho dessas soluções.

Ao considerar as alternativas de código aberto, há três produtos no valor de vendo:

- MonetDB (www.monetdb.com) É um produto desenvolvido em holandês Livre Universidade de Amesterdão. Tem uma pegada muito pequena (aproximadamente 5 MB) e é projetado para funcionar principalmente na memória principal, o que explica a impressionante resultados de benchmark, mesmo para conjuntos de dados um pouco maiores de 10 e 20GB. MonetDB é recomendado para uso quando houver a necessidade de tempos de resposta rápidos sólo com data warehousing e business intelligence em mente. conjuntos de dados menores (<10GB).
- LucidDB (www.luciddb.org) É um banco de dados de colunas concebido a partir do

Ele também inclui recursos de ETL com extensões SQL especiais para construir transformações. Os desenvolvedores do Pentaho e LucidDB trabalhar em estreita colaboração
juntos, o que torna esta uma alternativa viável para pequenos data warehouse ou dados implementações mart. LucidDB é um armazenamento de dados preferido para a
Mondrian e tem suporte embutido para o designer Pentaho agregado. Em Além disso, Pentaho Data Integration contém um conector LucidDB nativas a granel e carregador.

- Infobright (www.infobright.com) Lançou uma edição de sua comunidade banco de dados analíticos em setembro de 2008 sob a licença GPL. O mais característica notável do Infobright é que ele age como um mecanismo de armazenamento do MySQL,
assim você pode usar o produto para aumentar a capacidade de armazenamento de dados
do MySQL sem ter de executar ou aprender novas ferramentas. Infobright é capaz de lidar com volumes de dados de até 30 Terabytes e, portanto, é um grande companheiro para a suíte de BI Pentaho. Um ponto de cautela, no entanto: o versão de código aberto não oferece suporte a manipulação de dados SQL INSERT, UPDATEE DELETE declarações. Dados precisa ser recarregado toda vez que se você precisa para fazer as atualizações, tornando-se dificilmente utilizável como um armazém de dados
solução. A versão comercial não tem essas limitações.

Nenhum dos três produtos mencionados aqui têm o processamento paralelo (MPP) capacidades, daí não pode haver limitações de escalabilidade. No entanto, os dados armazéns fora do alcance Terabyte 10 ainda não são muito comuns, de modo a maioria dos casos, essas alternativas podem ajudar na criação de um elevado desempenho de arquitetura de dados do warehouse utilizando as soluções de código aberto.

Armazém de Dados Eletrodomésticos

O desenvolvimento final visível no mercado de data warehouse é o aumento
da assim chamada aparelhos. Nós todos sabemos que os aparelhos, como torradeiras ou rádios
que só têm de ser conectado a uma tomada e você está pronto e funcionando.
Aarmazém de dados do aparelho é uma solução plug-and-play que consiste em hardware e software e tem como objectivo tornar o armazenamento de dados tão fácil como torrar uma fatia de
pão. Porque este é um mercado emergente, com apenas alguns fornecedores estabelecidos e um monte de empresas que entraram recentemente no mercado, você deve tomar cuidado especial quando se considera uma destas soluções.

Várias empresas que começaram como vendedores de banco de dados analíticos agora parceiro
com empresas de hardware como a Sun (agora Oracle) ou a Hewlett Packard
entregar um aparelho completo. Um fenômeno interessante é a adoção
da tecnologia open source dentro destes aparelhos. PostgreSQL, especialmente os analisador de SQL, é utilizada por muitos dos novos fornecedores. Greenplum, por exemplo, não apenas com base a toda a solução no PostgreSQL, mas também incorpora o aberto biblioteca de fontes estatísticas Projeto de R para análise avançada de dados e banco de dados
mineração.

Uma das últimas novos operadores no mercado é o aparelho Kickfire (www.kickfire.com), Que é de particular interesse porque é arquitetada como um mecanismo de armazenamento do MySQL pluggable. Isso permite que os armazéns de dados desenvolvido para MySQL com InnoDB ou MyISAM a ser migrado sem problemas a um aparelho Kickfire. A empresa é um dos poucos fornecedores de aparelho ter publicado resultados de avaliação de desempenho em www.tpc.org e é um líder do preço / desempenho nos 100 e 300 Gigabyte benchmark TPC-H. TPC-H é uma referência no setor de medição de desempenho de consulta de BI típica consultas como para determinar a quota de mercado ou de volume de vendas de comunicação para um determinado período.

Em Data Warehousing Demand

Com o advento da chamada computação em nuvem soluções [com Amazon Elastic Cloud Computing (EC2) da infra-estrutura sendo provavelmente o mais conhecido, mas praticamente a única solução no mercado], ele está se tornando viável para acolher um armazém de dados ou até mesmo uma solução de BI completa fora do firewall corporativo. Soluções para hospedagem de aplicativos em um centro de dados off-site já estão disponíveis desde o final dos anos noventa e foram oferecidos por Application Service Providers (ASPs). Para distinguir os mais recentes desenvolvimentos do conceito ASP original, o prazo Software como Serviço ou SaaS foi introduzido, e às vezes as pessoas mesmo falar de Banco de Dados ou Data Warehouse como um serviço (DaaS, DWaaS[3]). A grande diferença é a arquitetura compartilhada, onde todos os clientes de um serviço partes uma solução de infra-estrutura comum, não só a nível técnico, mas também a o nível de aplicação. Ao dividir os recursos disponíveis de forma inteligente (Também chamado de multi-tenant solução), muitos usuários podem compartilhar os mesmos servidores, pedidos e licenças, tornando-a muito custo modo efetivo de operação. Desenvolvimentos nesta área estão avançando muito rápido por isso não podemos cobrir qualquer fornecedor específico ou oferecendo aqui (não são simplesmente demasiado muitos deles), mas na luta por uma visão completa precisávamos ressaltar a disponibilidade de estas soluções também.

Resumo

Este capítulo apresenta uma ampla gama de conceitos e tecnologias relacionados com a armazenamento de dados. Nós mostramos-lhe as seguintes coisas que vamos construir para o resto do livro:

- O que é um data warehouse e como ele pode ajudá-lo a atingir seus objetivos de negócio

3 Qual é engraçado, uma vez que a palavra holandesa Dwaas""significa"Louco."

- A arquitectura global de um armazém de dados e todo o edifício que constituem blocos que podem ser identificadas dentro desta arquitetura
- O objetivo de cada um dos blocos de construção do armazém de dados
- Os principais desafios que enfrentam na construção de um armazém de dados: dados qualidade, dados de desempenho, volume, capturando os dados alterados e adaptação a evolução das necessidades

Também destacamos a principal diferença entre os dois líderes do autor-
dades em modelagem de dados do armazém, Ralph Kimball e Bill Inmon, e apontou
por que nós escolhemos modelagem tridimensional e arquitetura de barramento para a nossa
exemplo de data warehouse.

Por fim, destacamos as principais tendências em armazenamento de dados e ilustrado
como o Pentaho BI Suite, o banco de dados MySQL, e outros relacionados com fonte aberta
tecnologias caber dentro

CAPÍTULO

7

Modelagem de Negócios Usando esquemas Star

Ao trabalhar com os dados do data warehouse, é provável que esquemas em estrela são usados para entregar os dados para o usuário final. Não diretamente, no entanto, normalmente, uma ferramenta de relatório ou análise é usado para acessar os dados, enquanto os mais avançados os usuários podem usar uma ferramenta de consulta diretamente. No entanto, é importante observar que esquemas estrela é o veículo de escolha se trabalhar com um estilo Kimball arquitetura de dados de ônibus ou uma fábrica de informações Inmon estilo corporativo.

NOTA Há uma exceção a esta regra. Um analista avançado ou minerador de dados muitas vezes precisa de acesso ao conjunto completo de dados no armazém de dados, assim contornar os esquemas estrela criada em um data warehouse Inmon estilo solução. Essa discussão, entretanto, será limitada a estrela acesso esquema.

O que é um esquema em estrela?

A primeira coisa que você pode fazer é: Por que esses modelos de banco de dados chamado"Star esquema?" Provavelmente porque o diagrama de entidade-relacionamento deste tipo de esquema se assemelha a uma estrela. O centro da estrela consiste em uma tabela fato grande e os pontos da estrela estão as tabelas de dimensão. A maioria dos usuários primeiro encontro de uma estrela esquema em um data mart de vendas com clientes, produtos, lojas, promoções e tempo, como mostrado na Figura 7-1.

NOTA Apesar de usarmos cinco pontos para tornar o modelo se parecer com uma estrela, não é de meios necessários para usar cinco pontos. Na verdade, mesmo uma tabela fato com apenas uma ou duas dimensão é também chamado de esquema estrela.

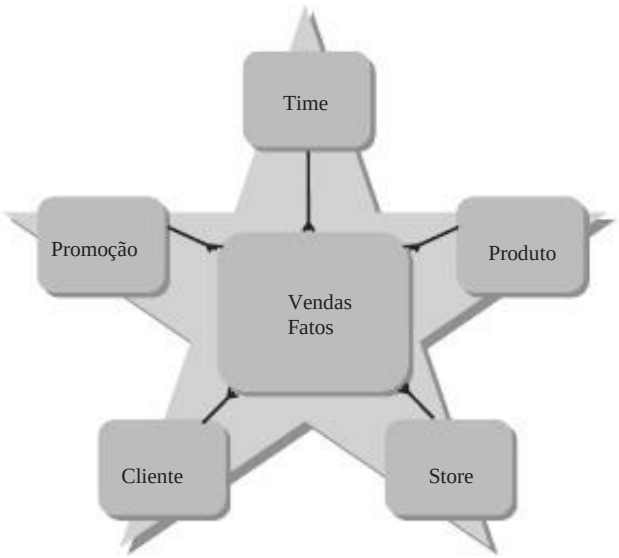


Figura 7-1: Star diagrama de esquema

É a técnica de modelagem que conta, não o número de tabelas de dimensão utilizado. Um dos benefícios óbvios de usar um esquema como este é a sua simplicidade e compreensão para os usuários finais. Muitas vezes, durante a fase de concepção do um armazém de dados, esquemas estrela são usados para desenhar a tradução inicial de questões de negócio em banco de dados diagramas lógicos.

NOTA Um diagrama simples e lógico ainda pode levar a uma maior complexidade técnica projeto que resulta em um modelo que é menos compreensível para o usuário final que o projeto da placa inicial branco.

As tabelas de dimensão e tabelas de fato

A próxima pergunta é, naturalmente: Qual é a diferença entre a dimensão e tabelas de fato, eo que torna algo uma tabela de fatos? Como você vai ver depois, a distinção nem sempre é clara, e há mesmo ocasiões em que uma tabela de dimensão em um esquema estrela pode tornar-se uma tabela de fatos de outra estrela esquema. Uma explicação simples é que as tabelas de dimensão conter informações sobre as entidades de negócio (clientes, produtos, lojas) e as tabelas de fato sobre eventos de negócios (vendas, transferências, ordens). A diferença mais notável é a as colunas mensuráveis, tais como receitas, custos e itens, que fazem parte da as tabelas de fatos. Além disso, todos os diferentes ângulos e atributos que são necessários para resumir estes fatos são armazenados nas tabelas de dimensão. É realmente muito simples, se você traduzir um pedido de relatório típicos, tais como''Mostre-me o total valor da ordem por mês, por grupo de produtos''em um modelo tridimensional, como apresentado na Figura 7-2. O item calculado de interesse (soma do valor do pedido) É um facto, o mês é um tempo""atributo pertencente à dimensão de tempo e

"Grupo"produto é um produto de atributos pertencentes à dimensão do produto.
A tabela de fatos, portanto, contém apenas as chaves estrangeiras apontando para a dimensão
tabelas e atributos que podem ser agregados (elementos quantitativos). Dimensão tabelas contêm todos os atributos que descrevem uma perspectiva certa organização (Elementos qualitativos).

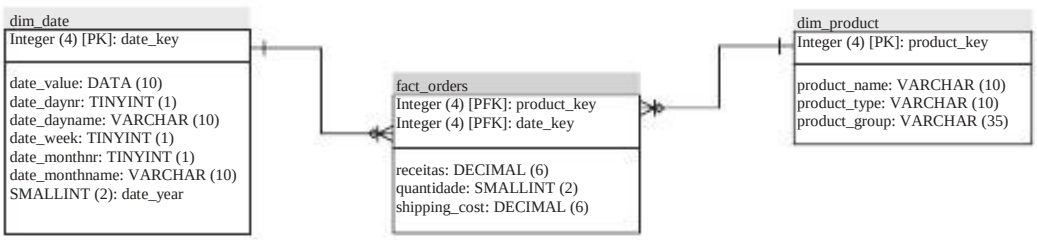


Figura 7-2: Star exemplo de esquema

Esta data mart lhe dá muito pouca informação, pois contém apenas algumas medições e duas dimensões. No entanto, é um bom ponto de partida ponto para ilustrar como funciona exatamente um modelo tridimensional:

- Todas as linhas de fato são armazenados no mais baixo possível granularidade nível. Por granularidade, queremos dizer a granulação dos dados, uma tabela de fatos no nível de data
- A granularidade da tabela de fatos é determinada pela combinação do granularidade das tabelas de dimensão. No exemplo, isto significa que para cada combinação de produto e data de há um possível fato linha. Nem todos os combinações devem estar presentes, é claro!

- Todas as medidas na tabela de fatos pode ser enrolada ou agrupadas pelos elementos das tabelas de dimensão, de modo que este data mart pouco pode calcular sua receita mês, ano e grupo de produtos ou por tipo e ano do produto, ou qualquer outra combinação desejada.

Você poderia argumentar que uma ordem é também uma entidade empresarial e apenas a ordem
linhas com as operações específicas são fatos. De certa forma, essa suposição é correta. Este é um caso onde a atenção especial é necessário na modelagem dos dados marts. Mais adiante neste capítulo vamos explicar o que é necessário uma atenção especial e descrever outros tópicos mais avançados de modelagem dimensional. Por agora, Basta manter a definição simples na mente.

Tabela de tipos de Fato

As tabelas de fatos nem sempre são baseados em transacções sozinho. Transacional tabelas de fatos
são o tipo mais comumente usado, mas há um par de variações que precisa estar ciente. A primeira é a acumulando instantâneo periódico tabela. Isso soa mais complicado do que realmente é, a única diferença entre

uma tabela de fatos regular e este são os campos de data vinculada a um processo específico que
precisa ser atualizada quando novos eventos ocorrem. Tomemos, por exemplo, World Class processo de locação do filme, onde os DVDs são ordenados em uma data do pedido, enviado sobre uma data de lançamento, e retornou em uma data de retorno. Porque não quero esperar até que o processo está concluído para começar a digitar os dados, vamos começar por adicionar
a linha de fato para o data warehouse, quando o pedido é feito e, posteriormente, acumular esta linha com as datas corretas como eles se tornam conhecidos.

Um tipo bem diferente da tabela de fatos é o periódica instantâneo, que é como um foto tirada a cada dia, semana ou mês. instantâneos periódicos são usados quando é preciso congelar os dados que estão apenas indiretamente relacionadas com a transacção. Um bom
exemplo, no armazém de dados WCM é o inventário de DVD. Os níveis de estoque mudam constantemente e não há maneira de controlar essas longo do tempo por usando uma tabela de fatos transacional. A única opção aqui é periodicamente tomar uma instantâneo que permite que você relate sobre o aumento de nível de estoque ou diminuição ao longo do tempo.

Outra maneira de olhar para as tabelas de fatos de transacção e é pelo instantâneo tipos de medidas que eles contêm. Medidas na tabela de transação regular, como receita de vendas, pode ser resumida em todas as dimensões, incluindo o tempo. Estes são referidos como fatos aditivo ou medidas. As medidas em um periódico tabela instantâneo não pode ser facilmente resumida, pelo menos não quando vários períodos estão envolvidos. Resumindo os níveis de estoque semanal de DVD não lhe dar uma estoque total anual! Estas medidas são chamados semi-aditivo, o que significa que poderá acrescentá-los juntos (por exemplo, o inventário total de todos os filmes que starring Tom Cruise), mas precisa incluir um período de tempo ou de filtro para cortar os dados.

O último tipo de medida é o não-aditivos fato. Um bom exemplo desta é a temperatura ambiente. Não faz sentido para resumir a temperatura de salas diferentes ou períodos diferentes, apesar de você sempre pode calcular médias.

NOTA A linguagem SQL não oferece suporte semi-aditivos e não aditivos medidas, mas alguns servidores OLAP pode ser configurado para acomodar exceção de agregação. Pentaho Mondrian (abordados no Capítulo 15) não tem suporte interno para medidas não e semi-aditivos.

A próxima parte do capítulo aborda o SQL básicos necessários para a obtenção de dados de esquemas em estrela com tabelas transacionais e de fato se acumulando. Porque esses dois tipos de tabela de fatos contém transações, você pode resumir com segurança os números que estão aí.

Consultando esquemas Star

Até agora, você viu diagramas ligando tabelas de dimensão e de fato juntos, mas em um banco estas são apenas as tabelas. As relações nos diagramas são chave estrangeira

restrições, o que significa que não será possível inserir uma chave de produto em a tabela de fatos ordens se essa chave não existir no dim_product dimensão tabela. Obtendo dados do banco de dados é outra coisa. Para isso, você precisa uma linguagem de consulta chamada Structured Query Language, ou SQL. Se você não está familiarizado com SQL, não se preocupe, o SQL que você precisa para recuperar informações a partir de um esquema em estrela não é muito complicado. E, embora este livro não é um texto SQL, cobrimos o básico suficiente aqui para que você possa começar a escrever consultas para recuperar dados de data marts dados dimensionais. Tudo o que precisamos para isso é
o SELECT declaração, uma forma de juntando os diferentes quadros juntos, e um compreensão de como grupo os resultados da consulta. Para começar, tem um olhar para os blocos básicos de construção de SQL para selecionar dados:

- SELECT -Uma lista de colunas, constantes e expressões que deseja recuperar a partir do banco de dados.
- DA -As tabelas e exibições que contêm os dados que você precisa, incluindo as relações entre essas tabelas e exibições.
- ONDE -As restrições que você deseja aplicar aos dados no selecionado tabelas, exibições e expressões, excluindo as agregações.
- GROUP BY -Especifica o nível de resumo da consulta. Todos os não-agregado colunas e expressões na SELECT lista deve ser parte do GRUPO POR declaração.
- TENDO -Contém condições de resumir expressões, por exemplo, Tendo soma (receitas)> 100000.
- ORDER BY -Indica a ordem em que os resultados serão exibidos.

Se apenas a receita total é necessário que a tabela de fatos, não há necessidade de usar JOINS ou GROUP BY cláusulas, apenas o seguinte:

```
SELECT SUM (receita)
FROMfct_orders
```

Esta consulta pouco não lhe diz muito, assim você pode adicionar a data dimensão para o cálculo da receita por ano:

```
SELECT      date_year, SUM (receita)
DA          fct_orders
JOIN        dim_date
ON          dim_date.date_key fct_orders.date_key =
GROUP BY    date_year
```

Agora você tem uma consulta simples que é a base para todas as consultas de outros possíveis
que pode ser executada contra um modelo de dados dimensional. A primeira linha com o SELECT cláusula diz que o banco de dados do conjunto de resultados será semelhante. Em seguida,
você precisa dizer de onde os dados são recuperados usando o DA cláusula.

Você também quer JOIN uma outra mesa e você precisa dizer que o banco de dados colunas de junção será baseada, neste caso, o date_key na dimensão e tabelas de fato. Finalmente, os resultados são agrupados por todos os elementos do SELECT declaração de que não fazem parte de uma função agregada. Por exemplo, quando o SELECT declaração parecido com este:

```
date_year SELECT, product_type, product_name, SUM (receita), SUM (quantidade)
```

o GROUP BY declaração deve ser

```
GRUPO PELA date_year, product_type, product_name
```

Para concluir esta consulta tudo que você precisa para adicionar são os DA e JOIN partes:

```
SELECT product_type, date_year, product_name,
       SUM (receita), SUM (quantidade)
FROM fct_orders
JOIN dim_date
ON fct_orders.date_key dim_date.date_key =
JOIN dim_product
ON fct_orders.product_key dim_product.product_key =
GRUPO PELA date_year, product_type, product_name
```

Os exemplos do SQL utilizado até agora ainda não o uso de apelidos. Um alias é outra nome que você pode usar para se referir a uma tabela, coluna ou expressão. Embora eles sejam nem sempre é necessário, é boa prática sempre usá-los. Adicionando o alias total_revenue à expressão SUM (receita) dá mais sentido à coluna do resultado, e o uso de aliases para nomes de tabela permite uma notação abreviada das condições de junção. Compare o seguinte JOIN cláusulas para entender o que queremos dizer com isso:

```
fct_orders DA
JOIN ON dim_date fct_orders.date_key dim_date.date_key =
```

versus

```
fct_orders FROM como f
JOIN dim_date SOBRE AS d f.date_key d.date_key =
```

Há casos em que o uso de apelidos não é opcional, mas necessário. A primeiro caso concreto em que os atalhos são, pelo menos, é muito útil quando o mesmo nome da coluna aparece em mais de uma tabela do conjunto utilizado na consulta. Suponha que no exemplo anterior você quisesse incluir date_key na conjunto de resultados. Como esta coluna é parte da realidade como a tabela de dimensão, o analisador de SQL não é possível determinar qual a tabela para escolher a coluna, assim você

necessidade de incluir a referência da tabela em seu SELECT declaração, por exemplo, dim_date.datekey SELECT. Agora é imediatamente óbvio que o uso de uma alias curto, como dvai poupar muita digitação. A instrução de consulta completa, incluindo o uso de apelidos, agora é a seguinte:

```
SELECT d.product_type, d.date_year, p.product_name,
       SUM (f.revenue) AS total_revenue, SUM total_quantity (f.quantity) AS
FROMfct_orders AS f
JOINdim_date como d = ONf.date_key d.date_key
SOBRE AS JOINdim_product p = f.product_key d.product_key
GRUPO PELA d.date_year, p.product_type, p.product_name
```

O segundo caso, onde alias não são apenas práticos, mas é necessária quando a mesma tabela é usada em diferentes papéis. Um exemplo é uma pesquisa que pede para todos
receitas por tipo de produto com data de fim em 2007 e uma data de lançamento em 2008:

```
SELECT    p.product_type, SUM (f.revenue) AS total_revenue
DA        fct_orders AS f
JOIN      dim_date SOBRE AS od f.order_date_key od.date_key =
JOIN      dim_date como SD ON f.ship_date_key sd.date_key =
JOIN      p dim_product SOBRE AS f.product_key d.product_key =
ONDE      od.year = 2007
E         sd.year = 2008
GROUP BY  p.product_type, d.date_year, p.product_name
```

A ONDE parte desta consulta é abordado em mais detalhes em breve.

Junte-se a tipos de

Nos exemplos usados até agora, usamos a cláusula JOIN para combinar diferentes tabelas. Embora isso seja tecnicamente correta, JOIN é um atalho para a lista completa declaração INNER JOIN. O termo interior significa que as entradas das colunas deve tem um jogo em duas colunas. Isso nem sempre é o caso, há, talvez, produtos que não foram vendidos ainda. Usando um INNER JOIN entre os produtos e ordens, portanto, gerar um conjunto de resultados que contém apenas os produtos para que existe uma ordem. Para contornar essa limitação, você também pode usar exterior junta-se, que irá retornar as linhas vazias também. Os três OUTER JOIN tipos são ESQUERDA,DIREITOE FULL, E há também um especial CROSS JOIN que combine qualquer valor da coluna de associação de tabelas com qualquer valor da tabela associada. O conjunto de resultados de uma CROSS JOIN também é conhecido como um produto cartesiano. Para melhor
compreender os diferentes tipos de junções, você pode usar os exemplos a seguir como uma referência. Todos os exemplos são baseados em duas tabelas, TABLE_A e TABLE_B, Cada com duas colunas, id e valor.

```
uma tabela
+----+-------+
| Valor | id |
+----+-------+
| 1 | A |
| 2 | B |
| 3 | C |
| 4 | D |
+----+-------+
```

```
b tabela
+----+-------+
| Valor | id |
+----+-------+
| 1 | Vermelho |
| 3 | Blue |
| 5 | Amarelo |
+----+-------+
```

Um regular INNER JOIN entre as duas tabelas irá produzir um conjunto de resultados com apenas os valores comuns de ambas as tabelas:

```
SELECTa.value AS a, b b.value AS
AS FROMtable_a um
INNER JOIN TABLE_B SOBRE AS b a.id = b.id
```

```
+-------+-------+
| A | b |
+-------+-------+
| A | Red |
| C | Blue |
+-------+-------+
```

A LEFT OUTER JOIN irá exibir todos os valores da tabela da esquerda na junção instrução e exibe uma NULL valor para os valores na tabela à direita sem correspondência id:

```
SELECTa.value AS a, b b.value AS
AS FROMtable_a um
LEFT OUTER JOIN TABLE_B SOBRE AS b a.id = b.id
```

```
+-------+-------+
| A | b |
+-------+-------+
| A | Red |
| NULL | B |
| C | Blue |
| D | NULL |
+-------+-------+
```

A RIGHT OUTER JOIN exibe todos os valores da tabela da direita na cláusula de junção e exibe uma NULL valor para os valores na tabela à esquerda sem correspondência id:

```
SELECTa.value AS a, b b.value AS
AS FROMtable_a um
RIGHT OUTER JOIN TABLE_B SOBRE AS b a.id = b.id
```

```
+-------+--------+
| A | b |
+-------+--------+
| A | Red |
| C | Blue |
| NULL | Amarelo |
+-------+--------+
```

A FULL OUTER JOIN declaração não está disponível no MySQL, mas outros bases de dados como MS SQL Server e Oracle produziria o seguinte saída de um FULL OUTER JOIN:

```
+-------+-------+
| A | b |
+-------+-------+
| A | Red |
| NULL | B |
| C | Blue |
| D | NULL |
| NULL | Amarelo |
+-------+-------+
```

A última opção, o CROSS JOIN, Não necessita de colunas de junção de todos:

```
SELECTa.value AS a, b b.value AS
AS FROMtable_a um
CROSS JOIN TABLE_B AS b
```

```
+-------+--------+
| A | b |
+-------+--------+
| A | Red |
| A | Blue |
| A | Amarelo |
| B | Red |
| B | Blue |
| B | Amarelo |
| C | Red |
| C | Blue |
| C | Amarelo |
| D | Red |
| D | Azul |
| D | Amarelo |
+-------+--------+
```

Cruz junta são ótimos para criar relatórios de tabela cruzada em que você quer ter uma grade completa com todos os grupos de produtos no eixo y, e todos os meses no x eixo, independentemente de dados para uma combinação específica existe ou não.

DICA Recomendamos que você use sempre a completa adesão cláusulas; uso INNER JOIN em vez de JOIN e LEFT OUTER JOIN em vez de LEFT JOIN.

Restrições aplicáveis em uma consulta

Na maioria dos casos, uma questão de negócio é mais específica do que as receitas de todos os produtos, todos os clientes, e todos os anos. E quando data marts conter milhões de linhas de dados, também é inviável para selecionar dados sem colocar alguns restrições na consulta. Por isso, a SELECT declaração também pode ser estendida com uma ONDE cláusula para filtrar apenas os dados de interesse. O mais simples ONDE condição é apenas um filtro em um determinado valor, por exemplo, date_year = 2008 ou date_year> = 2008. A maioria dos operadores de comparação que você já está familiarizado com pode ser usado para definir as seleções, como =,<,>,<=,> = e <> ou ! =. Esses operadores comparam o conteúdo da coluna restrito ou expressão, com um único valor. Suponha que você precisa para filtrar todos os clientes que tem um sobrenome com três caracteres ou menos. Então você pode usar o Char_length função para calcular o número de caracteres em uma string e compare isso com o valor 3:

```
ONDE char_length (c.last_name) <= 3
```

As opções do passado, vamos apresentar aqui são os EM e Entre ... e operadores. Usando EM, A parte esquerda da restrição pode ser comparado a mais de um único valor sobre o direito, por exemplo, para selecionar todos os clientes cujos nomes começar com um A,E,IOu Q:

```
WHERE SUBSTRING (c.last_name, 1,1) IN ('A', 'E', 'I', 'Q')
```

O lado direito do EM comparação não precisa ser uma lista predefinida, mas pode ser uma nova consulta, bem como, que nesse caso é chamado de subconsulta. Subconsultas pode, naturalmente, também ser usado em combinação com todos os outros operadores de comparação.

A Entre ... e operador pode ser usado para definir um limite inferior e superior a uma restrição. Entre ... e não exatamente se comportam da maneira como você pode esperar que isso tenha cuidado ao usar este. Quando Entre ... e é utilizado, como no a condição P. PRODUCT_PRICE entre 1 e 5, Os valores 1 e 5 são incluídos na restrição.

ONDE cláusulas não podem ser colocados em qualquer lugar do SELECT declaração, há regras muito rígidas para sua utilização. As condições devem ser colocados logo após o

DA mas antes GROUP BY, Como no exemplo a seguir:

```
SELECT p.product_type, d.date_year, p.product_name,
       SUM (f.revenue) AS total_revenue, SUM total_quantity (f.quantity) AS
FROM fct_orders AS f
INNER JOIN dim_date como d = ONf.date_key d.date_key
INNER JOIN dim_product SOBRE AS p = f.product_key d.product_key
WHERE CHAR_LENGTH (c.last_name) <= 3
GRUPO BY d.date_year, p.product_type, p.product_name
```

Combinando múltiplas restrições

Se você quiser usar mais de uma condição em uma ONDE cláusula, você pode combinar estes com E e OU operadores. Para aplicar uma lógica mais complexa ou determinado grupo declarações em conjunto, abrindo e fechando parênteses são necessários. Você já visto o Entre ... e operador, que é na verdade uma abreviação da usando duas comparações utilizando > = e <=. Nesses casos complexos, você provavelmente vai querer usar parênteses para agrupá-los também. O seguinte é um exemplo em que o uso de parênteses é obrigatório:

```
(D.year = 2007 d.month_nr E = 12) ou (d.year = 2008 d.month_nr E = 1)
```

Sem os parênteses, o conjunto de resultados estariam vazios, porque não existe uma única linha que satisfaz D. MONTH_NR = 12 e MONTH_NR D. = 1. Quando o parênteses são usados, o conjunto de resultados inclui os dados de dezembro de 2007 e Janeiro de 2008.

Restringir resultados agregados

Até agora temos usado restrições para filtrar os dados que foi armazenado no banco de dados. A ONDE cláusula só pode filtrar constantes, expressões, e colunas, não agregados. Há situações em que os valores resultantes de agregação precisam ser filtrados após a GROUP BY cláusula foi aplicada. Por exemplo, quando você está interessado apenas em clientes que têm uma receita total de pelo menos 100 dólares, é necessário resumir primeiro e filtrar depois. Isto pode ser conseguido por usando o TENDO cláusula:

```
SELECT      c.first_name, total_revenue c.last_name, SUM (f.revenue) AS
DA          fct_sales AS f
INNER JOIN  dim_customer AS C ON c.customer_id = f.customer_id
GROUP BY    c.first_name, c.last_name
TENDO       SUM (f.revenue)> = 100
```

Ordenação de Dados

A menos que você aplicar uma ordem de classificação específico, os dados retornados é exibido como um
lista, sem qualquer ordenação, o que torna difícil de navegar através de um resultado definido. Ordenação de dados é bastante simples, basta adicionar um extra ORDER BY cláusula em
consulta e adicionar as colunas que pretende ordenar os dados. A ordenação é aplicada Da esquerda para a direita na indicação da ordem. A ordenação padrão é crescente, mas os dados podem ser requisitados descendente, bem, adicionando o DESC ordenação especificação:

```
SELECT      c.last_name c.first_name, sum (f.revenue) AS total_revenue
DA          fct_sales AS f
INNER JOIN  dim_customer AS C ON c.customer_id = f.customer_id
GROUP BY    c.first_name, c.last_name
TENDO       SUM (f.revenue)> = 100
ORDER BY    c.last_name SUM (f.revenue) DESC,
```

NOTA O MySQL permite que você use os nomes de alias na TENDO e ORDER BY cláusulas bem, mas a maioria dos bancos de dados não suportam isso.

É isso aí. Você ganhou bastante conhecimento de SQL para ser capaz de consultar o data marts iremos apresentar no restante do livro. Se você quer aprender mais sobre o SQL, existe um conjunto abundante de livros e sites disponíveis. Mais informações específicas sobre a sintaxe SELECT do MySQL pode ser encontrada na guia de referência on-line em http://dev.mysql.com/doc/refman/5.1/en/select . Html.

A arquitetura de barramento

A arquitetura de barramento de dados do armazém foi desenvolvido por Ralph Kimball e é amplamente descrita em seus livros O conjunto de ferramentas de Data Warehouse e Os dados
Warehouse Lifecycle Toolkit. Ambos os livros são publicados pela Wiley Publishing e cobrem o ciclo completo da construção, modelagem e manutenção de dados armazéns. O termo ônibus refere-se ao fato de que a data marts diferentes
data warehouse são interligados por meio dimensões conformadas. Um simples exemplo, pode explicar isso. Suponha que você tenha as tabelas de dimensão para os clientes,
fornecedores, e dimensões de produtos e pretende analisar os dados sobre vendas e transações de compra. No caso das operações de compra, o cliente é
ainda desconhecido por isso não é muito útil para incluir a dimensão do cliente no estrela esquema de compra. Para as operações de venda, a situação é ligeiramente diferente: Você precisa de informações sobre o cliente que comprou um produto e os

fornecedor o produto foi adquirido. O desenho resultante para esta pequena armazém de dados de exemplo é mostrado na Figura 7-3.

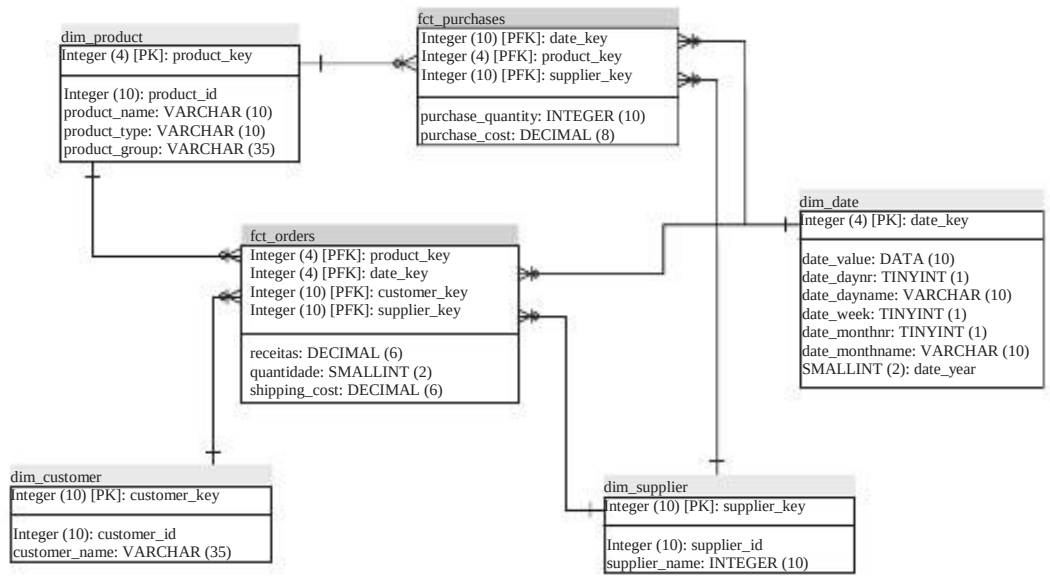


Figura 7-3: Exemplo de arquitetura de barramento

É melhor começar com uma arquitetura de barramento de alto nível antes da matriz de dados processo Mart projeto é iniciado. Figura 7-4 mostra um exemplo da matriz, onde todos fatos negócio identificadas são colocados nas linhas e todas as dimensões identificadas nas colunas. O ônibus""é formado pelo processo de negócio principal, ou o fluxo natural dos acontecimentos dentro de uma organização. No nosso caso, que seria requisitar dos fornecedores, armazenagem e movimentação de estoque, recebimento de clientes encomendas, envio de DVDs, e gestão de devoluções. Dentro de um tal negócio principal processo é fácil marcar todos os relacionamentos entre dimensões e fatos, o que torna o processo de design fácil de gerenciar e pode também ser usado para comunicar com os usuários de negócios quanto à integralidade dos dados armazém.

	Data	Time	Cliente	Lançamento DVD	Distribuidor	Armazém	Empregado	Promoção	Website
As ordens de compra	X	X		X	X		X		
Inventário	X	X		X	X	X	X		
Os pedidos dos clientes	X	X	X	X		X	X	X	X
Retorna	X	X	X	X		X	X	X	

Figura 7-4: Ônibus folha matriz

Usando a arquitetura de barramento com dimensões adequadas é o que permite a coleção de data marts a ser tratado como um verdadeiro Enterprise Data Warehouse. Cada tabela de dimensão é concebido e mantido em apenas um local, e existe um único processo para carregar e atualizar os dados. Isso contrasta fortemente com uma coleção de dados independentes marts, onde cada data mart individuais é projetada, construída e mantida como uma solução pontual. Nesse caso, cada um dos mart contém suas próprias dimensões e cada dimensão não tem relação às dimensões semelhantes em outros dados marts. Como resultado desta forma de de trabalho, você pode acabar tendo de manter cinco ou mais produtos diferentes e as dimensões do cliente. Nós opomos a isso fortemente tipo de''arquitetura''! Nossa conselho é sempre começar com o desenvolvimento e concordando com o alto nível ônibus matriz para identificar todas as entidades de interesse para o data warehouse. Só após a conclusão desta etapa pode projetar o detalhado para a dimensão individual e tabelas de fato ser iniciado.

Princípios de Design

Assim que você começar a trabalhar com esquemas estrela de apoio às empresas comuns perguntas que você vai se adaptar rapidamente aos princípios de design subjacente. É como dirigir um carro com uma alavanca, o que parece complicado no começo, mas sente muito natural uma vez que você pegar o jeito dele. Adaptar-se ao projeto básico princípios de modelagem esquema estrela pode ser ainda mais simples do que aprender a unidade de uma alavanca. As próximas seções introduzir as seguintes princípios de design que permitem a concepção de data marts de classe empresarial:

- As chaves substitutas
- convenções Nome e tipo
- Granularidade e agregação
- colunas de Auditoria
- Data e hora
- Desconhecido chaves dimensões

Usando chaves substitutas

Cada tabela em um banco de dados geralmente tem uma chave primária, que é o identificador único de um registro. Esta chave pode consistir de uma ou mais colunas, e qualquer tipo de dados pode ser usado para essas colunas. As bases de dados usado como um sistema de origem para um armazém de dados freqüentemente contêm as chaves primárias consiste em várias colunas. Estas chaves primárias vêm em todos os tipos e são ou banco de dados gerados ou

fornecidos pelo usuário. O termo usado para se referir a essas chaves do sistema de origem é natural
chave, em oposição ao termo artificial ou substituto chave usada em um data warehouse. chaves naturais geralmente contêm informações sobre a natureza do registro estão referindo. Uma chave de produto pode, portanto, consistem em várias partes indicando coisas como departamento, número de revisão do modelo, número e tipo de produto. Quando uma chave é composto por uma combinação dessas peças ea chave já é suficiente para revelar a um usuário que os dados se trata, é também referido como um chave inteligente.

Do ponto de vista de data warehouse, não há nada sobre um smart smart fundamental, pois eles ocupam espaço desnecessário e são difíceis de construir e manter índices. Em um data warehouse, chaves substitutas deve ser utilizado, que é talvez o mais importante princípio do design na construção de um data mart com esquemas estrela. Uma surrogate key é um banco de dados gerado pelo identificador, sem qualquer
significado inerente. Sua única finalidade é identificar um registro de dimensão usando o menor tipo de dados possíveis. A chave primária de uma tabela de dimensão portanto, é sempre composto por uma única coluna. Isto é importante porque os fatos registros podem ser identificados normalmente pela combinação das chaves primárias das as tabelas de dimensão.

Quando você olhar para o diagrama na Figura 7-1, você vai ver que cada um de vendas fato (uma operação de venda individual) tem cinco chaves estrangeiras, para o produto, tempo cliente, loja e promoções. Agora, suponha que o tempo""você usa um tipo de dados datetime e para todas as outras chaves seu sistema de fonte original chave primária. Suponha também que essas outras chaves são""inteligente e ocupam 15 caracteres, que converte a 15 bytes. Você acaba com uma chave de 68 bytes para o tabela de fatos (4 ×15 mais 8 bytes para a data e hora). Com 100 milhões de linhas verdade, você
necessidade de aproximadamente 6,5 gigabytes para armazenar essas informações. Usando o substituto
chaves, isso pode ser reduzido abaixo de 20 bytes por chave (cinco inteiros), resultando em 1,9
GB. Isso é 4,6 GB de espaço em disco para ocupar menos e também (o que é mais importante)
menos disco I / O, quando uma consulta é executada.

As chaves substitutas têm outras vantagens também:

- Há sempre apenas uma chave única coluna para cada tabela de dimensão para que o resultando índice de chave primária será menor.
- Integer índices são geralmente muito mais rápido do que os índices de caracteres ou datetime.
- Eles permitem o armazenamento de múltiplas versões de um item em que o item mantém a sua chave da fonte original, mas é atribuída uma surrogate key nova.
- Eles permitem lidar com as relações opcional, valores desconhecidos e dados irrelevantes, que, portanto você pode evitar o uso de associações externas em seu
consultas.

As chaves substitutas podem ser gerados de duas maneiras: usando a funcionalidade de banco de dados
(Auto-incremento valores ou seqüências) ou usando a ferramenta ETL para gerar

próximo valores fundamentais. Nós preferimos o último porque algumas ferramentas ETL necessidade especial
configuração para lidar com chaves de banco de dados gerado.

Naming e Convenções Tipo

A produção do armazém de dados pode conter várias tabelas e exibições, todos com muitas colunas de diferentes tamanhos e tipos. Como uma prática recomendada, use significativa
nomes, prefixos e sufixos para todos os objetos de banco de dados. Além disso, instituído orientações para os tipos de dados utilizados no armazenamento de dados para aplicar nor-padronização e evitar a possível perda de dados devido à conversão ou truncamento de valores. Muitas organizações já possuem um conjunto de diretrizes de design de banco de dados
prontamente disponíveis e você pode melhorar ou desenvolver os seus próprios. Em mínimo, você deve seguir as convenções de nomenclatura prevista se forem disponíveis.

Para a nossa demonstração de projectos de data mart, usaremos o seguinte conjunto de regras:

- Todas as tabelas de obter um prefixo (seguido por um sublinhado), indicando o seu papel
  ea função no armazém de dados:
  - STG_ para o estadiamento tabelas
  - HIS_ para tabelas de arquivo histórico
  - DIM_ para as tabelas de dimensão
  - FCT_ das tabelas de fatos
  - AGG_ de tabelas de agregados
  - LKP_ para tabelas de pesquisa
- Toda a dimensão colunas de chave são nomeados depois que a tabela a que pertencem, com-
  o postfix e obter um _Casos postfix (para a coluna de chave dim_product é nomeado product_keyE assim por diante).
- Todas as colunas de dimensão-chave é do tipo menor inteiro não assinado possível. O MySQL tem cinco tipos inteiros diferentes, que variam de pequenos a grandes. Inteiros podem ser definidos como ASSINADO ou UNSIGNED, Indicando se tomam valores negativos e positivos, ou valores positivos sozinho. Para a tecla colunas, use inteiros sem sinal.
  - TINYINT -1 Byte, 2 8 valores (0-255)
  - SMALLINT -2 Bytes, 2 16 valores (0-65532)
  - MEDIUMINT -3 Bytes, 2 24 valores (0-16,777,215)
  - INT ou INTEGER -4 Bytes, 2 32 valores (0-4,294,967,295)
  - BIGINT -8 Bytes, 2 64 valores (0-18,446,744,073,709,551,615)

- Usar nomes significativos para as colunas. Tentamos de todas as colunas de prefixo nomes com o nome da tabela, a menos que se torna impraticável fazê-lo devido a extremamente longos nomes de coluna.

- Usar nomes padrão para auditoria colunas. Essas são as colunas indicando quando e quem ou o processo que inseriu o registro ou fizeram a última atualização.

- Evite o uso de palavras reservadas para objetos de banco de dados como tabelas, colunas,
e pontos de vista. Utilizando palavras reservadas como grupo,tempo,vista,fim,campo, atualizaçãoE assim por diante torna mais difícil trabalhar com esses objetos em consultas porque eles devem ser citados para distingui-los as mesmas palavras
na linguagem SQL.

NOTA nteiros em um banco de dados MySQL são ASSINADO Por padrão, o que significa que assumir valores negativos. Certifique-se de definir explicitamente inteiros para a chave colunas como UNSIGNED para acomodar uma maior variedade de valores. Se você omitir o UNSIGNED palavra-chave, uma TINYINT aceitará valores de -128 a 127. Porque nós geralmente começam (auto-) A numeração de 1, quase a metade dos valores, então, permanecer não utilizados.

Granularidade e Agregação

Por granularidade, referimo-nos ao nível de detalhe no qual os dados são armazenados nos dados
armazém. A regra de ouro aqui é para armazenar os dados no menor nível de detalhe possível. Para uma empresa de varejo, isto significa que a operação de venda individual nível, para uma operadora móvel, é o nível recorde de detalhes de chamadas. Nos primeiros dias
de armazenamento de dados, espaço em disco era caro e limitado poder computacional, mas com o estado atual da tecnologia de armazenamento e consulta, terabytes de dados estão ao alcance da maioria das organizações.

Um dos equívocos sobre esquemas estrela é que as tabelas de fatos deve ser sempre pré-agregados. Agregação pode ser uma coisa boa para aumentar a consulta desempenho, mas só depois de carregar os dados no nível mais baixo de detalhe. É muito fácil de ver se esse nível mais baixo é realmente carregado: Se você precisa fazer umsum () função quando o carregamento de registros de fatos, você não está carregando o menor
nível de detalhe. O raciocínio por trás do projeto de data warehouse para capturar o menor nível de detalhe é muito simples. Você pode sempre agregar dados quando os detalhes estão disponíveis, mas quando os dados não estão disponíveis, é impossível para adicioná-los sem ter que reconstruir o armazém de dados inteiro.

A agregação pode ter um impacto dramático no desempenho. Uma tabela que dados agregados por mês, região e categoria de produto para exibir em um relatório de gestão contém, provavelmente, mais de 1.000 vezes menor do que os dados o menor nível da tabela de fatos transação. Nós testemunhamos a velocidade aumenta consulta
de uma média de 30 minutos a um par de milissegundos, usando agregado técnicas. Embora estes resultados são espetaculares, você não deve esquecer que

o modelo de dados exposto ao mundo exterior (os usuários finais) é a detalhada granular. A existência de tabelas de agregados deve ser invisível ao fim usuários, enquanto o redirecionamento de consultas a partir das tabelas de detalhe para os agregados deve ser manuseado por um ser inteligente governador consulta.

Como explicado no capítulo 6, este mecanismo de consulta que regem ainda não está disponíveis em bancos de dados open source, mas, felizmente, você pode usar o agregado designer de Mondrian, pelo menos parcialmente beneficiar da criação automatizada e uso de tabelas de agregados em sua solução de data warehouse. Usando o agregado designer, juntamente com um banco de dados em colunas, como LucidDB ou Infobright, você pode conseguir resultados muito bons consulta combinada com a disponibilidade de execução dados de fatos.

Auditoria Colunas

É altamente recomendável incluindo colunas de auditoria no data warehouse. Estes colunas permitem rastrear dados de sua fonte original, e dizer-lhe quando uma certa linha foi introduzido ou modificado e quem ou qual o processo executou a operação. Sob operação normal, o processo de ETL cuida de todas as alterações aos dados no data warehouse, mas às vezes pode necessárias para corrigir manualmente alguma coisa. Quando isso ocorre em um cuidado construídos tabela de dimensão, é importante ter o registro da tabela estas alterações automaticamente. Por razões de auditoria recomendamos o uso do quatro colunas (pelo menos nas tabelas de dimensão):

- Inserir hora
- Insira processo em lote
- timestamp Update
- Processo de atualização em lote

Normalmente os dados em um data warehouse é carregado e atualizado em lotes. Para Nestas situações, recomendamos também usando uma tabela de lotes distintos, que armazena informações sobre o horário de início e final do lote, número de registros processado, a máquina o processo a correr, o nome ea versão da ETL ferramenta que processou os dados, eo perfil do sistema (desenvolvimento, teste e produção) utilizados. Pentaho Data Integration permite-lhe criar estes log em ambas as tabelas de trabalho eo nível de transformação.

A combinação de colunas de auditoria e um lote de identificação ajuda a encontrar e corrigir problemas de dados no armazém de dados com mais facilidade. Eles também podem ser usados para demonstrar a veracidade dos dados ou, ainda mais importante, a carga correta de dados incorretos dos sistemas de origem. Em 99 por cento dos casos, problemas de dados são devido a erros no sistema de origem, não de erros no processo de ETL ou o armazém de dados em si. Quando confrontados com as discussões inevitáveis sobre o fiabilidade dos dados no data warehouse, as colunas de auditoria são sua rede de segurança.

Modelagem de Data e Hora

Data e hora parecem assuntos triviais no início, mas quando você pensa sobre isso por um alguns segundos há muito mais do que isso. Por exemplo, pense em como você seriam responsáveis por diferentes fusos horários da sua organização opera dentro Ou pense sobre como lidar com os anos fiscais que estão fora de sincronia com o calendário regular ano. A semana de numeração na Europa é baseada na ISO 3306, enquanto que no semana EUA numeração começa em 1 a 1 de Janeiro, causando possíveis diferenças na resultados, quando uma subsidiária da UE relatórios de resultados semanais a uma cabeça norte-americana
escritório. ISO também tem um cálculo de anos diferentes, o que significa que o ano ISO para uma data pode ser diferente do ano civil, da mesma data. E como é que que data do modelo e do tempo? Qual a granularidade que você deseja usar? Damos-lhe nossos pensamentos sobre o assunto aqui e um par de dicas e as melhores práticas para obter
você começou, mas finalmente as decisões de projeto devem ser feitas por você com base no caso em apreço.

Tempo de granularidade da dimensão

Às vezes quando as pessoas começam com a concepção de um modelo dimensional, que acho que a data ea hora devem ser armazenados em uma tabela única dimensão. Este suposição talvez não seja totalmente errado, mas geralmente não o melhor um para fazer. Suponha que você queira ser capaz de referência a dimensão de data e hora por segundos. Há 24 ×60 ×60 segundos em cada dia. Durante um ano, consistindo de 365 dias, você precisaria de 31.536.000 linhas na tabela de dimensão, e porque você geralmente armazenam 10 ou mais anos, que somam mais de 315 milhões linhas. Não se parece com uma abordagem sensata. A alternativa mais radical é a criar uma tabela de dimensão por parte de data e hora. Nesse caso, você teria que criar tabelas de dimensão para o ano, mês, dia, hora, minuto e segundo. Estas tabelas Seria muito pequeno, mas seriam necessários seis chaves estrangeiras na tabela de fatos. Este também não é uma aproximação muito boa. É melhor ficar em algum lugar no meio entre esses dois extremos, criando uma data e uma dimensão de tempo.
Mesmo no caso em que você precisa para suas transações referência específica segundo, a dim_time tabela só contém 86.400 linhas. Nossa amostra WCM armazém de dados utiliza uma dimensão de tempo por minuto, o que só é 1.440 linhas.

Hora local Versus UTC

Ao lidar com o tempo, você também deve considerar os desafios apresentados quando a sua organização abrange vários fusos horários. Nesse caso, pode ser benéfica para adicionar o tempo local e UTC para a tabela dimensão tempo para qualquer transação pode sempre ser visto sob diferentes perspectivas.

HORA UTC

UTC é um momento de padrão internacional e anotação do tempo regido pela Norma ISO 8601. A sigla é um compromisso entre o TUC francês ¨ '(Universel Temps Coordonne) e do Corte Inglês (Coordinated Universal Time). Qualquer pessoa familiarizada com a terminologia militar deve ter ouvido falar de"Zulu tempo", que é o mesmo que UTC. fusos horários internacionais são apontados como UTC mais ou menos um número de horas. Este deslocamento pode mesmo variar ao longo do tempo, quando países alterar o verão começar poupança ou data de término, ou quando um governo (Por exemplo, Venezuela) decide afastar-se as normas internacionais. A deslocamento fixo em nossa tabela de dimensão de tempo é, portanto, uma simplificação do mundo real.

Data Smart Keys

Há uma exceção para o uso do sentido surrogate keys-chave para a dimensão de data. Nós preferimos usar um inteiro no formato AAAAMMDD para duas razões. Em primeiro lugar, essa chave pode ser facilmente produzida a partir de uma data de entrada e, portanto, salva uma operação de pesquisa na tabela de dimensão de data. Segundo e provavelmente mais importante é o fato de que este esquema de numeração pode ser usados para dividir suas tabelas de fato. As partições de tabela também pode obter um nome com um Número de extensão da data para identificar facilmente, como para todos os P_200808 agosto 2.008 transações. Para além destas duas razões, a utilização de uma data-chave inteligente é proibida e não pode, em circunstância alguma, ser usado diretamente em consultas sobre o tabela de fatos sem participar da tabela dimensão correspondente data.

Handling Time Relativa

Em muitos casos, não estamos interessados em um determinado período de tempo, como um semana ou um mês, mas quero saber o que é um período específico de tempo parece em relação ao outro. A questão é como configurar a dimensão de tempo para lidar com o tempo relativo, como na semana passada, no mês passado, ou mesmo mês do ano passado.
Claro, você pode escrever o SQL para recuperar as linhas correspondentes da banco de dados, mas depois você nem precisa ajustar a declaração para cada novo período, ou necessidade de usar as funções e expressões para fazer uma dinâmica de afirmação. A exemplos nos blocos de consulta a seguir mostram o que queremos dizer. A primeira consulta recupera o resultado do mês passado e atual, mas as declarações são codificada:

```
SELECT      d.year4 d.month_number, f.sum (receitas), enquanto receitas
DA          fact_sales AS f
```

```
INNER JOIN dim_date d ON f.order_date_key d.date_key =
WHEREd.year4 = 2008 E d.month_number entre 7 e 8
```

Uma versão dinâmica poderia ser algo como isto:

```
SELECT      d.year, f.sum, d.month (receitas), enquanto receitas
DA          fact_sales AS f
INNER JOIN  dim_date d ON d.date_key f.order_date_key =
ONDE        d.year4 = EXTRACT (ano de NOW ())
E           d.month_number
ENTRE       EXTRACT (mês de agora ()) -1
E           EXTRACT (mês de agora ())
```

Isto irá funcionar bem, exceto quando o mês atual é de Janeiro. Para evitar codificação complexa como esta é melhor adicionar colunas adicionais com o tempo tabela de dimensão que podem ser atualizadas durante o processo de ETL. Como um exemplo,
nós vamos usar duas colunas: current_month e last_month. Quando aplicável, os conterá o valor 1, caso contrário, 0. A mesma técnica pode ser aplicada a indicar o ano até à data, mês a data, e assim por diante.

Outra técnica que gostaríamos de mencionar é o uso de compensações para ser capaz de executar aritmética de data de uma forma simples. Um bom exemplo disso é o uso das datas do calendário juliano, que são valores inteiros consecutivos. Isto torna muito fácil para filtrar os últimos 30 dias ou 30 dias antes da data escolhida. Novamente, você pode expandir esta a semanas, meses e anos também. Ao definir o atual semana, mês ou ano a 0 e em contagem decrescente, é possível ter uma indicador do período atual e um meio de filtragem dos últimos três meses, a última seis quartos, etc

A Tabela 7-1 mostra uma coleção parciais de linhas e colunas a partir do momento dimensão para ilustrar esta técnica. A data atual neste exemplo é julho 1, 2009, com uma semana a partir de segunda-feira.

Nesta tabela, você pode ver três maneiras de lidar com o tempo relativo:

- Número de Seqüência-A data do calendário juliano é um número inteiro que permite a simples
  aritmética, por exemplo,"últimos 30 dias."Qualquer data pode ser usado como o deslocamento (a partir
  ponto).
- True / false indicadores-As colunas de semana passado e atual são atualizado a cada semana. Recuperando a semana em curso é uma questão de adicionar
  current_week = 1 a uma consulta.
- Seqüência com offset 0-Isto combina as duas primeiras opções em um coluna. semana em curso é sempre 0, na semana passada -1, e assim por diante.

A terceira opção pode parecer a solução ideal para lidar com tudo relativo questões de tempo, mas há um prendedor. A corrente extra e última semana, com colunas apenas 1 e 0 permitem fazer cálculos também. Em muitos casos, é necessário

lado para ver a última semana em curso e de receitas a lado, e isto pode ser facilmente realizado usando uma declaração como a seguinte:

```
SELECT p.product_type,
       SUM (d.current_week f.revenue *) AS current_week,
       SUM (d.last_week f.revenue *) AS last_week
FROM fact_sales AS f
INNER JOIN dim_date d ON f.order_date_key d.date_key =
INNER JOIN p dim_product ON f.product_key p.product_key =
WHERE d.week_sequence entre -1 e 0
```

Tabela 7-1: Relativo colunas de hora

date_key	date_value	date_julian	current_week	last_week	week_seq
20090620	20-jun-09	2455002	0	1	-2
20090621	21-jun-09	2455003	0	1	-2
20090622	22-jun-09	2455004	0	1	-1
20090623	23 jun-09	2455005	0	1	-1
20090624	24-jun-09	2455006	0	1	-1
20090625	25 jun-09	2455007	0	1	-1
20090626	26 jun-09	2455008	0	1	-1
20090627	27-jun-09	2455009	0	1	-1
20090628	28-jun-09	2455010	0	1	-1
20090629	29-jun-09	2455011	1	0	0
20090630	30-jun-09	2455012	1	0	0
20090701	1-Jul-09	2455013	1	0	0
20090702	2-Jul-09	2455014	1	0	0
20090703	3-Jul-09	2455015	1	0	0
20090704	4-Jul-09	2455016	1	0	0
20090705	5-Jul-09	2455017	1	0	0
20090706	6-Jul-09	2455018	0	0	1

Desconhecido chaves de dimensão

Além das vantagens já mencionadas para a utilização de chaves substitutas
(Ea explicação programados neste capítulo da história da manipulação dos dados
warehouse), há outra boa razão para usar chaves substitutas. Às vezes
os dados serão carregados em uma tabela de fatos e nenhuma tecla correspondente
dimensão pode
ser encontrada com base na chave natural e / ou validade da dimensão
registro. Para esses casos, você precisa de um mecanismo que armazena os dados de
qualquer maneira
para se certificar de que nenhum dado seja omitido durante o processamento. Por esta razão,
recomendamos ter um registro de desconhecidos em cada tabela de dimensão. Key
geração de uma dimensão geralmente começa em 1, o que deixa o número 0 como um
candidato perfeito para a chave de dimensão desconhecida registro. Este registro deve
ter um valor de""desconhecidos em todos os campos de atributo. Tabela 7-2 mostra uma
parcial
exemplo do registro desconhecido em uma tabela de dimensão do cliente.

Tabela 7-2: registro de dimensão Desconhecido

chave	source_id	nome	endereço	telefone
0	0	Desconhecido	Desconhecido	Desconhecido

Usamos o valor Desconhecido para informar os nossos utilizadores empresariais que estes dados não é
disponíveis. Esta é uma alternativa muito melhor do que permitir que NULL valores, que
muitas vezes confunde os usuários. Além disso, quando NULL Os valores são usados em um cálculo que
pode provocar resultados errados. Um teste simples pode ilustrar isso; apenas o começo
MySQL Query Browser e digite 'Resultado' SELECT. Isto irá retornar o texto
»resultado '. Agora altere a declaração em 'Resultado' SELECT + NULL e ver o que
acontece.

Em alguns casos, um único indicador desconhecido não é suficiente. Às vezes, a
dimensão não é relevante para um fato específico. Isso pode ocorrer quando a tabela de fatos
contém diferentes tipos de fatos. Um exemplo disso é uma tabela, clique fato fluxo.
Nós gostaríamos de armazenar todos os cliques em conjunto para poder calcular facilmente
total de páginas
pontos de vista, mas nem todos os cliques são os mesmos. Alguns são apenas cliques de
navegação, alguns
confirmar a transação, e outros o download de um arquivo. No caso de uma navegação
clique, uma referência a um arquivo de download não é relevante, mas porque nós não
permitimos
NULL valores em nossa tabela de fatos colunas de chave estrangeira, acrescentamos um
substituto especiais
chave com o valor não relevante à tabela de dimensão para além da
desconhecido chave.

Tratando alterações Dimensão

Dimensional alterações de dados constantemente. mover clientes, novos produtos são
introduzidas, os funcionários terão um aumento salarial, armazéns obter novos gestores, e

assim por diante. Um dos principais desafios na construção de um data warehouse é a forma de
adaptar a estas mudanças de modo a que toda a história relevante é capturado e
todas as operações ligadas à correta registros de dimensão histórica. Alguns dos
essas mudanças ocorrem com freqüência, outros apenas ocasionalmente. Algumas alterações afetam
um único registro, e algumas envolvem tabelas completas, como a introdução de um novo
cliente classificação. E algumas mudanças são relevantes para armazenar em um historicamente
maneira correta para a organização, enquanto outros podem simplesmente substituir dados antigos.
Por todas estas situações, uma estratégia diferente pode ser aplicada, e juntos
eles formam as diferentes formas de manipulação dimensões de alteração lenta (SCDs). Em
sua primeira edição O conjunto de ferramentas de Data Warehouse, Ralph Kimball introduziu três
diferentes estratégias SCD: substituir, pilha e adicione. Essas estratégias são
respectivamente chamado SCD tipo 1, 2 e 3, serão descritas em breve. Ao longo dos
anos, muitas outras pessoas da comunidade de modelagem de banco de dados adicionados novos
tipos e variações do tipo original do SCD, a atual abordagem abrange
a manipulação de todos as possíveis mudanças que podem ser aplicados a dados de origem e
como lidar com essas mudanças de forma mais eficaz em um data warehouse.

O termo dimensões de alteração lenta é um pouco enganador, no sentido
que você pode ter a impressão de que estes tipos devem ser aplicadas ao
tabela de dimensão completa. Esta é uma suposição incorreta de que nós gostaríamos
para resolver antes de irmos. Cada coluna em uma tabela dimensão deve ser
tratado individualmente. Por isso, é possível que, para algumas colunas do tipo 1 é utilizado,
enquanto para outros o tipo 2 abordagem será a mais adequada. Outras estratégias
ou combinações também podem ser usados, mas identificar qual estratégia deve ser
aplicada a cada coluna é uma parte importante do processo de concepção. Nós vamos dar
algumas dicas de design no próximo capítulo sobre como fazer isso e documentá-lo.

A história não começa com o tipo 1, no entanto. Existe também um tipo 0,
significando tanto não fazer nada""ou"substituir completamente."Em ambos os casos, trata-se
não a uma estratégia recomendável, excepto se se tratar tabelas de dimensão estática. O mais
exemplo proeminente destes são as dimensões Data e Hora. É muito
improvável que o valor do ano da data de 20 de janeiro de 2008 nunca vai mudar. O que pode
mudança, no entanto, é uma classificação Ano Fiscal, mas, nesse caso, os valores normalmente
Ser substituídas (tipo 1) ou armazenado em uma coluna adicional (tipo 3).

NOTA Nem todos os atributos de dimensão serão utilizados para fins de análise e, portanto, nem todos são de interesse para correção histórica. Quais são os atributos analíticos e quais são os detalhes difere por setor e organização. Em alguns casos, um telefone número é apenas um detalhe da dimensão do cliente, que pode ser substituído quando um novo valor é extraído do sistema de origem. Em outros casos, o número de telefone é sendo usado para fins analíticos e as mudanças precisam ser controladas. Quando nós analisar as vendas, é geralmente suficiente para o grupo pela cidade, não necessariamente pela rua ou endereço. Nesse caso, a coluna cidade é um atributo analítico e da coluna endereço é um detalhe simples.

SCD Tipo 1: Substituir

Um tipo 1 dimensão mudando lentamente é o mais básico e não
exigir qualquer modelagem especial ou campos adicionais. tipo SCD 1 colunas apenas
ser substituídas por novos valores, quando eles vêm para o armazém de dados.
Figura 7-5 mostra o que um registro de dimensão parece antes e depois da aplicação
Neste caso, o cliente mudou-se para um novo endereço
em uma determinada data.

Situação existente

_Casos Cliente	Cliente _id	Cliente _Nome	Cliente _City
1	22321	Humphries	Toronto

Nova situação

_Casos Cliente	Cliente _id	Cliente _Nome	Cliente _City
1	22321	Humphries	Vancouver

Figura 7-5: SCD tipo 1 exemplo

Sem o conhecimento adicional, você não tem nenhuma maneira de identificar quando esta mudança ocorreu, e mesmo se você inserir uma coluna extra com uma modificada pela última vez
timestamp você só sabe que alguma coisa mudou dentro do registro, não
a coluna que a mudança ocorreu. Um tipo 1 substituir abordagem é usada
para as colunas que são de interesse para os nossos usuários (caso contrário, as colunas não estive lá no primeiro lugar), mas apenas o estado atual dos dados é
relevantes, mesmo quando se olha para as transações mais antigas. Lembre-se que quando você
substituir uma coluna como esta e executar uma consulta, todos os resultados vão mostrar o conteúdo
da coluna como é agora, não o valor anterior.

SCD Tipo 2: Adicionar linha

Tipo 2 não é"o"próximo passo em termos de funcionalidade e complexidade, mas é
na verdade, uma categoria própria.
tipo de suporte 2 SCDs com assistentes, macros e outros suplementos nos últimos
par de anos, então hoje você vai ser duramente pressionado para encontrar uma ferramenta sem que isso
apoio. Claro, Chaleira / PDI também tem essa funcionalidade na Dimensão
Pesquisa / passo Update. Mas o que exatamente ele faz? SCD tipo 2 é a história
preservação e permite que uma organização para captar alterações em uma dimensão
tabela para recuperar o histórico de dados corretos ao consultar o data warehouse.
Figura 7-6 mostra o mesmo exemplo do parágrafo anterior, mas agora
Você pode acompanhar as mudanças através do tempo, acrescentando alguns campos extras. Há
múltiplas formas de modelar este ou armazenar várias versões, a mais básica
uma é a de adicionar apenas um valid_from hora para o registro de dimensão. Omitindo
correspondente valid_to timestamp adiciona complexidade extra ao tentar
recuperar a versão correta da entrada, por assim marcar este como um imperativo

campo para este tipo de mudança também. Duas outras colunas extras podem muitas vezes ser
encontradas no tipo 2 mesas de apoio: a current_record coluna indica a
versão atual do registro de dimensão, e um número de seqüência ou versão
que é incrementado cada vez que uma nova versão do registro é adicionado.

Agora você pode fazer muitas coisas interessantes com esses dados. Suponha que o Sr. ou Sra. Humphries é um cliente regular e ordens algo cada mês.
O que acontece em sua tabela de verdade, quando estas operações são carregados é que o processo de ETL olha para o registro de cliente válido para cada cliente em particular
no momento do embarque. Isso significa que toda a ordem linhas de fatos para o cliente
com o ID 22321 (a origem do número de clientes do sistema) irá armazenar customer_key 1até 01 de maio de 2008, ea utilização customer_key 2 a partir desse dia até a próxima
mudança para este cliente é aplicada. A tabela de fatos exemplo é exibido na
Figura 7-7.

Situação existente

Customer_key	Customer_id	Customer_Name	Customer_City	Valid_from	Valid_to	Current_record
1	22321	Humphries	Toronto	1900-01-01	9999-12-31	1

Nova situação

Customer_key	Customer_id	Customer_Name	Customer_City	Valid_from	Valid_to	Current_record
1	22321	Humphries	Toronto	1900-01-01	2008-04-30	0
2	22321	Humphries	Vancouver	2008-05-01	9999-12-31	1

Figura 7-6: SCD tipo 2 exemplo

fatos Vendas

Customer_key	Date_key	Product_key	Itens	Receita
1	20080123	123	1	5
1	20080208	221	2	10
1	20080315	332	1	5
1	20080421	334	1	5
2	20080511	221	2	10
2	20080609	432	3	15
2	20080729	554	1	5
2	20080817	101	2	10

Figura 7-7: Fatos com SCD tipo 2

Agora, quando você quer saber como a receita foi gerada em muito Toronto em 2008, e você executar essa consulta, em setembro, a condição é onde customer_city = 'Toronto'. O registro de dimensão com o valor 1 para customer_key é o único registro que satisfaça essa condição. E porque a associação está em customer_key, Apenas as quatro primeiras linhas da tabela de fatos são recuperados para este cliente. Quando a condição é onde nome_cliente = "Humphries ', Ambos customer_id 1 e 2satisfaz a condição e todos o fato de linhas serão retornadas.

PROCESSAMENTO DE DADOS TIPO 2 SCD

Um processo de ETL, se a mão codificados ou codificados com uma ferramenta, precisa ter vários recursos para tratar alterações do tipo 2 dimensão. Primeiro de tudo, o dados de entrada devem ser comparados com os dados que já existe na Dimensão da tabela. A saída deste processo é o conjunto de dados de entrada com bandeiras acrescentado para novas linhas (I para Inserir) a serem inseridos, as linhas existentes que necessitam de ser atualizado (U para atualização), e possivelmente até mesmo excluído linhas que não existe mais no sistema de origem (D para Excluir). Com base na I U, ou D bandeiras, em seguida, o processo deve atribuir novas teclas de substituto no caso de inserção ou atualizações. Inserções são fáceis: a nova linha pode ser adicionada com o padrão definições para valid_to e current_record. Para obter uma atualização adicional transformação é necessária: a linha existente precisa ser detectado (por isso o current_record indicador é muito útil), o valid_to timestamps necessidade a ser definida para o valor correto, e os current_record bandeira tem de ser definido para um valor de 0, N, ou qualquer lógica que você desenhou para ela. Em seguida, uma nova registro com os dados atualizados terá que ser criada, uma surrogate key gerada, ea data e hora apropriada e current_record bandeira precisa ser definido. Como excluir dados dimensional é fora de questão (que provavelmente ainda é referenciada por fatos existentes), os registros com uma bandeira excluir obter a sua valid_to timestamp definido. Além disso, outro indicador pode ser usado para marcar o registro como excluído. Uma possível solução é usar um valor diferente de 0 ou 1 para o current_record pavilhão.

Claro que há mais para captar a história em um data warehouse, que é apenas a versão simples de correção histórica. Por exemplo, pensar sobre o seguinte questão: o momento em que nós usamos para o valid_from e valid_to timestamps? É o momento de inscrição da alteração no sistema de origem? A vez que os dados foram carregados no data warehouse? O tempo real no real mundo, quando ocorreu o evento? Ou o momento em que fomos notificados da presente evento? A terceira opção parece ser a versão mais adequada ea melhor representação do evento real, mas como vamos controlar isso? Em alguns indústrias, tais como o negócio de seguros, todas estas horas devem ser e armazenados no data warehouse deve ter uma explicação completa para o história da história, como é chamado, também. Uma discussão mais aprofundada do problema está além do escopo deste livro, mas queríamos a levantar a questão para ser completo aqui.

Para o restante deste livro, vamos usar a versão padrão da história conservação, como mostrado nos exemplos mostrados nas Figuras 7-6 e 7-7. Nós também adicionar uma coluna de versão extra para capturar o número de alterações feitas em um registro de origem particular.

SCD Tipo 3: Adicionar Coluna

O tipo 3 A estratégia exige pelo menos uma coluna extra na tabela de dimensão.
Quando os dados para um tipo de três alterações de coluna, o valor existente é copiado para o extra _old coluna enquanto o novo valor é colocado na coluna regular.
Figura 7-8 mostra um exemplo disso.

Situação existente

_Casos Cliente	Cliente _id	Cliente _Nome	Cliente _City	Cliente _City_Old
1	22321	Humphries	Toronto	Toronto

Nova situação

_Casos Cliente	Cliente _id	Cliente _Nome	Cliente _City	Cliente _City_Old
1	22321	Humphries	Vancouver	Toronto

Figura 7-8: SCD exemplo do tipo 3

Isso pode parecer uma estranha forma de manter os valores anteriores à primeira vista, e na maioria dos
casos ela é. Só é possível lidar com uma versão anterior. Armazenamento adicional mudanças requer uma coluna extra para cada versão que você deseja manter. Mas imaginar que as alterações da estrutura da organização completamente, ou completamente nova estrutura do grupo produto é introduzido. Nesses casos, onde todos os registros mudança, ao mesmo tempo, faz sentido usar uma coluna extra. Manuseamento essas grandes mudanças com um cenário do tipo 2, duplica o número de registros em sua tabela de dimensão e na maioria dos casos, apenas a nova estrutura dos dados é relevantes. A versão antiga é mantida apenas para efeitos de referência ou como uma tradução
tabela.

SCD Tipo 4: Mini-Dimensões

modelagem dimensional Kimball introduz o termo mini-dimensão. Alguns
fontes afirmam que este é um cenário do tipo 4, outros usam o termo"4"tipo de
outros fins. As noções de tipo 4 e 5 SCDs foram introduzidas em 1998
por Michael Schmitz, um renomado especialista em armazenamento de dados e dimensionais modelagem. Estamos em conformidade com essa classificação aqui, em contraste com outras fontes
como a Wikipedia, que usa uma classificação diferente.

Mini-dimensões resolver dois problemas específicos com a mudança de dimensão tabelas. Um problema ocorre quando as tabelas de dimensão ficar realmente grandes, digamos, um
dimensão do cliente, com 150 milhões de linhas (elas existem!). O segundo problema ocorre quando as mudanças acontecem com muita freqüência, causando a tabela de dimensão
dobrar ou triplicar de tamanho a cada ano. O truque aqui é a primeira a identificar quais atributos analíticos mudam muito freqüentemente e colocá-los como um grupo em um ou mais tabelas de dimensão em separado. O resultado disso é um ou mais extra chaves de dimensão na tabela de fatos.

Aqui está um exemplo. Suponha que você tenha uma tabela de clientes com grande dimensão
os atributos cidade,região,país,gênero,birth_dateE renda. O primeiro
três campos podem ser classificados como dados geográficos, os três últimos têm mais
natureza demográfica. Naturalmente, a descrição do gênero não muda muito
muitas vezes, mas é provavelmente um dos principais atributos para fins analíticos. Estes seis
campos pode ser posta em duas dimensões diferentes mini, um dim_geography
e um dim_demography.

Mini-dimensões fazer honra ao seu nome: Eles geralmente são muito pequenas,
não só no número de registros, mas também no número de atributos, como no presente
exemplo. Há, no entanto, uma troca envolvidos, a fim de manter este número
de registros tão pequenos quanto possível. Quando um atributo como a renda é usada em
uma mini-dimensão, é impossível para armazenar todos os possíveis valores diferentes, assim você
necessidade de trabalhar com valores em faixas ou intervalos. Lembre-se que quando se utiliza campos
A, B e C para uma mini-dimensão, o número de registros é determinada pela
multiplicando o número de possíveis valores de A, B e C. Então, se cada idade a partir de
0-100 é usado, o multiplicador é já 101. Faz mais sentido definir
idade e faixas de renda, talvez 10 de cada, resultando em 100 registros. Multiplicado
pelos três valores possíveis para o gênero (masculino, feminino e desconhecidos) a
número de linhas no mini-dimensão será de 300, que é muito pequeno, por qualquer
padrão. Mini-dimensões, com 100.000 linhas não são incomuns e com o
estado atual da tecnologia também não um problema. Se o mini-dimensões obter qualquer
maior do que isso, é aconselhável para redesenhar a tabela de dimensão e talvez dividir
-lo novamente em dimensões menores. Figura 7-9 mostra um exemplo de modelo de dados
com uma tabela de fatos e três tabelas de dimensão das quais duas são mini-dimensões.

A fim de fazer este trabalho é necessário identificar a dimensão correta mini-
chave quando o carregamento de dados, o que requer uma sobrecarga extra. Por exemplo, para
determinar a faixa etária pertence a um cliente, é preciso calcular a idade
com base na data de carga e data de nascimento do cliente. O mesmo vale para a renda
banda e todos os outros atributos necessários para determinar a dimensão correta mini-
chave. A recompensa, porém, é enorme. De repente, muitas poucas mudanças podem ocorrer
a sua tabela de dimensão''principal''. A história está perfeitamente coberta no fato de
tabelas usando o mini-dimensão chaves estrangeiras, e ainda por cima, você também
acrescentou a chave de dimensão mini-curso para a tabela dimensão principal. Este último
Além disso permite ao usuário utilizar o mini-dimensões em conjunto com o
dimensão principal, sem a necessidade de tabela de consulta fato. Na verdade, este modelo
serve a dois propósitos: ele dá o valor do registro atual dimensão, que é
muito útil para selecionar os grupos-alvo, por exemplo, para marketing direto
campanhas, mantendo a história completa dos atributos de análise na
tabela de fatos.

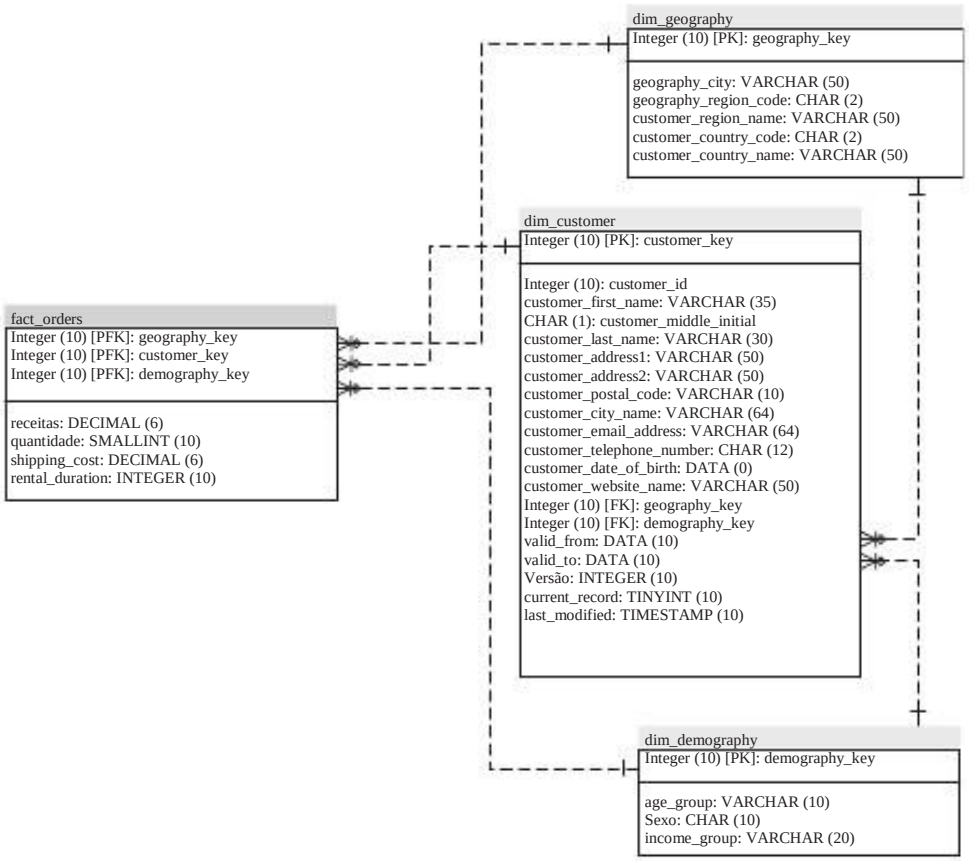


Figura 7-9: Mini-dimensões

SCD Tipo 5: Tabela de histórico separada

Até agora, as mudanças de dimensão têm afetado a forma como você olha para sua consulta resultados com relação ao fato de as linhas correspondentes. As diferentes estratégias permitem identificar qual versão de um registro de dimensão era válido no tempo de uma operação, seja uma venda, uma movimentação de estoque, ou algum outro evento empresarial. Tipo 5 é um pouco diferente, pois não pode ser usado para executar consultas analíticas que utilizam a tabela de fatos também. Com o tipo 5, uma em separado tabela de histórico é criado para uma tabela de dimensão com o único propósito de informar corretamente capturar todas as alterações a todos os atributos na tabela de dimensão. Um tipo 5 Estratégia É, portanto, além dos tipos existentes e SCD deve ser utilizado em conjunto com um ou uma combinação das outras estratégias. Tipo 5 história tabelas não devem ser utilizadas para consultas analíticas envolvendo tabelas de fatos. Figura 7-10 mostra um exemplo deste tipo de mesa junto com o pai Dimensão da tabela.

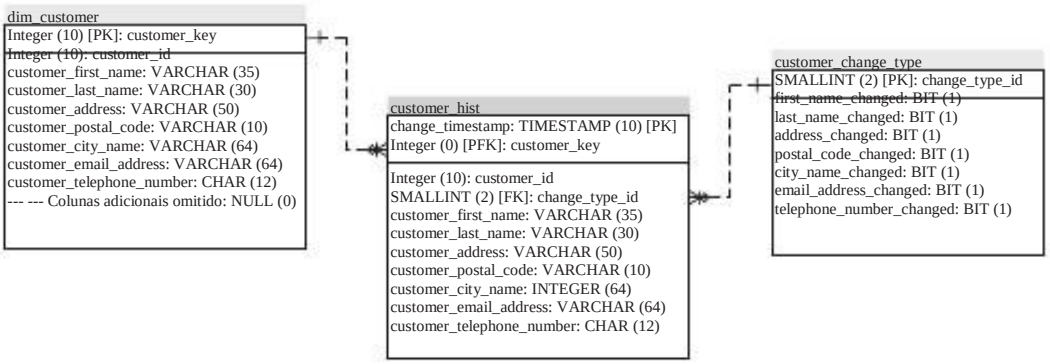


Figura 7-10: Tipo 5 tabela de histórico

Isto quase parece uma tabela de fatos, e realmente tem alguns dos mesmos características como uma tabela de fatos. No entanto, não existem fatos reais: Não há nada para resumir, porque não existem elementos mensuráveis. O único disponível opção é contar o número de clientes ou de mudanças, mas que poderia ser uma atividade bastante interessante: Quantas vezes os clientes recebem um novo e-mail endereço? Qual é a taxa média móvel?

Observe também que o esquema apresentado na Figura 7-10 contém uma tabela de tipo de mudança que serve como uma tabela que indica exatamente o que os valores mudaram. Tabela 7-3 mostra os primeiros registros desta tabela para que você possa ter uma sensação de O que estes dados se trata.

Tabela 7-3: Cliente tabela tipo de alteração

change_type_id	first_name_changed	last_name_changed	address_changed	postal_code_changed	city_name_changed	email_address_changed	telephone_number_changed
1	falsa	falsa	falsa	falsa	falsa	falsa	verdade
2	falsa	falsa	falsa	falsa	falsa	verdade	verdade
3	falsa	falsa	falsa	falsa	falsa	verdade	falsa
4	falsa	falsa	falsa	falsa	verdade	falsa	falsa

tipos de mudança são também um indicador perfeito para as questões de qualidade de dados. Olha, para
exemplo, no número de tipo 4, onde só o nome da cidade será alterado. É muito improvável que isso aconteça sem a mudança de endereço, código postal e possivelmente o número de telefone também. Independentemente de saber se você substituir valores existentes em sua dimensão de clientes, as mudanças devem ser controladas em tabela de histórico. O efeito final é uma pista de auditoria completa das mudanças na dimensão
de dados. Lembre-se, contudo, que este tem uma função diferente do que o normal tabelas de dimensão e de facto utilizados para relatórios e análise.

O exemplo descrito anteriormente é apenas uma das muitas soluções possíveis para a modelagem de uma pista de auditoria completa de dados. Outra abordagem comum é imitar
as tabelas de alteração ou registro usado por muitos sistemas de gerenciamento de banco de dados para
replicação e sincronização de dados entre dois ou mais bancos de dados. Além da dados alterados, essas tabelas de alteração conter a chave de registro, alterar data e hora, eo indicador de mudança de tipo I, U ou D para Insert, Update e Delete.
Às vezes antes e depois imagens estão presentes nesses quadros, comparáveis para um tipo de cenário SCD 3.

SCD Tipo 6: Estratégias Híbridas

Tipo 6 realmente não existe, mas às vezes é descrito como 1 +2 +3 = 6, indicando que é uma mistura de diferentes estratégias aplicadas a uma tabela única dimensão. Kimball dá um exemplo disso em O Data Warehouse Lifecycle Toolkit. A pedido de negócios comuns é"Eu quero ver todas as receitas do produto de acordo com a versão atual do grupo de produtos."Esta consulta não pode ser respondida se um grupo de produtos é tratada como um atributo do tipo 2, que é geralmente o caso. Um maneira de resolver isso é adicionar um atributo do produto extra para a tabela de dimensão onde a versão atual do grupo o produto é armazenado para todas as versões do produto. Tabela 7-4 ilustra esta característica.

Tabela 7-4: Hybrid estratégia SCD

product_key	product_id	product_name	product_group	product_group_curr	valid_from	valid_to	current_rd
101	ABC	AcidRip	Software	Open Source	1-1-1900	12-3-2002	falsa
102	ABC	AcidRip	Freeware	Open Source	13-3-2002	31-12-2008	falsa
103	ABC	AcidRip	Open Source	Open Source	1-1-2009	31-12-9999	verdade

Usando esta técnica de modelagem, você ainda pode armazenar a versão correta do produto em sua tabela de fatos para correção histórica, mas essa técnica também permite ao usuário olhar para esta informação, como se toda a história foram baseados em a classificação do grupo atual do produto. Isto é, no entanto, não é muito flexível solução e funciona apenas para um ou poucos atributos. Outra maneira de resolver este problema é juntar a tabela mesma dimensão na identificação de origem, e restringir a tabela extra juntou no registro atual. Ao fazer isso, todos os atributos podem ser usado como"é"e, ao mesmo tempo recuperar o histórico completo do fato tabela com base nas chaves SCD. A última opção é adicionar uma dimensão adicional tabela com apenas os valores de corrente na mesma. Esta tabela dimensão atual, em seguida, será obter sua própria chave na tabela de fatos. Uma desvantagem desta solução é que você precisa de um coluna de chave extra na tabela de fatos para cada dimensão atual de adicionar este caminho.

Advanced Concepts Modelo Dimensional

Projetando um armazém de dados pode ser uma tarefa difícil, especialmente quando parece que nenhuma das técnicas descritas até agora resolver o seu problema específico. O que você deve fazer quando, por exemplo, a sua dimensão de cliente é realmente, realmente grande, como 150 milhões de discos? Essa é a questão das dimensões monstro. Ou o que se pode existir vários valores para uma entrada única dimensão? O WCM banco de dados tem um dim_dvd_release dimensão do produto, mas cada filme tem múltiplos atores em campo. Como você se este modelo? E o que dizer quando tiver terminado de modelagem e há alguns atributos que não sobra parecem se encaixar em algum lugar? As próximas seções mostram-lhe os conceitos avançados necessárias para resolver estes problemas e explorar a modelagem dimensional técnica.

Dimensões Monster

Mencionamos os 150 milhões de tabela de clientes registro antes. Talvez o seu dimensão do cliente não é tão grande, mas mesmo com 5 milhões de registros, mantendo-se a tabela de dimensão em uma janela de lote limitado pode ser bastante difícil. Nós mostramos uma das soluções para fazer uma melhor dimensão do monstro o desempenho da consulta gerenciáveis e otimizar: a mini-dimensões. Ainda assim, quando a sua dimensão monstro contém uma grande quantidade de colunas de detalhe que não será utilizadas para fins de análise de qualquer maneira, pode ser uma boa idéia de dividi-las em uma tabela separada, também chamado particionamento vertical. A velocidade de consulta em uma linha-base banco de dados como o MySQL não é determinada apenas pelo número de linhas, mas também pelo tamanho total de bytes de cada linha. Quando um banco de dados baseado em colunas, como Infobright, LucidDB ou MonetDB é usada, esta penalidade é eliminada pelo fato de que os dados já estão armazenados em forma de coluna por coluna, o que eliminando a necessidade de particionar a tabela mesmo.

Uma terceira opção é particionar a tabela horizontalmente. Este é um bem comum técnica para as tabelas de fatos que normalmente são divididos por período de tempo, mas não é
muito utilizados para as tabelas de dimensão. O problema com o particionamento horizontal é que uma chave de partição deve ser definida de tal forma que o particionamento do faz sentido. O benefício de particionamento horizontal reside no fato de que, quando uma consulta é analisado, o otimizador pode determinar quais as partições de usar e que não, com base na chave de partição. A maioria das consultas envolvem uma data / horário
atributo, por isso quando se usa um esquema de particionamento baseada no tempo, isto é fácil. Se a sua
de dados é particionada por mês ea consulta afeta apenas as vendas do mês passado, o otimizador pode selecionar uma partição para recuperar os dados.

Com os dados do cliente não é tão simples. Você não pode usar um campo de data para partição porque nenhuma data disponível é relevante para suas consultas analíticas. Outros atributos que podem servir como uma chave de particionamento pode ser candidato uma geográfica ou demográfica entidades, mas a escolha dessas entidades só
permite uma forma de cortar os dados, o que não pode ser mais relevante em
dos casos. Somado a isso, você tem que levar em conta que o desenvolvimento ea manutenção de um esquema de particionamento não é uma tarefa trivial. Exige avançada habilidades de administração de banco de dados para configurá-lo e integrar a compartimentação
regime com o carregamento de dados de processos para o data warehouse. Para domar dimensões monstro, a combinação de particionamento vertical eo uso de
adicional de mini-dimensões Parece, portanto, a melhor alternativa disponível. Para determinar quais colunas se separou para o mini-dimensão (s) você precisa
olhar para um par de coisas:

- Similarity-Que colunas contêm informações similares ou informações
  que é logicamente agrupados, tais como atributos demográficos.
- Cardinalidade-How muitos valores diferentes podem ocorrer em uma única coluna? As colunas de baixa cardinalidade (como gênero) são candidatos ideais para mini-dimensões.
- Como volatilidade frequentemente valores em uma coluna de mudança ao longo do tempo? Nomes
  não mudam com muita freqüência; atributos demográficos, como idade ou renda grupo fazer.

Lixo, heterogêneo, e dimensões de degeneração

Ao desenvolver o modelo dimensional, você provavelmente vai acabar com alguns atributos que não se encaixam em uma das tabelas acordadas dimensão. Este
é geralmente o caso com os atributos que têm um significado em um processo como ordem ou transferência sinalizadores de status, tipo de ordem, ou condições de pagamento. O possível
soluções para estes atributos, como deixá-los na tabela de fatos, que se deslocam lhes dimensões separadas, ou deixá-los por completo todos têm os seus
desvantagens específicas. A tabela de fatos deve ser tão estreito quanto possível, para adicionar

colunas de texto é uma má idéia. Modelagem de cada atributo como uma dimensão separada é uma idéia tão ruim, e omitindo estes atributos significa que eles não podem ser usada para qualquer análise. A melhor solução é agrupar esses atributos em uma tabela de dimensão separada, denominada lixo dimensão não lixo no sentido literal mas como uma coleção de restos que precisam ser tratados de uma maneira elegante. Na verdade, você já viu vários exemplos da forma como essas dimensões são modelados no tipo SCD n º 4. A única diferença entre um mini e uma dimensão de lixo é o fato de que este último contém muitas vezes alheios Considerando que os atributos de atributos em uma mini-dimensão regular têm uma relação de algum tipo e pode ser nomeado em conformidade.

dimensões Heterogêneos são uma variação sobre este tema, onde diferentes tipos de itens são agrupados em uma tabela única dimensão. Porque a nossa demo empresa de classe mundial de filmes, tem apenas um único tipo de produto no estoque não serve como um bom exemplo. Para este caso, no entanto, podemos facilmente encontrar uma bom exemplo no supermercado local. Produtos em um supermercado pode pertencer em diferentes categorias, como alimentos, não-alimentares e bebidas. Os alimentos podem também ser classificados em várias categorias com características muito diferentes. Quando esses produtos são armazenados em uma única dimensão do produto, você acaba com um mesa onde a maioria dos atributos será irrelevante para um determinado produto. Indicando a data de vencimento ou o valor calórico de uma vassoura simplesmente não faz um monte de sentido. O desafio aqui é encontrar o equilíbrio certo entre a tabela de produto único, com muitos atributos inúteis e dezenas de produtos diferentes tabelas de dimensão que estão perfeitamente adaptados para uma categoria de produto específico. Em Neste caso, não há melhor resposta, mas tudo depende da situação na mão.

dimensões Degenerada são um tipo ligeiramente diferente de raça. Estes são dimensões que não existem realmente, mas deve conseguir um lugar no modelo dimensional de qualquer maneira. Um bom exemplo é o número de ordem. Números de ordem pode ajudar a traçar algumas das informações contidas no data warehouse de volta para o sistema de origem, mas não há nenhuma ordem real dimensão. Todos os atributos da linha da ordem e da ordem são modeladas no nível mais baixo de granularidade, que é o fato da linha fim. Os atributos que fazem parte de um pedido, como data do pedido e do cliente, são já se mudou para os fatos da ordem. No final deste processo, o que fazer com o número de ordem, então? Não há nenhum ponto na criação de uma dimensão de ordem porque todos os atributos relevantes já são empurradas para o fato mesa e dimensões relacionadas. Nestes casos onde você acaba com um único atributo, basta adicionar o atributo à tabela de fatos. Esse atributo não é uma medida objeto e também não é uma chave estrangeira para uma tabela de dimensão, pelo que é chamado de degenerar dimensão.

Dimensões de Interpretação de Papéis

Isto não é sobre as dimensões realização Romeu e Julieta, mas destina-se a indicam que a mesma dimensão pode ser usado para atuar como múltiplos, semelhantes

dimensões. O exemplo óbvio é a dimensão de tempo, que pode ser usado, ou melhor, deve ser utilizado para acomodar várias ocorrências de data e hora. Olhando para um quadro típico de vendas verdade, você vê data do pedido, data de envio, receber
data, data de retorno, e data de pagamento e cinco datas, uma dimensão. Fisicamente falando, todas as datas apontam para a mesma tabela dimensão data, logicamente, de mais claro, eles não fazem, porque você iria acabar com um conjunto de resultados vazio em dos casos em que se aplicam restrições de datas múltiplas. Tomemos por exemplo a consulta para recuperar o valor total da ordem de DVDs ordenado em Dezembro de 2007, que não foram devolvidos no prazo de cinco dias após a expedição. Esta consulta envolve três tipos de datas, e quando todas as restrições data exigida estão sendo definidas no tabela de dimensão mesma data, você acaba com um conjunto de resultados vazio ou, pelo menos um
conjunto de resultados que não responder corretamente a pergunta. O SQL para ilustrar a solução correta para isso é semelhante à consulta foi utilizado quando se discute apelidos de tabela (que é basicamente o mesmo):

```
SELECT      SUM (f.revenue) AS ordervalue
DA          fact_sales AS f
INNER JOIN  SOBRE AS dim_date o f.order_date_key o.date_key =
INNER JOIN  dim_date AS s ON f.shipping_date_key s.date_key =
INNER JOIN  dim_date AS r ON r.return_date_key r.date_key =
ONDE        o.year4 = 2007
E           o.month_number = 12
E           r.julian_date s.julian_date> + 5
```

Agora é fácil ver que se você aplicar as últimas duas restrições para o mesmo tabela, isso causaria um conjunto de resultados vazio, não é possível para uma data a ser maior que na mesma data, mais cinco dias.

Outro exemplo relacionado ao banco de dados WCM é o ator e diretor da informação. Podíamos combinar estas duas entidades em um único role-playing artista ou movie_person dimensão. Um monte de atores tornam-se mais tarde na administração sua carreira e às vezes até mesmo o oposto ocorre. Tomemos, por exemplo Quentin Tarantino, que começou sua carreira como diretor e mais tarde poderão ser encontradas na qualidade
em seus próprios filmes também.

Multi-valued dimensões e tabelas de Ponte

Uma das mais difíceis de resolver problemas de modelagem dimensional é a vários valores de dimensão. Novamente, a tabela de ator é um exemplo perfeito: o nosso clientes comprar ou alugar um único DVD, mas este item tem geralmente múltiplos atores aparecendo no filme. Dependendo do tipo de informação que gostaríamos de recuperar a partir do armazém de dados, existem duas soluções possíveis para resolver este problema. O primeiro é a lista dos atores em um campo de texto a nível de cinema, como na Tabela 7-5.

Tabela 7-5: Ator lista como atributo

dvd_key	dvd_id	dvd_name	dvd_actors
101	AA332	Lagarto Wars	Harry Protz; Pine Neill, Will Grant

Isso é bom para fins informativos e responde à pergunta"O que atores jogado no filme"X?, mas não pode ser utilizado para outros, mais interessantes, perguntas como"Em que filmes jogou ator Y?"ou"Quais são os top 10 atores com base na receita de aluguel?"Para resolver este último problema, você precisa de uma maneira relacionar fatos múltiplos valores de dimensão múltipla. E porque você não pode diretamente criar relacionamentos muitos-para-muitos em um banco de dados você precisa de uma ponte tabela para executar a tarefa. Figura 7-11 mostra a parte do depósito de dados modelo com o ponte tabela.

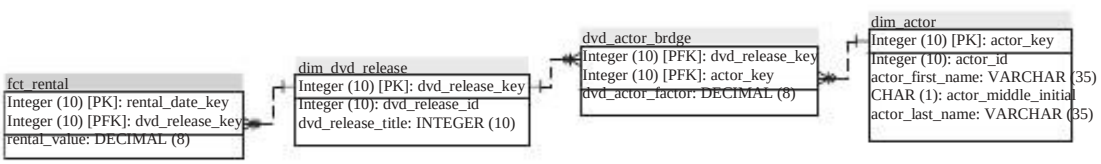


Figura 7-11: tabela da ponte dvd_actor

Note que neste diagrama, a tabela dvd_actor_bridge contém um dvd_actor_factor de campo. Este é um complemento necessário à tabela da ponte para vigor SQL para retornar os resultados corretos. Se omitir o fator em nossos cálculos envolvendo os atores, a receita será multiplicado pelo número de atores ligado a um DVD específico. Debate é, naturalmente, aberto ao se este deve ser uma divisão em partes iguais (10 atores, fator 0,1) ou que o protagonista recebe uma maior fator (10 atores, dos quais um é Brad Pitt, Brad conta para 0,55, para os outros 0,05 cada). Isso representa um problema quando você quer uma resposta para a pergunta "Como muita receita que geramos com o cinema, estrelado por Brad Pitt?"Em Nesse caso, apenas um ator é selecionado fazendo com que o fator para retornar um valor que é muito baixa de modo que o fator de cálculo deve ser omitido ou definido como o valor 1 neste caso. As coisas ficam realmente ruins quando queremos conhecer o nosso top 10 estrelas de cinema com base na receita total dos filmes que jogou dentro de alguns esses atores podem ter co-estrelou em um ou mais filmes no passado, causando resultados incorretos. A mensagem aqui é: Tenha cuidado quando estes tipos de modelagem de relacionamentos e ter certeza de que você fornecer acesso apenas a ponte relações quando o usuário ou o analista trabalhar com os dados sabe exatamente o que ele ou ela está lidando. Também é possível usar uma camada de abstração, como a camada de metadados Pentaho onde os objetos de cálculo podem ser criadas várias, uma para cada caso específico, uma definição estrita documentado. Capítulo 12 irá mostrar-lhe como configurar isso.

Criação de hierarquias

As hierarquias são instrumentos muito úteis para navegar apesar dos seus dados. A hierarquia permite que um usuário para iniciar em um alto nível de agregação (por exemplo, pro-
duto categoria) e suporta a perfuração nos detalhes de uma dimensão particular.
A maioria das hierarquias será implicitamente modelado dentro das tabelas de dimensão. Boa Exemplos disso podem ser encontrados dentro da dimensão de data com a hie-time Chies trimestre do ano-mês-dia ou na semana do ano, ou a dimensão do cliente com país-região cep endereço. Essas são simples, hierarquias fixas, onde todos os folhas na hierarquia têm a mesma profundidade''.''Ela começa a ficar interessante quando você
necessidade de modelar hierarquias de profundidade variável também. A forma mais comum de
construção de hierarquias de profundidade diferentes em um sistema de origem é ter os registros
em uma tabela de referência outros registros na mesma tabela. Pense, por exemplo, de um tabela de funcionários, onde cada trabalhador registro aponta para um gerente, que também é um empregado. Nesse caso, as referências da tabela em si, pelo que é na maior parte se refere
como um auto-associação. Oracle SQL contém um conectar-se antes declaração, que pode atravessar essas árvores relacionamento. Isto também é chamado recursão, mas este não é um
Instrução SQL ANSI padrão para a maioria dos bancos de dados, incluindo MySQL, não suporte a isso.

Felizmente, você também pode usar tabelas ponte aqui. Usando a tabela da ponte para hierarquias desequilibrada é opcional, sem a tabela da ponte, a dimensão
tabela podem ser unidas à tabela de fatos, como de costume. A tabela a ponte está lá apenas para
ajudar na navegação de hierarquias adicionais. É por isso que estes quadros são ainda por vezes referido como ajudante tabelas. Figura 7-12 mostra o banco de dados resultante diagrama quando utilizando uma tabela de hierarquia de ponte.

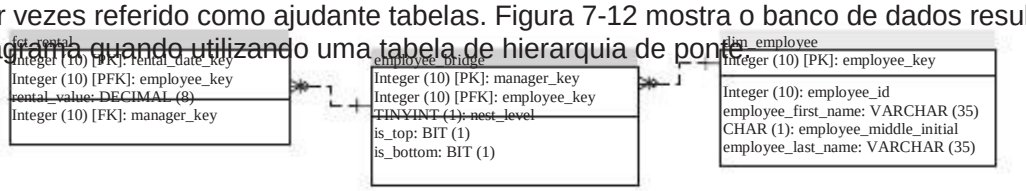


Figura 7-12: Hierarquia ponte

Tabela 7-6 mostra o que os dados dessa tabela parece que quando os dados para o relações empregado-gerente é adicionado, como mostrado na Figura 7-13.

Esta tabela ponte permite acumular os dados com base em qualquer questão você gostaria de perguntar. O cuidado é necessário, no entanto, se você não adicionar todos os necessários
restrições, há um risco de dupla contagem de alguns dos valores. Suponha que você quero a receita total do empregado 2 filtrando a chave do empregado. Sem filtros adicionais sobre nest_level o conjunto de resultados é dobrado porque employee_key 2 é feita duas vezes. Esta é também a grande desvantagem de uma ponte de hierarquia tabela: Cada caminho de cada item para qualquer outro item na mesma árvore é armazenado em um registro separado. Como resultado, a tabela da ponte fica muito maior do que o tabela de dimensão a que pertence.

Tabela 7-6: Empregado conteúdos tabela da ponte

manager_key	employee_key	nest_level	is_top	is_bottom
1	1	0	Y	N
1	2	1	N	N
1	3	1	N	N
1	4	2	N	Y
1	5	2	N	Y
2	2	0	N	N
2	4	1	N	Y
2	5	1	N	Y
3	3	0	N	Y

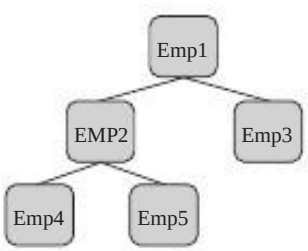


Figura 7-13: hierarquia desbalanceada

Uma alternativa ao uso de tabelas de ponte para o modelo de hierarquias é desequilibrada para forçar o achatamento da hierarquia. As manchas em branco no diagrama são simplesmente preenchido, repetindo os valores do nível acima. Tabela 7-7 mostra que este princípio.

Tabela 7-7: Achatada hierarquia desbalanceada

employee_key	MANAGER_ID	boss_id
1	1	1
2	1	1
3	1	1
4	2	1
5	2	1

Agora você pode criar uma estrutura de navegação do patrão para o empregado nível em todos os níveis conter uma relação com o nível acima. O número de colunas que você precisa depende do número de níveis na hierarquia, há há necessidade de a sobrecarga adicional de uma mesa de carteado. Existe um potencial risco envolvido aqui: Uma hierarquia achatada assume um número fixo de níveis. Se um outro nível é adicionado aos dados, a tabela de hierarquia resultante tem de ser reestruturado novamente.

Não importa qual é utilizado para construir o modelo de hierarquia, a transformação dos dados a partir de uma auto-junção relação ao posicionamento correto em ponte ou tabelas de hierarquia não é uma tarefa trivial.

Flocos de neve e dimensões de agrupamento

Até agora, temos sido quase sempre falando sobre a desnormalização de dados nos dados dimensões mart. Como resultado, a consulta a juntar os caminhos para a tabela de fatos são apenas um
nível de profundidade. A única excepção até agora foi o uso de tabelas para a ponte multi-valorizados colunas de dimensão. Você viu que a regra geral do polegar quando data marts modelagem é desnormalizar as tabelas de dimensão. Quando você levar isso para o extremo, você pode desnormalizar ainda mais. A derradeira modelo de dados desnormalizada consiste em uma única tabela, pelo menos a partir do usuário
ponto de vista. autor holandês e consultor Dr. Harm van der Lek descritos este como o Um conjunto de atributos Interface (AHV) conceito, em que todos os não-chave atributos em um esquema estrela são publicados para o usuário final e / ou consulta como uma ferramenta
lista única. No outro extremo da escala, você pode encontrar o totalmente normalizada modelos de dados que são usados principalmente em sistemas de operação. dados dimensionais
marts são posicionados no meio do caminho entre esses dois extremos.

Utilizando a normalização do esquema em estrela é normalmente chamado snowflaking, para
indicar a semelhança deste tipo de esquema com um floco de neve real. Figura 7-6 no capítulo anterior apresentou um exemplo deste para o cliente-região do país-relacionamento. Como acontece com qualquer modelo de data warehouse
técnica, há defensores e opositores da utilização de flocos de neve em dimensionais de data marts. Ralph Kimball opõe-se firmemente com flocos de neve com uma única exceção, explicada na seção seguinte. Gostaríamos a outra lista de exceção, que é chamado clustering. Este conceito é descrita em um artigo pelo Dr. Daniel Moody, que pode ser baixado em http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-28/paper5.pdf.

O artigo original é de 2000, e no verão e outono de 2003 Moody escreveu dois artigos subseqüentes para o Data Warehouse Institute (TDWI), que ainda digno de leitura. Estes artigos última gota o termo clustering e introduzir o termo starflake, que se resume a mesma coisa. A questão na mão é causada por múltiplas referências à mesma tabela normalizada em um data mart. No nosso exemplo WCM, temos esta situação com os clientes,

armazéns, empregados, fornecedores e todos eles de referência da mesma região
e mesa país em seus campos de endereço. Em um esquema em estrela com rigor, nós
necessidade de construir quatro transformações desnormalização, uma para cada dimensão.
Neste caso, a Moody aconselha a se aglomerar na região mesa país / e fazer este
uma subdimensão partilhada por todas as quatro tabelas de dimensão. A regra de ouro é que
assim como uma chamada garfo aparece no modelo de dados, a tabela de pesquisa não é
desnormalizada mas é usado como uma mesa de cluster. Uma bifurcação significa que dois
candidatos
referência as tabelas de dimensão da mesma tabela de pesquisa, como você pode ver na
Figura 7-14.
O diagrama mostra um exemplo de uma solução estritamente normalizado à esquerda
e estrela de um cluster ou esquema starflake à direita.

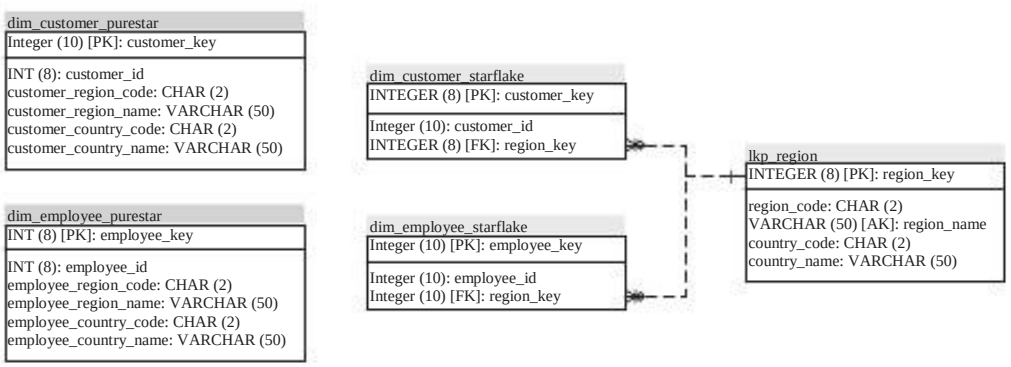


Figura 7-14: esquema Starflake

Esta abordagem tem várias vantagens. Primeiro de tudo, é uma questão menor, mas o
tabela de dimensão fica um pouco menor. De maior importância é a manutenção da
tabelas starflake: As mudanças ocorrem apenas em uma mesa eo processo de ETL tem
apenas
para atualizar uma tabela em vez de dois ou mais. Há desvantagens, bem como,
é claro. Você precisa criar exibições extras (quando o modelo de solução em
antecedência) ou usar apelidos na sua consulta extra, porque você não pode fazer referência
a
mesma tabela de consulta em uma consulta, onde clientes e funcionários estão envolvidos.
A maior desvantagem, entretanto, é que você está criando dependências no seu
ETL processo. Você precisa ter certeza de que a região / país tabela de pesquisa
é processado antes de as dimensões que usar essa tabela ou você corre o risco de
inconsistências em seu data warehouse.

Se você quiser aderir a um modelo rigoroso esquema em estrela, o que significa um
nível máximo de 1 para a junta entre fatos e dimensões, há também
outra solução. Em vez de snowflaking as tabelas em cluster, você pode tratar
-los como dimensões regulares. Isto significa que as chaves de dimensão farão parte do
a tabela de fatos também. Inconvenientes para essa forma de modelagem são que você
sempre
necessidade de percorrer a tabela de fatos para chegar à região / país de um cliente,
e é claro que você precisa teclas extras em sua tabela de fatos, o que pode torná-lo
desnecessariamente larga.

Estabilizadores

Há, na opinião de Kimball, um único caso permitido para o uso do
técnica de floco de neve. Ele usa o termo guiga tabelas para descrever esta parti-
ular tipo de snowflaking, assim que nós gentilmente adotar a mesma palavra para isto para evitar
qualquer confusão. Primeiro, vamos explicar o que entendemos por um outrigger: Suponha que você
tem um conjunto de atributos que são dependentes de um dos maiores da dimensão
nível de atributos. Um exemplo são os dados Zipcensus, que contém um grande conjunto
de atributos analíticos. CEP é um atributo de nível superior ao cliente (você
pode ter vários clientes compartilhando o mesmo código de endereçamento postal) e todos os Zipcensus
atributos são dependentes CEP. Se você armazenar todas estas colunas adicionais
na tabela de clientes, você é a desnormalização um monte de dados e desordenar o
tabela de dimensão com um monte de atributos extra, neste caso, 112 atributos extras
além do código postal próprio. Figura 7-15 mostra um exemplo do Zipcensus
tabela estabilizadores em combinação com a dimensão do cliente.

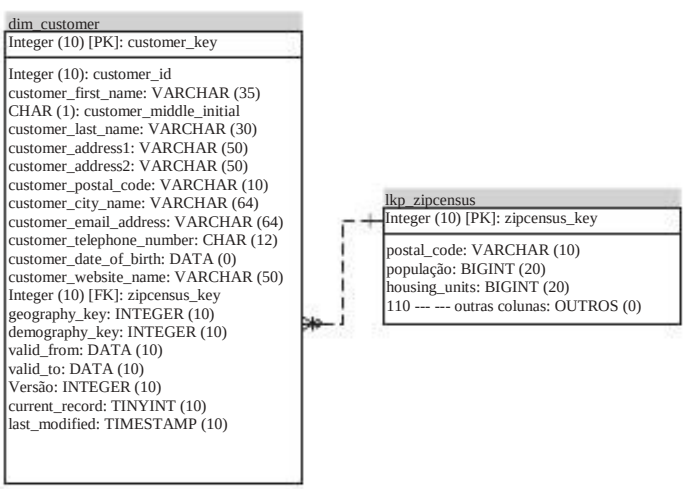


Figura 7-15: exemplo Outrigger

Tabelas Consolidação multi-grão

A última questão de modelagem é um pouco diferente daquelas descritas assim
agora, uma vez que não aborda a modelagem de uma única dimensão. O caso é o seguinte:
suponha que você tenha uma tabela de fatos contendo dados de nível de transação e você quer
comparar estes valores reais para um orçamento ou previsão. A maioria das organizações não
criar orçamentos para o cliente individual e nível de produto, mas fazer isso, por
exemplo, no mês e de nível de grupo de produto, omitindo o cliente
e outras dimensões. Assim como você acomodar essa diferença no
granularidade? Uma coisa é muito clara: você não pode comparar um valor mensal para
um dia, sem primeiro resumindo os valores diários do nível meses. Assim

A primeira coisa necessária é uma tabela resumida, fato que pode ser feito tanto em
A (tabela extra plus processo de carga) de forma física ou virtual (ver). O último
É mais fácil criar, mas pode ser proibitivo devido aos problemas de desempenho. A
tabelas de dimensão que acompanham deverão estar disponíveis no mesmo nível, bem como,
quer pela criação de uma tabela separada ou criando um ponto de vista. (Lembre-se que,
quando
se juntar a uma tabela de fatos no nível do mês com uma tabela de dimensão a nível dia,
os resultados são multiplicados pelo número de dias no mês!)

Figura 7-16 mostra um exemplo com base no armazém de dados WCM que
contém tanto o orçamento e os dados reais a um nível consolidado de granularidade.
É por isso que Ralph Kimball chama esse consolidada tabelas de fatos.

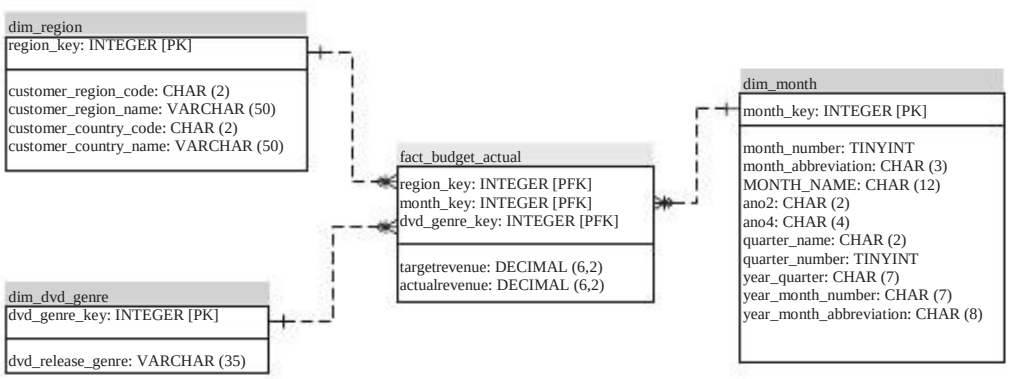


Figura 7-16: Real e orçamento consolidado

As tabelas da Figura 7-16 podem ser criados diretamente a partir das tabelas existentes no
o data warehouse. Observe também que quando os dados estão disponíveis em uma tabela
de dimensão
em um nível mais alto de granularidade, ele é mantido na tabela de dimensão derivada como
também. Este é o caso para as colunas de trimestre e ano na dimensão mês,
e as informações do país na dimensão região.

Resumo

Este capítulo apresenta uma grande coleção de técnicas de modelagem para trabalhar
com os dados dimensionais. Nós abordamos os seguintes assuntos:

- esquema em estrela terminologia básica e modelagem
- Uma introdução ao SQL necessários para consultar esquemas em estrela
- Aplicando a arquitetura de barramento para colar todos os data marts para formar um dimensional data warehouse empresarial
- Um conjunto de princípios de design para usar na construção de um data warehouse
- As várias estratégias para captar a história do modelo tridimensional

- Conceitos avançados que podem ser usados para a construção de dados dimensional armazéns

Como mencionado, recomendamos o Data Warehouse Toolkit série por Ralph Kimball e companhia sobre o assunto, e também recomendo vivamente uma visita ao o site www.kimballgroup.com, Onde você pode encontrar muitas dicas mais em desenho dimensional.

CAPÍTULO

8

O Data Mart Processo de Projeto

Capítulos 6 e 7 apresenta-lhe o armazenamento de dados, a tecnologia disponível e as técnicas de modelagem predominante usado. Estes capítulos constituiu uma base sólida fundamentação teórica que você pode construir em vez de realmente começar a desenvolver e
criação de sua primeira soluções reais. Agora é hora de explicar os passos necessários para obter essas soluções no lugar. Como explicado no capítulo 6, um data warehouse é um processo, não um projeto. Um processo consiste em mais do que apenas tecnologia e especialmente no caso de uma solução de business intelligence, envolvendo o seu fim usuários no processo é essencial para seu sucesso.

Antes que você possa iniciar a construção de um data warehouse ou data mart primeiro o data warehouse, você precisa saber o que você quer colocar nele. Em maioria dos casos existe uma demanda existente por uma solução que irá entregar a insights e relatórios necessários para melhor gerir uma organização. O primeiro passo no processo do projeto é identificar essas demandas, muitas vezes latente e converter los em requisitos concretos para a solução de business intelligence. Com o necessidades identificadas, você pode começar a projetar e construir um data warehouse solução que pode atender a essas demandas que oferecer à sua organização valor real.

Análise de Requisitos

O ponto de partida de qualquer projeto de data warehouse é definir claramente o requisitos do ponto de vista empresarial. Quantos dados serão provenientes para a organização, desde que as fontes, e em que condição? Que tipo de informações, em formatos que, a sua organização precisa? Que se afastem-mentos vai executar relatórios, quantas vezes, e em que dados? Quanto técnica

conhecimento que os usuários tenham o seu negócio, eles precisam de um pacote de relatórios ou
eles vão executar consultas ad hoc? E o mais importante, a informação que
ajudar os usuários a identificar se o negócio está ainda na trilha para alcançar
os objectivos definidos na estratégia corporativa? Com base nesses requisitos, uma
de negócios adequado deve ser desenvolvido. O negócio é, em última análise o
justificativa para o início do projeto de data warehouse, em primeiro lugar. Sem
um caso de negócio, há obviamente nenhuma razão aparente para embarcar em um
frequentemente projeto caro e demorado como o desenvolvimento de uma base de dados warehous-
ing solução para sua organização. Assim como você reunir os requisitos e
desenvolver um business case (BC) para um projeto de data warehouse?

Primeiro, como explicamos nos capítulos anteriores, um data warehouse e
solução de inteligência de acompanhamento de negócios são um meio para terminar. Este fim""
é o que é sobre tudo e deve ser descrito no BC em termos de negócios
beneficia de uma perspectiva de negócio. Em outras palavras, você não quer para justificar
um investimento em tecnologia pela tecnologia, mas o valor pode
trazer para a organização. A melhor maneira de se começar a envolver um ou mais chave
stakeholders e criam em conjunto o plano de negócios e requisitos.

Isso nos leva a um dos aspectos mais importantes, ou talvez o mais
aspecto importante, de uma equipe de projeto de data warehouse. A equipe deve ser
multi-disciplinar e consistem de duas pessoas tecnicamente qualificados e de negócios
usuários. Escusado será dizer que, em todos os casos, o departamento de TI devem estar envolvidos, bem
porque você vai precisar de uma infra-estrutura de TI para executar o banco de dados e
relatórios. A próxima seção ajuda a identificar os usuários de negócios que serão
mais útil na identificação de suas necessidades.

Obtendo o direito de usuário Envolvidos

Ausuário é alguém na organização que irá trabalhar com o front-end-Pentaho
ferramentas para dashboards, relatórios e análise. Normalmente, existem quatro tipos
de usuários:

- One-click (consumidor) do usuário podem apenas consumir informação que está sendo empurrada em um formulário pré-formatado, geralmente por e-mail ou através de um portal.
- Dois cliques (reciclagem) do usuário-Usos um portal para procurar um painel ou abre relatórios predefinidos. Pode abrir e atualizar documentos sob demanda, mas pede-se preencher a interatividade somente envolvidos.
- Multi-clique (construtor) do usuário podem trabalhar de forma interativa com os apresentados
  informação e pode criar relatórios adicionais e análise com vistas
  Ad-Hoc componente Relatório e Mondrian frente JPivot.
- Power User (analista), é possível criar relatórios usando o Designer de Relatórios e cria novos modelos de Mondrian. Os usuários mais avançados também podem funcionar com a bancada de mineração de dados Weka.

Curiosamente, estes tipos de usuários mapa para diferentes partes e os níveis da organização. Para o sucesso do seu projeto, sua primeira prioridade é conseguir o um-clique os usuários a bordo. Estas são as pessoas que alocam orçamentos e pode puxar as cordas direito de você ir. Eles também são chamados gerentes, e porque um projeto de data warehouse tende a exigir um orçamento considerável e tem um forte impacto e visibilidade dentro da organização, seu desafio é convencer pelo menos um executivo de nível C para patrocinar o projeto de data warehouse. Com este patrocínio, é fácil de atrair outros usuários-chave para se envolver. Sem ele, você terá um tempo difícil convencer a empresa que todo o esforço ea despesa é de valor.

O segundo grupo de pessoas geralmente é maior. Eles têm a capacidade de puxar a informação a pedido de um cliente de software de portal ou especiais pedido. Apenas um pouco de treinamento é necessário para o uso dessas ferramentas, geralmente um
Basta breve introdução para obter os usuários vão. Normalmente, a combinação dos usuários de um e de dois cliques é chamado usuários finais.

O terceiro grupo e os que você estará trabalhando com a forma mais frequente base pode ser encontrada no grupo de usuários clique com o botão multi. Esses usuários podem ajudá-lo
na definição de"que"do armazém de dados em termos da produção. Este grupo geralmente é composto por pessoas que já fornecem informações para o organização sob a forma de planilhas Excel ou outros formatos similares, e normalmente trabalham em departamentos como finanças ou marketing. Não é coincidência que a maioria dos projetos de data warehouse começar de qualquer uma instituição financeira, vendas ou marketing
perspectiva, e essas pessoas podem explicar-lhe que o resultado final deve aparência.

O quarto grupo é composto dos usuários avançados ou analistas. Eles podem fazer sua vida muito fácil, porque eles provavelmente sabem onde os dados vem, que a qualidade dos dados é, e como ela pode ser integrada. Eles também podem tornar a sua vida complicada, porque eles têm altas exigências que são normalmente não abrangidos durante a primeira iteração do data warehouse. Você precisa deles na equipe, mas tem que gerir as suas expectativas de antecedência. Além disso, ser muito cuidado em fazer promessas para este grupo sobre o que vai e não vai ser possível na primeira fase do data warehouse.

Coleta de Requisitos

O processo de levantamento de requisitos requer a disponibilização de uma chamada analista de negócios. Em muitos casos, esta tarefa cabe a um consultor externo. Tanto como nós gostaríamos que você vai contratar uma consultoria externa, há uma armadilha possível aqui
que você deve estar ciente. Não importa o quão experientes são essas pessoas, elas Não sei todos os cantos e recantos da organização e da empresa
cultura. Os projetos de inteligência de negócios mais bem sucedidos trabalham com insiders, pelo menos em um papel de apoio. Um analista de negócios deve não apenas entender o negócio, mas também precisa ser capaz de traduzir os requisitos de negócio em

a solução certa. Ele ou ela também deve ser capaz de fazer essa tradução dos outros caminho de volta, para explicar questões técnicas da terminologia empresarial.

Uma maneira comum de recolha de requisitos é a entrevista de usuários potenciais, gestores e membros da equipe. Essas entrevistas devem ser realizadas por dois pessoas, onde o analista de negócios tem o papel principal e é geralmente acompanhada por alguém da equipe do data warehouse que tem discernimento na disposição dados e funcionalidades das ferramentas que serão utilizadas. Explicando a entrevista processo em profundidade está fora do escopo deste livro, mas o já mencionado Data Warehouse Lifecycle Toolkit por Ralph Kimball é um excelente recurso para neste tópico.

Em muitos casos, já existe alguma forma de comunicação no local e estes relatórios existentes, muitas vezes fazer um excelente conjunto de requisitos, ou pelo menos uma boa
ponto de partida. Se este for o caso, ainda é uma boa idéia para realizar entrevistas para afirmar o envolvimento do usuário no projeto, e determinar os pontos fortes e fracos dos relatórios atuais.

Quer entrevistas ou outros meios de coleta de informações é utilizado, o resultado final deve ser um conjunto bem definido de requisitos, anotadas na planície Inglês. Os requisitos devem ser agrupadas por actividade principal processos porque seu data warehouse será construída em torno de processos de negócios principais, tais como vendas, retenção de clientes, gestão de armazéns e finanças. O documento deve conter no mínimo os seguintes dados:

- Tópico-A principal área ou processo a exigência pertence.
- Público-Quem é a solução para?
- Dona-Quem será o proprietário do negócio da solução?
- demanda do usuário em texto explicando o que os usuários precisam e como eles irão utilizar a solução descrita.
- Questões respondidas-A questões de negócios que serão respondidas por a solução. Capítulo 5 contém vários exemplos destas questões.
- Benefícios para a empresa, que será o ganho de organização com a construção do parte específica do armazém de dados?
- Como mecanismo de entrega as informações serão disponibilizadas para os usuários? Isso pode ser qualquer coisa, desde uma simples lista enviada por e-mail para um
painel análise interativa.
- As fontes de informação, onde é obter informações adicionais sobre este requi-mento disponível, o que as pessoas possam ser feitas?
- As fontes de dados, que sistemas ou bancos de dados podem ser usados para obter a dados?
- A cobertura dos dados Indicação- da integridade dos dados disponíveis para responder às questões de negócios.

- A estimativa de custo indicação aproximada do tempo e investimentos necessários para desenvolver esta solução.

Cada um desses documentos serve como um caso de pequenas empresas por conta própria e
deve se encaixar no caso de negócios global do projeto de data warehouse. Ao preencher em todos os tópicos da lista, a coleta de requisitos também pode ser usado para priorizar incrementos projeto.

Análise de Dados

A fonte para cada solução de data warehouse é um dado. Os dados podem ser obtidos de muitas fontes e cada fonte pode ter desafios específicos de recuperação e transformar os dados para ser utilizado no data warehouse. Os seguintes lista serve como um guia para ajudar você a selecionar as fontes de dados e endereço certo alguns dos desafios envolvidos:

- A maioria dos sistemas ERP- organizações de hoje executar um ou mesmo vários Enterprise Resource Planning (ERP) como SAP, Oracle Financials, COMPI `re ou OpenERP (só para citar alguns). Estes sistemas se esforçam para supporte o processo de negócios completos e abrangem tudo, desde contabilidade a comprar para a fabricação de RH. Isto também significa que esses sistemas são notoriamente complexo, o amplamente utilizado sistema ERP SAP R / 3 contém mais de 70.000 mesas! Estes quadros, muitas vezes não podem ser acessados diretamente, mas expor seu conteúdo através de uma API ou camada outros metadados. Como consequência, não é o suficiente para ser capaz de se conectar ao banco de dados, que é frequentemente proibidas pelas licenças de qualquer maneira. Você precisa de um software especializado para ser capaz de ler os metadados do sistema e, como você já deve ter adivinhado já, muito poucas soluções open source já estão disponíveis que oferecem este funcionalidade. Para Pentaho Data Integration, um comercial do plugin A empresa alemã ProRatio está disponível. A solução de ETL francês Talend Open Studio tem um nativo de código aberto conector SAP, mas se você executar E-Business Suite da Oracle, JD Edwards EnterpriseOne, ou qualquer dos outros suítes principais ERP, você enfrentará alguns desafios na obtenção de dados corretos.
- sistemas de Home-grown-se a sua organização desenvolveu a sua própria sup sistemas de portabilidade, você pode ter uma vantagem sobre as soluções pré-construídas porque o conhecimento sobre as estruturas de dados subjacente já está dis-capaz. Cuidado, porém, que nestes casos a documentação não pode ser up-to-date, as pessoas que inicialmente desenvolveu o sistema pode ter deixado da empresa, ou você simplesmente não estão autorizados a acessar os dados dos sistemas diretamente.
- sistemas mainframe Grande corporações, como bancos ou seguros empresas ainda dependem fortemente do mainframe para a sua informação essencial

necessidades de processamento. Isso significa que você precisa para obter os dados do mainframe, que pode ser uma tarefa desafiadora. A maneira mais fácil de obter esses dados é que ele seja entregue à área de carga do armazém de dados em um padrão de caracteres ASCII de arquivos separados.

- Planilhas-A apenas um bom conselho quando uma organização quer usar planilhas como uma fonte de dados é: não! Planilhas são freqüentemente usados como subsistemas completa, especialmente nos departamentos de finanças e contabilidade. Sua grande vantagem é a flexibilidade ea facilidade de uso, e é exatamente por que você deve se recusar a aceitar as planilhas como uma fonte. Eles também usam formatos de exibição que mostra os dados de uma forma formatada, mas ao ler os dados diretamente contém algo completamente inútil. Assim, mesmo que o PDI é capaz de planilhas de leitura, não torná-los parte do processo. Se nenhuma fonte alternativa está disponível (típico de orçamentos e previsões), têm o proprietário da planilha, exportar os dados em uma pré-definidos e acordados formato como um arquivo de texto ASCII.

- bases de dados dos computadores de mesa Veja a entrada anterior para planilhas. Não há uma regra muito simples para os quais fontes de dados podem ser aceites para os dados armazém e que não pode: Se o sistema não é suportado pela TI departamento ou há apenas um usuário a manutenção do sistema, não perca muita energia sobre ele.
- Estruturado de dados externos Estruturada significa que os dados sejam entregues em um formato bem definido por um terceiro, que disponibiliza esses dados, ser on-line ou em qualquer outra forma. Bons exemplos destas fontes de dados são os dados Zipcensus define que usamos para WCM ou os dados do mercado de varejo que pode ser obtido a partir de Nielsen.
- -Mais de dados on-line e mais informações podem ser obtidas diretamente na web e consumido sob a forma de um serviço Web ou feeds RSS. Pentaho Integração de dados contém todos os recursos para trabalhar com dados on-line.
- Weblogs Esta- é um caso especial de dados estruturados, pois você vai precisar um pouco de expressões regulares customizadas para rasgar as linhas de log separados e tornar a informação útil fora dele.
- XML O Extensible Markup Language (XML) tornou-se a lingua franca do mundo da informática. Todos os produtos da Pentaho armazenar suas (meta) informações em um formato XML, e muitos outros produtos são capazes de exportação de dados em formato XML também. A entrega das informações é não limitado a arquivos, a maioria dos sistemas de mensagens baseados em fila ou entregar a sua mensagens em formato XML também. A grande vantagem do XML é que ele é basicamente um arquivo de texto, apenas como um ASCII regular por vírgula ou por tabulação arquivo. Isto significa que os arquivos XML podem ser abertos com qualquer editor de texto para visualizar o conteúdo. O desafio com arquivos XML, porém, é que é assim flexível que pode ser difícil de transformar a estrutura aninhada a uma utilizável

formato relacional utilizado no data warehouse. Felizmente, ao contrário de muitos outros ferramentas de integração de dados, o PDI tem um passo de entrada muito forte e pode XML
RSS, leia também alimenta diretamente (que é um formato XML).

Estivemos envolvidos em projetos de data warehouse desde o início dos anos 90 e descobriu que cada atribuição e caso é diferente. Qualquer organizações de tamanho razoável,
nização tem uma mistura de todas as fontes de dados descritas. Às vezes a parte do ERP da equação é preenchida por um único sistema financeiro ou de RH, e vários
sistemas de home-grown ou feitos sob medida rodeiam este. Em nenhum dos projetos para agora poderia todas as informações necessárias ser encontrada em um único sistema integrado.
Todos os dados estruturados é normalmente armazenado em um banco de dados relacional, o que faz
facilmente acessível, pelo menos do ponto de vista técnico. O mais comum
dados adicionais não são armazenados em bases de dados consiste de orçamentos, estimativas, previsões,
e planos de conta e é normalmente mantida em planilhas. O desafio é
para decifrar o conteúdo, estrutura, qualidade e significado de todos esses dados, e isso é o que você precisa de documentação (se disponível, do contrário ela precisa ser criado) e dados de perfil para.

Data Profiling

Dados perfis é o processo de recolha de estatísticas e outras informações sobre os dados disponíveis nos sistemas de origem diferente. A informação obtida é inestimável para o design ainda mais do seu armazém de dados e processos de ETL. perfis de dados também é uma parte importante de qualquer iniciativa de qualidade dos dados, antes
qualidade pode ser melhorada, uma linha de base tem de ser estabelecida, indicando quais o estado atual dos dados é. Perfil pode ser executada em três diferentes
níveis:

- Perfil de Coluna-Coleta estatísticas sobre os dados em uma única coluna
- Dependência do perfil Cheques para as dependências de uma tabela entre diferentes colunas
- Junte-se do perfil Cheques para as dependências entre tabelas diferentes

O ponto de partida para definição de perfil é sempre o perfil do nível de coluna, que gera informações úteis sobre os dados em uma coluna, incluindo, mas não
se limitam a:

- Número de valores distintos, como muitas entradas exclusivas que a coluna contém?
- Número de NULL e valores vazios-How muitos registros não têm valor ou um valor vazio?
- valores máximos e mínimos, não apenas numéricos, mas também para tex- dados contratuais.

- Numérica soma, média, mediana e desvio-padrão Vários cal-
  Os cálculos feitos sobre os valores numéricos e distribuição de valor.
- String e padrões de comprimento são os valores armazenados corretamente? (Para exemplo, alemão códigos postais deve conter cinco dígitos).
- Número de palavras, o número de caracteres maiúsculos e minúsculos
- contagens de freqüência, que estão no topo e no fundo Nitens em uma coluna?

A maioria dos dados de perfis ferramentas podem fornecer essas informações e às vezes até
mais. Ele fica mais complicado quando você olha os perfis dentro de uma tabela única para identificar correlações e interdependências. Exemplos disso são as combinações
de código postal para cidade, a cidade-região, e, região para outro país. Obviamente, uma cidade
nome é dependente de um código postal, o nome da região sobre a cidade, e os
país na região. A existência destas dependências violar o terceiro
forma normal, então quando encontrar essas relações em uma terceira forma normal sistema de origem, você deve tomar cuidado extra, principalmente em relação ao endereço da informação. Às vezes as relações não são muito claras ou são mesmo confusas,
o que torna difícil a distinção correta de entradas incorretas. Esta é
exatamente a razão pela qual tantas caras endereço de correspondência e limpeza
As soluções existem. Tomemos, por exemplo, a combinação cidade-região: Há mais de dez estados nos Estados Unidos com uma cidade chamada Hillsboro. Sem conhecimento adicional dos códigos de país ou zip, é difícil dizer se
um registro contém informações erradas ou não. Para esses casos, você vai precisar informações externas para validar os dados contra.

As relações inter-table são mais fáceis de perfil, é simplesmente uma questão de ava-
equacionar a possibilidade de uma relação é aplicada corretamente. Em um sistema de entrada de ordem,
não deve ser possível encontrar um número de cliente na tabela para que não
não existe na tabela de clientes. O teste mesma relação pode ser usada para encontrar quantos clientes estão na tabela de clientes, mas não (ainda) na ordem
tabela. O mesmo se aplica aos produtos e detalhes dos pedidos, estoques e fornecedores, e assim por diante.

Usando DataCleaner eobjects.org

Atualmente, o Pentaho BI Suite não contém dados de perfis de capacidades,
por isso vamos usar uma ferramenta chamada DataCleaner desenvolvidos pela comunidade open source,
Comunidade eobjects.org. O software pode ser obtido a partir de http://datacleaner
. Eobjects.org e é muito fácil de instalar. No Windows, basta descompactar
o pacote e iniciar datacleaner.exe. Em uma máquina Linux, depois descompactar-
ção da tar.gz arquivo que você primeiro precisa fazer o datacleaner.sh shell script executável para iniciar o programa. Se você estiver usando o ambiente desktop GNOME ente, isto é muito fácil: basta o botão direito do mouse no arquivo e abra as propriedades. Em seguida, vá para a aba Permissões e marque a caixa de seleção antes de permitir a execução

arquivo como programa. Agora você pode clicar duas vezes sobre o datacleaner.sh arquivo eo programa será iniciado. Se você quiser uma maneira mais conveniente de iniciar o programa da próxima vez, você pode criar um atalho (no Windows) ou um lançador (no GNOME).

DataCleaner prevê três tarefas principais:

- Perfil-Tudo as tarefas coluna perfis descritos anteriormente. A idéia aqui é para obter insights sobre o estado dos seus dados. Assim, você pode usar o perfil tarefa, sempre que você quiser explorar e tirar a temperatura do seu banco de dados.

- Validar-To criar e testar as regras de validação com os dados. Estes regras de validação podem posteriormente ser traduzido (por mão) em Pentaho Data validação de medidas de integração. O validador é útil para a aplicação das regras sobre os dados e controlar os dados que não estejam em conformidade com esses regras.

- Compare-To comparar dados de diferentes tabelas e esquemas, e verificar a consistência entre eles.

A partir dessas descrições, é claro que não DataCleaner fornecer recursos de profiling tabela intra-como uma opção direta, mas há outros maneiras de fazer isso com a ferramenta, como mostraremos mais tarde.

A primeira coisa que você precisa, é claro, é uma conexão com o banco de dados que você quer ao perfil. Para cada tipo de banco de dados, DataCleaner necessidades dos correspondentes driver, que permite a comunicação entre a ferramenta ea base de dados.

Antes de explicar como adicionar drivers e conexões, vamos dar uma primeira olhada nas funções disponíveis. DataCleaner começa com o painel de tarefas Novo aberta, que permite que você escolha uma das três opções principais: perfil, validar, e Compare. Clique no perfil para iniciar uma tarefa de perfis novos. Você verá um tela vazia quase dois painéis com algumas opções e os dados selecionados no indicação no painel à esquerda (ver Figura 8-1).

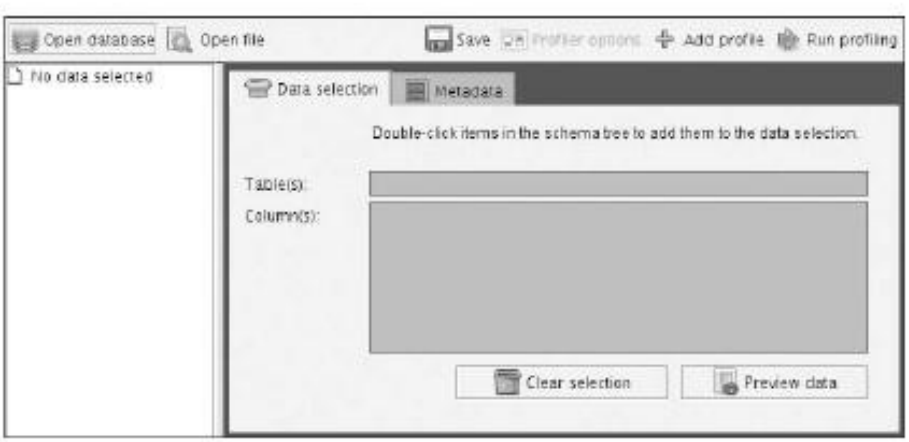


Figura 8-1: tarefa Criação de Perfil

Agora selecione Abrir banco de dados, selecione a entrada de SampleData DataCleaner Na lista de conexão Named drop-down e clique em Conectar ao banco de dados. Todos outros campos têm sido definido. Quando você abre o nó da árvore PÚBLICO à esquerda, clicando no +sinal, a lista com as tabelas é exibida. Cada tabela pode ser aberto individualmente, o que exibe as colunas disponíveis. Para adicionar um coluna para a seleção de dados, basta clicar duas vezes nele. Você notará que a tabela nome é acrescentado ao campo de tabela (s), ea coluna para o campo de coluna (s). Para remover uma coluna da seleção, clique duplo-lo novamente ou utilize Limpar Seleção para remover completamente a tabelas e colunas selecionadas. O Preview opção mostra uma amostra dos dados selecionados, o número de linhas a serem recuperadas pode ser ajustado após clicar no botão. O valor padrão geralmente é suficiente para obter uma primeira impressão do conteúdo dos dados. Cada tabela começa seu selecionados colunas exibidas em uma janela separada.

Ao lado da guia de seleção de dados é o guia Metadados. Ao clicar nesta, o metadados técnicos das colunas selecionadas é exibida. O tipo de campo, campo comprimento e, especialmente, a indicação Nullable lhe dar uma primeira impressão de o tipo de dados que podem ser esperados.

Adicionando tarefas perfil

Depois de selecionar algumas colunas para o perfil, você pode adicionar perfis diferentes. Dados
Cleaner contém as seguintes opções de perfil padrão:

- As medidas padrão-Row contagem, o número de valores nulos, valores vazios, valor maior e menor.
- análise String Porcentagem de caracteres maiúsculos e minúsculos, por cento idade de caracteres não-letra, número mínimo ou máximo de palavras, eo número total de palavras e caracteres na coluna.
- Time-análise Menor e maior valor de data, acrescido de número de registros por ano.
- análise do número-Maior, menor soma, média, média geométrica, nor- desvio padrão e variância.
- Localiza-finder Padrão e conta com todos os padrões em uma coluna de caracteres. Geralmente usado para números de telefone, códigos postais, ou outros campos que devem
  obedecer a um padrão específico alfa-numérico. exemplos padrão é 9999 aa (4 dígitos, espaço 2 caracteres), (hífen três personagens, três dígitos) aaa-999.
- Dicionário matcher-Jogos as colunas selecionadas contra o conteúdo de um arquivo externo ou outra coluna de banco de dados (um dicionário"").
- Regex matcher-Jogos colunas contra uma expressão regular.
- Data matcher máscara Jogos colunas de texto contra padrões de data, esta não pode ser usado com campos de data, apenas com campos de texto contendo a data e / ou tempo de informação.

- distribuição de valor Calcula o topo ea base valores de N em um coluna com base na sua freqüência, ou classifica o número de ocorrências e calcula a porcentagem de freqüência para cada valor. O valor para Npode ser qualquer número entre 0 e 50, o padrão é 5.

A coleção de perfis em uma tarefa é muito flexível, é possível adicionar perfis do mesmo tipo a uma única tarefa. Cada tarefa pode ser salvo, bem como, mas isso só vai salvar os perfis de conexão e de tarefas, não os resultados de perfil. Esta última opção é uma função distinta e salva os resultados em um arquivo XML, que infelizmente é uma rua de sentido único; DataCleaner não pode ler esses arquivos de volta. Persistindo resultados de perfil faz parte do roteiro para futuros lançamentos.

Adicionando conexões de banco de dados

Uma das primeiras coisas a fazer, ao estabelecer o ambiente de dados de perfis É para adicionar os drivers de banco de dados e ao armazenamento correto das ligações para o seu próprio
bases de dados para facilitar a seleção. A primeira tarefa é bastante simples e, no DataCleaner tela principal, selecione Arquivo, Registro driver de banco de dados. Há dois maneiras de adicionar um novo driver. A primeira é baixar e instalar automaticamente elas. Esta opção está disponível para MySQL, PostgreSQL, SQL Server / Sybase, Derby, e SQLite. A segunda maneira de fazer isso é para registrar manualmente uma Jar. arquivo com os drivers. Para ajudar você a encontrar os drivers, DataCleaner contém a opção de visitar o site do driver para os drivers de banco de dados mais comuns, tais como os de Oracle ou IBM DB2. Depois de baixar um driver, você precisará fazer referência a ela, selecionando o arquivo e da classe driver correto. Para o MySQL, nós usará o download automático e instalar opção.

DICA Se você já instalou o driver JDBC do MySQL, não há necessidade de download -lo novamente, basta registar o seu vigor Jar. arquivo.

Adicionando a conexão para que você possa selecioná-lo na lista drop-down na Abra a caixa de diálogo banco de dados é um pouco mais complicado. Para isso, precisamos de alterar
o arquivo de configuração DataCleaner, que pode ser encontrado no DataCleaner pasta e é chamado datacleaner-config.xml. Para editar arquivos XML, é melhor usar um editor de texto simples que compreende a sintaxe XML. Para o Windows plataforma, o Notepad + + de código aberto pode ser usado; em uma máquina Linux, apenas direito do mouse no arquivo e abra com o editor de texto. Olhe para a parte do arquivo que diz:

<-! Conexões nome. Adicione suas próprias conexões aqui. ->.

Abaixo desta linha há uma entrada vazia para o drop-down list, basta deixar que onde ele está. A segunda entrada é a conexão com os dados da amostra. Cópia parte da amostra de dados que começa com feijão < e termina com </ Bean>, Incluindo

o começo eo fim feijão tags. Cole-o logo abaixo da marca de fechamento da amostra entrada de dados e ajustar a informação para refletir suas próprias configurações. Abaixo está a entrada, ele deve procurar a conexão com o banco de dados WCM no seu máquina local:

```
class="dk.eobjects.datacleaner.gui.model.NamedConnection"> <bean
  <property name="nome_qualquer" value="WCM MySQL database" />
  <Nome da propriedade = "connectionString" value = "jdbc: mysql: / / localhost:
    3306 "/>
  <property name="username" value="yourusername" />
  <property name="password" value="yourpassword" />
  name="tableTypes"> <property
   <list>
     <value> TABELA </ value>
   </ Lista>
  </ Property>
</ Bean>
```

Para ter DataCleaner também ligar para o catálogo correto, no nosso caso o Catálogo de WCM, uma linha extra deve ser adicionada abaixo da propriedade senha linha, como este:

```
valor <property name="catalog" = "wcm" />
```

Não é recomendável armazenar senhas em arquivos de texto simples, na verdade, nós se opõem fortemente a fazê-lo, e neste caso você pode deixar o campo de senha vazia também. Nesse caso, você precisará fornecer a senha cada vez que você criar um novo perfil de tarefa.

Para usar DataCleaner com outras fontes de nosso banco de dados WCM, você pode encontrar exemplos do elemento XML de feijão para outros bancos de dados populares no documentação DataCleaner online.

Fazer um perfil inicial

O Profiler DataCleaner foi otimizado para permitir que você faça uma vez rápida e, ao mesmo tempo perspicaz perfil com pouco esforço. Para começar com perfil, você pode adicionar as medidas padrão, String Análise, Número Análise e Análise Tempo perfis clicando repetidamente o perfil Adicionar botão no canto superior direito da janela de tarefa perfil. Você pode aplicar esses perfis para todas as colunas de seu banco de dados para obter a percepção inicial.

Trabalhando com Expressões Regulares

As expressões regulares ou expressões regulares são uma forma de mascarar e descrever os dados,
principalmente para fins de validação, mas também para encontrar certos padrões em um texto. Vários
livros foram escritos sobre como trabalhar com expressões regulares, assim nós nos referimos

a informação existente aqui. Além disso, o capítulo 11 contém alguns exemplos de como você pode usar expressões regulares para analisar dados de websites. Dados Limpador contém matcher uma regex como um dos perfis, bem como uma regex validação como parte do validador. Antes que você pode usar expressões regulares, você precisará adicioná-los ao catálogo Regex DataCleaner na tela principal. Inicialmente, este catálogo está vazio, mas é fácil de adicionar regexes. Quando você clica regex Nova, três opções aparecem. A primeira é criar um homem novo regex vamente eo último é conseguir uma regex do propriedades. arquivo. A segunda é a opção mais interessante: Ao selecionar Importar da RegexSwap, uma biblioteca on-line é aberto com uma grande coleção de expressões regulares existentes para escolher a partir de. É também possível contribuir com a sua própria regexes ao RegexSwap em http://datacleaner.eobjects.org/regexswap para outros (re) utilização. Depois importação de uma regex do RegexSwap, você pode abri-lo a mudar seu nome e a própria expressão, e não há uma opção para testar a expressão, introduzindo strings que você deseja validar. Se o RegexSwap não satisfazer as suas necessidades, uma vasta número de expressões regulares estão disponíveis em sites da internet de outros também. O site http://regexlib.com, Por exemplo, contém regexes por telefone nos EUA números e códigos postais. Outro site muito interessante, sobretudo se você quiser aprender a sintaxe de expressão regular, é www.regular-expressions.info. Use o seguimento ing etapas para tentar uma para os números de telefone na nossa base de clientes WCM.

1. Uma expressão que irá coincidir com a maioria dos números, com ou sem extensão parecido com este:

   (((\ D {3} \)) |?. (\ D {3} [- \ .]))? \ d {3} [-] \ d {4} (\ s (x \ d + ?)) {0,1} $

   Clique regex novo na tela principal DataCleaner e criar um novo expressão. Dê o nome de expressão e digite a expressão acima, ou copiá-lo a partir de amostras de WCM no site da companhia.

2. Salvar a expressão (você pode, naturalmente, testá-lo primeiro com um telefone real número) e começar uma nova tarefa. Abra o banco de dados WCM e clique duplo sobre o telephone_number campo na tabela de clientes.

3. Adicionar um perfil matcher Regex, ative a guia matcher Regex e selecione somente o nome da expressão que você criou. DataCleaner, por padrão, seleciona todos os regexes disponíveis por isso, se este é o único disponível é já selecionado. Caso contrário, clique em nenhum Selecione primeiro e depois ativar os EUA Número de telefone. Se o número de telefone é a única coluna nessa tarefa, não há diferença entre a aplicação do jogo, para todos os dados selecionados, mas é melhor selecionar explicitamente a telephone_number coluna como um subconjunto de dados.

Sua tela deve agora olhar como o exemplo na Figura 8-2.

Se você não selecionar o número do telefone como um subconjunto de dados e adicionar mais colunas mais tarde, eles serão automaticamente digitalizados com a mesma regularidade expressão, que não faz muito sentido na maioria dos casos.

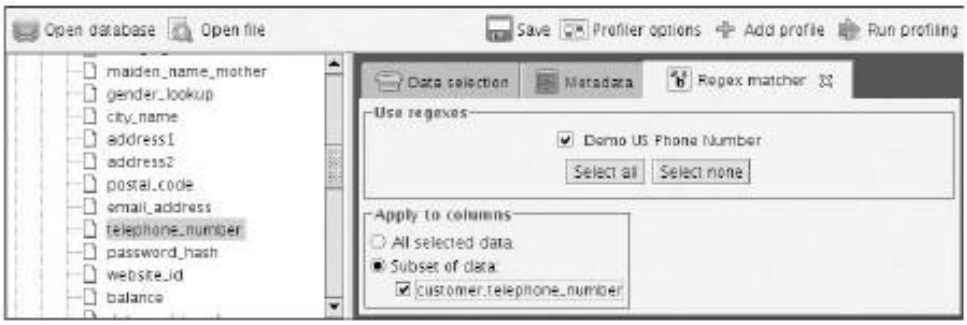


Figura 8-2: Regex matcher definição

A caracterização e exploração de resultados

Agora você está pronto para o perfil dos números de telefone, então clique em Executar de perfil para
iniciar o processo. DataCleaner irá exibir uma tela de status, onde você também pode acompanhar o andamento do processo perfil. Quando o perfil é concluído, um
guia resultados é adicionado à tela, um para cada tabela que continha perfilado colunas. Abra a guia ao cliente, ea tela de resultado deve ser algo
como na Figura 8-3.

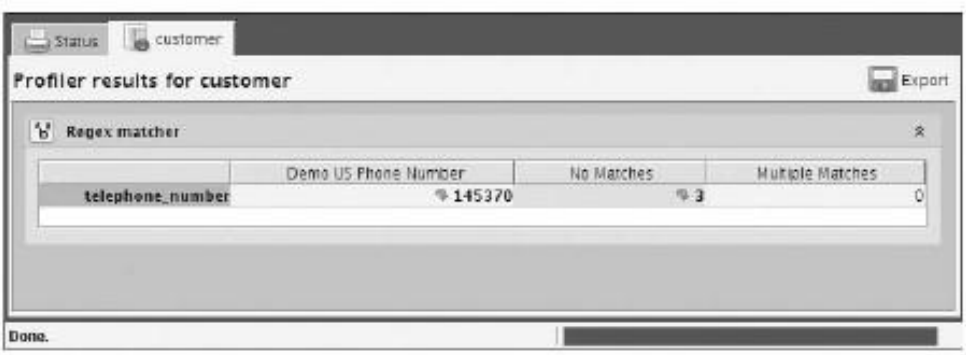


Figura 8-3: resultados Profiler

Nós deliberadamente alterou um pouco os números de telefone aqui para obter algumas exceções
e ser capaz de mostrar outro recurso interessante DataCleaner: perfuração até a detalhes. Clicando na seta verde ao lado do abre três exceções encontradas
a tela mostrada na Figura 8-4.

Figura 8-4: Exceções

Este não é o fim, porém. Quando você clica direita, você verá dois de exportação opções: uma para as células seleccionadas e uma para a tabela inteira. A última opção

também irá adicionar o cabeçalho da coluna para a área de transferência, o primeiro de cópias apenas o
dados selecionados. As células selecionadas não precisam ser adjacentes. Ao usar a tecla Ctrl
Você pode, por exemplo, selecionar o código do cliente e número de telefone e apenas copiar as colunas para a área. Depois disso, você pode facilmente colar os dados em uma planilha ou outro arquivo para posterior exploração.

Validação e comparação de dados

Validação funciona de forma semelhante como a tarefa de perfil, mas acrescenta algumas capacidades. Você pode verificar os valores nulos ou fazer uma verificação do intervalo de valor para descobrir
se as entradas de coluna em uma queda entre um valor inferior e superior. A característica mais avançada é a avaliação de JavaScript, que permite usar qualquer JavaScript expressão para avaliar os dados. A diferença é a saída: o val-
tarefa idation exibirá apenas as entradas que não passarem nos testes com uma contagem dos registros. O roteiro inclui DataCleaner futuros planos para integrar o perfil e validação de tarefas e oferecer uma única interface integrada para ambas as tarefas.

Na comparação dos dados permite a comparação de bases de dados diferentes ou esquemas, ou comparar os dados com um arquivo. Portanto, essa tarefa pode ser usado para verificar se todos os clientes no ordens tabela também existem no cliente tabela
e as funções idênticas comparação.

Usando um dicionário para verificações de dependência Coluna

DataCleaner não fornece uma solução out-of-the-box para verificar a combi-
nações de colunas, ou se uma coluna dependente contém um inválido entrada com base nas informações em outra coluna. Há, no entanto, uma forma para fazer essas análises usando um dicionário, combinadas com vistas banco de dados. A DataCleaner dicionário é um arquivo de texto contendo os valores que podem ser usados para
validar os dados em uma tabela de banco de dados. Por exemplo, você pode baixar a ISO mesa país, armazenar os valores em um arquivo de texto, e usar este arquivo de texto como um catálogo
para verificar as entradas em uma coluna país. Se você tomar um passo adiante, é também possível armazenar vários campos concatenados por linha e criar uma visão no banco de dados, que concatena as colunas a serem validados no mesmo caminho. Agora, a visão pode ser perfilado usando o dicionário com a concatenadas entradas, cada linha que não corresponde com os valores corretos no arquivo de texto será reconhecido pelo DataCleaner. Como uma alternativa ao uso de arquivos de texto, também é
possível utilizar uma"verdadeira"banco de dados do dicionário. Esta base de dados do dicionário necessidades
a ser adicionada ao arquivo de configuração DataCleaner como explicado no"Adicionar Conexões de banco de dados"seção.

Soluções Alternativas

Muito poucas alternativas open source existem para autônomos ou de dados incorporados profiling. A ferramenta de modelagem de dados de SQLPower, que apresentamos em breve,

tem alguns recursos básicos de perfis, e oferece uma Talend Profiler dados também.
Se qualquer uma dessas ferramentas de trabalho para você, basta usá-los. Outra
frequentemente utilizados
alternativa para criação de perfis de dados é criar scripts personalizados para perfis de dados
SQL
finalidades. Recomendamos que isso só se você tiver muito especializadas
requisitos que não são fornecidos fora da caixa por DataCleaner. Embora
está fora do escopo deste livro, é possível estender a DataCleaner
funcionalidade com suas tarefas de perfil do próprio cliente, que lhe dá uma
solução mais rápida, mais confiável e mais flexível do que totalmente a partir de
zero.

Desenvolvimento do Modelo

Depois que as exigências são claras, as fontes de dados corretos foram identifi-
cados, e os dados de perfis processo tiver fornecido suficientemente detalhada
informações sobre o conteúdo ea qualidade dos dados de origem, você pode começar por
desenvolvimento global de data mart modelo de dados. Antes de criar o real
modelo de dados detalhados em uma ferramenta de modelagem de dados, primeiro você
precisa identificar quais
temas e medidas que você irá criar na sua data mart inicial. É aí que
os requisitos criado anteriormente entram em jogo. Uma exigência que cobre
"Cliente do produto de análise de vendas"claramente precisa de ter os assuntos cliente,
produto,e tempo adicionado com as medidas receita de vendas e número de
itens. Sem descrever estas entidades em detalhes, é possível já a desenhar
o modelo de alto nível, como mostrado na Figura 8-5.

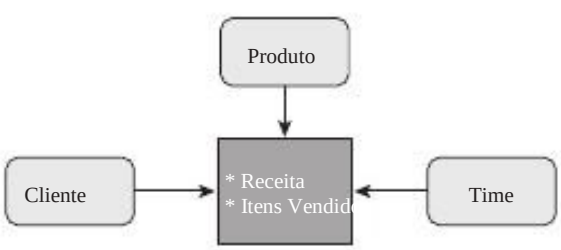


Figura 8-5: Alto nível do modelo estrela

Este é um ponto de partida perfeito para especificar melhor o conteúdo exato da sua
tabelas de dimensão e de fato, pois este modelo é fácil de entender, mesmo para um
usuário de negócios não-técnicos. O próximo passo no refinamento do modelo é a
determinação
quais atributos devem fazer parte de cada dimensão. Ela ajuda a diferenciar
entre os atributos de análise e de detalhe. Um atributo é um campo de análise que
será usada para informar ou agrupar os dados sobre como um grupo de clientes, cidade,
gênero,
e mês. Detalhe atributos são em sua maioria de elementos descritivos, tais como o cliente
nomes, números de telefone e endereço. Você pode até considerar
deixando os detalhes atributos fora do armazém de dados completo. Isso não é

sempre uma boa idéia, no entanto. Suponha que você queira gerar listas de discussão directamente a partir do armazém de dados com base em algumas análises, ou processo de seleção;
você vai precisar das informações de endereço, números de telefone, e-mail endereços.

DICA Nos quadros de pequenas dimensões, a distinção entre detalhe e analítica atributos é desprezível, mas quando a tabela de dimensão fica muito grande, isso distinção auxilia na determinação de quais atributos podem ser movidos para uma mini-dimensão.

Para cada atributo, você precisa definir que tipo de dados e comprimento, estar. Também é uma boa idéia para adicionar uma descrição de cada campo, o que pode parecer
complicado no início, mas este metadados adicionados é muito útil tanto para final usuários e pessoas que serão a manutenção do sistema. Tabela 8-1 contém uma exemplo parcial de uma dimensão de cliente. Uma coluna muito importante é a com o título SCD (por tipo de Dimensão de Alteração Lenta). Isso determina como as atualizações para essas colunas serão tratadas pelo processo de ETL.

Tabela 8-1: dimensão Cliente

CAMPO	ANÁLISE	TIPO	DISTÂNCIA	SCD	DESCRIÇÃO
Customer_key	N	INT	4		Substituto chave de dimensão
Customer_id	N	INT	4		Fonte original chave do sistema
Customer_name	N	VARCHAR	63	2	Nome completo (primeiro + média +passado)
Customer_city	Y	VARCHAR	64	2	Nome da cidade
Customer_phone _number	N	CHAR	12	2	Telefone número
Customer_register _date_key	Y	INT	4	1	Primeiro registo data de cliente

Mais adiante neste capítulo, vamos estender essa tabela com origem e transformação informações, bem como, inicialmente, mas para discutir e documentar a data mart mesas, o arranjo mostrado aqui funciona bem.

Uma observação final sobre comprimentos de campo: Pode acontecer que o sistema de sua fonte
esquema diz que o campo de nome da cidade é de 64 caracteres de largura. Quando você perfil os dados, você encontrar alguma coisa, ou seja, que o nome é a maior cidade apenas 30 caracteres (o que é verdadeiro para o banco de dados exemplo WCM: STE Marthe LA MADELEINE DU CAP). Por favor, não caia na armadilha de cortar a cidade

campo Nome na tabela de dimensão de 30 caracteres, nesse caso. O mesmo se aplica a outros campos descritivos, tais como descrição do trabalho, título do DVD, e assim em. Mesmo que as entradas atuais são todos menores que o comprimento especificado não significa que em algum momento alguém não vai entrar um valor que excede esse número. Isso pode levar tanto a uma entrada incompleta ou uma falha no ETL processo, que precisa ser tratada. Tome sempre o comprimento da fonte campos da tabela, no mínimo, para a meta de comprimentos de campo da tabela.

Especial atenção deve ser dada à tabela de fatos e, especialmente, o granularidade (nível de detalhe) dos fatos. Os usuários corporativos deveriam acordar sobre o significado de um único registro. Quando olhamos para os fatos ordem em nosso exemplo, a linha de ordem é o menor nível de detalhe e, assim, a granularidade da o modelo.

Modelagem de dados com Power * Architect

Até agora temos analisado os dados e trabalhou com aplicações de escritório para documentar
nosso projeto. Agora é hora de traduzir essa informação em um modelo de dados reais em um banco de dados. Porque Pentaho BI é um conjunto e não uma ferramenta de modelagem de dados,
irá usar uma solução de terceiros aqui também. Há alguns dados de código aberto ferramentas de modelagem disponíveis, mas a maioria deles têm capacidades limitadas. Às vezes
a limitação é em termos de funcionalidade, por exemplo, você não pode fazer engenharia reversa
bases de dados existentes e, por vezes, a limitação das bases de dados é a ferramenta pode trabalhar com ele. Por estas razões, nós não vamos usar o MySQL Workbench. Ele pode só podem ser utilizadas em conjunto com o MySQL ea versão de código aberto é um
de [illegible] edição do comercial.

> **DICA** Assumimos uma compreensão básica dos conceitos de modelagem de dados e teoria. Se este não é o caso, o livro *Dominando Data Warehouse Design* por Claudia Imhoff et al. (Wiley, 2003) tem uma excelente introdução sobre o assunto e também fornece um mergulho mais profundo em alguns dos conceitos de modelagem que introduzimos neste e nos capítulo anterior.

Ao olhar para uma ferramenta de modelagem de dados, certifique-se que pelo menos os seguintes
funcionalidade está disponível:

- Multi-banco de dados de suporte Talvez você deseja fazer engenharia reversa de uma exis-
  ção de banco de dados Oracle e gerar um banco de dados MySQL a partir de uma adaptação
  modelo. Mais tarde você pode decidir mudar para o PostgreSQL, mas se a ferramenta não suporta isso, você tem um problema.
- Encaminhar engenharia Gerar novos objetos de banco de dados a partir de um modelo.
- A engenharia reversa Gerar um modelo de um banco de dados existente.

- Modelo comparação Compare o modelo com o esquema de banco de dados (ou vice-versa) e gerar scripts de modificação.

As ferramentas mais avançadas no mercado também oferecem recursos como versão gestão e versão de comparação, a divisão entre as empresas, lógica e modelos físicos de banco de dados, recolher e expandir elementos gráficos, e características documentação. Os preços para ferramentas comerciais com um ou mais dos esses recursos avançados variam de US $ 200 a US $ 5.000, mas se você está projetando um sistema de banco de dados de missão crítica para uma grande empresa, esses custos não podem ser um problema. Nós estaremos usando o código-fonte aberto Power * Architect (P * A) ferramenta de SQLPower (www.sqlpower.ca), Porque é um dos poucos ou talvez mesmo a ferramenta de código aberto apenas disponível que suporta as quatro funções na lista anterior. Power * Architect está disponível para Linux, Windows e Mac e pode ser baixado do site SQLPower. A versão para Windows é um instalador executável, tal como qualquer outro programa do Windows, definindo o programa em Linux leva um pouco mais de esforço (mas não muito). Download o tar.gz arquivo e extraí-lo para um diretório de sua escolha. O programa é uma aplicação Java arquivo (Architect.jar), Que pode ser iniciado com o comando java-jar architect.jar na linha de comando dentro da pasta Architect. Você também pode criar facilmente um lançador ou um novo item de menu para o programa.

Podemos acrescentar muito pouco ao Poder excelente guia de usuário * Arquiteto desde por SQLPower. Como você vai descobrir muito rapidamente, a ferramenta inclui ainda um dados de perfis de opção. Para uma visão básica sobre os dados, essa funcionalidade é suficiente mas não existem validadores regex ou domínio, como em DataCleaner. Um outro ponto de cuidado aqui: O profiler pode trabalhar em uma tabela completa apenas, e não em colunas individuais. Perfil de tabelas muito grandes podem, portanto, ter uma longa tempo para completar.

Mais prós e contras de ARQUITETO POWER *

Alguns recursos que você pode encontrar em outras ferramentas, mas não em P * A é uma seleção de opções de exibição que permitem exibir somente o nome da tabela, ou somente a chave primária campos, ou todos os campos. E embora haja um zoom in e out facilidade, isso funciona para todo o modelo, não para uma área selecionada. Mas tudo isso são questões de menor importância e não nos impediu de usar e recomendar esta ferramenta. Na verdade, há uma Casal de características únicas não disponíveis em qualquer outra ferramenta, como a ETL e opções OLAP. Um gadget muito bacana (que pode realmente provar ser muito útil) é a possibilidade de copiar as tabelas de um banco para outro. Imagine você tem um banco de dados Oracle e querem alguns quadros copiados para um MySQL instância. Basta criar uma conexão com bancos de dados e utilize a tabela de cópia dados de opção no menu Ferramentas para selecionar uma tabela de um banco de dados de origem, clique em um esquema de destino e clique em OK.

Construindo o Data Marts WCM

Depois de ter consultado o Poder guia do usuário * Arquiteto e fez-se familiarizado com a ferramenta, é hora de começar a construir data marts WCM. Você precisa de duas conexões para fazer isso, uma ao banco de dados WCM e outro para WCM_DWH um novo banco de dados, que servirá o armazém de dados.

Em muitos casos, o data warehouse irá puxar os dados de mais de uma fonte do sistema. Às vezes você não vai mesmo ter acesso direto a estas bases de dados, ou pelo menos não quando você está projetando a data marts. Isso torna difícil para construir o modelo baseado em um ou mais sistemas de origem. Se você pode fazer e usar conexões diretamente aos bancos de dados fonte, existem algumas vantagens quando trabalhar com P * A. A ferramenta""lembra os mapeamentos que você faz quando arrastar uma coluna a partir de uma conexão de fonte para o painel do projeto. Você também pode
ver isso no editor de propriedades da coluna, onde as duas primeiras linhas logo acima do nome da coluna exibir o nome da fonte bases de dados e nome da fonte coluna da tabela. Uma coluna que é adicionada à mão irá exibir o texto''Nenhum ''Especificado para o banco de dados e nome da coluna.

DICA Para abrir rapidamente as propriedades de uma tabela ou coluna, clique na tabela ou nome da coluna e pressione Enter.

Você provavelmente já viu o Criar Chaleira Job opção no menu de ETL e questionou se esta é uma forma rápida de construir postos de trabalho de transformação. A resposta é: depende. Se todas as colunas são mapeadas a partir de uma fonte banco de dados, você pode considerar a geração de empregos como uma chaleira de início rápido para o seu
transformações. Em todos os outros casos, essa opção é inútil, porque um trabalho só pode ser criada se houver um mapeamento real.

DICA Com Power * Architect é fácil migrar parte ou um banco de dados inteiro do uma plataforma para outra usando a opção Criar Chaleira Trabalho. O trabalho pode ser Chaleira criado como uma coleção de arquivos (um arquivo de trabalho de mais um arquivo de transformação para cada
tabela) ou diretamente na Chaleira repositório. Uma solução ainda mais rápido é usar a cópia Quadros Assistente em Pentaho Data Integration (Spoon) diretamente.

O diagrama na Figura 8-6 exibe os pedidos do esquema de data mart, que é provavelmente a primeira vez que será criado. Usando esta data mart, uma multidão de questões comerciais podem ser respondidas:

- Quais gêneros de filme ou sites geram mais receita?
- Como eficazes são as promoções que o lançamento?
- Quem são nossos clientes mais rentáveis?
- Como é a nossa receita evoluindo ao longo do tempo?
- Qual é a relação entre os nossos sites e grupos demográficos?
- Em que hora do dia em que os clientes, o lugar mais pedidos?

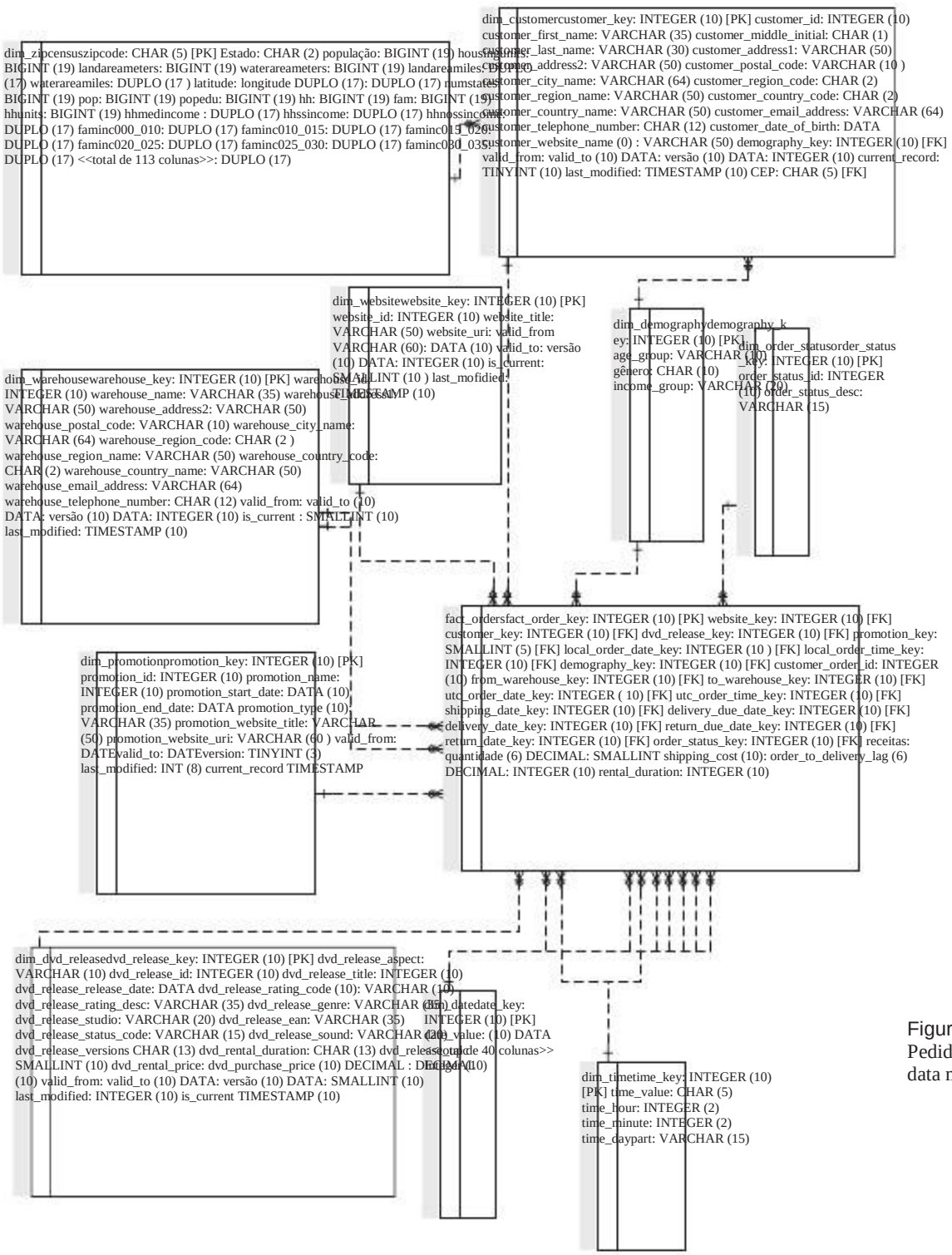


Figura 8-6: Pedidos de data mart

O diagrama contém a conclusão das encomendas data mart, mas, para maior clareza, nós abreviou o data dimensão, que é abordado no parágrafo seguinte. A dim_zipcensus tabela também está incluída em uma forma abreviada. Primeiro, vamos ter um olhar para o diagrama e explicar algumas das escolhas que fizemos.

- Fato tabela para evitar duplicações, nós explicitamente adicionado um banco de dados gerado
  fact_order_key como uma chave primária para a tabela. Há também um casal de entradas calculado: order_to_delivery_lag e rental_duration. Estes cálculos não estão definidos no banco de dados, mas fazem parte da ETL processo.
- Demografia dimensão Este é uma tabela de dimensões regulares, mas nós necessidade de os dados do cliente durante o processo de carregamento para determinar a
  valor correto. Por favor, note que existem dois caminhos possíveis de junção entre fact_orders e dim_geography: uma direta e uma via dim_customer.
  No modelo do usuário final, precisamos criar um alias extra para dim_geography para contornar isso.
- Região / país Baseada na nossa discussão no capítulo 7 sobre starflakes poderíamos ter snowflaked região e país à medida que ocorrem em ambos os dim_warehouse e dim_customer mas nós decidimos fazer o esquema em estrela como"pura"possível.
- Para colunas de Auditoria clareza que deixou as colunas de auditoria batch_insert, batch_update (Que em lotes inseridos ou atualizados o registro?), E dt_insert e dt_update (Quando o registro foi inserido e atualizado?).
- colunas SCD2-A tabelas de dimensão para o tempo, o status do pedido, data e demografia não contêm os campos de gestão de mudança lenta tipo 2 dimensões, porque estes quadros são considerados estáticos ou totalmente do tipo 1 (substituir).

DICA tabelas Power * Architect pode ser dada a sua própria cor na barra de título. Você pode usar este recurso para diferenciar as tabelas de dimensão e de fatos e fazer mesas especiais, tais como a dimensão da demografia e da mini-Zipcensus tabela outrigger facilmente reconhecível.

Gerando o banco de dados

Com P * A, a criação do modelo físico de dados é muito simples. Primeiro você precisa para se certificar de que o catálogo de destino ou o esquema existe e adicionar uma conexão para
neste catálogo. Com a opção do menu Ferramentas Forward Engineer, P * A pode gerar o SQL para a criação de tabelas e índices em vários dialetos SQL. Clique OK, para que a ferramenta gera o script e exibi-lo em um painel de visualização. Agora você tem duas opções: executar o script diretamente usando a conexão selecionada informação, ou copiar o conteúdo do painel para a área de transferência para uso em um ferramenta externa. Se você quiser modificar o script antes da execução, terá

usar a opção de copiar, porque não é possível alterar o script no
painel de visualização.

Dimensões gerar estática

Nosso projeto inicial mart de dados contém duas tabelas estáticas que não recebem os seus dados
dos sistemas de origem, mas será gerada usando scripts. A primeira é a
dimensão temporal, que contém uma linha para cada minuto em um dia (1440 registros).

A dimensão de data é especial e pode ser configurado para acomodar-resposta
eliminação de muitas questões relacionadas ao tempo, como você viu no Capítulo 7. Embora não
há regras rígidas para as colunas que devem estar presentes em uma dimensão de data,
Tabela 8-2 mostra a estrutura de uma dimensão de data que lhe dará uma grande
de flexibilidade quando se trabalha com questões relacionadas ao tempo. A dimensão de data
apresentado tem uma granularidade de um dia, por isso a cada dia no calendário traduz
em um único registro. Observe também que a tabela não é realmente estática como a data relativa
colunas precisam ser atualizados toda noite. Então é verdade que a geração da tabela
é um processo de um tempo, um outro processo é necessário para manter a tabela atualizada.

Você deve ter notado que o título contém a tabela''norte-PT''adição.
Nomes e abreviações para dias, meses e trimestres pode variar para cada
linguagem, e você precisa encontrar uma maneira de acomodar este em sua solução. Não
duas alternativas para implementar isso. A primeira é a de adicionar campos adicionais
para cada idioma na tabela de mesma data, eo segundo é a criação de separar
tabelas que contêm apenas as colunas descritivo para cada idioma. Nós
preferem a última vez que oferece mais flexibilidade. O parente colunas de data
só precisam ser armazenados e mantidos, uma vez que permanecem os mesmos
independentemente da linguagem utilizada.

Não é só uma questão de nomes diferentes em linguagens diferentes, alguns
partes da data pode variar também. O primeiro dia da semana é diferente
em países europeus (segunda-feira) do que nos Estados Unidos (domingo), e
a maioria dos países europeus aderem ao padrão ISO 8601 para a numeração
semana e ano. Isto significa que um número de semanas em que os EUA poderiam diferir
na mesma semana no Reino Unido Por esta razão, foi incluído o _ISO colunas
para a semana, ano, e na semana do ano. É importante quando você está construindo uma base de dados
armazém para uma organização multinacional de ter estas diferenças em
conta e tentar acordar um modo padrão de usar estes números. Outras
colunas que você pode precisar em sua própria situação são indicadores período fiscal.
Nem todas as organizações usam o ano civil como o seu ano fiscal, e quando esta é
o caso, você deve incluir o ano fiscal, o mês fiscal, e em seu trimestre fiscal
dimensão de data também. A lista de entradas possíveis é quase ilimitado; nós
indicadores de visita para o dia de trabalho e número de dias úteis no mês e
semana, os indicadores para as férias com referências a tabelas que contêm descrições
de viagens internacionais e religiosas, e assim por diante. Nós limitamos nossa concepção de
o padrão de campos de calendário e as respectivas colunas de tempo relativo, mas
sinta-se livre para experimentar outros conceitos em seu próprio data mart.

Tabela 8-2: Data dimensão (EUA-PT)

CAMPO	TIPO	DISTÂNCIA	DESCRIÇÃO	EXEMPLO
date_key	INT	4	dimensão Substituto chave	20091123
date_value	DATA	4	Data-valor do dia	23-11-2009
date_julian	INT	4	Arredondado data do calendário juliano	2455159
date_short	CHAR	12	valor de texto curto para a data	11/23/09
date_medium	CHAR	16	Valor médio de texto para data	23 de novembro de 2009
date_long	CHAR	24	Long valor de texto para a data	23 de novembro de 2009
date_full	CHAR	32	texto completo o valor para a data	Segunda-feira, 23 de novembro de 2009
day_in_week	TINYINT	1	Número de dias na semana	2
day_in_month	TINYINT	1	Número de dias em Mês	23
day_in_year	SMALLINT	2	Número de dias no ano	327
is_first_day_in _month	TINYINT	1	1 para o primeiro dia, 0 para outros	0
is_first_day_in _week	TINYINT	1	1 para o primeiro dia, 0 para outros	0
is_last_day_in _month	TINYINT	1	1 para o último dia, 0 para outros	0
is_last_day_in _week	TINYINT	1	1 para o último dia, 0 para outros	0
day_name	CHAR	12	Nome completo do dia	Segunda-feira
day_abbreviation	CHAR	3	Nome abreviado do dia	Seg.
week_in_year	TINYINT	1	Número da semana	47
week_in_month	TINYINT	1	Semana número no mês	3
week_in_year _ISO	TINYINT	1	Número da semana	47
is_weekend	TINYINT	1	1 para o Sat-Sun	0
is_weekday	TINYINT	1	1 para Mon-Fri	1

CAMPO	TIPO	DISTÂNCIA	DESCRIÇÃO	EXEMPLO
month_number	TINYINT	2	Número de meses em ano	11
MONTH_NAME	CHAR	12	Nome completo do mês	Novembro
Mês _abbreviation	CHAR	3	Nome abreviado do mês	Novembro
ano2	CHAR	2	indicador do ano Curta	09
ano4	CHAR	4	Long indicador do ano	2009
year2_iso	CHAR	2	Curta indicador do ano ISO	09
year4_iso	CHAR	4	Long indicador do ano ISO	2009
quarter_number	TINYINT	1	Número de quarto	4
quarter_name	CHAR	2	Texto valor do trimestre	Q4
year_quarter	CHAR	7	Valor Ano trimestre	2009-Q4
YEAR_MONTH	CHAR	7	valor de Ano-mês	2009-11
year_week	CHAR	7	semana do ano o valor	2009-47
year_week_iso	CHAR	7	ISO valor de semana do ano	2009-47
current_week_cy	TINYINT	1	*	1
current_month_cy	TINYINT	1	*	1
last_week_cy	TINYINT	1	*	0
last_month_cy	TINYINT	1	*	0
current_week_ly	TINYINT	1	*	1
current_month_ly	TINYINT	1	*	1
last_week_ly	TINYINT	1	*	0
last_month_ly	TINYINT	1	*	0
ytd_cy_day	TINYINT	1	*	0
ytd_cy_week	TINYINT	1	*	0
ytd_cy_month	TINYINT	1	*	0
ytd_ly_day	TINYINT	1	*	0
ytd_ly_week	TINYINT	1	*	0
ytd_ly_month	TINYINT	1	*	0

Continuou

Tabela 8-2 (Continuação)

CAMPO	TIPO	DISTÂNCIA	DESCRIÇÃO	EXEMPLO
current_year	TINYINT	1	*	1
last_year	TINYINT	1	*	0
week_sequence	INT	4	0 para a semana em curso, -1 para a anterior, uma para a próxima, e assim por diante	0
month_sequence	INT	4	0 para o mês atual, -1 para a anterior, uma para a próxima, e assim por diante	0

* Consulte a seção"Special Data Fields e Cálculos"mais adiante neste capítulo.

Uma última observação sobre a geração das tabelas de data e hora: Pode ser tentador incluir a definição destas tabelas diretamente em um Pentaho Data Integration SQL transformação, mas então você terá que criar um código de banco de dados específico para gerar a tabela. Ao utilizar uma ferramenta de modelagem de banco de dados dados independentes para isso, manter o projeto genérico, o que torna mais fácil alternar para outro banco de dados quando necessário.

Especial campos de data e Cálculos

Capítulo 7 já introduziu a current_week e last_week colunas. Em Tabela 8-2 e nossa amostra de data warehouse, estendemos esta construção com campos adicionais que são grandes para comparar diferentes períodos de tempo. Estes campos são todos do tipo TINYINT e só pode ter valores 0 ou 1, o que significa que podemos usá-los em cálculos. As abreviaturas cy e ly são para a corrente ano e no ano passado, ytd é para o ano até à data, ea última parte, dia,semanaOu Mês Indica os períodos cumpridos para comparar.

Neste caso, é segunda-feira da semana número 47, o que significa que todas as datas com um número de 46 semanas ou menos vai ter o valor 1 para ytd_cy_week. A coluna ytd_ly_week conterá o valor 1 para todos os dias do ano anterior (Neste exemplo, 2008) que têm um número de 46 semanas ou menos. Para relatar receita deste ano até a última semana em relação ao mesmo período do ano passado pode ser realizada facilmente com a seguinte consulta:

```
p.product_type SELECT
, SUM (d.ytd_cy_week f.revenue *) AS ytd_current_year
, SUM (d.ytd_ly_week f.revenue *) AS ytd_last_year
, ROUND (SUM (f.revenue d.ytd_cy_week *) /
      SUM (f.revenue * d.ytd_ly_week) * 100,0) como índice de
FROMfact_sales AS f
INNER JOIN dim_date SOBRE AS d f.order_date_key d.date_key =
INNER JOIN dim_product SOBRE AS p = f.product_key p.product_key
```

O que acontece na consulta anterior é que para cada linha na tabela fato, o valor da receita é multiplicado pelos valores dos campos ytd_cy (semana) e ytd_ly (semana). Em seguida, os resultados desses cálculos estão resumidos e agrupadas por tipo de produto. Tabelas 8-3 e 8-4 mostram um exemplo de como isso realmente funciona.

Tabela 8-3: Relativo cálculos de data

PRODUCT_TYPE	YTD_CY	RECEITAS YTD_LY	O VALOR YTD_CY	YTD_LY
Blu-ray	30	10	300		
DVD	40	1	0	40	0
DVD	25	0	1	0	25
Blu-ray	33	0	0	0	0
Blu-ray	21	1	0	21	0
DVD	56	0	0	0	0
Blu-ray	45	0	1	0	45
DVD	35	0	1	0	35

Tabela 8-4: resultado relativo data definida

PRODUCT_TYPE	YTD_CY	YTD_LY	ÍNDICE
Blu-ray	51	45	113
DVD	40	60	67

A grande vantagem do uso dessas áreas especiais é que você não precisa combinar várias consultas, cada um com seus próprios filtros, para gerar a necessária relatórios. De fato, é possível obter todos os resultados e as comparações de período o armazém de dados em uma única consulta, o que simplifica bastante a construção do consultas e relatórios. Outros elementos comuns em vendas e análise de receitas são movimento totaliza e médias. Para este tipo de cálculo, o total das últimos 12 meses (para um total anual ou em movimento MAT) é comparado com o total dos 12 meses anteriores aos 12 últimos. Para uma comparação mês a mês, você também pode calcular a MAT por mês para eliminar as influências sazonais os cálculos e obter uma melhor visão sobre a evolução das vendas reais. Para calcular um tapete, a seqüência de mês é uma grande ajuda, mas o SQL necessário é um pouco mais complexa, pois você vai precisar de um caso declaração para determinar se um período

devem ser incluídos no resultado final. A instrução a seguir irá calcular
o MAT e MAT - 1 ano:

```
SUM (CASE WHEN * f.revenue d.month_sequence entre -12 e -1
      Então 1 ELSE 0 END) AS mat1,
SUM (CASE WHEN * f.revenue d.month_sequence entre -24 e -13
      Então 1 ELSE 0 END) AS mat2
```

Capítulo 13 fornece vários exemplos de como você pode usar a data
e dimensão de tempo para simplesmente analisar os resultados sem ter que depender de fórmulas complexas no próprio relatório. Na maioria dos casos, é mais fácil para modelar esses
construções diretamente no banco de dados, o que economiza tempo mais tarde, no edifício o modelo de meta ou nos relatórios.

Fonte para Alvo Mapeamento

A etapa último projeto antes que possamos começar a construir os processos reais de ETL é identificar os mapeamentos do sistema de origem (s) para o data warehouse de destino. Uma boa maneira de começar é pelo primeiro desenho os fluxos de dados de alto nível, no nível da tabela
sem se preocupar muito com os detalhes. Figura 8-7 mostra um exemplo de
esse mapeamento de alto nível, onde nós fingimos as tabelas país ea região vêm
de outra fonte (banco de dados mestre). O diagrama mostra apenas uma parte
o armazém de dados completo, é claro, e exibe o mapeamento lógico, não
o físico. O mapeamento físico provavelmente incluirá intermediária
tabelas de teste que são usados para armazenar os dados extraídos dos sistemas de origem ser transformados e carregados em suas tabelas de dimensão destino final.

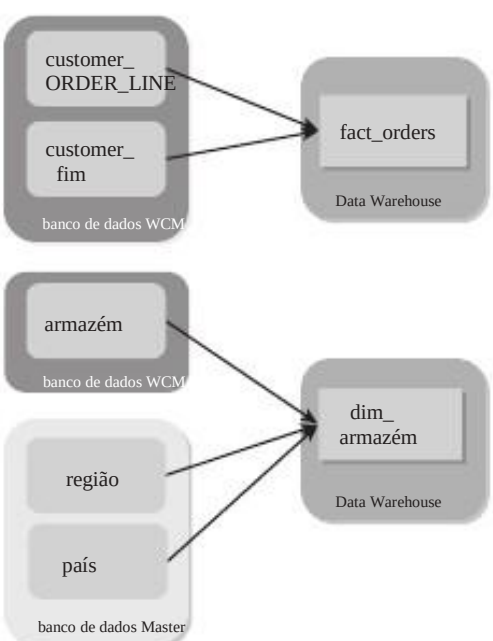


Figura 8-7: Alto nível de mapeamento

Agora que nós determinamos que a nossa data mart olhará como e feito um mapeamento global do sistema de origem para as tabelas de data mart. A próxima etapa é criar o mapeamento detalhado em nível de campo. Este processo não é de fonte para alvo, mas na verdade o contrário. Nós já sabemos que a nossa meta deve ser semelhante, então agora é hora de encontrar as fontes de direito e as transformações para obter os dados. Você trabalha a partir de destino para a fonte porque é possível que haja nenhuma fonte disponível para os campos de determinada meta. Quando você trabalha"para trás"da as tabelas de destino, essas lacunas são capturados mais cedo do que quando se trabalha para a frente"" a partir das fontes. Para cada campo de destino (veja a Figura 8-1 para um exemplo), precisamos identificar os seguintes elementos:

- Sistema-In nosso caso, a maioria dos campos são originários do banco de dados WCM, mas as chaves geradas e colunas de auditoria não tem um sistema de origem. Eles são preenchidos a partir da ETL processo, que pode ser inserido como o sistema de origem para estas colunas.
- Quando esquema aplicável, digite o nome do esquema do banco de dados ou catálogo.
- Quando a Objetos aplicável, digite o nome da tabela, exibição ou armazenados procedimento que serve como fonte para a coluna.
- Campo-A nome do campo no objeto de origem.
- O tipo de dados tipo de dados da coluna.
- Duração-O comprimento da coluna.
- Descreve-Transformação os passos necessários para converter o valor de o sistema de origem para o valor correto da meta. Isso pode (e deve!) incluem pesquisas com outras tabelas também.
- Comentários adicionais- observações que podem ajudar o programador de ETL para construir as transformações de dados correto.

DICA Com DataCleaner, você pode facilmente selecionar as tabelas de origem e cópia da Geração de mesa a partir da guia de metadados em uma planilha. Isto lhe dará o esquema, campo, objeto de dados, tipo e comprimento, que você pode usar para ampliar as informações de mapeamento.

Depois de concluído o mapeamento global e detalhada, você deve verificar a integralidade de seus esforços de modelagem com os usuários de negócios. É não quer magoar a ter uma análise técnica por um DBA, se tiver um ao redor. O mais importante é verificar se o modelo gerado pode ser utilizado para responder às questões de negócios que foram inicialmente apresentados pela comunidade de utilizadores. Se não, você deve identificar as partes que faltavam e voltar à prancheta de desenho para ajustar o modelo e mapeamento. Basta fazer certeza de que o modelo inicial é capaz de lidar com as prioridades do negócio as partes interessadas. Se o seu projeto físico e lógico acaba por ser completa, você está pronto para avançar com o desenvolvimento dos processos de ETL.

Resumo

Este capítulo introduziu-lhe duas peças importantes do software em seu ferramentas: eobjects DataCleaner e SQLPower Dados * Arquiteto. DataCleaner permite analisar os dados de origem e informar sobre as questões de qualidade que possam surgir, e Dados * Arquiteto ajuda na construção do modelo físico de dados para os data marts, e finalmente, o data warehouse. Como tal, este capítulo atua como uma ponte entre os capítulos mais teóricos anteriores e mais capítulos práticos que se seguem. Nós abordamos os seguintes itens para começá-lo em seu caminho com o seu primeiro data mart:

- Os diferentes grupos de usuários que você precisa para entender e trabalhar com
- A análise dos dados e os desafios envolvidos com acesso a diferentes fontes de dados
- O objetivo da caracterização de dados, com uma introdução ao uso de dados profiling ferramenta para avaliar a estrutura de dados, conteúdo e qualidade
- Uma introdução à modelagem de dados com dados * Arquiteto
- O projeto da ordem de data mart fatos que vamos utilizar no restante do livro para a prestação de ETL, análise e exemplos
- Um modelo mais refinado da dimensão de data com o benefício adicional de operações aritméticas em datas avançadas e simplificação do processo de comunicação
- As informações de mapeamento final necessário para criar as transformações ETL

Parte

III

Integração de dados e ETL

Nesta parte

CAPÍTULO

9

Pentaho Data Integração Primer

Quando o projeto do data warehouse é estabilizado, um processo deve ser projetado para preencher o data warehouse com dados. Nós usamos o termo geral integração de dados para descrever o conjunto de atividades que resultem ou contribuir para o preenchimento de o data warehouse. Pentaho oferece uma coleção de ferramentas conhecidas coletivamente como
Pentaho Data Integration que são projetados para suportar essa tarefa.

Este capítulo fornece algumas informações básicas sobre a integração de dados processo de imigração em geral. Nós fornecemos uma visão geral dos dados Pentaho ferramentas de integração, e explicar em detalhe como essas ferramentas podem ser usadas para projetar
e criar uma solução de integração de dados para carregar o data warehouse.

Visão geral de integração de dados

De um modo geral, a palavra"integração"denota um processo que constitui um todo de várias partes. O termo""integração de dados é geralmente entendida como a processo que combina dados de diferentes fontes para fornecer uma única compre-preensível vista sobre todos os dados combinados. Um exemplo típico de integração de dados seria uma combinação dos dados de um sistema de inventário do armazém com a de o sistema de entrada de modo a permitir o cumprimento fim de estar diretamente relacionado às mudanças
no inventário. Outro exemplo de integração de dados está se fundindo ao cliente e dados de contacto do gerenciamento de relacionamento com clientes do departamento separado
sistemas corporativos em um sistema de gestão de relacionamento com clientes.

Na introdução a este capítulo, afirmou que compreende a integração de dados as atividades que resultam no preenchimento de um data warehouse. Esta é uma considerável simplificada noção de integração de dados no que diz respeito à sua definição geral, mas

pode ser justificada, ao considerar que um devidamente concebidos data warehouse automaticamente fornece uma visão unificada dos dados de diferentes sistemas.

Nesta seção, vamos explicar algumas das características fundamentais dos dados processo de integração e de seus componentes para fornecer o contexto necessário para os leitores que são completamente novos para o tópico. Uma discussão detalhada sobre integração de dados está além do escopo deste livro. Felizmente, muitos bons livros estão disponíveis, que abrangem este tema, como O ETL do Data Warehouse Toolkit por Ralph Kimball e Caserta Joe (Wiley, 2004).

Atividades de Integração de Dados

Uma forma de compreender a integração de dados é decompor o processo de encher o depósito de dados em uma série de atividades distintas. Nesta subseção Nós descrevemos um número de tipos essenciais de atividades de integração de dados.

Em um nível muito alto, o problema do preenchimento de um armazém de dados consiste em
apenas três grandes tipos de actividades:

- Extração de aquisição dados de um ou mais sistemas de origem. Para exemplo, a obtenção e carregar todos os registros de clientes que foram adicionados ou alterados desde o último carregamento de dados.

- Transformação Mudar a forma e / ou conteúdo dos dados para adequá-lo na estrutura do data warehouse de destino. Por exemplo, olhando para cima estado e do país nomes para valores de chave.

- Carregando -Na verdade, o armazenamento de dados no data warehouse de destino.

Essas três atividades-Extração, Transformação e Carga são freqüentemente referidas pela sigla ETL.

Em alguns casos, o ETL termo é entendido literalmente, e levado para significa que as atividades individuais de extração, transformação e carregamento são realmente executados nessa seqüência. A partir desta escola de pensamento, o termos relacionados ELT (Extração, carga, transformar) e ETLT (Extração, transformação, carga,
transformação) foram introduzidos para fazer justiça ao fato de que os dados transformando-atividades podem ser aplicadas por qualquer um RDBMS (ELT), ou por um técnico especializado
ferramenta fora do RDBMS (ETL), ou ambos (ETLT).

Embora possa haver alguns contextos onde ela é vantajoso usar esses diferentes termos, a distinção não é explorado mais adiante neste livro. Em vez disso, preferem usar apenas o termo de ETL, sem limitar de antemão o que componentes de software são ocupadas com a transformação dos dados. Como você deve ver, Pentaho Data Integration é perfeitamente capaz de transferir certas tarefas o RDBMS, e não faz nada para impedir que o RDBMS de participar de dados transformação. Portanto, nós tratamos de ETL como um termo guarda-chuva que pode implicar
em vez de excluir ELT e ETLT.

ETL é apenas uma classificação muito ampla de atividades de integração de dados. Dentro cada um dos principais processos de Extração, Transformação e Carga, podemos identificar uma série de atividades de apoio. Alguns destes estão listados abaixo:

Para a extração podemos discernir:

- Change Data Capture-In extração muitos casos, é limitada à parte de fonte de dados que mudou desde a última extracção. O processo de identificar os dados alterados é chamado Change Data Capture.
- preparação de dados -Nem sempre é possível ou eficiente imediatamente transformar os dados extraídos. Muitas vezes, o extrato é armazenada temporariamente até que entra no processo de transformação. Isso é chamado dados de teste.

A transformação é um processo amplo e variado. Não é possível fornecer uma lista exaustiva das actividades descritas, mas algumas atividades típicas são:

- Os dados de validação de dados validação é o processo de verificar a fonte dados estão corretos, e possível filtrar dados inválidos.
- limpeza de dados -Data limpeza é o processo de correção de dados inválidos.
- Decodificação e renomear -Em muitos casos, os dados brutos do sistema de origem não é adequado para fins de notificação, pois contém os nomes obscuros e códigos. Uma grande parte do processo de transformação está ocupada com convertendo isso para um nome mais descritivo e user-friendly e etiquetas.
- Agregação-geral, aplicações de BI apresentar dados agregados para o utilizadores finais. Às vezes, os agregados são calculadas de forma antecipada como parte da processo de transformação.
- A geração ea gestão - Novas linhas de dimensão são excepcionalmente identificados por chaves substitutas, que devem ser gerados. Para armazenar dados de fatos novos, estas chaves devem ser procurados.

No processo de carregamento, podemos discernir duas atividades principais:

- Carregando as tabelas de fatos-geral, tabelas de fato crescer pela adição de novas linhas. Às vezes, as linhas existentes são atualizados para refletir um novo status.
- Carregamento e manutenção de tabelas de dimensão-Nova linhas de fatos podem dar origem a linhas de dimensão.

Vamos olhar para essas atividades em mais detalhes nas subseções a seguir.

NOTA Esta não é uma lista exaustiva, embora estas atividades são bastante típicos.

Você pode encontrar muito uma visão completa das atividades de ETL no artigo Ralph Kimball "Os 38 subsistemas de"ETL na http://www.intelligententerprise.com/showArticle.jhtml? articleId = 54200319.

Extração

A primeira etapa em todo o processo de integração de dados é a aquisição de dados. Este processo é geralmente chamado extração.

Muitas vezes, a fonte de dados é um sistema de banco de dados relacional que faz parte de trás
fim de algum aplicativo operacional. Nestes casos, pode ser possível
acesso à fonte de dados diretamente através de uma conexão de dados. Neste caso, extração pode ser uma tarefa relativamente fácil. No entanto, nem sempre pode ser viável para ter acesso direto ao sistema de banco de dados back-end.

Pode haver políticas em vigor que proíbem o acesso de banco de dados além dos aplicação operacional. Os sistemas operacionais são cada vez mais esperadas ou obrigado a estar sempre disponível, e o impacto de um processo de aquisição de dados pode simplesmente ser incompatível com estes requisitos. Nestes casos, os dados pode ter de ser extraídos dos dados do sistema de backup ou arquivos de log do banco de dados.

Outro fator que pode complicar a extração é o fato de que, muitas vezes, não todos os dados de origem podem ser adquiridos no mesmo local. Pode haver múltiplas sistemas operacionais em uso a partir do qual os dados devem ser adquiridos. De fato, o obrigação de comunicar através de múltiplos sistemas de informação é muitas vezes a condução
força por trás de BI e data warehousing projetos.

Change Data Capture

Change Data Capture (CDC) é o processo de rastreamento de alterações de dados na fonte sistemas, a fim de atualizar o armazém de dados em conformidade. Por um lado, CDC é um processo que conduz a extração, pois os dados precisam ser adquiridos somente até o ponto em que faz a diferença em relação ao estado actual da
o data warehouse. Por outro lado, o CDC tem também funcional e lógica
facetas, pois determina em que medida o armazém de dados é capaz de registrando a história do negócio.

Há uma série de métodos para implementar praticamente CDC:

- Gravação de dados"natural"caso dos sistemas operacionais. Em muitos casos, os principais eventos são registrados como parte do processo de negócios operacionais. Para
  exemplo, o registro do cliente, colocação de pedidos e expedição de ordem são normalmente registadas a nível da data no sistema operacional.
- Usando dados sequenciais chave nos sistemas de origem.
- Journaling com banco de dados gatilhos. Em sistemas de banco de dados de muitos, é possível
  adicionar gatilhos para o banco de dados de esquemas de aplicações operacionais.
- Lendo banco de dados de log.

Data Staging

Extração de dados pode ter um impacto considerável sobre o desempenho e dis-capacidade do sistema operacional de fonte. Muitas vezes, existe uma obrigação estrita de

manter a quantidade de tempo gasto na extração de um mínimo em uma tentativa de diminuir o impacto sobre as operações normais, como entrada de dados. Normalmente, esses limitações de tempo não deixa tempo suficiente para completar o processo todos os dados antes
armazenando-o no armazém de dados.

Para além da duração efectiva do processo de extração, a questão da
tempo também pode entrar em jogo. Em não poucos casos, dados de vários distintos sistemas operacionais podem precisar de ser combinado antes de ser alimentada em dados armazém. Muitas vezes não se pode confiar em todos os sistemas de origem para ser simultaneamente
disponíveis para a extração.

Para lidar com estes problemas, os dados são normalmente armazenados temporariamente em uma chamada
área de teste imediatamente após a extração. Esta abordagem permite a extração atividades a serem realizadas no quadro menor tempo possível, porque o tempo não é gasto à espera de tratamento posterior. Ele também permite a sincronização de processos que combinam dados de fontes distintas chegam em momentos diferentes.

Na maioria dos casos, a área de preparação de dados é simplesmente um banco de dados relacional que é
especificamente concebido para servir como um amortecedor entre os sistemas de origem e os
data warehouse. Isso nos leva a uma outra vantagem de preparação de dados. Porque os dados são armazenados em um sistema de banco de dados distintos, índices que podem ajudar a melhorar
o desempenho do tratamento dos dados podem ser livremente adicionadas sem alterando o sistema de origem.

Validação de dados

Uma vez que os dados são adquiridos (e possivelmente armazenado em uma área de teste), há geralmente
algum processo no local para avaliar a validade dos dados. Dados inválidos devem ser tratados de forma diferente do que os dados válidos, pois pode manchar a confiabilidade dos dados
armazém. Detectando dados inválidos é uma condição indispensável para o tratamento de forma diferente.

No contexto da ETL e validação de dados, os dados são considerados inválidos, quando ela contém erros lógicos. Isso ocorre quando são encontrados registros de origem que jamais poderia ter sido inscrito, se todas as restrições implementadas pelo aplicação de origem (e seu sistema de banco de dados subjacente) havia sido executada. Por exemplo, os dados para os campos obrigatórios podem estar faltando, ou valores em um campo
pode contradizer os valores em outro campo, como quando cai uma data de entrega antes da data da ordem correspondente.

Pode parecer desnecessário para verificar se há erros lógicos quando os dados são adquiridos
de aplicações e sistemas de banco de dados que são conhecidos para fazer cumprir rigorosamente
restrições. No entanto, a realidade é que não há nenhuma maneira de avaliar os dados validade que não sejam realmente verificar. Se a fonte de dados inválidos acidentalmente termina no armazém de dados, pode ser descoberto pelo usuário final. Isso pode levar à desconfiança geral do armazém de dados e apoio à integração de dados os processos.

Fazendo validação de dados uma parte dos dados de resultados do processo de integração benefício imediato. Se não houver dados inválidos for pego, ele oferece a paz de espírito

que o sistema de origem pode ser confiável. Por outro lado, se os dados inválidos são capturados pode solicitar apoio extra para o projeto de integração de dados, pois oferece uma oportunidade única para melhorar o sistema de origem.

De limpeza de dados

Em muitos casos, possíveis problemas com os dados de origem são conhecidos antecipadamente e
processos podem ser criados para ajudar a corrigir os dados que seriam inválidos. Isso é conhecido como limpeza de dados.

A solução mais simples para lidar com dados inválidos para descartá-lo. Embora esta impede que os dados inválidos de manchar a confiabilidade dos dados conhecidos correta isso geralmente não é uma opção aceitável. A melhor solução é manter conhecidos dados inválidos de lado, e, se possível, corrigi-lo.

Manter os dados inválidos, desde que ele está marcado e classificados em conformidade, tem muitas vantagens. No futuro, de alguma forma pode ser encontrada para corrigir ou caso contrário conciliar os dados inválidos para que ele possa ser carregado depois de tudo. Pode ser
vale a pena usar os dados inválidos como meio de prova, na tentativa de convencer qualquer responsáveis para reparar ou melhorar o sistema de origem. Por exemplo, nós
pode obter números de telefone a partir de um sistema de CRM em todos os tipos de formatos diferentes:
alguns podem conter um código de país, enquanto outros podem omitir isso. Alguns podem contêm um código de área, e outros podem omitir isso. Os códigos de país pode ser denotado usando um + outros podem utilizar um 0. Às vezes contêm números de telefone parênteses, traços ou caracteres de espaço para facilitar a leitura. Em muitos desses casos podemos analisar os números de telefone e padronizar a notação. Em alguns casos, podemos fazer um palpite com base no endereço do cliente para o preenchimento de um omitido código do país ou código de área.

Outra possibilidade é carregar os dados inválidos, depois de devidamente marcação dele. Como os dados inválidos esteja marcada como tal, podem ser incluídas ou excluídas da analisa a critério do usuário. Ele também permite que os usuários finais para inspecionar a natureza
dos dados inválidos, e que lhes permite fazer um julgamento informado com
em conta a qualidade dos dados. Essa abordagem também pode atuar como uma alavanca para corrigir os dados
problema de qualidade na origem, porque todos os interessados podem agora ver o impacto dos dados inválidos em seus relatórios por si.

Decodificação e Renomeando

Renomeando e decodificação estão entre as atividades de transformação mais básica. Embora humilde por natureza, estes são talvez os tipos mais ubíquo de transformações.

Decodificando ocorre quando os valores de um campo no sistema de origem são mapeados a outros valores no sistema de destino. Por exemplo, um campo de origem contendo a valores 1e 0pode ser descodificada em valores mais compreensível Sim e Não no sistema de destino.

Renomeando ocorre quando o nome de um determinado campo na fonte é dado um novo nome no sistema alvo. Por exemplo, um campo que é chamado zip na sistema de origem pode acabar como postal_code no sistema de destino.

É importante perceber que, normalmente, nem decodificar nem renomear adicionar qualquer informação no sentido formal. No entanto, essas atividades podem ajudar a tornar os dados mais acessíveis para o usuário final. Isto é particularmente verdadeiro quando o
sistema de origem usa abreviaturas de nomes de campo ou valores de campo.

Key Management

Capítulo 6 explicou que as tabelas do data warehouse não uso natural
chaves primárias. Todas as tabelas de dimensão têm substituto chaves primárias, e de facto tabelas são associadas às tabelas dimensão usando somente referências a estes substitutos chaves. Os valores para essas chaves não devem ser provenientes de sistemas de origem que alimentam o armazém de dados (com a possível exceção da dimensão data tabela). Em vez disso, eles devem ser gerados como parte do processo de integração de dados.

Agregação

Há vários casos em que o processo de integração de dados envolve a agregar o ção de dados.

As tabelas de dimensão podem conter atributos que são obtidos por agregação. Por exemplo, uma tabela de dimensão do cliente podem conter atributos, tais como montante total despendido. Carregando tabelas agregadas com pré-calculados, agregados métricas podem ser necessárias para melhorar o desempenho de determinados relatórios e ROLAP
cubos.

O Pentaho Analysis Server (Mondrian) é um servidor ROLAP que pode levar vantagem de tabelas agregadas. Pentaho fornece também a agregação Pentaho Designer, que é especialmente adaptado para a criação e manutenção de agregados tabelas para essa finalidade. Por esta razão, discutimos agregação no capítulo 15 ao invés de neste capítulo.

Dimensão e Manutenção de Tabelas Ponte

A maioria das tabelas de dimensão não são estáticos. Seu conteúdo necessidade de se adaptar de acordo
para adições e mudanças que ocorrem nos sistemas de origem.

Há exemplos simples, tais como novos produtos que precisam ser adicionados à tabela de dimensão do produto. Exemplos mais complexos incluem manipulação vários tipos de alteração lenta dimensão. Mesmo exemplos mais complexos ocorrer quando a tabela de dimensão necessita de apoio para a navegação ao longo de uma recursiva
relação pai-filho.

Guardar as alterações nas tabelas de dimensão é uma das responsabilidades nobre vínculos do processo de integração de dados.

Carregando Tabelas de fatos

Carregando as tabelas de fato é a atividade mais importante da integração de dados
ção do processo. O processo de integração de dados precisa superar uma série de
desafios ao carregar as tabelas de fatos.

Em muitos casos, a enorme quantidade de dados de fato é em si um desafio. Dimensão
chaves devem ser consultadas para cada linha que será armazenado na tabela de fatos,
eo desempenho é muitas vezes um problema, especialmente quando muitas tabelas de
dimensão
estão envolvidos, que tendem a ser grandes se (como de produtos e clientes
tabelas). Métricas devem ser armazenados corretamente, e às vezes, métricas adicionais
precisam ser calculadas, também. Para complicar ainda mais as coisas, alguns tipos de fato
tabelas exigem linhas existentes para ser atualizado para refletir as mudanças no estatuto
sistemas de origem.

Pentaho Data Integration Conceitos e Componentes

Soluções Pentaho Data Integration são construídas a partir de dois tipos diferentes de
objetos:

- Transformações
- Empregos

O coração do produto Pentaho Data Integration é formado pela Pentaho
dados do motor de integração. Este motor é um componente de software que é capaz de
interpretar e executar trabalhos e transformações, assim, realizar o
dados reais de tarefas de integração em conformidade. Além do motor, Pentaho
Integração de dados oferece uma série de ferramentas e utilitários para criar, gerenciar e
transformações lançamento e empregos.

Todas as ferramentas Pentaho Data Integration e componentes estão disponíveis para down-
carga como um único Zip. arquivo. Este arquivo pode ser encontrado na área de downloads da
a página do projeto Pentaho em sourceforge.net / projects / pentaho. Pentaho
integração de dados não requer um procedimento de instalação separada que não
descompactar o download.

Para uma visão geral de alto nível dos componentes de integração Pentaho de dados,
consulte
Figura 9-1.

Os componentes são descritos nas secções seguintes.

Ferramentas e Utilitários

Pentaho Data Integration compreende o seguinte conjunto de ferramentas e utilitários:

- Spoon-A IDE integração gráfica de dados para criar transformações
  e empregos
- Cozinha-A ferramenta de linha de comando para a execução
  de trabalhos
- Pan-A ferramenta de linha de comando para a execução de
  transformações

- Carte-A servidor para executar trabalhos leves e transformações em um host remoto

NOTA Como você pode ver, um tema culinário foi usado para nomear as ferramentas. Este esquema de nomeação de volta aos dias em que o software que agora é conhecido como Pentaho Data Integration foi criado. O produto original foi chamado Chaleira (Ou realmente chaleira, que é um acrônimo recursivo para Chaleira Extração, Transformação, Transporte, Meio Ambiente e de carga). A chaleira termos, Chaleira e Pentaho Data Integration podem ser usados alternadamente.

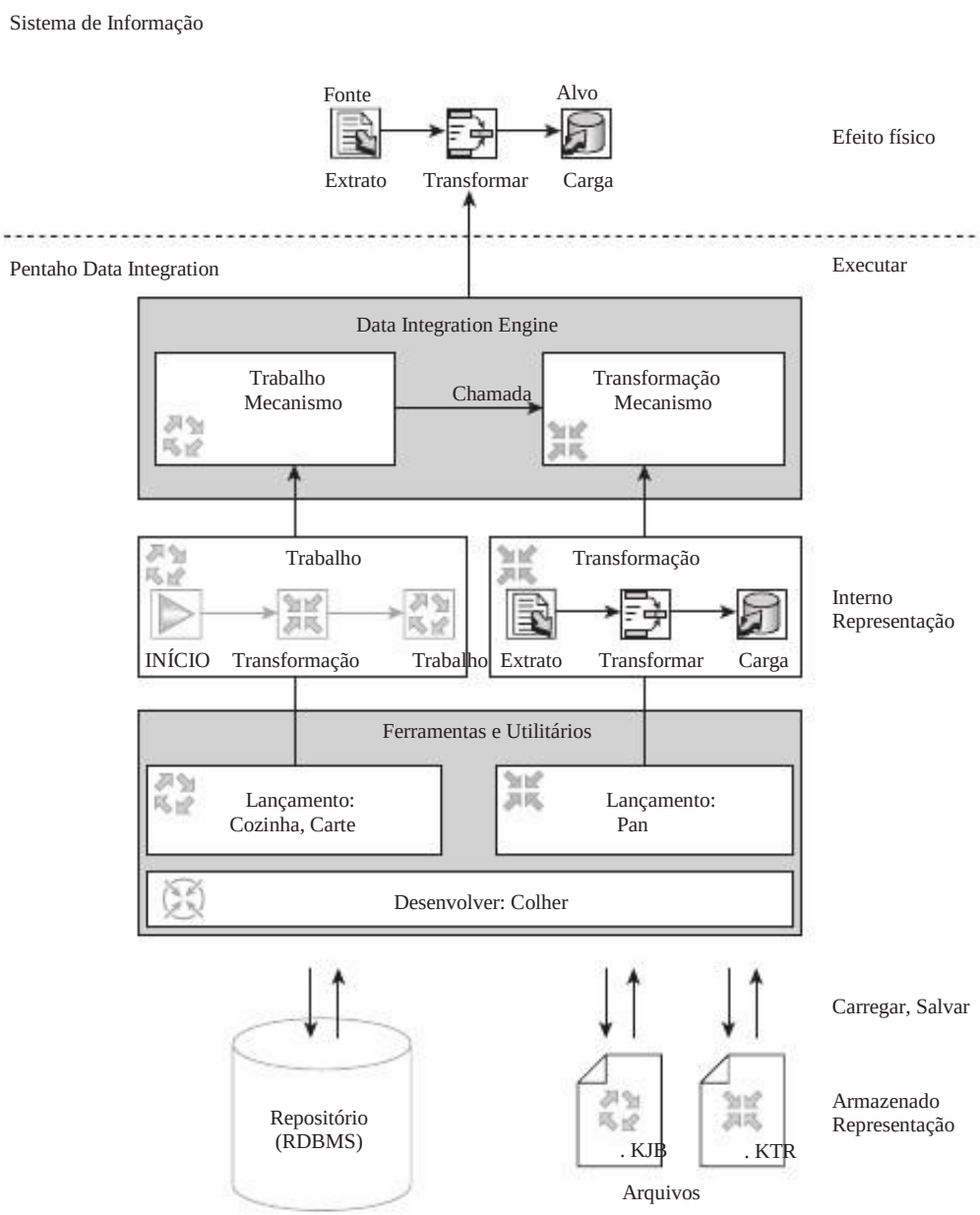


Figura 9-1: dados Pentaho ferramentas de integração e de componentes

O Mecanismo de Integração de Dados

O mecanismo de integração de dados é responsável por interpretar e executar os dados trabalhos de integração e transformações. Empregos e as transformações são realmente manipuladas por diferentes partes do motor. No entanto, porque os empregos podem conter transformações e execução de um trabalho pode resultar na execução de uma ou mais transformações, geralmente é mais conveniente considerar o mecanismo de trabalho e o mecanismo de transformação como um todo (o motor de integração de dados).

O mecanismo de integração de dados é fisicamente implementado como uma biblioteca Java.
As ferramentas de front-end usar uma API pública para que o motor de executar trabalhos e transformações em seu nome, em resposta a interação do usuário.

Usando o motor desta forma não se limita às ferramentas de front-end: O motor pode ser usado por qualquer aplicação. O mecanismo de integração de dados também é
incluído no servidor de BI Pentaho, permitindo que os trabalhos e as transformações a serem executada como parte de uma seqüência de ação.

Repositório

Empregos e as transformações podem ser armazenados em um banco de dados do repositório. O front-end
ferramentas podem se conectar ao banco de dados e transformações de carga e definição de postos de trabalho
ções armazenadas no repositório. Usando o repositório também oferece uma maneira fácil para múltiplos desenvolvedores a colaborar quando se trabalha em uma integração de dados solução.

Note-se que o repositório não é um requisito. Quando não
trabalho com o repositório, as transformações e os trabalhos são armazenados como arquivos em um
formato XML. Neste caso, um sistema de controle externo a versão, como subversão ou CVS pode ser utilizado para facilitar a colaboração.

Empregos e Transformações

Agora que explicamos as ferramentas e componentes, é hora de tomar uma análise mais olhar para os trabalhos e transformações.

Já mencionamos que o mecanismo de integração de dados e interpreta
executa trabalhos e transformações. Isto sugere uma semelhança entre os trabalhos e transformações, por um lado e um programa de computador os arquivos de código-fonte em o outro. No entanto, há boas razões para rejeitar essa noção.

Normalmente, o código de um programa de computador de origem é composta de um conjunto de mais
instruções literal, e cabe ao programador para garantir que esses instru-
ções juntos alcançar o efeito desejado. Para a maior parte, as transformações
e emprego não consistem de um conjunto literal de instruções. Pelo contrário, o emprego ea trans-
formações são declarativas de natureza: eles especificam um determinado resultado que é ser alcançado, e deixá-lo até o motor de descobrir uma maneira de conseguir
esse resultado. Outra maneira de colocá-la é dizer que o mecanismo de integração de dados é

orientada por metadados: transformações e trabalhos contêm informações sobre os dados, o sistema de origem, eo sistema alvo, e ao executar um trabalho ou transfor- informação, esta informação é utilizada para realizar as operações necessárias para alcançar o resultado. Isso é feito sem gerar o código do programa intermediário.

A diferença entre o computador de código-fonte do programa e do emprego e trans- formações também é evidente na forma como estes são desenvolvidos. O código fonte é normalmente baseados em texto, e desenvolvido por declarações de digitação de acordo com a
gramática da linguagem de programação (codificação). Em contraste, o emprego ea trans- formações são desenvolvidas de uma forma bastante gráfica. Eles são criados por arrastar e soltar elementos em uma tela e conectá-las
para formar um gráfico ou diagrama. Este processo é semelhante ao desenho de um fluxograma.
Transformações e empregos podem conter elementos que possam envolver o script, mas esta é uma excepção e não a regra.

Transformações e empregos são ambos constituídos por um conjunto de itens que são interligados por lúpulo (Veja a seção seguinte). A este respeito, há uma similaridade entre as transformações eo emprego. Esta semelhança torna-se especialmente resulta na representação visual, como ambas as transformações e os trabalhos são descrito como gráficos.

No entanto, um olhar mais atento as transformações e os trabalhos revela que a seme- dades são realmente poucos em número e bastante superficial na natureza. Há diferenças importantes nos conceitos subjacentes empregos e transformações, ea semântica dos seus elementos constitutivos e lúpulo são muito diferentes um do outro. Esperamos que isso ficou claro pela leitura do seguinte seções.

Transformações

Uma transformação Pentaho representa uma tarefa ETL em sentido estrito. Trans- formações são orientadas a dados, e sua finalidade é extrair, transformar e carregar dados.

A transformação consiste de uma coleção de etapas. Um passo denota uma particular operação em um ou mais registro de córregos. As etapas podem ser conectados por lúpulo. A hop é como um gasoduto através do qual os registros podem fluir de uma etapa para outro passo.

Um fluxo de registro é uma série de registros. Um registro é uma coleção de valores que está estruturado de tal forma que cada valor está associado com exatamente um de campo. A coleção de campos associados com todos os valores em um registro é chamado de registro tipo. Todos os registros em um fluxo de registro deve ser do mesmo tipo de registo.

Cada campo tem um nome que deve ser exclusivo dentro do tipo de registro. Campos definir propriedades, como tipo de dados eo formato que, coletivamente, descrever o natureza de qualquer valor associado com o campo. Essas propriedades formam o seu valor metadados (Dados sobre dados""). De maneira semelhante, o tipo de registro constitui a metadados do registro. Veja a Figura 9-2, para uma representação gráfica destes conceitos.

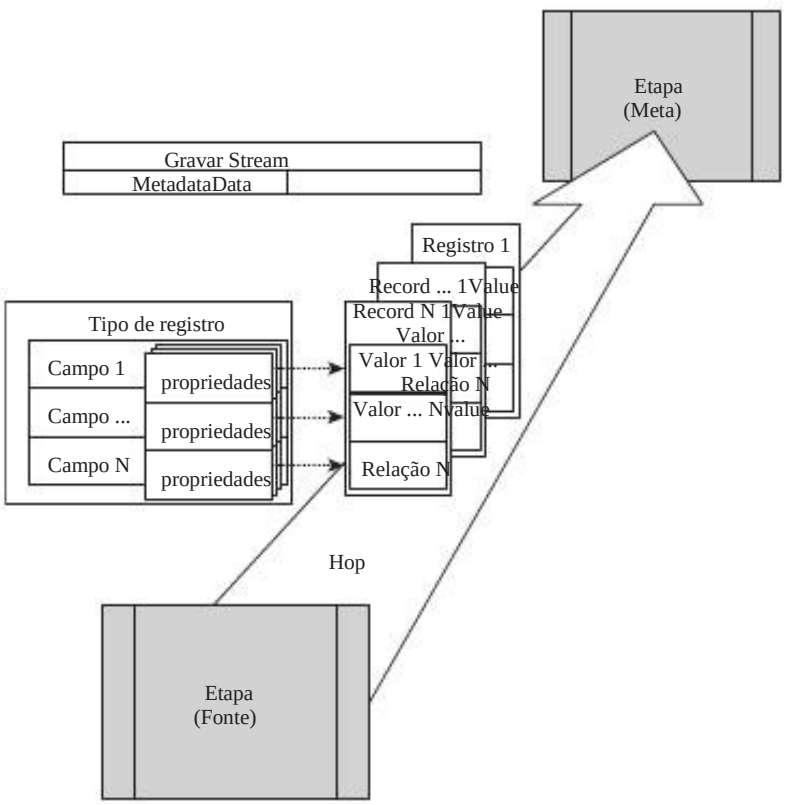


Figura 9-2: Passos, lúpulo e recorde córregos

etapas de transformação produzir registros de forma contínua. Registros
viajar através de saída da etapa do lúpulo imediatamente após serem produzidos
até chegar ao passo do outro lado do lúpulo. Lá, a chegada
registros estão na fila e esperar para ser consumido pela etapa de recebimento. Para cada
registro de entrada, a etapa de recebimento realiza algumas operações pré-definidas.
Normalmente, isso gera registros de saída, que são então empurrados para a etapa de
registro de saída de fluxo.

Uma coisa importante a perceber sobre todo o processo é que as etapas do trabalho
simultaneamente e de forma assíncrona. Ao executar uma transformação, as etapas
que sabem para gerar linhas baseado em alguma fonte de dados externos
basta começar a gerar filas até a fonte de dados está esgotada. Records
imediatamente o fluxo para os próximos passos a jusante, onde são processados o mais
rápido
possível que eles chegam: as etapas a jusante, não espere para o montante
etapas para terminar. Desta forma, os registros de gotejamento, através da transformação.

A operação realizada por uma etapa de transformação depende do tipo de
o passo ea forma como ele está configurado. Há um pouco de variedade na etapa
tipos. Alguns tipos de etapas de gerar um registro de saída única para cada entrada
registro. Outros tipos de etapas de agregação de um grupo de linhas de entrada em um único

saída de linha. No entanto, outros tipos de medidas podem dividir uma única linha de entrada em uma coleção
de registros de saída. Para alguns tipos de etapas, os fluxos de saída têm o mesmo
tipo de registro como os fluxos de entrada, mas outros tipos de etapa pode adicionar, remover ou
renomear campos a partir dos registros de entrada.

Algumas das etapas de transformação disponíveis são discutidos em detalhe no Capítulo 10.

Empregos

Normalmente, os trabalhos são compostos de uma ou mais transformações. Por exemplo, para
carregar um esquema em estrela, normalmente você construir uma transformação para fazer o
extracção em si, e construir uma transformação para cada tabela de dimensão, e
uma transformação para carregar a tabela de fatos. A Emprego seria usado para colocar todos esses
transformações na seqüência correta (primeiro extrato, em seguida, carregar todas as dimensões
tabelas, e então carregar a tabela de fatos). Como transformações, trabalhos consistem em uma
número de itens interligados, mas a semelhança termina aí.

Os trabalhos são processuais e orientado para a tarefa, em vez de orientadas a dados. Os itens
aparecendo em empregos são chamados entradas de emprego, e denotam a execução de um determinado
tarefa. A ligação entre as etapas de um trabalho denota a ordem seqüencial
dessas tarefas.

Execução de uma entrada de emprego sempre resulta em um estado de saída. Isto é usado para
saber se a tarefa de base foi executada com êxito. Dependendo
sobre o valor do status de saída, a execução do trabalho continua com a adequada
entrada de trabalho subseqüente.

Ao comparar a natureza das entradas trabalho para as etapas de transformação, a chave
diferença é que as etapas de transformação operar em fluxos de dados. entradas de Job
operar no estado de saída resultante da execução da tarefa.

entradas de emprego são tipicamente usados para executar transformações, mas também podem
ser usado para executar outras tarefas de apoio, tais como o esvaziamento tabelas de dados,
iniciar a transferência de arquivos de um host remoto, ou enviar uma mensagem de correio electrónico.
entradas podem ser implementadas separadamente do principal produto de software
componentes chamados plug-ins. Plug-ins podem ser carregados dinamicamente, sem
recompilar o núcleo.

Pentaho empregos são discutidos em mais detalhes nos Capítulos 10 e 11.

Plug-Arquitetura na

Pentaho Data Integration apresenta uma arquitetura plug-in. Passos, bem como trabalho

Atualmente, existem muito poucos de terceiros plug-ins para o Pentaho Data
Integração. Alguns deles são de código fechado e está disponível sob os termos de um
licença comercial. Muitos outros são licenciados sob alguma forma de código aberto
licença.

Começando com uma colher

Spoon é o nome do Pentaho Data Integration gráfica de desenvolvimento ambiente. Ele pode ser usado para criar transformações e empregos.

Nesta seção, você aprenderá como usar Spoon. Após iniciar o aplicativo, e falar brevemente sobre alguns dos principais elementos da interface do usuário, você vai rapidamente começar a construir uma"Olá, Mundo! transformação"para se familiarizar com a interface do usuário. Nós, então, apresentar a você alguns economizam tempo funcionalidades que podem ajudar você a solucionar problemas de forma eficiente. Em seguida, você vai aprender a trabalhar com conexões de banco de dados e aumentar a"Olá, mundo!" exemplo, para colocar esse conhecimento em prática.

Iniciando o aplicativo Spoon

Spoon pode ser iniciado executando o script de inicialização colher no inte-Pentaho gração diretório home. Para os sistemas Windows, o script é chamado Spoon.bat. Em sistemas Windows, você também pode iniciar Kettle.exe. Para sistemas baseados em UNIX, Este script é chamado spoon.sh.

NOTA Usuários de sistemas baseados em UNIX precisa habilitar permissões de execução para o script antes de executá-lo:

```
$ Chmod ug + x spoon.sh
```

Todos Pentaho ferramentas de integração de dados são iniciadas através de um script, assim que você poderia muito bem permitir que as permissões de execução para todos eles:

```
$ Chmod ug + x *. sh
```

Depois começa a colher, você verá uma tela inicial. Algum tempo depois, um diálogo aparece, solicitação de credenciais repositório. Vamos explorar as possibilidades do repositório no Capítulo 11. Por enquanto, você vai trabalhar sem um banco de dados repositório. Para continuar, clique no botão com a legenda no repositório. Opcionalmente, você pode limpar o Presente esta caixa de diálogo na inicialização para evitar a caixa de diálogo de aparecer na próxima vez que você abre Spoon.

A janela principal do aplicativo deve aparecer. Você pode ver uma colher dicas janela mostrando uma dica do dia, e, opcionalmente, você pode limpar o Show dicas na inicialização? caixa para impedir que ele aparecer na próxima vez que você Spoon aberto. Prima Fechar para fechar a janela de dicas Spoon.

A janela do aplicativo Spoon, mostrado na Figura 9-3, é dividido em uma das principais espaço de trabalho à direita e um painel lateral à esquerda. A principal área de trabalho pro-nece uma guia de interface / página para trabalhar com todas as transformações e abriu postos de trabalho.

Inicialmente, uma página de boas-vindas especiais podem aparecer automaticamente no espaço de trabalho logo após o aplicativo é iniciado. Também é possível que a página de boas-vindas aparece

em seu navegador de Internet. A página inicial contém links úteis
o funcionário Pentaho documentação de integração de dados, bem como para a comunidade
recursos e blogs de desenvolvedores.

Figura 9-3: A janela principal Spoon

Se você gosta, você pode impedir que a página de boas-vindas aparecer, limpando
Mostrar a página de boas-vindas na caixa de arranque na caixa de diálogo Opções. As Opções
diálogo pode ser aberto usando o menu principal (Menu Editar Opções).

Um mundo simples''Olá,!''Exemplo

Spoon oferece uma única interface de usuário para criar e projetar os postos de trabalho e transformações. Apesar de postos de trabalho e as transformações são tipos completamente diferentes
das coisas, há muitas semelhanças na maneira como eles são representados visualmente, e isso se reflete em uma interface quase idêntica para projetá-los.

A melhor maneira de se familiarizar com a interface do usuário é começar a construir uma Mundo muito simples''Olá!''Exemplo, o estilo de ETL. Isso é descrito em detalhes no
o restante desta seção.

Construindo a Transformação

As instruções seguintes descrevem em detalhes como construir uma transformação
que extrai os nomes das pessoas a partir de um arquivo de texto para gerar um''Olá, mundo!'' mensagem. Finalmente, as mensagens são armazenadas em um arquivo de texto novo.

1. Usando um editor de texto simples como o Notepad (Windows) ou vi
   (Sistemas baseados em UNIX), crie um arquivo texto com o texto na Listagem 9-1. Nome

o arquivo hello_world_input.txt e salvá-lo em algum lugar você encontra conveniente.

Listagem 9-1: O conteúdo do hello_world_input.txt

```
Nome
George Clooney
Daniel Lay Lewis
Johnny Depp
Tommy Lee Jones
Viggo Mortensen
```

2. Iniciar Spoon (se ele não estiver em execução) e criar uma nova transformação escolhendo Arquivo nova transformação a partir do menu principal. Um novo página é aberta automaticamente no espaço de trabalho, ea transformação também é adicionado à exibição em árvore no painel lateral, como mostrado na Figura 9-4.

   Além de utilizar o menu principal, você pode criar uma transformação por usando o atalho do teclado Ctrl + N, ou usando a barra de ferramentas: Barra de ferramentas
   Nova Transformação. O botão da barra de ferramentas é mostrada na Figura 9-4.

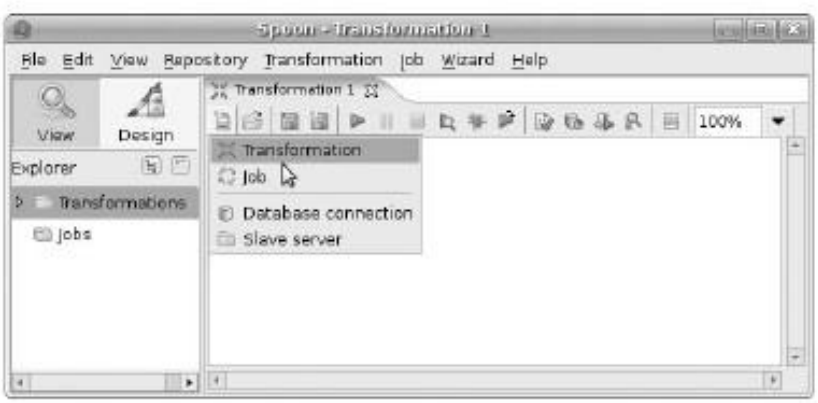


Figura 9-4: Usando a barra de ferramentas para criar uma nova transformação

3. No painel do lado, mudar de modo de visualização para o modo Design, clicando em o botão Design. No modo Design, o painel lateral exibe uma exibição de árvore contendo um número de pastas que representam categorias de tipo de etapa. Expanda a pasta chamada de entrada que fica no topo da árvore.

4. Arraste o item de input do arquivo de texto marcado a partir da pasta de entrada e solte-o Página da transformação para criar um novo passo, como mostrado na Figura 9-5.

   Quase todos os passos requerem alguma forma de configuração para controlar a sua comportamento. Passos são configurados através de sua janela de configuração, que pode ser levantada, clicando duas vezes o passo, ou escolhendo Editar etapa a partir do menu de contexto do passo.

PASSOS: tipos, criar, mover, E RETIRAR

Observe que o ícone você arrastou a exibição de árvore era um tipo degrau, vez do que um passo real. Abolindo o tipo de passo para a tela, você criou um passo real desse tipo.

Passos também podem ser criados usando o menu de contexto de transformação. Para exemplo, o texto de entrada em degrau de arquivos também podem ser adicionados clicando no Página de transformação e escolhendo Nova etapa de entrada de entrada de arquivo de texto.

Uma vez que uma etapa é criado, ele pode ser colocado em qualquer lugar da transformação página usando drag-and-drop. Basta colocar o ponteiro do mouse sobre a etapa e em seguida, pressione e segure o botão esquerdo do mouse. Movendo o ponteiro do mouse agora arraste a passo ao longo até que o botão do mouse é liberado.

Para excluir uma única etapa, botão direito do mouse para abrir o menu de contexto do passo; em seguida, escolha Excluir etapa.

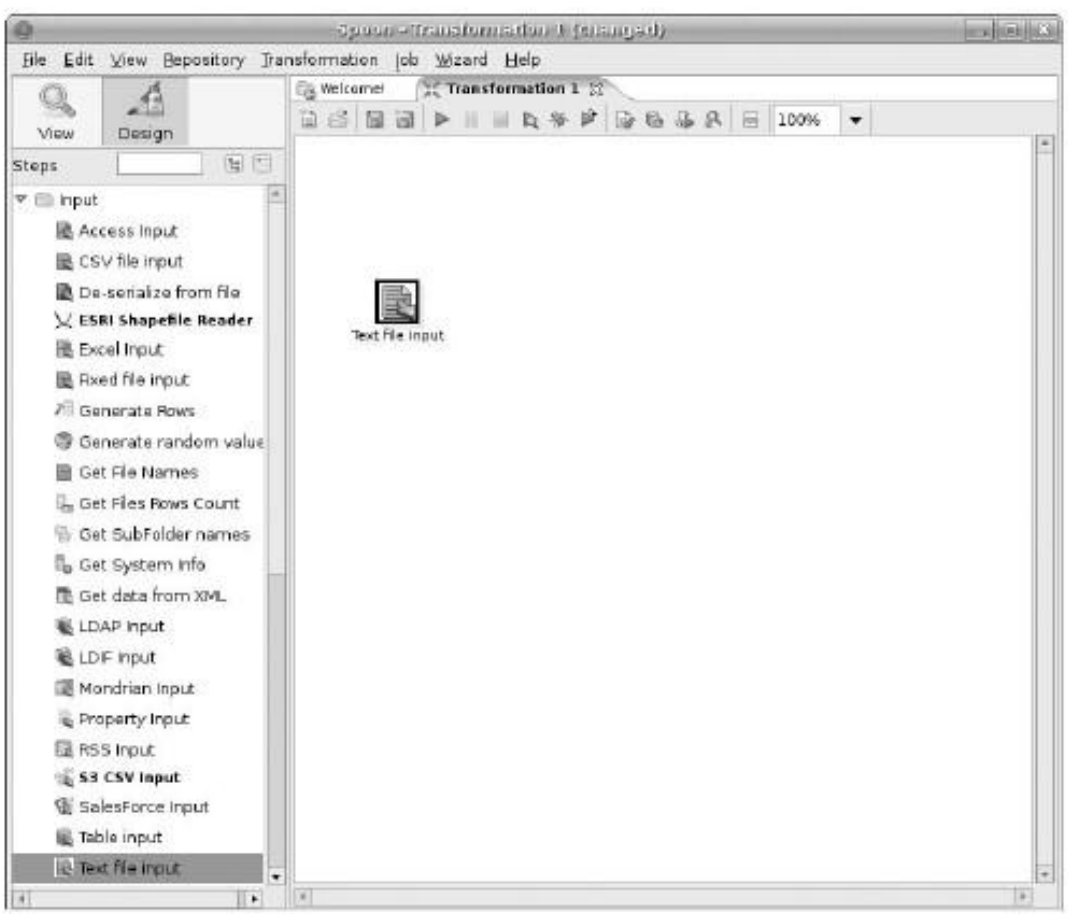


Figura 9-5: A página de transformação com o seu primeiro passo

Uma coisa que pode ser configurado para todas as etapas é o nome. Cada passo tem um nome que deve ser exclusivo dentro da transformação. O nome também é usado como uma legenda para a etapa na página de transformação. Na janelas de configuração, o nome da etapa aparece sempre antes de qualquer outra opções de configuração.

Por padrão, o nome da etapa é o nome do seu tipo de passo, possivelmente seguido de um número para garantir o nome é exclusivo. Por enquanto, você pode deixar o nome padrão (arquivo de entrada de texto), mas geralmente o padrão nome seria alterado para refletir a propósito da etapa dentro do transformação (em oposição a um rótulo que só descreve o funcionamento é executada). Especificando um nome descritivo para as etapas é fundamental para a concepção
transformações sustentável.

Outros elementos de configuração que se repetem com freqüência em diferentes etapas de
tipos são páginas de guia para as opções relacionados a um grupo e grades para configurar o
campos de dados da etapa córregos. No entanto, a maioria dos disponíveis opções de configuração depende do tipo de etapa.

5. Na página de transformação, clique duas vezes no arquivo de entrada de texto passo para abrir
sua janela de configuração. Na página da guia do arquivo da caixa de diálogo, use o botão Procurar
botão para selecionar o hello_world_input.txt arquivo. Em seguida, clique em Adicionar botão para atribuir o arquivo para a grade de arquivos selecionados.
Em seguida, clique na página guia Conteúdo e defina a lista suspensa Formato na metade inferior do formulário para misto.

6. Finalmente, ative a ficha de Campos, e clique no botão Get Campos. Este botão diz Colher para digitalizar os dados para descobrir os campos na entrada. Você primeiro será solicitado o número de linhas de varredura. Você pode deixar o número padrão (100) para agora. Após a confirmação, uma janela mostrando o resultados da verificação será exibida. Isso tudo é mostrado na Figura 9-6.

Note-se que um campo foi descoberto, e note o mínimo eo máximo valores. Feche essa janela, e observe o campo a ser adicionado à Campos grade. Feche a janela de configuração clicando em OK.

NOTA Após criar e configurar uma etapa, você sempre pode voltar a abrir o
janela de configuração e revisão de suas propriedades. No entanto, para rápida inspeção, Spoon oferece funcionalidades mais conveniente a partir do menu de contexto do passo.

Os campos de entrada Show e Show campos de opções de saída pode ser usado para inspecionar o tipo de registro da entrada e saída de fluxos, respectivamente. Estas opções de trazer um janela que exibe uma grade que contém uma linha para cada campo no córrego mostrando o nome do campo, tipo de dados e nome da etapa, onde o campo foi originalmente criado.

A etapa de entrada de arquivo de texto não espera um fluxo de entrada, e isso está devidamente indicado na janela pop-up. A janela para a opção Mostrar campos de saída é mostrado na Figura 9-7.

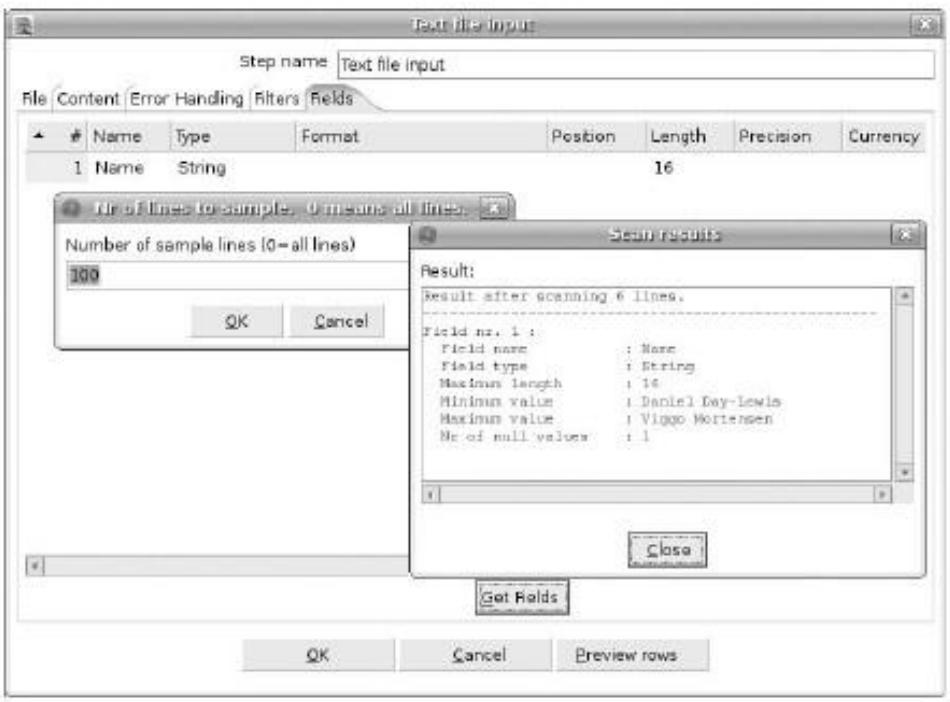


Figura 9-6: Definindo os campos com o botão Get Campos

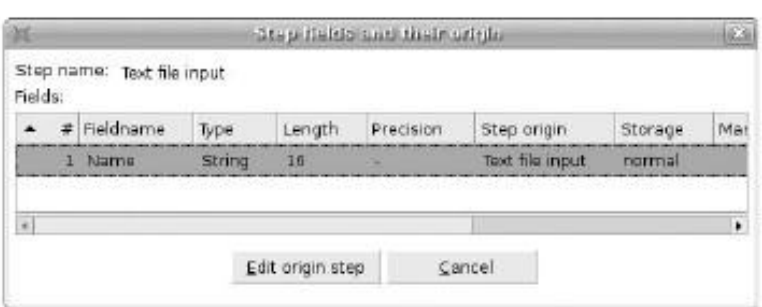


Figura 9-7: Os campos de saída para a etapa de arquivo de entrada de texto

7. Adicionar Encontre o tipo de etapa constantes na categoria Transform, e adicionar um para a transformação perto a etapa de introdução de texto de arquivo. Abra a sua janela de configuração e digite o seguinte duas linhas na grade de Campos (Ver Figura 9-8):

   Nome: mensagem; Tipo: String; Valor: Olá
   Nome: exclamação; Tipo: String; Valor:!

NOTA Spoon suporta seleções criação de uma ou mais etapas. Isso ajuda muito facilita o trabalho com várias etapas, porque isso lhe permite mover ou remover todas as medidas incluídas na seleção de uma vez. passos escolhidos são reconhecíveis pelos seus borda preta grossa.

Você pode selecionar uma etapa individual, basta clicar nele. Normalmente, isso desfaz a seleção anterior. No entanto, se você pressionar e segurar a tecla Ctrl e clicando em um etapa desmarcada irá adicioná-lo à seleção, e clicar em um já selecionado etapa irá removê-lo da seleção.

Você também pode selecionar todos os passos de uma área retangular ao apontar o mouse para um lugar vazio na página de transformação, pressionar e segurar o botão do mouse e arrastando o ponteiro do mouse para outro local vazio abaixo e à direita do ponto de partida. Uma faixa de borracha""aparece e quando o botão do mouse é lançado, todas as medidas que se situam entre os limites da faixa de borracha tornam-se a nova seleção.

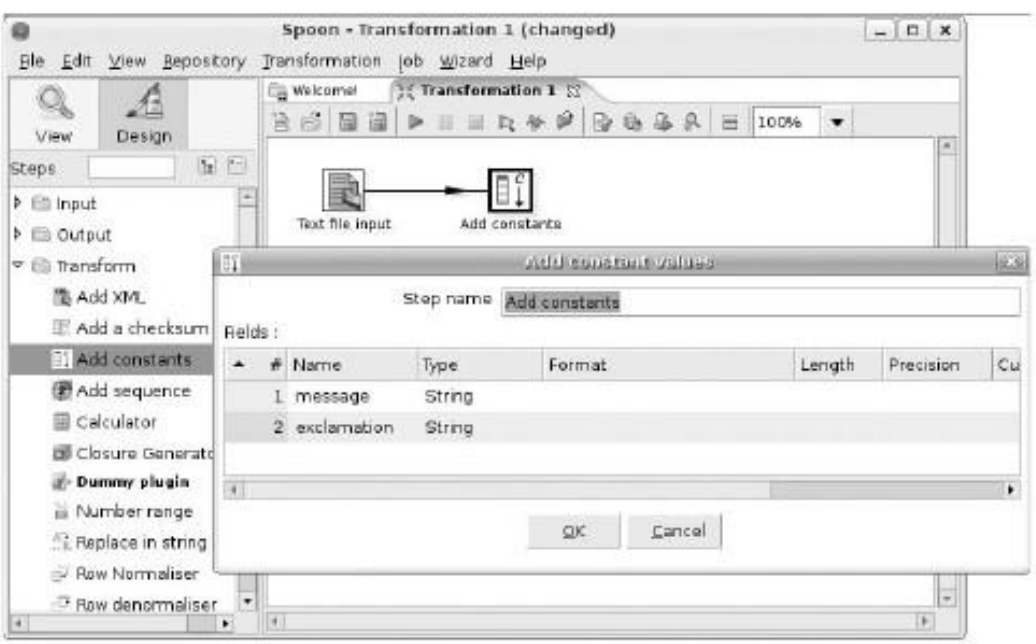


Figura 9-8: A etapa constantes Adicionar

O lúpulo pode ser criada para direcionar o fluxo de saída de uma etapa para a entrada fluxo de outro. Neste exemplo, você quer criar um salto a partir do Texto etapa arquivo de entrada para as constantes Adicionar passo para adicionar os campos Adicionar ao passo que os registros constantes do fluxo de saída do Texto arquivo de entrada em degrau.

Para criar um salto, segure a tecla Shift e arraste a etapa de origem para a etapa de destino. A tecla Shift pode ser liberado depois de deixar o

etapa. Quando não estiver segurando a tecla Shift, a etapa só é movido, mas não conexão é feita. Alternativamente, se você tiver uma roda do meio do mouse, basta pressioná-lo na etapa de origem, mantê-lo pressionado e mover para a alvo etapa. Existe ainda uma outra maneira de criar um pulo: ativar o Vista modo no painel lateral e se a visualização em árvore nó correspondente para a transformação é expandido. Em seguida, botão direito do mouse no nó do lúpulo abrir o menu de contexto. No menu de contexto, clique em Novo para trazer o diálogo de lúpulo. Neste diálogo, você pode simplesmente ligar-se passos selecionando as etapas de origem e destino nas caixas de lista.

Um pulo é exibido como uma linha traçada a partir da etapa de origem para o destino etapa. A linha tem uma ponta de seta um pouco além do centro da linha, apontando para fora da etapa de origem e de destino para a etapa.

8. Dê um salto a partir da etapa de introdução de texto de arquivo para a etapa de constantes Adicionar
arrastando o arquivo de entrada de texto passo para o passo constantes Adicionar enquanto
segurando a tecla Shift. Isso também é mostrado na Figura 9-8.

DICA Normalmente, quando você adicionar mais passos e mais e ao lúpulo, a transformação pode começam a parecer um pouco confuso. Para evitar isso, você pode alinhar horizontal ou verticalmente etapas. Para alinhar verticalmente várias etapas, selecione as etapas e pressione para cima chave de seta (para alinhar os topos das etapas selecionadas para o início da etapa que é mais próximo do topo da tela) ou a tecla seta para baixo (para alinhar a parte inferior da etapas selecionado para o fim da etapa que está mais próximo do fundo do tela). Da mesma forma, você pode alinhar os passos horizontalmente usando a esquerda e direita teclas de seta.

9. Em seguida, crie um texto passo a saída do arquivo (da categoria de saída) e adicione uma
hop para ele a partir do Passo constantes Adicionar. O resultado é mostrado na Figura 9-9.

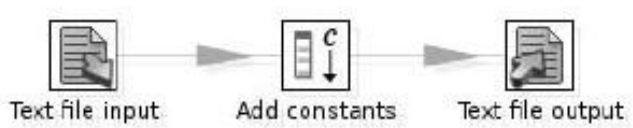


Figura 9-9: O exemplo a transformação de três etapas

10. Abra a janela de configuração para a etapa de produção de texto do arquivo e defina Nome da propriedade em um arquivo chamado hello_world_output na mesmo diretório que o hello_world_input.txt arquivo. (Nota da Extensão propriedade é definida por padrão para txt, Razão pela qual não incluiu os . Txt extensão ao especificar a propriedade Filename).

Ative a ficha de Campos, e pressione o botão para obter campos automat- camente preencher a grade Campos. Depois disso, pressione o botão de largura mínima e
ajustar o comprimento dos campos de comprimento variável para gerar mais compacto

saída. Note que a ordem dos campos corresponde exatamente à seqüência de dos passos de onde se originam: o primeiro Nome campo gerado por Texto etapa arquivo de entrada, então o Mensagem e exclamação campos adicionados pela etapa de constantes Adicionar.

Botão direito do mouse sobre a linha que define o Nome campo e no menu pop-up menu escolher o item Mover para baixo para que a ordem final dos campos é Mensagem, Em seguida, NomeE, em seguida exclamação. Alternativamente, você pode selecionar a linha da grade correspondente ao Nome campo e pressione o cursor para baixo. Clique em OK para fechar a janela de configuração.

O texto passo a saída do arquivo e seu diálogo campos são mostrados na Figura 9-10.

GRADES DE CAMPO

etapas de transformação Muitos exigem múltiplos campos a serem configurados. Em muitos casos, o diálogo de configuração oferece uma grade como você só viu em Campos guia da etapa de saída de arquivo de texto. Normalmente, um botão Get Campos está presente a auto- automaticamente preencher a grade. Ao preencher um grid não-vazia neste homem ner, você será solicitado a fazer uma escolha para lidar com os campos existentes. A Adicionar novas escolhas são, adicione todos e claras e adicionar todos, com o comportamento óbvio para cada escolha.

Linhas na grade podem ser selecionados clicando neles, e vários seleção pode ser feita mantendo a tecla Ctrl pressionada enquanto seleciona as linhas. Linhas são adicionados, basta digitar em uma nova linha após o último existente. Excluir linhas selecionadas é feito usando a tecla DEL. Manter linhas selecionadas (excluindo todos os desmarcada linhas) é feito usando Ctrl + K.

A lista dada aqui não é exaustiva, é aqui apenas para dar uma idéia. Se você quer um panorama completo das possibilidades, clique com o botão direito sobre uma grade para ver a sua menu de contexto.

11. Está feito agora com o actual edifício da transformação. Esta é uma bom momento para salvar a transformação, usando o arquivo item Salvar do menu principal. Salve a transformação como hello_world_transformation em mesmo diretório que você salvou o arquivo de entrada. O arquivo deve ser au- camente obter o . KTR extensão, que é a abreviação de"Chaleira Transformação."

NOTA Note-se que um botão Save também está disponível na barra de ferramentas. Alternativamente, você também pode usar o atalho do teclado Ctrl + S.

Executando a Transformação

Você pode executar a transformação activa de dentro Spoon através dos seus principais menu com a opção MenuTransformationRun. Alternativamente, você pode usar o atalho do teclado F9 ou o botão "Run (o verde-cheia seta apontando para a direita).

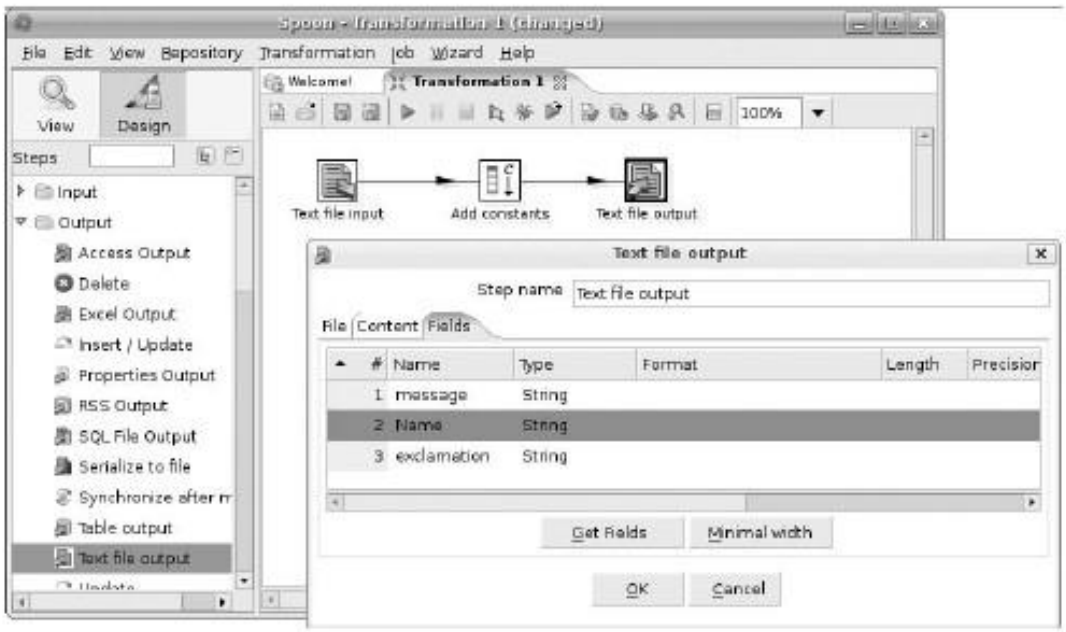


Figura 9-10: O texto passo a saída do arquivo e os seus campos de diálogo

Ao executar a transformação, um diálogo com o título executar um transfor-
informação aparece em primeiro lugar. Vamos discutir este diálogo em detalhes mais adiante neste
capítulo e, novamente, no capítulo 11, mas agora você pode simplesmente pressionar o lançamento
botão na parte inferior da janela.

A Execução Painel de Resultados

Depois de lançar a transformação, a execução de painel de resultados torna-se visível na parte inferior do espaço de trabalho. Este painel fornece um número de páginas de guia que oferecem funcionalidade útil para monitorar uma transformação. Por padrão, o Passo Métricas página da guia é ativado e, por enquanto, vamos nos limitar a maior parte usando esta página guia, bem como a página de registo. Todas estas páginas de guia será discutidos em mais detalhes no Capítulo 11.

Se tudo funcionasse como deveria, o painel de Execução deve agora olhar semelhante à Figura 9-11. Na grade de Métricas etapa, você verá exatamente três linhas (uma para cada etapa da transformação).

Todas as linhas devem indicar que estão com acabamento na coluna de ativos (em do lado direito da grade). Se uma ou mais linhas na grade tem um fundo vermelho, ea coluna Active mostra o estado parado, ocorreu um erro enquanto executar a transformação. A seção"Verificando a Transformação" mais adiante neste capítulo pode lhe dar algumas pistas sobre o que correu mal.

NOTA

superior no sentido vertical. Para fazer isso, coloque o ponteiro do mouse exatamente entre os superior da barra de título do painel e na parte inferior do espaço de trabalho. Você pode encontrar o

""Pega colocando o ponteiro do mouse direito no texto da barra de título do painel e lentamente movendo-se em direção ao fundo da área de trabalho até que o ponteiro do mouse ícone muda para um ícone de redimensionamento.

Você pode alternar a visibilidade do painel usando Hide / Show alternar Resultado da Execução botões encontrados na barra de ferramentas. Os ícones destes botões são mostrados na Figura 9-12. Você também pode controlar a visibilidade usando a janela padrão ícones que aparecem na do lado direito na barra de título do painel.

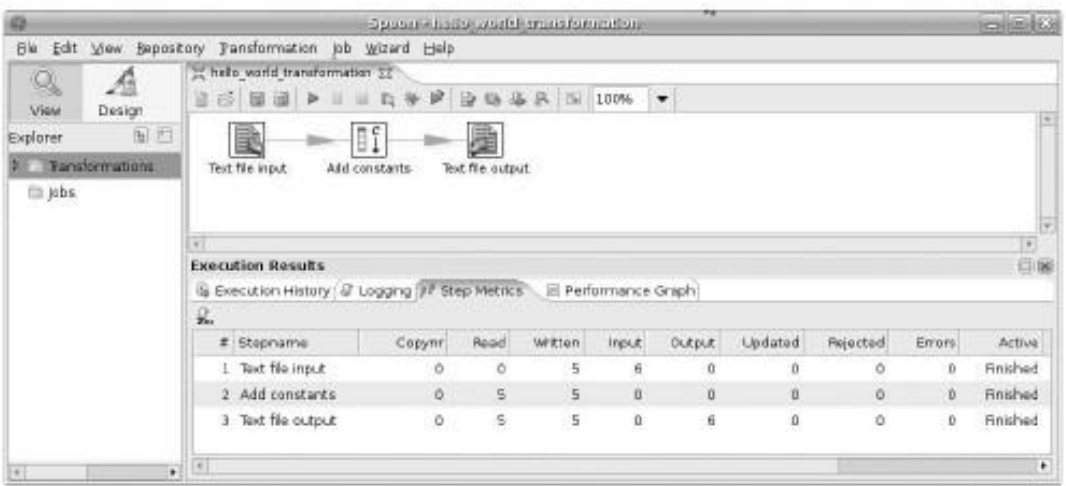


Figura 9-11: A transformação final

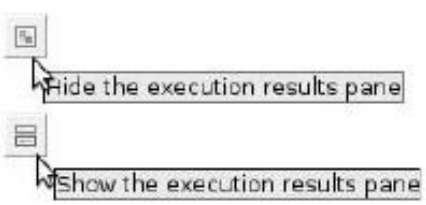


Figura 9-12: Alternar a visibilidade do painel de Resultados da Execução

A saída de

A hello_world_output.txt arquivo que você criou anteriormente agora deve conter Olá, <name>! mensagens como aquelas mostradas na Listagem 9-2.

Listagem 9-2: O conteúdo do hello_world_output.txt

```
messageNameexclamation
Olá, George Clooney!
Olá, Daniel Day-Lewis!
Olá, Johnny Depp!
Olá, Tommy Lee Jones!
Olá, Viggo Mortensen!
```

Verificação de consistência e Dependências

Você só fez uma jogada mais caprichosa de imediato executar o seu trans-
formação, sem sequer realizar os controlos mais básicos. Concedido, o
transformação é muito simples, mas mesmo assim, há um número impressionante
de coisas que poderiam dar errado. As seções a seguir apresentam algumas informações básicas
verificações que você sempre deve executar.

Consistência lógica

A coisa mais básica é verificar se a transformação é logicamente
consistente. Por exemplo, se um passo opera em um campo particular, segue-se que
que o campo deve estar presente no fluxo de dados de entrada, e é considerado um
erro se este não for o caso. Outro tipo de erro ocorre quando o registro de fluxos de
incompatíveis com layouts são"misto."Isso ocorre quando duas recorde córregos
ter um layout de registro diferentes estão conectados à mesma etapa. Há mais
variedades de questões lógicas, como as etapas com saída nem entrada nem
saltos e os passos que estão ligados de uma forma circular.

Dependências de recursos

Outra coisa a verificar é se todos os recursos que são usados pelo trans-
formação estão realmente disponíveis. Por exemplo, o nosso"Olá, Mundo! Exemplo"
se refere a um arquivo de entrada e um arquivo de saída. transformações do mundo real são susceptíveis
a depender de conexões de banco de dados e sobre a disponibilidade dos objetos de banco de dados
como tabelas e, talvez, seqüências ou procedimentos armazenados. Obviamente, há
será um problema de execução tal transformação se esses recursos não são
disponíveis.

Verificando a Transformação

A opção Verificar no menu de transformação pode ser usado para verificar lógica
erros, bem como a disponibilidade de recursos como arquivos e banco de dados
tabelas. Escolhendo esta opção aparece uma janela intitulada Resultados da transformação
verifica mostrando cada problema potencial como uma linha em uma grade.

utilizar linhas de cores de semáforo para indicar a gravidade da questão: linhas verdes
indicar a seleção está bem, as linhas vermelhas indicam os erros, e as linhas amarelas indicam
avisos. Selecionando a opção "Exibir resultados na parte inferior da
janela de resultado revela os resultados de todos os cheques que passaram sem nenhum problema,
fornecendo uma visão geral de tudo o que poderia ter dado errado. Não
também podem ser linhas observação no resultado, que não tem nenhuma cor especial na
todos. Estas linhas são relatados pelos controlos, que atualmente não podem ser executadas.

Veja a Figura 9-13 para obter um exemplo do que a validação dos resultados para o"Olá,
Mundo"transformação pode parecer.

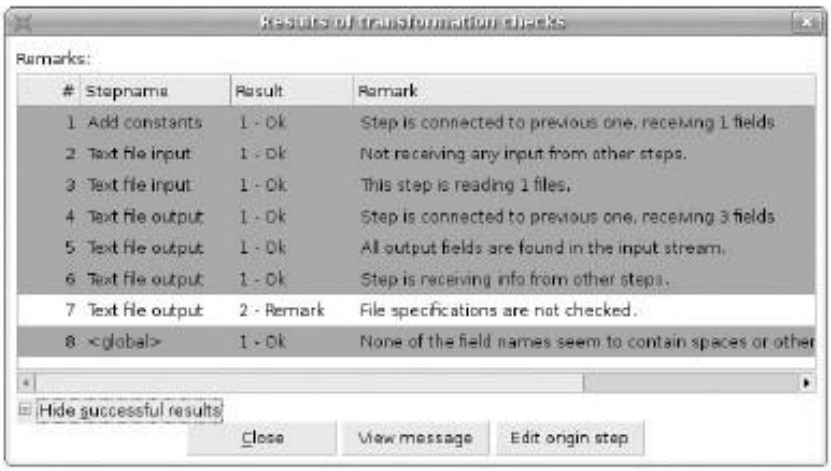


Figura 9-13: Verificando a transformação

Ao verificar o"Olá, Mundo!"Transformação exemplo, você pode encontrar uma linha de observação no relatório, como mostrado na Listagem 9-3.

Listagem 9-3: Uma observação no resultado da verificação

Stepname	Resultado	Observação
Texto de arquivo de saída	2 - Observação	Especificações de arquivo não são verificadas.

Este resultado ocorre se o arquivo de saída ainda não existe no momento da validação a transformação. Spoon gostaria de verificar se é possível gravar o arquivo de saída, mas não é possível concluir que a verificação até que o arquivo realmente existe.

Trabalho com o Banco de Dados

acesso RDBMS é uma das funcionalidades básicas de qualquer produto de integração de dados.
Pentaho Data Integration oferece suporte incorporado para mais de 30 diferentes produtos de banco de dados. Isso inclui todos os produtos conhecidos RDBMS, tais como IBM DB2, Microsoft SQL Server, MySQL, PostgreSQL e Oracle, e muitas menos conhecidos, como Kingbase, Gupta, MonetDB e outros.

JDBC ODBC e conectividade

Ao longo do Pentaho BI Suite, conectividade de dados é baseado em JDBC e Pentaho Data Integration não é excepção. Além do suporte JDBC nua, Spoon oferece suporte adicional para uma grande coleção de produtos RDBMS, protegendo o usuário de detalhes específicos do driver, como o formato exato da a seqüência de JDBC contato, apresentando frequentemente utilizados opções em um amistoso maneira. Mesmo se não houver direta suporte embutido para um RDBMS especial, ainda é possível se conectar a ele usando uma conexão JDBC genéricos, oferecendo uma driver JDBC compatível está disponível para esse RDBMS particular.

JDBC é um padrão muito bem estabelecido, ea maioria dos vendedores de RDBMS fornecer drivers JDBC para o seu RDBMS. Mas, mesmo se nenhum driver JDBC está disponível, é muitas vezes ainda é possível se conectar usando a conectividade ODBC.

ODBC conectividade é fornecida por uma ponte JDBC-ODBC, que é essencialmente um driver JDBC que pode atuar como um proxy para ODBC motorista. Isso permite o acesso a qualquer RDBMS para o qual existe um driver ODBC disponíveis. Como a conectividade ODBC adiciona uma camada extra, deve ser evitado se possível. Dito isso, pode ser a única opção no caso de haver simplesmente não existe um driver JDBC.

Por padrão, Pentaho integração de dados usa ponte Sun JDBC-ODBC para ODBC conectividade. Este driver está incluído na edição padrão do Sol JDK. No entanto, é sempre possível recorrer a qualquer terceiro JDBC-ODBC ponte. Qualquer driver JDBC compatível sempre pode ser utilizada com um genérico JDBC de conexão, e isso inclui todas as pontes de terceiros JDBC-ODBC.

Criando uma conexão de banco de dados

Conexões de banco de dados são definidos dentro de transformações e de emprego. Para criar uma ligação da barra de ferramentas, clique em Novo e escolha a opção de conexão. A janela de conexão do banco aparece.

NOTA Existem algumas maneiras de criar uma conexão de banco de dados além de utilizar a barra de ferramentas.

Para criar uma conexão usando o painel lateral, certifique-se o painel lateral está no modo Visualizar e trazer o menu de contexto da pasta Conexões abaixo do nó que representa a transformação em curso. No menu de contexto, escolha Novo.

As conexões também podem ser criados a partir de dentro as janelas de configuração de qualquer banco de dados relacionados com as etapas de transformação. Essas etapas conter uma caixa de lista para a escolha uma conexão existente, e você pode usar o novo e editar botões, respectivamente, criar ou modificar a conexão configurada.

Finalmente, você pode usar o atalho F3 para iniciar uma conexão com o banco assistente, que oferece um processo passo-a-passo para preencher dados de conexão propriedades.

O lado esquerdo da janela de conexão do banco contém uma lista de categorias dos tipos de coisas que podem ser configurados. Na verdade, especificando essas propriedades é feito no preenchimento de formulário no lado direito da janela. O conteúdo deste forma são sincronizados de acordo com a categoria selecionada no lado esquerdo da diálogo. As categorias são:

- General-Basic propriedades, tais como o tipo de banco de dados, nome do host, porta número, e assim por diante são configurados aqui. O nome da conexão deve também ser especificado aqui.

- Avançado-Este categoria pode ser usado para especificar algumas opções que afetam como identificadores de banco de dados são tratados por todas as etapas usando esta conexão. É
também pode ser usado para especificar uma ação personalizada, sempre que tal ligação
é inicializado.
- Opções A JDBC padrão define um caminho comum para configurar o JDBC propriedades específicas do driver. Estes podem ser especificados aqui.
- Connection Pooling- agrupamento de opções.
- Clustering-Estes opções podem ser usadas para criar um grupo de conexões que são usados em um ambiente de cluster e também com o particionamento.

Veja a Figura 9-14 para obter um exemplo do diálogo Conexão de Banco de Dados.

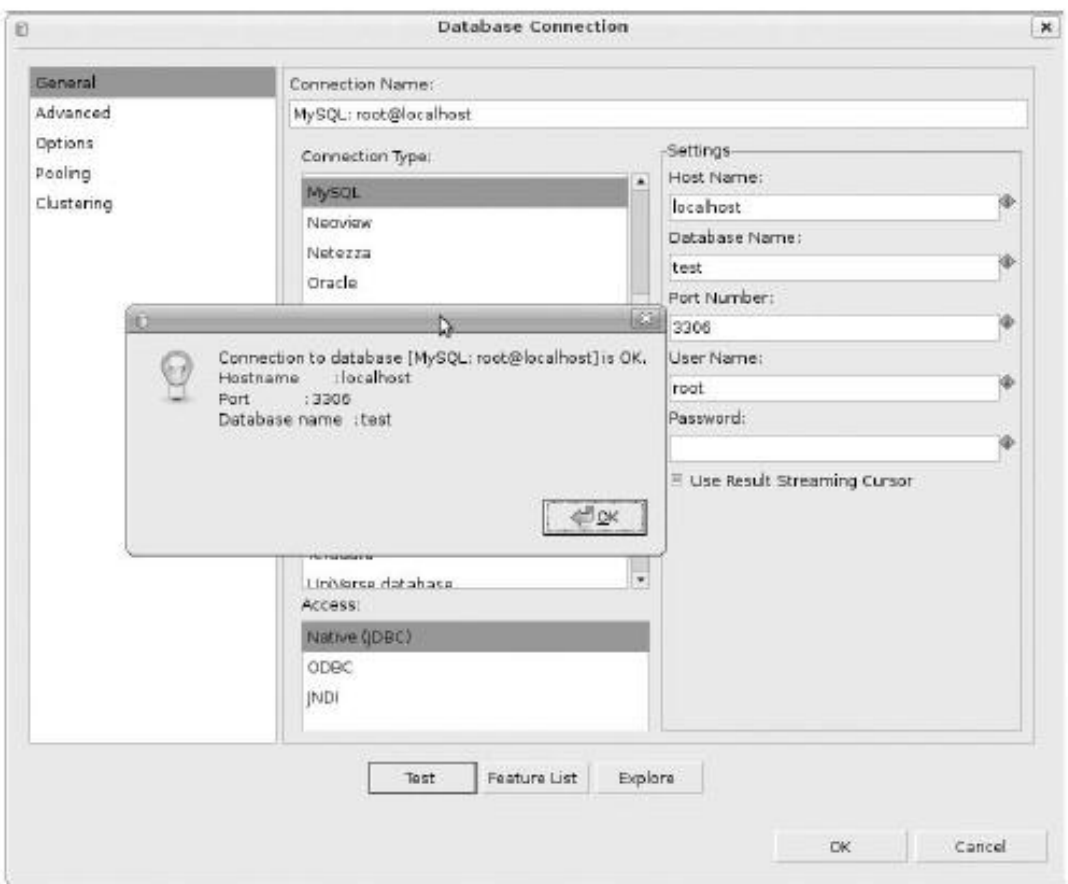


Figura 9-14: Típico opções de configuração para conexões JDBC

Por padrão, a categoria Geral é selecionada. As propriedades nesta categoria são suficientes para todos os casos de uso comum. Iremos descrever as outras categorias

mais adiante neste capítulo. Os seguintes itens são encontrados no lado esquerdo da formulário de propriedades em geral, e deve ser preenchido para cada ligação:

- Nome da conexão-In neste campo, um nome deve ser especificado que identifica neste contexto, dentro da transformação. Passos que exigem uma conexão de banco de dados referem-se a uma conexão usando esta-ção nome.

- Tipo de conexão-Este lista é usada para escolher um dos muitos banco de dados produtos para conectar-se.

- Access-Após você especificar o tipo de conexão, esta lista fornece a métodos de ligação disponíveis para o RDBMS especificado. Na maioria dos casos, um JDBC ODBC e uma opção é mostrada. Para todos os tipos de conexão que suporte JDBC, esta lista também contém uma opção de JNDI. A opção será JNDI ser explicada no capítulo 11.

O lado direito do formulário de propriedades exibe um quadro de configurações que contém opções de configuração que são específicos para o tipo de conexão especificado e método de acesso. conexões JDBC para um RDBMS, que é diretamente suportado geralmente requerem as seguintes propriedades:

- Nome do Host-A nome de domínio ou endereço IP do computador onde o servidor de banco de dados reside. Para um servidor de banco de dados local, este deve normalmente localhost ou 127.0.0.1, mas você também pode tentar deixar o campo em branco.
- O número de porta- host TCP / IP onde o servidor de banco de dados está escutando os pedidos de conexão. Em muitos casos, o valor padrão para o RDBMS's porta é automaticamente preenchidos

- Nome de Usuário e Senha-A credenciais para fazer logon no banco de dados servidor.

Para RDBMSs muitos, isso é tudo o que é necessário para estabelecer uma conexão JDBC. Pentaho Data Integration usa essa informação para carregar o driver apropriado, e para gerar a seqüência de JDBC adequado contato para estabelecer a conexão.

Para conexões ODBC, propriedades como host e credenciais já estão configurada na fonte de dados ODBC. Portanto, as conexões ODBC requerem somente o nome de uma já existente fonte de dados ODBC (DSN).

Às vezes, as opções extras estão disponíveis para um tipo de conexão particular. Para exemplo, para uma conexão Oracle, a forma também fornece caixas de texto para especificar que tablespaces para uso de tabelas e índices. Outro exemplo aparece na Figura 9-14, que mostra o resultado Use Streaming direito checkbox Cursor abaixo da caixa de texto de senha.

Testando conexões de banco de dados

Após configurar a conexão, é uma boa idéia testá-lo. O botão de teste
na parte inferior da caixa de diálogo Conexão de banco de dados pode ser usada para descobrir se
é pelo menos possível para Pentaho Data Integration para se conectar ao especificado
banco de dados. Após pressionar o botão de teste, uma caixa de mensagem aparece. Se o teste
for bem sucedida, a caixa de mensagem deve ser semelhante ao mostrado na Figura 9-14.
Se o teste não for bem sucedida, a caixa de mensagem irá exibir um rastreamento de pilha grande,
que pode ser usado para analisar o problema.

Embora o rastreamento de pilha pode ser assustadora, os problemas se resumem a
erros de ortografia nas credenciais ou erros de digitação no nome do host ou número da porta.
Geralmente, o topo da StackTrace exibe uma mensagem útil fornecer uma justa
chance de resolver o problema.

Como as conexões de banco de dados são usados

conexões de banco de dados são consultados por todas as etapas de transformação ou entradas de emprego
que precisam trabalhar com um banco de dados. As janelas de configuração dessas etapas
e entradas de emprego fornecer alguns meios de estabelecer a conexão, explorando seus
banco de dados associado ou até mesmo criar um novo.

conexões de banco de dados especificado no Pentaho Data Integration são realmente
mais apropriadamente denominado de banco de dados descritores de conexão. Este é um importante
momento porque uma conexão''tal''pode muito bem traduzir em múltiplas
muitas vezes, as ligações reais do ponto do servidor de banco de vista.
Por exemplo, pode haver muitos passos em uma transformação que se referem a
o descritor de conexão mesmo. Quando essa transformação é executada,
todas essas etapas, normalmente aberto sua própria instância da conexão,
cada um dos quais corresponde a uma''ligação''real ao nível RDMBS.
Isto significa que, por padrão, você não pode confiar em mudanças de estado dentro de um
conexão. Um exemplo típico são as operações de banco de dados: você não pode sim-
ply confirmar ou reverter uma transformação, pois cada etapa abre o seu próprio
conexão, e cada um vai usar a sua própria transação, que é completamente inde-
independente das operações mantidas pela conexões abertas pelo
outras etapas.

Agora que identificamos a diferença entre o Pentaho Data Inte-
gração ligações e conexões de banco de dados''real'', não vamos ser exigente em
nossa terminologia. Nós iremos usar a conexão de expressão''''para se referir a uma conexão
objeto do descritor no Pentaho Data Integration, e vamos fazer uma explícita
distinção se faz sentido fazê-lo para o tópico em questão.

Usando transações

Em princípio, as conexões Pentaho Data Integration são apenas descritores-recipes que podem ser utilizados repetidamente por vários passos para estabelecer um real, físico conexão com o banco. Isto significa que uma transformação pode ser ligada muitas vezes ao mesmo tempo usando o descritor mesma conexão. Múltiplas conexões permitem que o trabalho a ser paralelizado, o que é bom para desempenho.

Às vezes, várias conexões físicas não são desejáveis. Por exemplo, Tendo uma conexão física separada para cada passo significa que o estado da sessão (Como o status da transação) não se traduz em várias etapas: cada etapa, ou seja, cada ligação física, tem seu próprio estado (e, portanto, suas próprias transação).

Se você realmente precisar, você pode garantir que cada descritor de conexão realmente corresponde exatamente a uma conexão física. Para fazer isso, verifique a fazer o caixa transformação banco de dados transacional na página separador Diversos da caixa de diálogo Propriedades de Transformação. Para abri-lo, selecione Menu Transforma- Configurações ção.

Um banco de dados habilitado"Olá, Mundo!"Exemplo

Agora que você aprendeu a criar uma conexão de banco de dados, você pode melhorar a"Olá, Mundo! transformação"e adicionar algum suporte de banco de dados. Para manter as coisas simples, você simplesmente adicionar uma etapa que grava o"Olá" mensagens para uma tabela do banco, além do arquivo de saída.

1. Abra o"Olá, Mundo! Transformação exemplo"se ele não estiver aberto. Escolha Menu Transformação Configurações para abrir a Transformação De diálogo Propriedades. Na caixa de texto Nome de Transformação, tipo hello_ Salvar arquivo comoworld_transformation_with_db. Em seguida, escolha Menu e salvar a transformação como hello_world_transformation_with_db . KTR no mesmo diretório onde a transformação original está armazenado.

NOTA Estritamente falando, poderíamos ter criado a cópia da transformação sem editar o nome na janela de configuração de transformação. Dito isto, é uma boa idéia fazer isso porque usa o nome Spoon entrou aqui como a exibição título para a transformação. Ao editar o nome, você pode evitar qualquer confusão quanto ao transformação que você está modificando.

2. Criar uma conexão de banco de dados chamado Target Database. Escolha o SQLite como o tipo de conexão, e Nativo (JDBC) como método de acesso. Na

Configurações frame, tipo <path> / hello_world.sqlite onde <caminho> stands para o caminho do sistema de arquivos, onde a transformação está armazenado. Especifique -1 para
do Porto. Deixe o Nome do Host, Nome de Usuário e Senha campos em branco. Clique no botão Test para verificar a conexão do banco funciona corretamente.

NOTA Optamos por criar uma conexão de banco de dados SQLite para este exemplo. SQLite é um banco de dados incorporado leve relacional. A principal razão para esta escolha é para permitir uma configuração rápida para aqueles leitores que não têm um servidor de banco de dados criado.
SQLite exige apenas a capacidade de criar e modificar um arquivo.

Se você tem um servidor de banco de dados criado, você pode tentar se conectar a esse lugar.

Note que usamos a barra para separar o caminho do banco de dados nome do arquivo hello_world.sqlite. Mesmo que a barra não é normalmente usados em caminhos do Windows, isso, de fato, funciona.

3. Arraste a tabela tipo de etapa de saída da categoria na saída do lado painel e mova-o sobre o salto entre as constantes de Adicionar e arquivo de texto etapas de saída. Se posicionados à direita, a linha usada para representar o salto será engrossar, como mostrado na Figura 9-15, e você será perguntado se deseja dividir o salto. Solte a passo lá. Isto irá inserir a etapa de saída de mesa entre as constantes e as etapas Adicionar Texto de arquivo de saída. A saída hop a partir de Adicionar etapa constantes é automaticamente conectado à entrada de a etapa de saída de mesa, eo hip entrada da etapa de produção de texto do arquivo é igualmente ligado à saída da etapa de saída de mesa, assim, a divisão do hop original. Confirme que você deseja dividir o salto e, opcionalmente, verifique o Não perguntar de novo opção para evitar ser solicitado no futuro.

NOTA A etapa menu de contexto contém a opção Detach, que é o exato oposto da divisão saltos: destaca a etapa da sua lúpulo, sem descartar se o lúpulo.

Figura 9-15: Dividindo uma hop existentes

4. Abra a janela de configuração da nova etapa de saída da tabela. Nota que a caixa de listagem de conexão já está definido para o nome do banco de dados

conexão. Na caixa de texto tabela de destino, tipo de hello_world_output. Em seguida, pressione o botão SQL na parte inferior da janela de configuração. A Simple SQL Editor abre contendo uma instrução SQL CREATE TABLE declaração, como mostrado na Figura 9-16, para criar o hello_world_output tabela. Note-se que você pode editar o mapa gerado para o conteúdo do seu coração. Pressione o botão Botão Executar para criar a tabela. Outra janela se abre, o que deve indicar a instrução foi executada com sucesso. Fechá-lo, e também perto do SQL janela do editor.

DICA Você também pode chamar o editor SQL diretamente a partir do contexto da conexão menu, que pode ser levantada com o botão direito do mouse a entrada de conexão no barra lateral.

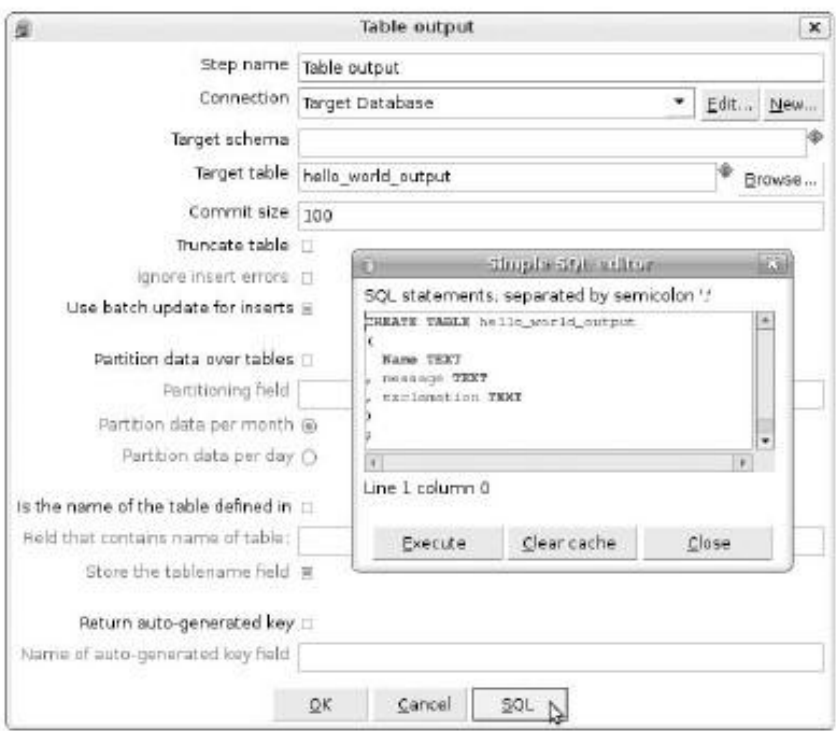


Figura 9-16: Gerando declarações CREATE TABLE

5. Executar a transformação. Verifique o painel de resultados para a execução de quaisquer erros.
   A tabela agora deve conter linhas.
6. Na barra lateral, alternar para o modo de Exibir e clique com o alvo Banco de dados de conexão. No menu de contexto, escolha a opção Explorar. Isto abre a janela Database Explorer. Na vista de árvore, abra o Nó Tables e clique no hello_world_output tabela para selecioná-lo. Em seguida, clique na primeira prévia de 100 linhas botão para examinar o conteúdo da tabela. Uma janela de visualização é aberta, e você deve ver algo semelhante ao que é mostrado na Figura 9-17.

DICA O SQL Explorer também pode ser chamado de dentro da configuração passo diálogos de todos os passos dados relacionados.

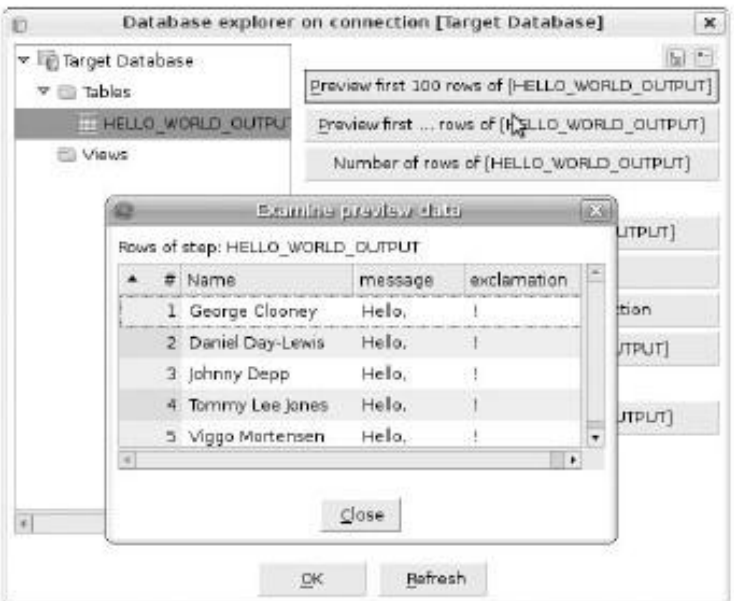


Figura 9-17: Visualizando dados da tabela através do Database Explorer

Banco de Dados Gerenciamento de Configuração de Conexão

conexões de banco de dados fazem parte de transformações e de emprego. Isso significa que uma conexão pode ser reutilizado dentro, mas não através de transformações e / ou postos de trabalho. Isso representa algo como um problema de manutenção na concepção de uma base de dados
solução de integração que consiste em várias transformações e emprego (que é quase sempre o caso).

Felizmente, existe uma maneira de lidar com isso. Transformações e os trabalhos são associado a um arquivo compartilhado de objetos. O compartilhamento de arquivos objetos é armazenado no local
máquina, e pode ser usado para armazenar conexões de banco de dados (entre outras coisas).
Todos os objetos no arquivo de objetos compartilhados são automaticamente disponíveis para todos os trabalhos e
transformações.

NOTA Por padrão, Spoon usa um único arquivo de objetos compartilhados por usuários do sistema operacional. Este arquivo é
chamada shared.xmlI e armazenados no . Chaleira diretório abaixo home do usuário diretório. O arquivo compartilhado de objetos pode ser configurado na página da guia Misc Propriedades de transformação ou propriedades diálogo Job, respectivamente.

Para armazenar uma conexão de dados no arquivo de objetos compartilhados, abra o confio Bluetooth menu de contexto e escolha a opção Compartilhamento. O nome da conexão agora aparece em negrito na barra lateral. Neste ponto, você deve salvar o transformação.

ATENÇÃO Infelizmente, o ato de compartilhar a conexão não é em si considerado como uma mudança. Isso significa que, se não houver alterações não salvas antes compartilhando a conexão, logo após salvar a transformação partilha será não guardar a conexão com o arquivo de objetos compartilhados.

Você pode ver se acha o trabalho de colher ou transformação mudou. Uma forma de descobrir é olhar para a barra de título da aplicação. Se que exibe o texto"(alterado)" após a legenda documento, as alterações serão salvas. Outra maneira é verificar se o rótulo na aba da página é exibido em negrito.

Se você estiver usando o banco de dados do repositório, o problema é discutível. O repositório
lojas de conexões de banco de dados global. Todas as transformações e do emprego num repositório pode acessar automaticamente todas as ligações, mesmo se elas foram criadas de outro emprego ou transação.

Conexões de banco de dados genéricos

Se você deseja se conectar a um RDBMS para as quais não suporte embutido está disponível na colher, você pode tentar estabelecer uma conexão de banco de dados genéricos. Neste caso,
primeiro você deve obter um driver JDBC para o RDBMS. Estes geralmente podem ser obtidos a partir do fornecedor de banco de dados, muitas vezes, sem nenhum custo extra.

Os drivers JDBC são normalmente distribuídos como Jar. arquivos (arquivos Java). Antes você pode usar o Jar. arquivo (ou melhor, o seu conteúdo), você deve garantir Pentaho Integração de dados pode carregá-lo. Isto é conseguido através de cópia do Jar. arquivo para o
libext / JDBC diretório, que está localizado abaixo do diretório de instalação do Pentaho Data Integration.

Para criar a conexão, abrir a janela de conexão do banco, como é costume maneira. Selecione a entrada de genéricos no banco de dados lista Tipo de conexão, e selecionar Nativa (JDBC) na lista de acesso. As seguintes propriedades são disponíveis no quadro de configurações:

- Customer Connection URL-Este é onde você especifica o JDBC contato string (chamado de URL JDBC terminologia). O formato deste URL é jdbc: string> <driver específicas. Consulte a documentação do driver específico que você deseja usar para descobrir exatamente o que o formato é de a seqüência específica do driver.

- Custom Class Name Driver-A driver JDBC real é uma classe Java. A nome completo qualificado (nome do pacote, seguido por um ponto, seguido do nome da classe) desta classe devem ser especificados aqui. Como o formato do URL, essa informação também é específica do driver. Você deve procurá-lo na documentação fornecida com o motorista.

- Nome de usuário e senha credenciais para fazer logon no RDBMS. Estes dados não são específicas do driver. Eles correspondem a uma conta em seu banco de dados do sistema.

Figura 9-18 mostra um exemplo de uma conexão JDBC genérico para LucidDB.

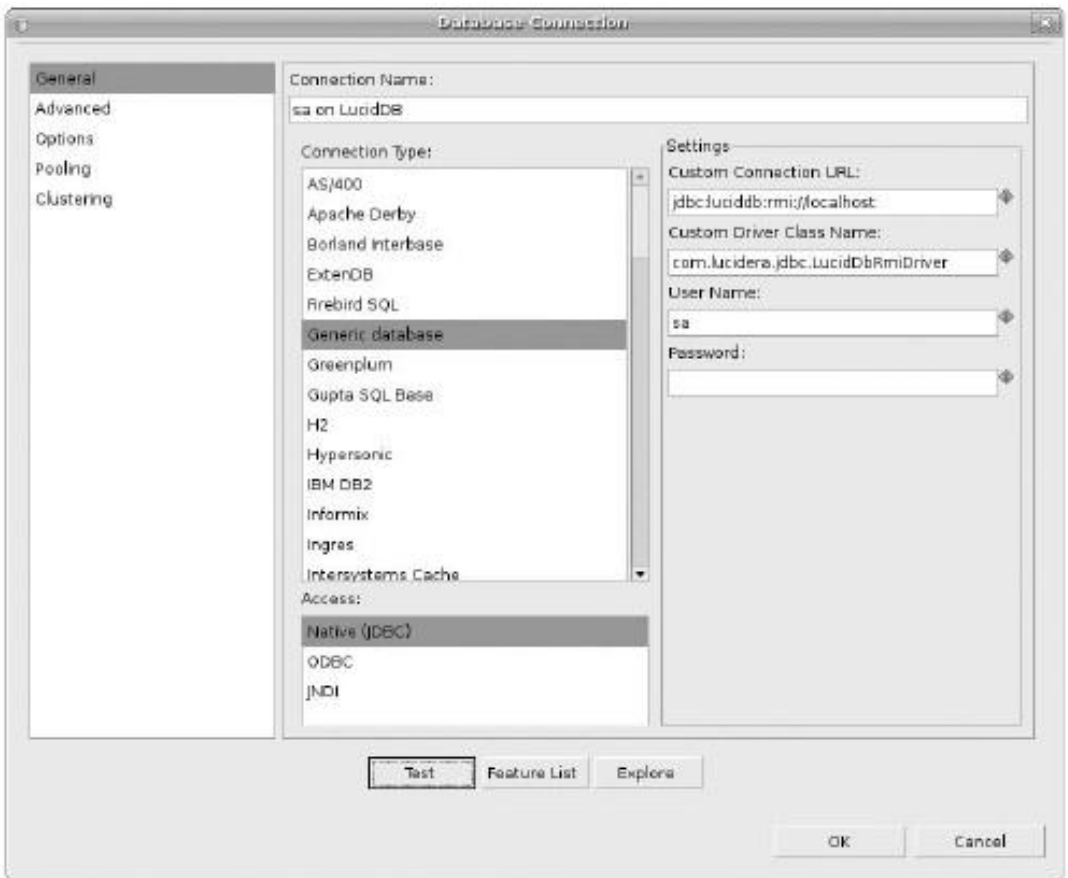


Figura 9-18: Uma conexão JDBC genérico para LucidDB

Também é possível utilizar o ODBC como um método de acesso para o banco de dados genéricos
conexão. Isso requer apenas o nome de um dos dados já configurado ODBC
fonte.

Resumo

Neste capítulo, você aprendeu sobre os processos de ETL e atividades envolvidas no preenchimento do data warehouse. Além disso, você aprendeu sobre um número de actividades de apoio, tais como a captura de mudança de dados, validação de dados, e decodificação e renomeação.

Introduzimos também Pentaho Data Integration ferramenta, também conhecida como Kettle. Você aprendeu que o PDI consiste em um mecanismo de integração de dados que podem executar tarefas
e transformações. Além disso, o PDI apresenta uma série de ferramentas e utilidades para criar e executar trabalhos e transformações.

Transformações são orientadas a dados e consiste de etapas, que são operados no registro córregos. etapas de transformação podem ser conectados através de saltos, permitindo que os registros de fluxo de uma etapa para a outra. Os trabalhos são processuais na natureza, e consistem em entradas de emprego, que são executados seqüencialmente tarefas.
Normalmente, os trabalhos de conter transformações, entre outros tipos de medidas.

Você também aprendeu os conceitos básicos de colher, que é a ferramenta para criar PDI emprego e transformações. Você construiu o seu primeiro Mundial''Olá,!''Transformação em Spoon. Depois disso, você aprendeu sobre as conexões de banco de dados, e prorrogado o''Olá Mundo! transformação''para armazenar o resultado em um banco de dados.

Você deve agora ter uma base razoável para o Capítulo 10, em que você vai aprender a usar esse conhecimento e habilidades para encher o World Class Filmes data warehouse.

CAPÍTULO

10

Pentaho Data Projetando Soluções de Integração

No capítulo anterior, apresentamos um conjunto de ferramentas e utilitários que fazem até Pentaho Data Integration (PDI). Nós demos um pouco de atenção extra para os dados integração Spoon ferramenta de design, e deu duas orientações para familiarizar o leitor com sua interface de usuário. Neste capítulo, você aprenderá como colocar essas habilidades para utilizar para construir as transformações e os trabalhos que são usados para preencher o Mundo
Classe armazém de dados de filmes.

Este capítulo oferece uma abordagem prática, com foco em problemas comuns e soluções práticas. Apesar de usar o World Class data warehouse Filmes como exemplo, acreditamos que a maioria dos problemas encontrados são comuns a a maioria dos armazéns de dados. Os métodos utilizados neste capítulo são de modo algum a única solução possível. Ao contrário, optaram por uma série de abordagens para familiarizá-lo com os recursos mais necessários do Pentaho Data Integration.

A melhor maneira de ler este capítulo é para baixar todas as transformações PDI e emprego a partir do site deste livro em www.wiley.com / go / pentahosolutions assim você pode facilmente abri-los, verificar todos os detalhes, execute um preview, ou mais.

NOTA Além dos exemplos e exercícios discutidos neste capítulo, você está encorajados a dar uma olhada nas amostras incluídas com PDI. Você pode encontrá-los em o amostras diretório que está localizado no diretório home do PDI. A amostras próprio diretório contém um empregos e um transformações subdiretório contendo amostra de trabalhos e transformações. A amostras diretório também contém um mapeamento diretório que ilustra o uso de subtransformations.

Para além destes exemplos, você também deve ter um olhar para os dados Pentaho Integração documentação. A referência a Passo wiki.pentaho.com / exposição / EAI / Pentaho Data Integration + + + Passos ea referência de trabalho na entrada

mostrar wiki.pentaho.com / EAI / Pentaho Data + + + Integração Trabalho
+ Entradas são recursos especialmente grande para obter novas idéias para a construção de PDI soluções.

Tabela Gerando Dimension Data

Quase todos os projetos de data warehouse contém tabelas de dimensão poucos que podem, em grande medida, ser preenchida com os dados gerados. Um exemplo bem conhecido é a dimensão de data. Também conhecida, mas menos utilizadas são o tempo e dimensões da demografia.

Nesta seção, primeiro discutir brevemente a prática de utilizar banco de dados armazenados procedimentos para carregar esses tipos de tabelas de dimensão. Nós, então, rapidamente mover-se sobre
e explicar como fazer isto usando o Pentaho Data Integration para esta tarefa.
Isso proporciona uma grande oportunidade para conhecer melhor Spoon e um
algumas de suas etapas de transformação de base. Isto irá estabelecer uma base importante para
seções subseqüentes.

Usando Stored Procedures

Um método comum para gerar dados para uma dimensão de data é a de criar um banco de dados
procedimento armazenado que usa funções de data do RDBMS alvo.

NOTA Uma desvantagem de usar um procedimento armazenado do banco de dados é que ele precisa ser
escrito mais e mais para cada RDBMS destino diferente, como há geralmente
diferenças sintáticas com relação a ambos os idiomas o procedimento armazenado ea
built-in funções de data e hora. Por esse motivo (e outros explicado mais adiante neste capítulo), na verdade nós não recomendamos essa abordagem, mas por uma questão de integralidade que inclua o tema de qualquer maneira.

A título de exemplo, um procedimento simples MySQL armazenados para carregar uma dimensão data
Sion é mostrado na Lista 10-1. Neste caso, a função interna DATE_FORMAT
(Em negrito) é usado para gerar múltiplas representações do valor de data.

Listagem 10-1: Um simples (parcial) MySQL procedimento armazenado para carregar uma dimensão de data

```
CREATE PROCEDURE p_load_dim_date (
    p_from_date DATA
, P_to_dateDATE
)
BEGIN
    DECLARE v_date DATE DEFAULT p_from_date;
    ENQUANTO v_date <p_to_date DO
        INSERT INTO dim_date (
            date_key
        , Data
```

```
      , Date_short
      , ...
      ) VALUES (
         v_date + 0
      , V_date
      ,DATE_FORMAT (v_date, '% Y-%% c-d')
      , ...
      );
      SET v_date: = v_date + intervalo de 1 dia;
   Fim Enquanto;
END;
```

Carregando um Dimension Data Simples

Todos os problemas de compatibilidade causados pelo uso de procedimentos de banco de dados armazenados podem
ser superado simplesmente evitá-los. Isso pode parecer mais fácil do que realmente é, mas para o carregamento de uma dimensão de data que o processo é bastante simples e fácil de implementar diretamente no Pentaho Data Integration.

Para evitar complexidade desnecessária, primeiro iremos usar um pouco simplificado versão da tabela de dimensão de data. Sua CREATE TABLE declaração é apresentada na Lista 10-2. Como você pode ver, esta tabela de dimensão de data simplesmente fornece uma
número de diferentes formatos para uma data.

Listagem 10-2: CREATE de uma tabela simplificada data dimensão

```
CREATE TABLE dim_date (
  date_key                INTEGER       NÃO  NULL,
  date_value              DATA          NÃO  NULL,
  date_short              VARCHAR (12)  NÃO  NULL,
  date_medium             VARCHAR (16)  NÃO  NULL,
  date_long               VARCHAR (24)  NÃO  NULL,
  date_full               VARCHAR (32)  NÃO  NULL,
  day_in_week             SMALLINT      NÃO  NULL,
  day_in_year             SMALLINT      NÃO  NULL,
  day_in_month            SMALLINT      NÃO  NULL,
  is_first_day_in_month   VARCHAR (10)  NÃO  NULL,
  is_last_day_in_month    VARCHAR (10)  NÃO  NULL,
  day_abbreviation        CHAR (3)      NÃO  NULL,
  day_name                VARCHAR (12)  NÃO  NULL,
  week_in_year            SMALLINT      NÃO  NULL,
  week_in_month           SMALLINT      NÃO  NULL,
  is_first_day_in_week    VARCHAR (10)  NÃO  NULL,
  is_last_day_in_week     VARCHAR (10)  NÃO  NULL,
  is_weekend              VARCHAR (3)   NÃO  NULL,
  month_number            SMALLINT      NÃO  NULL,
  month_abbreviation      CHAR (3)      NÃO  NULL,
  MONTH_NAME              VARCHAR (12)  NÃO  NULL,
  ano2                    CHAR (2)      NÃO  NULL,
```

```
    ano4                        CHAR (4)        NÃO  NULL,
    quarter_name                CHAR (2)        NÃO  NULL,
    quarter_number              SMALLINT        NÃO  NULL,
    year_quarter                CHAR (7)        NÃO  NULL,
    year_month_number           CHAR (7)        NÃO  NULL,
    year_month_abbreviation     CHAR (8)        NÃO  NULL,
    PRIMARY KEY (date_key),
    UNIQUE KEY (date_value)

);
```

Figura 10-1 mostra uma transformação tão simples. A transformação é dividido em quatro partes distintas: Preparar, entrada e saída de Transformação, e na Figura 10-1 você pode ver algumas notas para indicar isso.

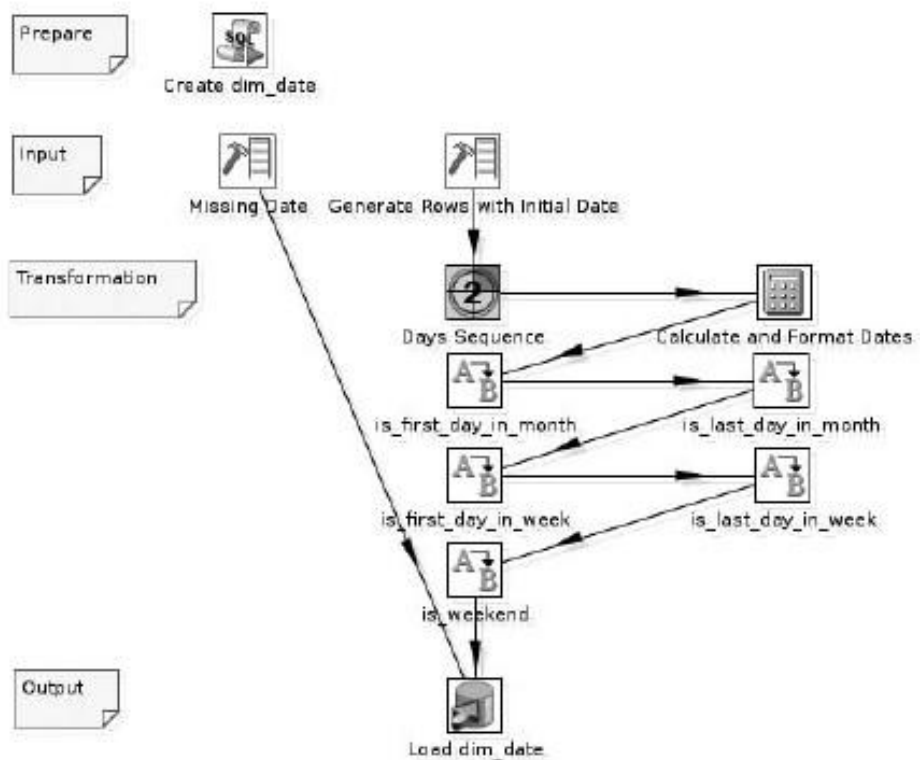


Figura 10-1: Uma transformação simples para carregar uma tabela de dimensão de data

NOTA Você pode colocar anotações em qualquer lugar na tela de transformação ou de trabalho. Embora as notas não são um substituto para a documentação, eles são um conveniente meios para esclarecer a intenção da transformação.

Para adicionar uma nota, clique com o botão direito sobre a tela para abrir o menu de contexto. Escolha Novo
Observe no menu de contexto. Uma caixa de diálogo na qual você pode digitar o texto para sua nota. Quando terminar de editar, clique em OK para fechar a janela e coloque a nota sobre

a tela. Uma vez que a nota é colocada na tela, você pode usar o mouse para arrastar para um novo local. Para editar uma nota existente, clique duas vezes nele. Para removê-lo, botão direito do mouse sobre a nota e escolha nota Excluir no menu de contexto.

Aqui está um resumo do que acontece nessa transformação:

1. Criar dim_date: Questões as instruções SQL necessárias para o alvo banco de dados para criar a dim_date Dimensão da tabela.

2. Faltando Data: Gera uma linha que representa uma data especial que é usado para representar todas as datas em falta ou inaplicável.

3. Gerar linhas com Data Inicial: Gera uma linha por dia de calendário. As linhas têm um campo que especifica a data inicial.

4. Dias Seqüência: Adiciona um campo inteiro para incrementar as linhas geradas pela etapa anterior.

5. Calcular e Formato Datas: O inteiro incrementado é adicionado à data inicial, resultando em uma seqüência de dias de calendário.

6. is_first_day_in_month, is_first_day_in_week, is_last_day_in_month, is_last_day_in_week e is_weekend: Essas etapas mapa de dados numéricos ao texto que está a ser armazenado na tabela de dimensão.

7. Carga dim_date: As linhas de fluxo de entrada são inseridos no dim_date tabela.

Os capítulos seguintes analisam as etapas de transformação e suas configuração em detalhes.

CREATE TABLE dim_date: Utilizando o Execute SQL Script Etapa

Na Figura 10-1, a etapa Criar dim_date é do tipo Execute o script SQL. Este tipo de medida pode ser encontrado na categoria Scripts; seu ícone (mostrado na parte superior da
Figura 10-1) aparece como um pergaminho com uma etiqueta de SQL.

A etapa de obras, executando os comandos SQL especificado no script propriedade contra a conexão especificada. Você pode usar a propriedade script para especificar uma ou várias instruções SQL, separados por ponto-evírgula. Não há nenhuma restrição no que diz respeito à forma ou tipo de instruções SQL que compõem o roteiro, enquanto eles são sintaticamente válida de acordo com a RDBMS subjacente a conexão especificada. Etapas do tipo Executar SQL são especialmente úteis para a execução de comandos DDL.

Figura 10-2 mostra como a etapa foi configurado para a amostra de transção mostrado na Figura 10-1. Como você pode ver na figura, o script contém duas demonstrações: uma DROP TABLE declaração a cair na tabela em caso de ele já e existe uma CREATE TABLE declaração para definir a tabela. A lógica por trás

soltando e depois de (re) criar a tabela é que isso torna mais fácil adicionar ou remover colunas.

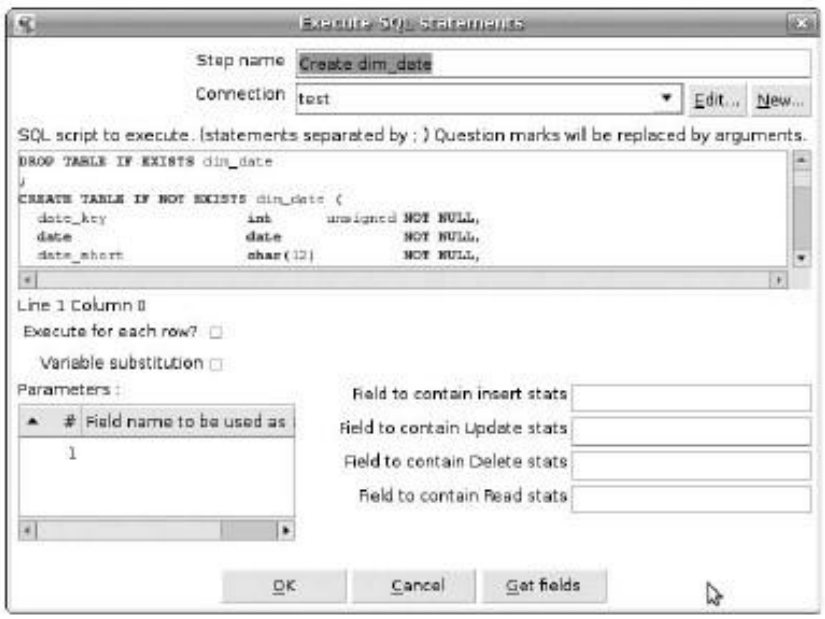


Figura 10-2: Configurando o SQL Execute passo para executar múltiplas instruções SQL

A etapa de execução de script SQL é um pouco mais de um pato estranho quando comparado a outros
tipos de etapas de transformação, porque não é particularmente orientadas a dados. Por padrão, as etapas deste tipo executar apenas uma vez antes de qualquer das outras etapas são executados.

NOTA Ao executar uma transformação, todos os passos são primeiramente inicializados. Isso é chamado o inicialização fase. As acções preparatórias, como a abertura de arquivos e banco de dados conexões, mas também substituindo as variáveis (ver Capítulo 11 para mais informações sobre as variáveis) e preparar declarações de banco de dados geralmente são executadas na inicialização fase. Após todos os passos foram iniciados, as etapas entre a fase de execução e iniciar linhas de processamento.

Por padrão, o Executar etapa SQL executa instruções SQL apenas uma vez no inicialização fase. Por esta razão, a transformação de exemplo mostrado na Figura 10-1 não precisa de nenhum hop para conectá-lo ao restante da transformação: o SQL será executado antes de outras etapas do transformação iniciar linhas de processamento de qualquer maneira. É possível conduzir a execução do SQL
usando dados de um fluxo de entrada. Um cenário como este é discutido em detalhe mais tarde neste capítulo na seção"pesquisa"Staging Valores.

Observe que as etapas entrar na fase de inicialização de forma assíncrona, e não particular da ordem. Isso significa que você deve garantir que o trabalho feito na inicialização fase de um passo em particular não dependem de concluída a fase de inicialização

de outro passo. Por exemplo, o DROP e CREATE declarações no Criar
etapa dim_date na Figura 10-1 não pode ser executado separadamente cada um no seu próprio
etapas de execução do SQL, pois não haveria garantia de que o DROP declaração
seria executado antes do CREATE declaração.

Falta de data e gerar linhas com a data inicial: O Gere Passo Linhas

Na Figura 10-1, faltando as etapas de data e gerar linhas com data inicial são da Gerar linhas do tipo degrau. Esse tipo é encontrado na categoria de entrada e seus ícone aparece como um martelo ao lado de uma escada (ver a linha de entrada da figura 10-1).

Sem surpresa, o objetivo desta etapa é gerar linhas. As etapas deste tipo de são normalmente configurados definindo um valor para a propriedade para especificar o limite número de linhas para gerar. Opcionalmente, um número de campos com constante Os valores podem ser especificados. Figura 10-3 mostra como as linhas com Gere inicial Data passo da transformação mostrado na Figura 10-1 foi configurado.

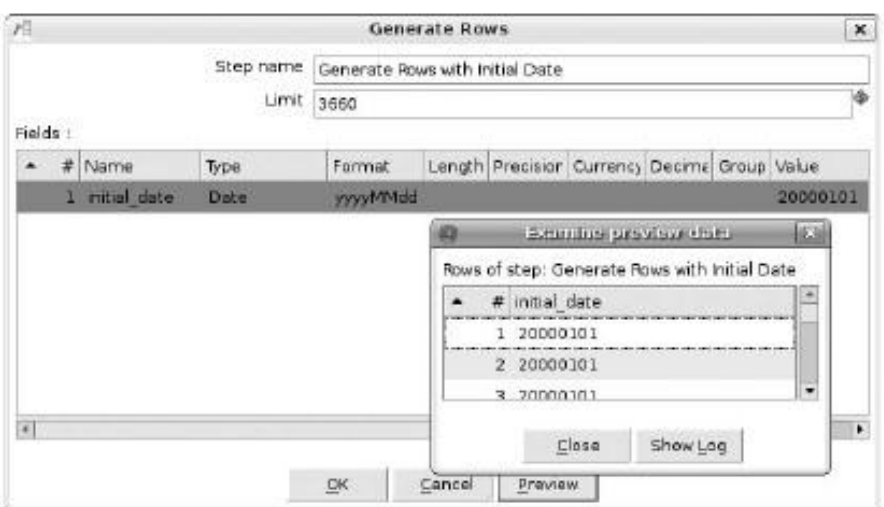


Figura 10-3: Configurando o passo Gerar linhas para obter 10 anos de valor dos dias

Para as linhas com a etapa Gerar data inicial, que especificou um limite de 3660 (Que é um pouco mais de 10 anos no valor de dias). Na grade de Campos, que especificado um único campo chamado initial_date, Que serve como a data de início para a dimensão de data. Observe que especificamos o Data como tipo, o valor 20000101E formatar o aaaammdd. O valor representa a data do calendário de 1 de janeiro de 2000. O tipo e formato são necessárias para explicar a Spoon como interpretar esta data. Você vai encontrar muitos outros exemplos de data formatação na subseção"a etapa Selecionar Valores"mais tarde nesta seção.

Para a etapa de Data Desaparecidas, que especificou um limite de 1 (um). Em vez de especificar apenas um campo para a data inicial, nós adicionamos campos que representam um

data em falta ou inaplicável. Para o date_value e date_key domínios, foi utilizado os valores 0001-01-01 e 10101 , respectivamente. Quando você carrega as suas tabelas de verdade,
Você deve se lembrar de usar essa chave para qualquer NULL datas. Para os campos de texto desta data especial utilizou-se o rótulo Não Aplicável,N / AOu NA dependendo sobre o espaço disponível.

DICA Digitar todos os campos para a etapa de Data ausentes podem ser um pouco entediante. Você pode
poupar tempo e esforço por criar a primeira parte genérica do transformação e, em seguida, acrescentando os passos para inserir essas linhas especiais.

Depois de ter definido o destino final de carga dim_date etapa, você pode clicar com o botão direito sobre ele
e escolha a opção Mostrar campos de entrada no menu de contexto. Uma grade aparece com os campos de fluxo de entrada. Clique na primeira linha da grelha para selecioná-lo e, em seguida selecionar todas as linhas usando as teclas Ctrl + A atalho de teclado. Em seguida, copiar a grade para o clipboard usando Ctrl + C. Você pode então colar que na grade de Campos do desejado Gerar etapa linhas usando Ctrl + V.

Este método, pelo menos, ajudar a preencher os nomes de campo e tipos. Você pode precisar para remover alguns dados colados das colunas restantes e ainda precisa digite os valores campo. No entanto, ele geralmente lhe poupar algum tempo e esforço, e mais importante, evitar erros, porque este método garante que o layout de registro o fluxo de saída irá coincidir com a etapa de destino.

Depois de configurar a etapa Gerar linhas, você pode clicar no botão Preview Para ver um exemplo das linhas que serão gerados na visualização Examine dados de diálogo (veja a Figura 10-3).

Dias Seqüência: A etapa Seqüência Adicionar

Os Dias etapa seqüência na Figura 10-1 é de Dê seqüência tipo. Este tipo de do passo pode ser encontrado na categoria Transform. As etapas deste tipo de trabalho por adição de um campo inteiro de novo incremento para as linhas de fluxo de entrada. Sua ícone aparece como um círculo número 2 (ver Figura 10-1).

Na transformação da amostra, o objetivo deste campo é para gerar os dados para calcular uma seqüência de datas no calendário. (O cálculo real é discutido na próxima subseção.) A etapa seqüência Adicione pode ser configurado para trabalhar em qualquer uma das seguintes formas:

- Para desenhar o próximo valor a partir de uma seqüência de base de dados. Alguns RDBMSs (tais
  como Oracle e PostgreSQL) oferecem um esquema seqüência especial de objeto que é projetado especialmente para proporcionar substituto valores-chave, que pode ser utilizados pela etapa de Seqüência Adicionar.
- Para incrementar um contador mantido pelo mecanismo de transformação.

Nosso uso do tipo de seqüência Adicionar os dias passo seqüência, usa o último opção. A configuração da etapa de seqüências de dias é mostrado na Figura 10-4.

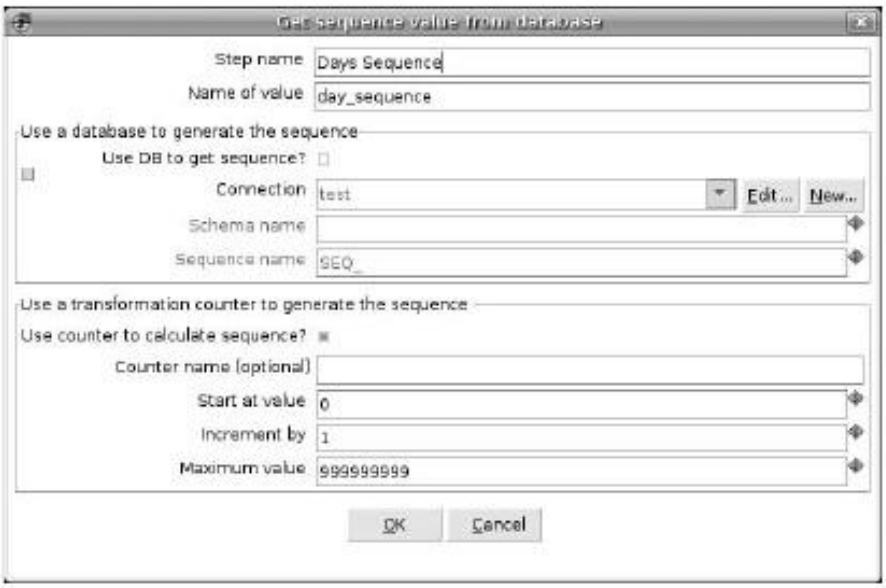


Figura 10-4: Configuração da etapa de seqüências de dias

Em Nome do valor da propriedade, o nome day_sequence é fornecido para o campo que irá conter o número de seqüência. O Uso contador para gerar o caixa sequencial é selecionado para usar um contador mantido pela transção do motor. Na transformação da amostra, a propriedade nome opcional Counter é deixado em branco. Se necessário, poderá ser preenchido para forçar a seqüência de múltiplos Adicionar
passos para tirar um e mesmo balcão. Para o início de valor, o valor 0(Zero) foi especificado explicitamente (em vez do padrão de um) para assegurar que a data inicial especificada na etapa Linhas Gere é realmente a data de início para a dimensão de data.

Calcular e formatar datas: a Etapa Calculadora

Na Figura 10-1, a calcular e formatar passo Datas usa um Calculadora passo para gerar todas as representações data para carregar a dimensão de data. A calculadora passo é encontrada na categoria de Transformação. Não é novidade que o ícone representa uma calculadora de mão.

A etapa Calculator permite que você escolha a partir de um número de précálculos. Dependendo do tipo de cálculo, que utiliza até três campos como argumentos, gerando um novo campo para o resultado, que é adicionado à saída de fluxo. Os campos de argumento pode ser retirado do fluxo de entrada, mas Também a partir dos campos de saída gerado na etapa calculadora mesmo, permitindo-lhe a pilha de cálculos em cima uns dos outros.

Na transformação da amostra, o passo da calculadora está configurado como mostra a Figura 10-5.

Na grade de Campos, uma série de expressões são criadas para calcular várias representações da data do calendário.

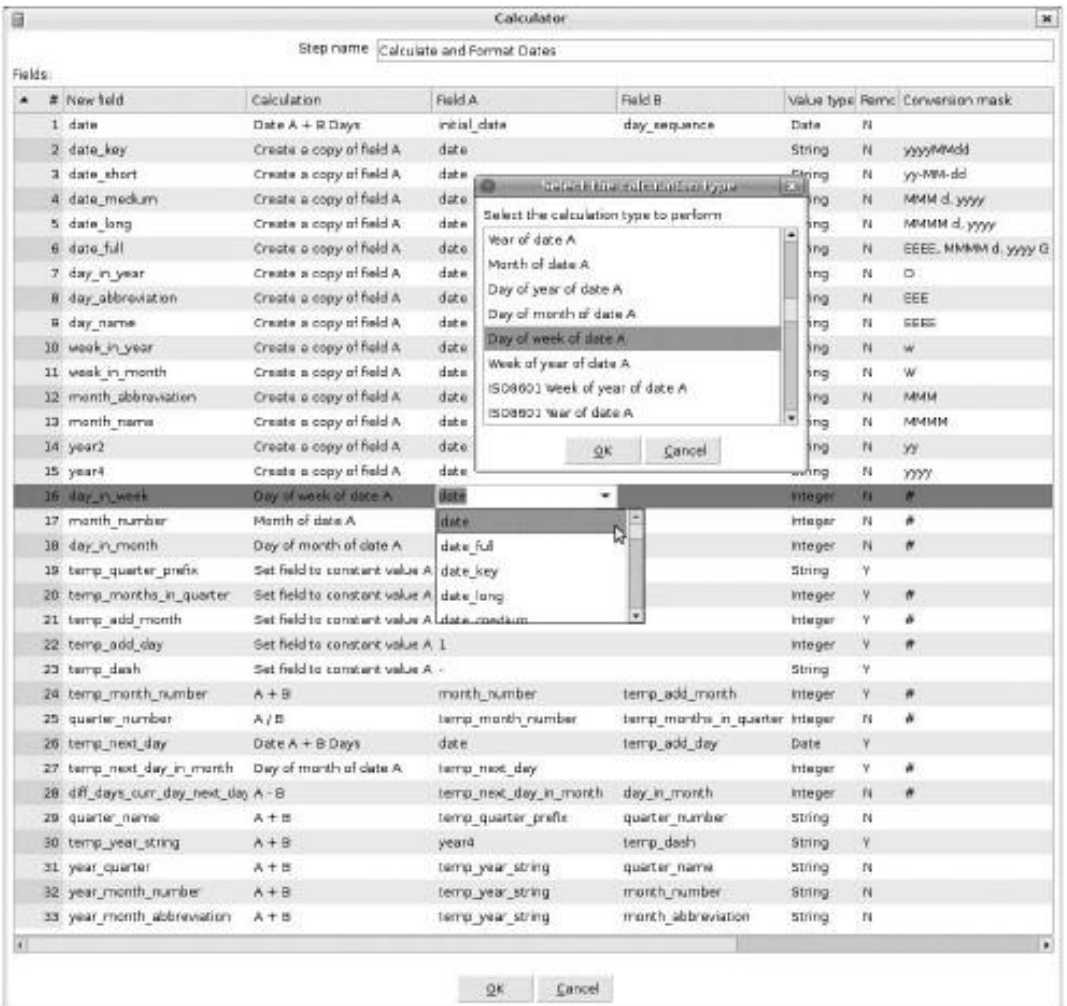


Figura 10-5: Usando a calculadora para calcular passo dias

Cálculo # 1 usa a expressão Dias Data A + B para adicionar o valor do day_sequence campo gerado pela etapa de Seqüência Adicionar ao initial_date campo que foi especificado na etapa Linhas Gerar. Isso gera um novo campo chamada data, Que representa a data do calendário. Este data campo é usado pelos cálculos a seguir para gerar representações data alternativa. Para integralidade, a Listagem 10-3 mostra como este cálculo foi configurado.

Listagem 10-3: Adicionando a seqüência até a data inicial para a obtenção dias

```
Nova fieldCalculationField aField BTYPE
---------------------------------------------------- --------------------
Data date_value A + B Daysinitial_dateday_sequenceDate
```

Cálculos # 2 a # 15 na verdade, não fazer um cálculo adequado.
Pelo contrário, estes usam o Crie uma cópia de um campo de expressão para duplicar a date_value campo que é o resultado de cálculo # 1 e aplicar um determinado formato para convertê-lo para um valor String. Isso gera uma série de alternativas

representações da data do calendário, aos quais são atribuídos um nome de acordo com o nome da coluna de destino. A seqüência de formato é celebrado a máscara de conversão coluna da grade de Campos. Para conversões de data para string, a máscara deve ser especificada usando a notação também utilizado para o java.text.SimpleDateFormat Java classe. A configuração desses campos é mostrada em detalhes na Lista 10-4.

Listagem 10-4: Usando seqüências de formato para formatar datas

```
Nova fieldCalculationFld máscara AConversion
------------------------------------------------- ----------------------
date_keyCreate uma cópia de um campo date_value aaaammdd
date_shortCreate uma cópia de um campo date_value aa-mm-dd
date_mediumCreate uma cópia de um campo date_value MMM yyyy d,
date_longCreate uma cópia de um campo date_value MMMM yyyy d,
date_fullCreate uma cópia de um campo date_value EEEE, MMMM d, yyyy G
day_in_yearCreate uma cópia de um campo date_value D
day_abbreviationCreate uma cópia de um campo de EEE date_value
day_nameCreate uma cópia de um campo de EEEE date_value
week_in_yearCreate uma cópia de um campo w date_value
week_in_monthCreate uma cópia de um campo date_value W
month_abbreviation Criar uma cópia de um campo de date_value MMM
month_nameCreate uma cópia de um campo date_value MMMM
year2Create uma cópia de um campo aa date_value
year4Create uma cópia de um campo aaaa date_value
```

DICA Para obter mais informações sobre a notação das máscaras de formato de data, dê uma olhada http://java.sun.com/j2se/1.5.0/docs/api/java/text / SimpleDateFormat.html.

Cálculo usa o # 16 Dia da semana da data de um expressão para calcular um valor inteiro correspondente ao dia útil para gerar os valores para o day_in_week coluna da tabela de dimensão de data. Para garantir o resultado está formatado como um número inteiro, definir a máscara de conversão para o símbolo hash (#). Não faria sentido calcular isso simplesmente aplicando uma adequada seqüência de formato, tal como fizemos para expressões # 2 a # 16. Infelizmente, a seqüência padrão de formatação apropriadas não parece existir, razão pela qual nós usamos um cálculo para fazer o trabalho.

Cálculo # 17 calcula o month_number de campo, utilizando a expressão Mês da data de um sobre o data de campo. Embora esse valor poderia ter sido obtida através da conversão para o tipo String e aplicando a seqüência de formato M, o Mês da data de um expressão recupera o valor como uma expressão inteira, o que nos permite usá-lo para calcular o número trimestre em cálculos # 24 e # 25. Seria bom para calcular trimestres, também, usando uma seqüência de formato, mas assim como o day_in_week número, este não é suportado.

Cálculo # 18 calcula o day_in_month campo utilizando a expressão Dia do mês da data de um. Como o cálculo de n º 17, que poderíamos ter feito isso, basta Convertendo para string e aplicando o dseqüência de formato, mas usando este expressão, obtemos o valor como um inteiro, o que nos permite (indiretamente)

descobrir se o dia atual passa a ser o último dia do mês em cálculos # 26 a # 28.

A configuração dos cálculos de # 16, # 17 e # 18 é mostrado na Lista 10-5.

Listagem 10-5: Obtendo o dia da semana eo número de meses usando predefinidos cálculos

```
Nova fieldCalculationField máscara ATypeConversion
---------------------------------------------------- --------------------
day_in_weekDay da semana da data Adate_value Integer #
month_number mês da data Adate_value Integer #
Dia day_in_month do mês da data date_value A Integer #
```

Os cálculos""# 19 a # 23 use o Defina o campo de valor constante Aexpressão para definir uma série de valores constantes que são utilizadas no final alguns cálculos. Porque nós só precisamos temporariamente estes cálculos, vamos definir o Remover propriedade para Y, O que impede os campos resultado destes cálculos a ser adicionado ao fluxo de saída. Você pode rever a configuração destes cálculos na Lista 10-6.

Listagem 10-6: Definindo constantes na etapa Calculadora

```
Nova fieldCalculationFld ATypeRmv
---------------------------------------------------- ----------------------------
temp_quarter_prefixSet campo para o valor constante AQStringY
campo Definir temp_months_in_quarter a constante valor A3Integer Y
temp_add_monthSet campo para o valor constante A2Integer Y
campo de valor constante temp_dashSet A-Barbantinho
```

Cálculos # 24 e # 25 juntos compõem o cálculo do quar-Número ter. Cálculo # 25 faz o cálculo real, realizando um Integer divisão do temp_month_number por temp_months_in_quarter. A temp_month_number de campo foi obtido pela soma dos dois constante à month_number. Assim, em janeiro deste ano, irá avaliar a (1 + 2) / 3 = 1, Em Fevereiro deste será (2 +2) / 3 = 1E assim por diante até abril de onde obtemos (4 +2) / 3 = 2 e assim por diante. Note que não há cálculo em separado predefinidos para a divisão de número inteiro. divisão de número inteiro é realizada automaticamente quando se aplica o A / B expressão de argumentos inteiros e configuração o tipo de resultado também inteiro. A configuração dos cálculos de # 24 e # 25 é mostrado em detalhe na Lista 10-7.

Listagem de 07/10: Calculando o número trimestre usando divisão inteira

```
fieldCalculation Novo Campo aField B
---------------------------------------------------- --------------------
temp_month_number A + Bmonth_numbertemp_add_month
quarter_numberA / Btemp_month_numbertemp_months_in_quarter
```

Cálculos # 26 a # 28 faz um cálculo que pode ser usado para ver se o valor atual da data campo passa a ser o último dia do mês. Isto é feito subtraindo o valor do dia no mês do data atual daquele do dia seguinte. Se os dois dias consecutivos Acontece que no mesmo mês, a diferença será de 1 (um). Se no dia seguinte acontece de mentir no próximo mês, o resultado será um número negativo entre 27 (28/01) e 30 (1-31). Cálculo # 26 calcula o dia seguinte adicionando a uma constante para o valor da data de campo. No cálculo # 27, o dia no mês de que no dia seguinte é calculado e, finalmente, no cálculo # 28, a diferença é calculada. Os resultados dos cálculos # 26 e # 27 são intermediários e são descartados, especificando Remover Y =. O resultado final é mantido na diff_days_curr_day_next_day campo, que será utilizado fora este passo para calcular o valor da coluna atual. A configuração desses cálculos é mostrado em detalhe na Lista 10-8.

Listagem de 08/10: Calculando o último dia do mês

```
Nova fieldCalculationField aField B
------------------------------------------------------ ----------------------
next_dayDate A + B Daysdate_valueadd_day
next_day_in_monthDay do mês da data Anext_day
diff_daysA - B next_day_in_monthday_in_month
```

O # 29 # 33 cálculos por meio do uso A + B expressão para concatenar algumas das seqüências previamente calculado para obter representações para trimestre _name,year_quarter,year_month_numberE year_month_abbreviation. Como com a divisão inteira, não há nenhum operador de concatenação de string separadas, mas aplicação da expressão A + B aos campos do tipo de dados String tem exactamente nesse sentido. Os detalhes dos cálculos finais, # 23 e # 29 são mostrados na Listagem 10-9.

Listagem de 09/10: Cálculo do trimestre e concatenando strings

```
Nova fieldCalc.Field aField B
------------------------------------------------------ ------------------------
quarter_nameA + Btemp_quarter_prefixquarter_number
temp_year_stringA + Byear4temp_dash
year_quarterA + Btemp_year_stringquarter_name
year_month_numberA + Btemp_year_stringmonth_number
year_month_abbreviationA + Btemp_year_stringmonth_number
```

A Etapa Mapper Valor

Os passos seguintes são todas baseadas no Valor Mapper tipo de etapa:

- is_first_day_in_month
- is_last_day_in_month

- is_first_day_in_week
- is_last_day_in_week
- is_weekend

O valor do passo Mapper pode ser encontrado na categoria de Transformação. Sua ícone mostra uma seta conectando as letras A e B (veja a Figura 10-1).

O valor traduz etapa Mapper valores de um campo nos fluxos de entrada com os valores especificados em uma lista. O valor da tradução pode ser usado para substituir o valor original do campo no fluxo de entrada ou gravados em um novo campo, que é adicionado ao fluxo de entrada. Opcionalmente, um valor padrão pode ser especificado em nenhum caso de os valores na lista corresponde ao valor da fluxo de entrada. Um exemplo muito simples é mostrado na Figura 10-6, que retrata a configuração do passo marcado is_first_day_in_month.

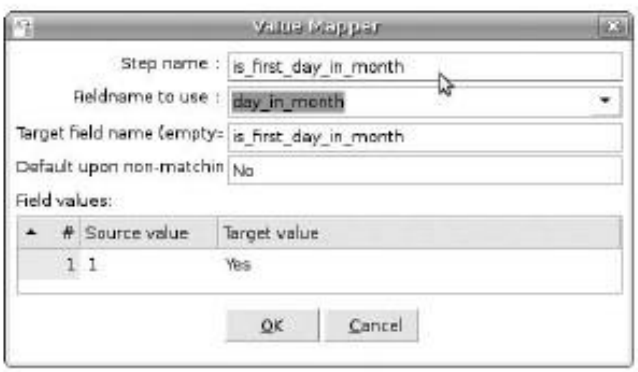


Figura 10-6: Mapeamento de dados inteiro para rótulos de texto com o passo Mapper Valor

Conforme mostrado na Figura 10-6, esta etapa levará o valor do day_in_month campo do fluxo de entrada e cria um novo is_first_day_in_month campo para manter os valores traduzidos. Existe apenas um valor para mapa: Somente quando o day_in_month campo acontece a ter o valor 1 (um) caso a string Sim ser devolvido. Especificando o valor Não no padrão sobre os não-correspondência propriedade, o valor Não será devolvido em qualquer outro caso.

A configuração do is_first_day_in_week e is_last_day_in_week medidas é inteiramente análogo ao da etapa is_first_day_in_month. Ambos Essas etapas usam o day_in_week campo do fluxo de entrada e retornar o valor Sim no caso de um valor específico (1 e 7 para is_first_day_in_week e is_last_day_in_week respectivamente), é casado, e Não em qualquer outro caso. Novamente, as duas etapas escrever o valor para um novo campo do fluxo de saída (is_first_day_in_week e is_last_day_in_week respectivamente).

A configuração da etapa is_weekend é quase exatamente o mesmo. Este etapa converte os valores da day_in_week campo do fluxo de entrada para uma nova is_weekend campo no fluxo de saída, mas desta vez, dois valores são

mapeada: ambos os valores 1 e 7 são mapeados para Sim, E novamente Não é retornado por padrão.

dim_date Carga: A etapa de saída de mesa

O passo final na transformação mostrada na Figura 10-1 é rotulado de carga dim_date e com base no Saída de mesa etapa. Esta etapa de obras, inserindo o registros do fluxo de entrada em uma tabela do banco de dados. Você pode encontrar a Tabela
Saída na categoria de saída. Seu ícone é uma seta apontando para um tambor verde cilindro (Figura 10-1).

Na etapa dim_date carga, a etapa de saída da tabela é usada para inserir os dados criados nesta transformação em dim_date tabela que foi criada em
O primeiro passo da transformação, a etapa de execução do SQL. Você já encontrou a etapa de saída de Mesa em nosso banco de dados de transformação walk-através do capítulo anterior. Não, não incomoda a configuração do mapeamento entre os campos do fluxo de entrada e as colunas da tabela porque os campos no fluxo de entrada já foi combinado as colunas da tabela. Em esta transformação, as coisas são um pouco diferentes. O fluxo de entrada contém ainda uma alguns campos que não correspondem a qualquer colunas da tabela, como initial_date, day_sequenceE diff_days_curr_day_next_day. Outra coisa é que o data campo do fluxo de entrada deve ser mapeado para o date_value coluna da dim_date tabela. Figura 10-7 mostra parte da configuração.

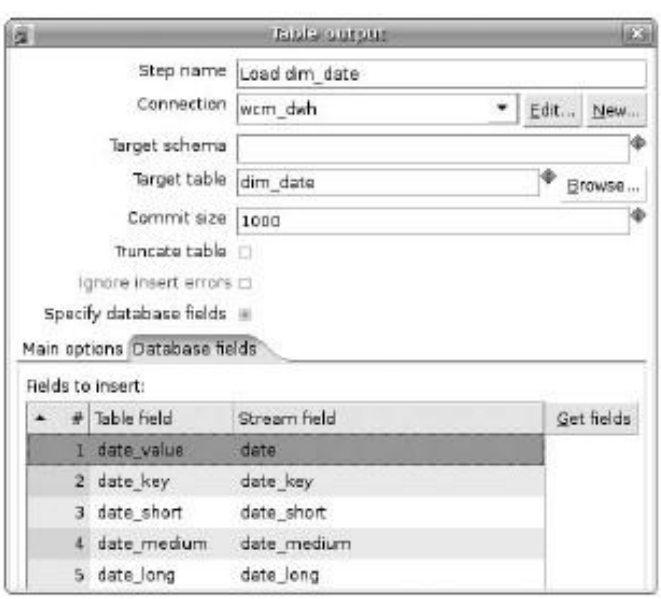


Figura 10-7: Mapeamento de colunas específicas do fluxo de entrada nas colunas da tabela

A fim de mapear campos para colunas, primeiro você tem que selecionar a especificar campos de banco de dados caixa de seleção. Você pode então ativar o banco de dados ficha de campos

e use o Get campos de página para preencher a grade com todos os campos presentes no fluxo de entrada. Você pode selecionar e remover os campos que você não precisa, e escreva o nome da coluna correspondente na coluna de campo Tabela, conforme necessário.

Características Dimension Data mais avançada

Agora que você aprendeu a preencher uma dimensão de data simples, você pode estender o projeto e adicionar mais passos para personalizar a dimensão de data para o seu gosto. O design dimensão de data a partir do capítulo 8 menciona um par de possíveis melhorias, tais como ISO incluindo o ano e os números da semana especial campos de data, e cálculos para facilitar a comparação entre o ano em curso com o ano anterior, e com suporte a locales e línguas diferentes.

ISO semana e ano

A etapa Calculator (mostrado na Figura 10-5) fornece o Semana de ISO8601 Uma data e ISO8601 Ano da Data A expressões. Usando esses, é trivial estender a tabela de dimensão de data com a norma ISO semana e atributos ano.

Passado e atual Indicadores Ano

Manter os indicadores ano passado e atual discutido no Capítulo 8 implica uma actualização regular a tabela de dimensão de data. Alternativamente, você pode integrar os cálculos necessários extra diretamente na transformação que gera as linhas de dimensão de data e simplesmente truncar e em seguida atualize a tabela de dimensão de data em sua totalidade. A atualização da tabela faz sentido se data de sua dimensão contém informação que é difícil de gerar (como feriados). Neste caso, você iria construir uma transformação separada para fazer o atualizações. Truncado e dimensão atualizar a data na sua totalidade é mais simples, porque você só precisa manter uma transformação.

Calculando os valores dos indicadores envolve algumas etapas extra. Em primeiro lugar, você precisa de um passo para introduzir a data atual para que você possa comparar com o datas geradas pela transformação. A maneira mais fácil de fazer isso é por adicionar um Get System Info etapa. Esta etapa permite definir novos campos. Ele oferece uma caixa de listagem com uma série de itens de informação diferentes do sistema, incluindo o atual data. Neste caso específico, o item que você precisa é chamado sistema de data (fixa), que irá preencher o campo com o sistema de data / hora como determinado na etapa de inicialização fase.

Usando uma calculadora passo, você pode dissecar a data do sistema em partes, como data ano, mês, semana e dia para que você possa compará-los com os correspondentes campos na tabela de dimensão de data. Da mesma forma, você pode usar a calculadora para a etapa calcular o ano anterior, mês e semana.

Para, finalmente, calcular os valores para os campos de bandeira pode utilizar a fórmula etapa, encontradas na categoria scripts. A etapa da Fórmula permite que você use fórmulas com uma sintaxe semelhante à utilizada em programas de planilhas como o Microsoft

Excel ou do OpenOffice.org Calc. Por exemplo, para calcular o valor do current_year coluna, você deve criar um novo campo na etapa da Fórmula chamado current_year e especificar uma fórmula como esta:

```
IF ([ano4] = [current_year_number], 1, 0)
```

Observe que a fórmula não é precedido por um sinal de igual, como seria o caso em programas de planilhas. Neste exemplo, ano4 é o ano4 coluna do dim_date mesa, e current_year_number é a parte do ano da data do sistema tal como calculado pela etapa anterior calculadora. Observe que você precisa para incluir campos entre colchetes. IF () é uma das funções internas fornecidas pela Fórmula etapa. Esta função recebe três argumentos: expressão de um primeiro Boolean (No nosso caso, a comparação ano), em segundo lugar, o valor a retornar se o primeiro argumento é verdadeiro e, finalmente, o valor a retornar se o primeiro argumento é falso. Note-se que os argumentos da função são separados com ponto e vírgula.

Usando essa técnica, você pode adicionar quantos campos quiser para calcular o bandeiras.

Internacionalização e suporte de locale

Java oferece suporte embutido localidade. Porque PDI é programado em Java, é relativamente fácil de tocar para o sistema local de Java e usá-lo para formatar datas e os números de uma maneira dependente de localidade. Nós descrevemos o método em poucas palavras.

IDP oferece a modificação Javascript tipo de etapa de Valor na categoria scripts. Usando este tipo de etapa, você pode usar o JavaScript para datas processo gerado por sua transformação. A modificação Javascript passo valor é baseado na o Mozilla Rhino motor de JavaScript, que permite criar uma instância e acessar objetos Java. Desta forma, você pode criar um java.util.Locale objeto para a localidade desejada e usar essa data para o formato usando o java.text.SimpleDateFormat.

O procedimento é descrito em detalhes no http://rpbouman.blogspot.com/2007/04/kettle-tip-using-java-locales-for-date.html.

Carregando uma dimensão de tempo simples

A dimensão do tempo aparece no cliente e esquemas Pedidos de estrelas o World Class data warehouse Filmes. Assim como a dimensão de data, a maioria dos dados da dimensão de tempo podem ser gerados, seja com um banco de dados procedimento armazenado ou um Pentaho Data Integration transformação. Nós já discutida a desvantagem de banco de dados usando procedimentos armazenados para este tipo de trabalho. Por esta razão, nós preferimos fazer isso usando uma transformação PDI.

Uma transformação simples para carregar o dim_time tabela na wcm_dwh banco de dados é mostrado na Figura 10-8.

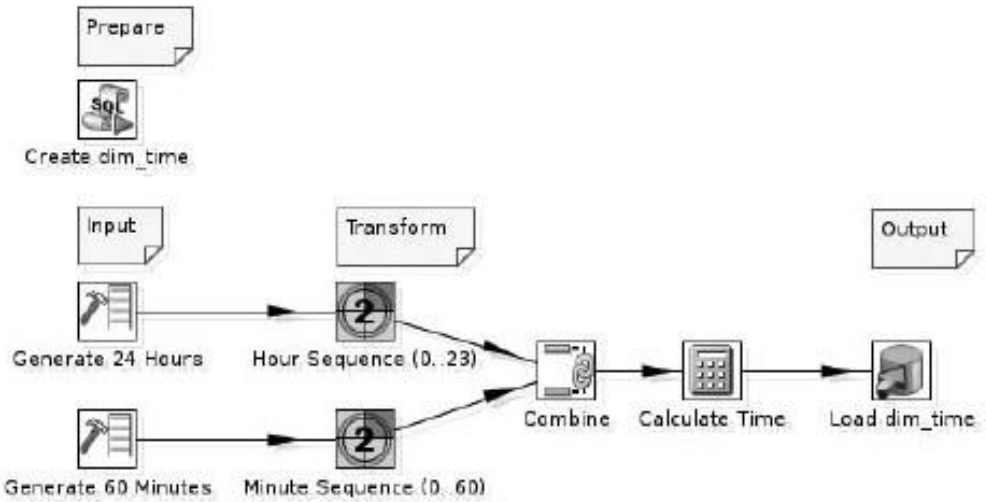


Figura 10-8: Uma transformação simples para carregar a dimensão de tempo

Aqui está um resumo dos trabalhos dessa transformação:

1. Criar dim_time: Este é um passo executar script SQL para criar o alvo tabela de dimensão de tempo.

2. Gerar 24 Horas e gerar 60 Minutos: Estes são os dois Gerar Linhas etapas que criam as linhas que compõem as horas e minutos de a dimensão de tempo. O Limite da Gerar etapa Horas 24 está definido para 24 porque há 24 horas por dia. Da mesma forma, o limite do Gere 60 Minutos passo é definido como 60, porque há cerca de 60 minutos por hora.

3. Seqüência horas (0 .. 24) e Minuto Seqüência (0 .. 60): Estes são os dois Adicionar Seqüência de passos, que adicionam um campo inteiro para incrementar as linhas de entrada para representar as horas e minutos, respectivamente. Em ambos os casos, o início no valor for definido para 0 (zero) e o incremento é deixado para o padrão de 1 (um).

4. Combine: Esta é uma junção de linhas (produto cartesiano) etapa. Esta etapa reúne as linhas de hora e minuto linhas de sua entrada córregos.

5. Calcule o tempo: Nesta etapa, calculadora, a hora / minuto são combinações analisado em um valor Date, que é então convertido em um valor String usando uma seqüência de formato de reter apenas parte do tempo.

6. dim_time Carga: Este é um passo de saída da tabela para inserir as linhas geradas no dim_time tabela de destino.

Ao olhar para a Figura 10-8, você vê que, exceto para a etapa Join Linhas esta transformação usa os mesmos elementos básicos como a transformação de carga a dimensão de data mostrado na Figura 10-1.

Em comparação com a transformação para carregar a dimensão de data, não há algo radicalmente diferente sobre a forma como esta transformação gera a sua

de dados. A transformação para carregar a dimensão de data foi totalmente linear, e os dias em que conduziu a transformação foi obtida pela adição de dias para a data inicial. Nesta transformação, as linhas Gerar e Sequência etapas são independentes uma da outra. O tempo real do dia é obtida pela combinando os dados dos dois córregos e, em seguida, analisar uma data fora do combinados valor / hora, minuto. Esta abordagem seria menos prático para carregar uma dimensão de data: embora você pode configurar fluxos separados para gerar anos e meses, isso não é tão fácil para a parte do dia a contar da data porque o número de dias depende do mês e se o ano é um ano bissexto.

Combine: a etapa Join linhas (produto cartesiano)

A etapa Combine na Figura 10-8 é de Junte-se linhas (produto cartesiano) tipo. Isto é encontrado na categoria Associações. Seu ícone, visível na Figura 10-8, é um conjunto de elos da cadeia.

A associação de linhas (produto cartesiano) passo é funcionalmente análogo ao INNER JOIN operação em SQL, a diferença é que ela opera no registro fluxos de entrada e não as tabelas do banco.

NOTA Embora a associação de linhas (produto cartesiano) passo é funcionalmente análogo para um banco de operação de junção, deve-se muito cuidado para não considerá-lo como um substituição. Como regra geral, você não deve usar essa etapa para evitar a escrita SQL, em vez disso, você deve considerar usar este passo no caso de você não pode usar SQL. Para exemplo, quando você quer se juntar as linhas de duas tabelas que estão disponíveis no a mesma conexão, você provavelmente deve usar um passo de entrada de tabela e especificar o SQL adequadas para resolver o problema.

Um caso de uso típico para a junção de linhas (produto cartesiano) etapa é quando você quiser criar um produto cartesiano de conjuntos de dados entre servidores de banco de dados ou fontes de dados.
Embora você possa configurar uma condição de junção para a junção de linhas (produto cartesiano) etapa, você deve considerar usar o Merge Join passo no caso de você querer usar complexos tipos de junção e / ou condições.

A etapa de obras, combinando os registros de todos os seus fluxos de entrada em um registro de novo composto, que é enviado para o fluxo de saída. O homem-particular ner na qual os registros são combinados é conhecido como o produto cartesiano. Para dois fluxos de entrada, o resultado é criado por sorteio a cada um dos registros provenientes a partir do Horário de Seqüência (0 .. 23) a etapa com todos os registros vindos de Seqüência hora (0 .. 60) passo. O fluxo de saída resultante tem tanto uma hora e um minuto campo, e contém 24 ×60 (= 1440) linhas, que juntas formam acima de toda hora / combinações possíveis minutos, a partir das 00:00 através 23:59.

Para este uso particular da etapa Join linhas (produto cartesiano), apenas um pequena quantidade de configuração é necessária, mas para a integralidade, a configuração de diálogo é mostrada na Figura 10-9.

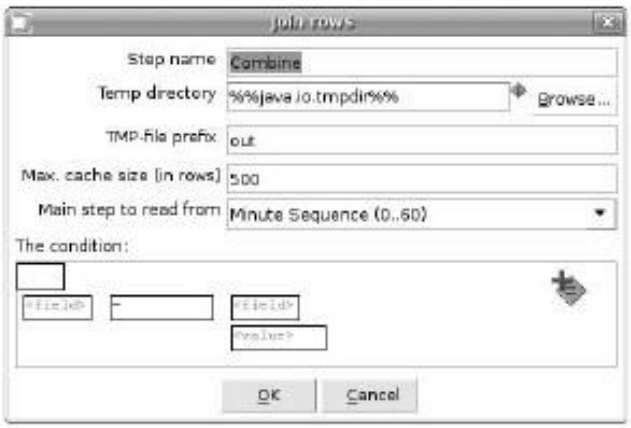


Figura 10-9: A janela de configuração da etapa Join linhas (produto cartesiano)

A associação de linhas (produto cartesiano) passo sempre pressupõe que uma entrada fluxo é a''corrente''principal. Você pode especificar esta etapa na etapa principal para ler da propriedade. Este fluxo de unidades de""da etapa: a etapa consome um registro da corrente principal, e em seguida, cria registros combinação com cada um dos registros da entrada de outras correntes. Quando todas as combinações são criadas, a procede passo com o próximo registro da corrente principal.

Para melhorar o desempenho, a junção de linhas (produto cartesiano) etapa usa na memória cache para todos os fluxos de entrada, exceto o fluxo de entrada principal. A tamanho do cache fluxo de entrada pode ser especificada definindo o número de linhas para cache no máx. cache propriedade de tamanho. Se o número de linhas em uma fluxo de entrada excede o tamanho de seu cache, as linhas são gravadas em um temporário arquivo. O padrão para o máx. propriedade tamanho do cache é de 500, o que excede o número de registros a partir de qualquer fluxo de entrada em nossa transformação, para que possamos
estar confiante de nenhum arquivo temporário serão necessários.

NOTA Os registros do fluxo principal não precisa ser armazenado em cache porque o Junte-se linhas (produto cartesiano) passo sempre olha exatamente um registro de que stream em qualquer momento. Por esta razão, você sempre deve especificar o fluxo de que contém a maioria das linhas como a etapa principal para ler. Em nosso transformação, vamos definir o passo principal para ler a seqüência de minuto (0 .. 60) passo nesse sentido.

Se necessário, você pode especificar em qual diretório os arquivos temporários devem ser armazenadas, definindo a propriedade de diretório Temp. O padrão para essa propriedade é %%%% Java.io.tmpdir, O que denota o valor de uma variável interna referente com o padrão de diretório de arquivos temporários Java. (Para mais informações sobre

variáveis, consulte a seção Usando variáveis""no capítulo 11). Se você gosta, você pode também especificar o prefixo para os nomes dos arquivos temporários, o que é útil principalmente para
fins de depuração.

Embora seja possível especificar uma condição de junção para a junção de linhas (cartesiana
etapa do produto), é recomendável usar um Merge Join etapa vez que esta permite melhor controle sobre a condição de junção e tipo de associação.

Calcular Tempo: Mais uma vez, a etapa Calculadora

Na transformação mostrado na Figura 10-8, a calcular passo de tempo é baseada na etapa de calculadora. Nós já discutimos este tipo de passo extensivamente em Calcular o título"e formatar datas: a Etapa calculadora"no início
neste capítulo. Para completar, vamos mostrar a configuração específica utilizada para calcular o tempo na Figura 10-10.

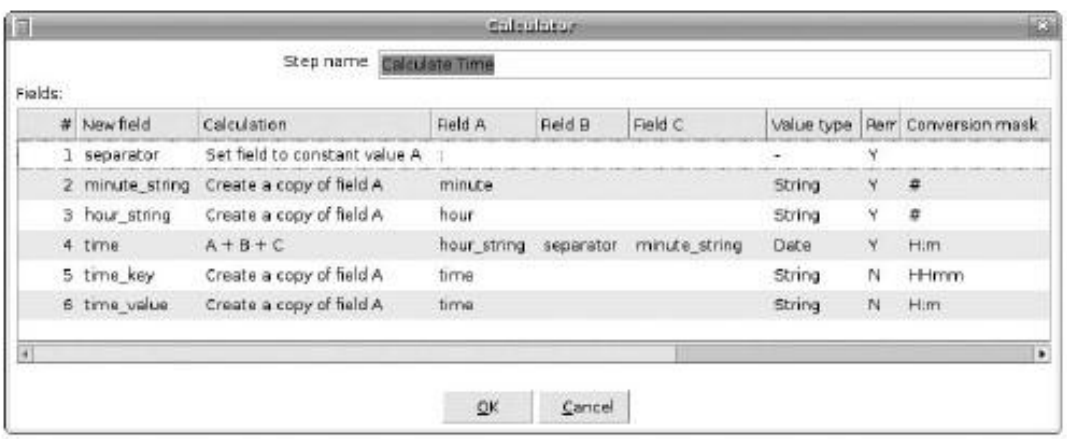


Figura 10-10: Usando o passo Calculadora para calcular a hora do dia

Cálculo # 4 é o mais importante, como ele concatena represen-corda tações do hora e minuto campos e converte o resultado para um campo do Data tipo usando o formato H: m. Ao reformatar novamente, o time_key e time_value campos são gerados, o que corresponde diretamente ao time_key e time_value colunas na dim_time tabela.

Carregando a dimensão Demografia

A World Class Movie Database usa uma demografia mini-dimensão para gestão da dimensão grande cliente. A transformação para carregar o dimensão que a demografia é mostrado na Figura 10-11.

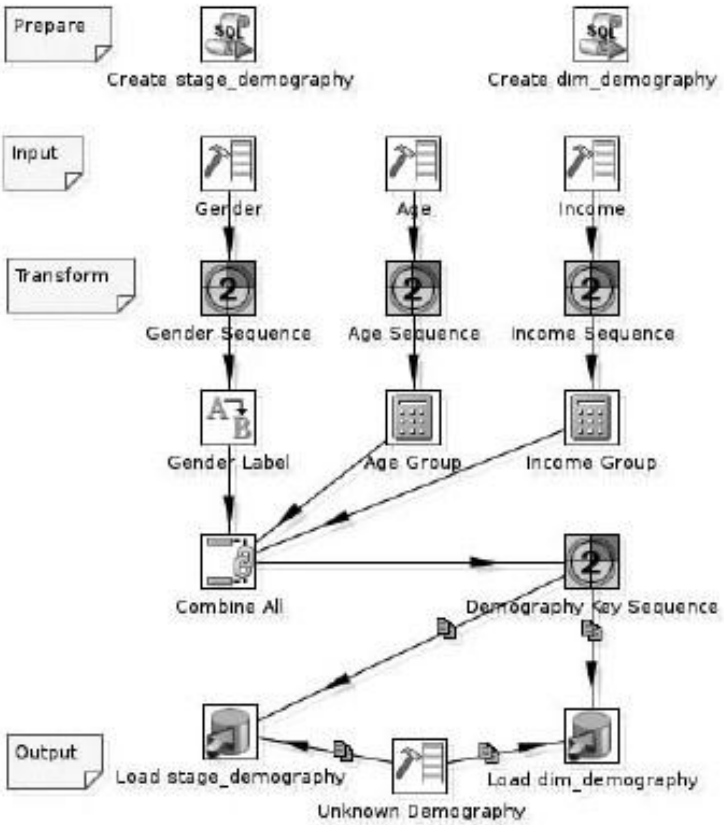


Figura 10-11: Uma transformação para carregar a dimensão da demografia

Aqui está um resumo do que está acontecendo nessa transformação:

1. Criar dim_demography: Esta etapa Execute SQL cria o real escurecer _demography tabela de dimensão na wcm_dwh banco de dados.

2. Criar stage_demography: Esta etapa Execute SQL cria uma estágio _demography tabela na wcm_staging banco de dados. Nós descrevemos o efeito desta tabela em detalhe na próxima seção.

3. Sexo e Género sucessão de idade, e Seqüência, idade e renda e Seqüência de Renda: Os pares de gerar linhas e etapas Adicionar Sequência gerar os dados em bruto que faz os possíveis valores para sexo, idade grupo e grupo de renda, respectivamente.

4. Sexo no rótulo: Esta etapa mapper Valor converte os valores da Sexo passo em seqüência Masculino e Feminino rótulos, que devem ser armazenados na gênero coluna das tabelas de destino.

5. Faixa Etária e Grupo de Renda: Essas etapas Calculadora gerar o adequados os valores limites para a idade e grupo de renda, e também criar etiquetas bem legível para identificar o grupo.

6. Combine Todos: Assim como fizemos durante o carregamento do dim_time dimensão, usamos
Junte-se a um passo de linhas (produto cartesiano) para fazer todas as combinações possíveis
dos registros de todos os fluxos de entrada.

7. Demografia seqüência de teclas: Este etapa Adicionar Seqüência é usado para gerar valores-chave que serão usadas para identificar as linhas da dim_demography tabela.

8. Demografia Desconhecido: Esta etapa Gerar linhas cria uma única linha que pode ser usado sempre que a linha demografia adequada não pode ser encontrado.

9. stage_demography carga e dim_demography carga: Estes Out Mesa colocar os passos executados em paralelo para carregar os dados demográficos a partir da entrada
fluxos de entrada da peça Demografia Seqüência e Desconhecido Demografia etapas.

Compreender o stage_demography e dim_demography Tabelas

As transformações que você viu até agora, sempre usou uma única tabela de destino. Este atos de transformação em duas tabelas:

- A dim_demography tabela na wcm_dwh banco de dados, que é o real demografia tabela de dimensão. Esta tabela é criado no Create-dim_de etapa mamografia e carregado na etapa dim_demography Load.

- A stage_demography tabela na wcm_staging banco de dados, que serve como uma tabela de pesquisa para procurar a chave da tabela de dimensão. Essa tabela é criado na etapa stage_demography Criar e carregado na carga stage_demography etapa.

O motivo para exigir duas tabelas, esperamos tornar clara quando olhar para a definição do dim_demography Dimensão da tabela. A CREATE TABLE declaração para a dim_demography tabela é mostrado na Listagem 1-10. Na listagem, você vê que, além da demography_key, Apenas os rótulos legíveis para o gênero,age_group e income_group são armazenados no dim_demography tabela.

Listagem 1-10: A instrução CREATE TABLE da tabela de dimensão dim_demography

```
CREATE TABLE dim_demography (
  demography_key SMALLINTNOT NULL,
  age_groupVARCHAR (10) não NULL,
  genderVARCHAR (10) não NULL,
  income_groupVARCHAR (20) não NULL,
  PRIMARY KEY (demography_key)
);
```

Agora, suponha que você queira carregar clientes para o dim_customer ou fact_customer tabelas. Em ambos os casos, você precisa para armazenar o valor apropriado para o demography_key como parte da linha do cliente. Mas supondo que você tem idade do cliente, renda e gênero, é um grande desafio para usar esses dados para a pesquisa dim_demography tabela.

Para superar nossa incapacidade de utilizar o dim_demography tabela diretamente para procurar o demography_key, Introduzimos o stage_demography tabela na wcm_staging banco de dados. A stage_demography tabela armazena os demography_key e gênero rótulo, bem como os valores inteiros reais que formam o limites das faixas etárias e de renda. A CREATE TABLE declaração para a stage_demography tabela é mostrado na Listagem 10-11.

Listagem 10-11: A instrução CREATE TABLE da tabela de dimensão dim_demography

```
CREATE TABLE stage_demography (
  demography_keySMALLINTNOT NULL,
  genderVARCHAR (10) NOT NULL,
  min_age_groupSMALLINTNOT NULL,
  max_age_groupSMALLINTNOT NULL,
  min_income_group INTEGERNOT NULL,
  max_income_group INTEGERNOT NULL,
  PRIMARY KEY (demography_key),
  UNIQUE (
     min_age_group, max_age_group,
     min_income_group, max_income_group,
     gênero
  )
)
```

Dado um valor para o sexo, renda e idade, é bastante simples procurar o demography_key no presente stage_demography tabela. Além exigindo um valor correspondente na gênero coluna, temos de exigir que o valor para a idade seja igual ou maior que o valor da min_age_group coluna, mas inferior ao valor do max_age_group coluna. Da mesma forma, o valor de renda devem estar entre os valores na min_income_group e max_income_group colunas, respectivamente. Iremos discutir o processo de pesquisa em mais detalhes posteriormente neste capítulo, quando descrevemos o carregamento do dim_customer e fact_customer tabelas.

Gerando Idade e Grupos de Renda

Os grupos de idade e renda são gerados usando um suplemento etapa Seqüência com um valor alterado para o incremento de propriedade. O incremento é definido como um por padrão. Para os grupos de geração de renda, nós usamos um incremento de 10.000. Para grupos de idade, usamos um incremento de 5 e um valor inicial de 15. (Classe Mundial Filmes não alugar ou vender DVDs para os menores).

Nas etapas seguintes, calculadora, o valor utilizado para o incremento é usada para calcular o limite (exclusive) superior do grupo, tendo o valor da seqüência se como o limite inferior (inclusive). As etapas da calculadora também gerar uma etiqueta legível para cada grupo. Por exemplo, o grupo de renda de R $ 10.000 a $ 20.000 é rotulado como R $ 10.000 - $ 20.000. Para integralidade, a configuração da etapa de Renda Calculadora Grupo é apresentada na Figura 10-12.

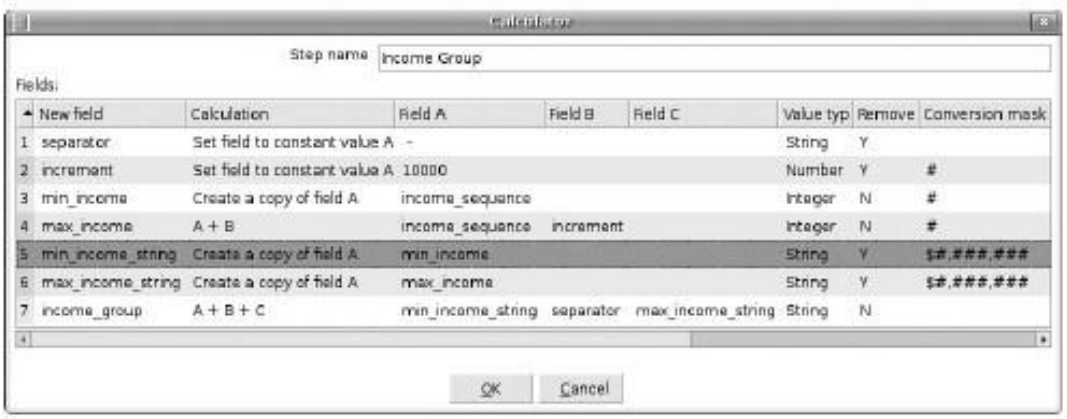


Figura 10-12: Calculando e formatação de grupos de renda

Múltiplos Fluxos de entrada e saída

A transformação na Figura 10-12 mostra uma característica interessante PDI. A Demografia seqüência de teclas e as etapas Desconhecido cada uma, duas Demografia saída de lúpulo. Para essas duas etapas, a saída de lúpulo são levados à mesa Saída stage_demography Carga passos e dim_demography carga, que consequentemente, tem duas entrada de lúpulo.

Você já viu um exemplo de um passo que pode aceitar a entrada de vários córregos, quando cobria a associação de linhas (produto cartesiano) etapa. No entanto, há uma diferença em relação à entrada do lúpulo Output Table etapas. A natureza da associação de linhas (produto cartesiano) operação simplesmente pressupõe pelo menos dois fluxos de entrada. Um produto cartesiano de uma única entrada não não faz sentido porque a operação é definido pela combinação de dois dados ou mais fluxos. Mas para a etapa de saída de mesa, um único fluxo de entrada faz sentido, e podemos legitimamente perguntar-nos que o significado é de um fluxo de entrada extra neste caso.

A maioria das etapas de transformação pode manipular um único tipo de fluxo de entrada. Mas não há nada errado com os registros de entrada aceitar múltiplos fluxos desde que tenham registro de layouts idênticos. Outra maneira de colocar é que para a maioria das etapas, não importa onde se originam os registros de entrada, desde como todos têm as mesmas propriedades. A entrada lúpulo simplesmente despejam seus registros no buffer de entrada da etapa de recebimento, independentemente da origem.

O mesmo vale para fluxos de saída. A maioria dos tipos etapa gerar registros de saída que todos tenham o mesmo layout. Não há nenhuma razão para que todos esses registros devem
ser enviados a um único salto de saída. Eles poderiam muito bem ser enviado vários fluxos. Note que isso não é totalmente simétrico em relação ao múltiplos fluxos de entrada. No caso de múltiplos saltos de saída, há duas diferentes maneiras de enviar os registros para fora:

- Os registos podem ser copiados, o envio de todos os registros de saída para todos os de saída de lúpulo.
  Isso é útil para enviar os mesmos dados ao longo de vários caminhos para paralelo transformação. Na Figura 10-11 você pode dizer que os dados estão sendo copiados porque este é indicado pelo meio pequeno ícone de cópia em todo o salto.
- Os registros podem ser distribuídos em forma de rodízio sobre todas saída lúpulo. Isto essencialmente as partições do fluxo de saída, o envio de apenas uma parte de todos os registros de saída para baixo cada salto.

O método de distribuição pode ser configurado através do menu de contexto do etapa. No menu de contexto, localizar o submenu circulação de dados. Lá, escolha quer distribuir dados para os próximos passos ou copiar dados para os próximos passos.

Na transformação mostrado na Figura 10-11, a seqüência de teclas Demografia e as etapas Desconhecido Demografia estão configurados para copiar todos os dados para todos
fluxos de saída. Isso faz sentido porque precisamos de todos os dados estejam disponíveis em
o data warehouse, bem como a área de preparo.

Carregando Dados da Fonte Sistemas

Até agora, nós só carregado as tabelas de dimensão com base nos dados gerados. No entanto,
A maioria das tabelas no data warehouse são preenchidos com dados provenientes de vários tabelas de banco de dados em um sistema de origem e, por vezes, os sistemas de múltiplas fontes.

Nesta seção, vamos dar uma olhada em algumas das considerações, questões e soluções que entram em jogo quando o carregamento de dados do armazém com as tabelas dados provenientes de sistemas de origem. Começamos por introduzir alguns conceitos e em seguida, continuar a descrever soluções Pentaho Data Integration para carregar o Mundo Classe armazém de dados de filmes.

Encenação valores de pesquisa

A World Class possui um banco de dados de filmes único valor_procurado tabela que serve como referência para as listas de relativamente pequeno e fixo de valores pré-definidos. Todos
Os valores que pertencem a uma lista têm o mesmo tipo de pesquisa. A lookup_type tabela é usada para armazenar o nome da tabela e da coluna para que a lista de valores se aplica.

No projeto do data warehouse, a maioria das referências à valor_procurado tabela de reaparecer como atributos de dimensão normalizada. Por exemplo, no dim_promotion tabela, o promotion_type coluna é carregado a partir dessas linhas

em valor_procurado que se refere ao uso do promotion_type_lookup coluna na promoção tabela.

Os dados do valor_procurado e lookup_type tabelas quase nunca muda. As alterações devem ser esperadas apenas se o sistema de origem é atualizado para uma nova versão.

Precisamos de os valores de pesquisa com tanta frequência, e as mudanças ocorrem muito raramente que faz sentido para puxar todos os dados do valor_procurado mesa para armazenamento permanente na área de teste. Enquanto estamos aqui, nós pôde apenas como também armazenar os valores para cada tipo de pesquisa em sua própria tabela. Cada lista individual de valores de pesquisa será muito pequena, o que deve fazer pesquisas e se junta como mais rápido possível.

O stage_lookup_data Trabalho

Para carregar os valores de pesquisa na área de preparação, criamos um Pentaho Data Integração do trabalho. Criando um novo emprego é muito parecido com a criação de uma nova transformação.

Você pode criar uma nova tarefa a partir do menu (escolhendo novo arquivo de trabalho) ou a partir da barra de ferramentas. Alternativamente, você pode usar Alt + Ctrl + N.

O trabalho de carregar os valores de pesquisa na área de teste é mostrado na Figura 10-13.

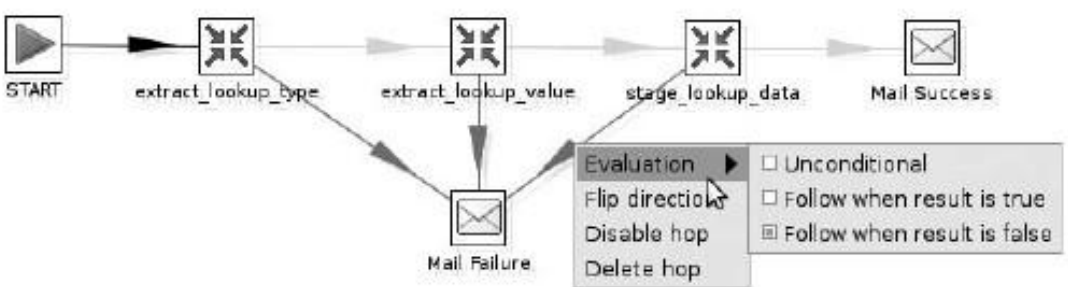


Figura 10-13: Um trabalho de carregar os valores de pesquisa na área de preparo

NOTA No capítulo 9, discutimos as diferenças mais importantes entre os trabalhos e transformações. Não vamos repetir essa discussão, mas gostaríamos de salientar vez que as transformações e empregos podem ser parecidos, mas são tipos muito diferentes das coisas. A principal diferença para se manter em mente é que os trabalhos consistem em entradas de emprego, que representam a execução de uma tarefa, enquanto as transformações consistem etapas de transformação que representam uma operação no registro córregos. Lúpulo entradas trabalho conectando representam uma seqüência de execução, enquanto que o lúpulo conexão etapas de transformação representa um fluxo de registros.

Aqui está um resumo do que está acontecendo no presente trabalho:

- START-Este é o ponto de entrada para execução do trabalho.
- extract_lookup_type e extract_lookup_value-By execução dessas transformações, os dados são lidos a partir do valor_procurado e lookup_type tabelas do sistema de origem e, temporariamente armazenados.

- stage_lookup_data-Este transformação lê os dados do tem-
  armazenamento temporário e carrega-lo para as tabelas na área de teste.

- Mail Failure Sucesso e-mail Estes entradas de emprego enviar uma notificação
  mensagens de e-mail para informar se o trabalho de falha ou sucesso.

As inscrições de trabalho estão ligados uns aos outros com lúpulo. O verde lúpulo (Que servem para ligar as cinco entradas de emprego ao longo da fileira de cima) indicam o normal caminho de execução: se uma entrada de trabalho é concluído com sucesso, a execução é reiniciada às a entrada de trabalho encontrada na outra extremidade da saída hop verde. Na Figura 10-13, o caminho de execução normal, forma uma seqüência de entradas de trabalho que começa na entrada de trabalho rotulado como Iniciar e termina com a entrada de trabalho marcado Mail Sucesso.

Todas as entradas, mas primeiro emprego e por último também tem um segundo, de cor vermelha de saída hop para a entrada de trabalho marcado Falha Mail. O vermelho lúpulo indicar um erro execução do caminho. O caminho de execução de erro é inserida sempre que uma entrada de trabalho não Não é possível executar com êxito. Se uma entrada de trabalho não tem um salto de erro de saída, execução sem êxito de que a entrada no trabalho fará com que o trabalho como um todo a ser abortada com erro.

Você pode alterar as propriedades de um salto com o botão direito do mouse e escolha de um ação a partir do menu de contexto. Para modificar se o salto será seguido em caso de sucesso ou fracasso, escolha a opção apropriada na avaliação submenu. Isso também é mostrado na Figura 10-13.

A entrada do início do trabalho

A entrada do início do trabalho é uma entrada especial de trabalho que denota o ponto de entrada de um emprego. Esta entrada de emprego encontra-se na categoria Geral, e seu ícone é um verde seta (ver figura 4-13).

Este tipo de passo pode ter apenas um salto de saída que é incondicionalmente seguido. início da execução do trabalho com a entrada de trabalho encontradas na outra extremidade do hip saída da entrada do início do trabalho. Todo trabalho deve ter exatamente um entrada de trabalho deste tipo.

A janela de configuração da entrada do início do trabalho contém uma série de opções para agendar a execução do trabalho. Detalhes sobre a programação de trabalho são discutidos em Capítulo 14.

Job Entradas Transformação

O trabalho extract_lookup_type entradas, extract_lookup_value e stage_lookup _data são todos Emprego entradas Transformação. Este tipo de entrada de trabalho é encontrado no Geral da categoria. O ícone (visível na Figura 10-13) mostra quatro setas que visa um ponto central.

Na janela de configuração, o arquivo de transformação que está a ser executado pode ser especificado no campo Nome da Transformação. Alternativamente, se você estiver conectado a um repositório, você pode usar o nome de Transformação e campos Repositório Directory para especificar uma transformação armazenados no repositório.

(Usando o repositório é coberto extensivamente no Capítulo 11.) Em ambos os casos, uma botão está disponível à direita para procurar transformações. Para rapidamente aberta a transformação especificada, botão direito do mouse e escolha a etapa do Open opção de transformação.

Figura 10-14 mostra a janela de configuração da entrada de emprego Transformação rotulados extract_lookup_type.

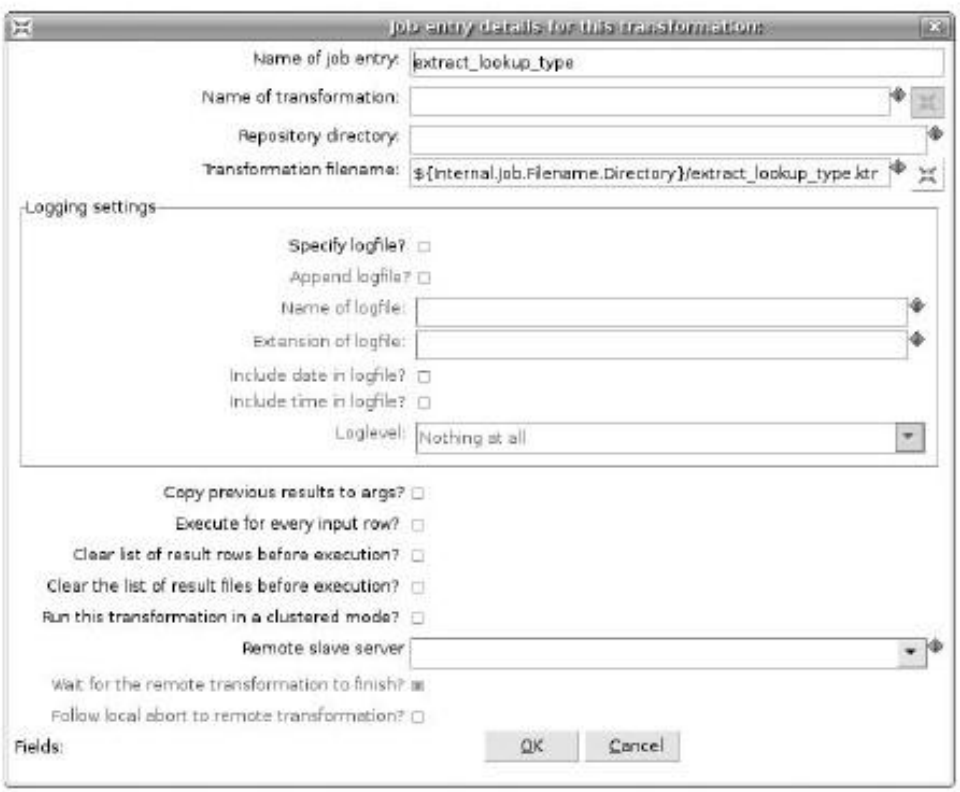


Figura 10-14: A transformação do emprego de diálogo configuração de entrada

Na Figura 10-14 o nome da transformação de arquivo é especificado como $ {Interno . Job.Filename.Directory} / extract_lookup_type.ktr. Isto denota o arquivo extract_lookup_type.ktr que residem no mesmo diretório que o arquivo de trabalho própria, conforme explicado em detalhe no Capítulo 11.

A janela de configuração contém muitos mais imóveis que poderiam ser configurada. Você vai encontrar alguns deles mais tarde neste capítulo.

Mail Failure Sucesso e-mail

Na Figura 10-13, as entradas de emprego rotulados Mail sucesso eo fracasso são do Correio o tipo de correio, indicado por um ícone de envelope. Isto pode ser encontrado no Correio categoria.

A entrada de emprego Mail é projetado para enviar uma mensagem de e-mail com o Simple Mail Transfer Protocol (SMTP). Seu objetivo principal é fornecer informações básicas notificação de status do trabalho de execução (fracasso, sucesso, ou o progresso).

Configuração da etapa e-mail não é particularmente difícil, embora o número de opções de configuração pode ser um pouco assustador no começo. A configuração diálogo contém quatro páginas guia. Na página endereços, mostrado na Figura 10-15, você deve especificar pelo menos um endereço de e-mail válido no endereço de destino propriedade. Opcionalmente você também pode configurar endereços de CC e BCC. Em adição para o endereço de destino, o nome do remetente e as propriedades endereço do remetente
são obrigados pelo protocolo SMTP e deve ser especificado. Você pode opcionalmente especificar um Endereço de resposta e alguns dados adicionais de contato, como o nome e número de telefone da pessoa de contato. Para o sucesso típica / insucesso notificações, seria enviar notificações para o suporte de TI, pessoal e especificar detalhes de um membro da equipe de integração de dados como o remetente. Figura 10-15 mostra a ficha de endereços.

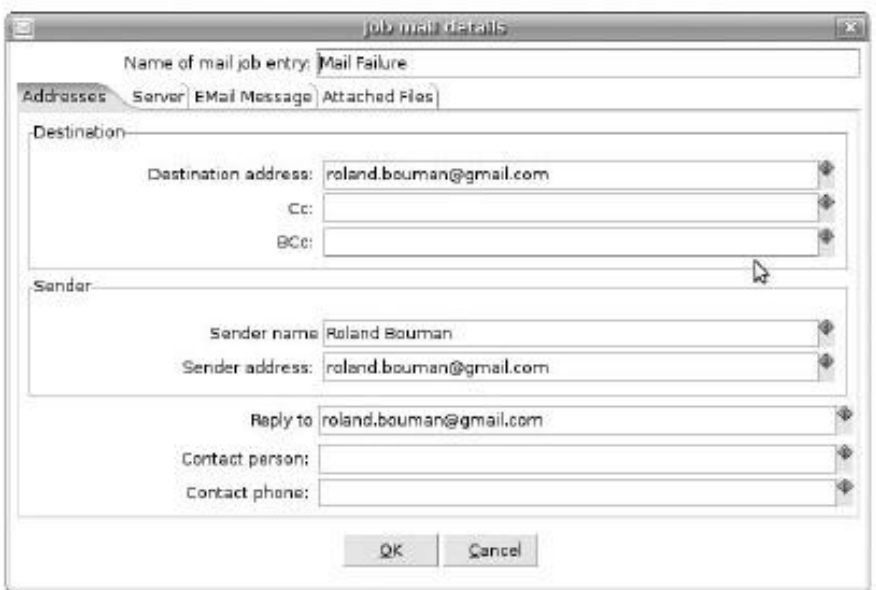


Figura 10-15: A ficha de endereços na janela de configuração da entrada de emprego Mail

Você deve especificar os detalhes do servidor SMTP na página do guia Servidor mostrados na Figura 10-16.

Você é obrigado a fornecer, pelo menos o nome do host ou endereço IP do seu servidor SMTP. Opcionalmente, você pode fornecer a porta para usar. Por padrão, a porta 25 (padrão SMTP) é usado. Na maioria dos casos, servidores SMTP requerem autenticação de usuário. Para habilitar a autenticação, verifique a autenticação de Uso caixa e fornecer o nome de usuário e senha na autenticação
usuário e campos de senha de autenticação, respectivamente. Mais e mais frequentemente, servidores SMTP requerem autenticação segura através de um protocolo, como SSL (Secure Sockets Layer) ou TLS (Transport Layer Security). Você pode especificar autenticação segura, selecionando a opção Usar autenticação segura e escolher o protocolo apropriado na caixa de listagem Secure tipo de conexão. Nota

que a comunicação de rede para um protocolo de autenticação segura em geral emprega uma outra porta. Para SSL, a porta padrão é 465. Contate seu local administrador da rede ou do sistema para obter esses dados.

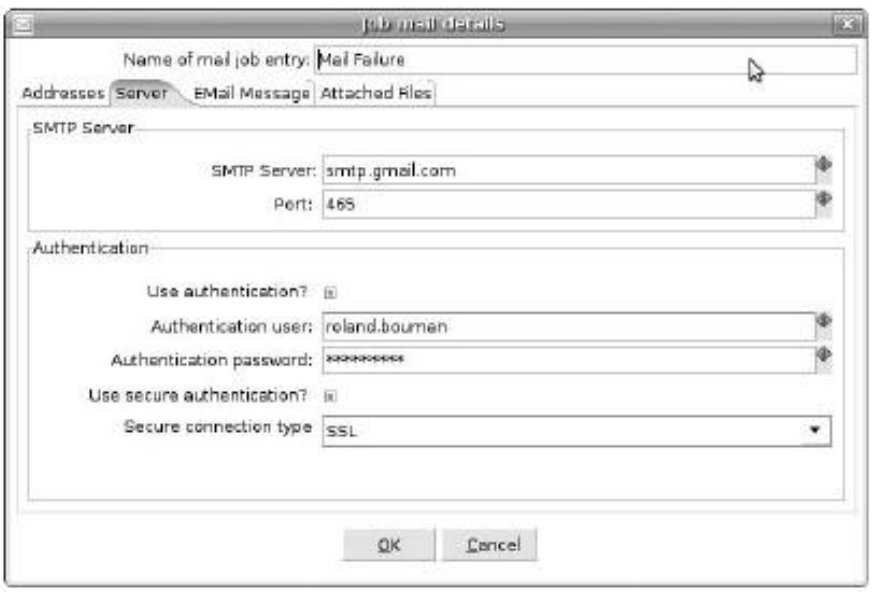


Figura 10-16: A página da guia Servidor na janela de configuração da entrada de emprego Mail

Você pode especificar o conteúdo real da mensagem na guia Mensagem de e-mail. Este ficha de registro é mostrado na Figura 10-17.

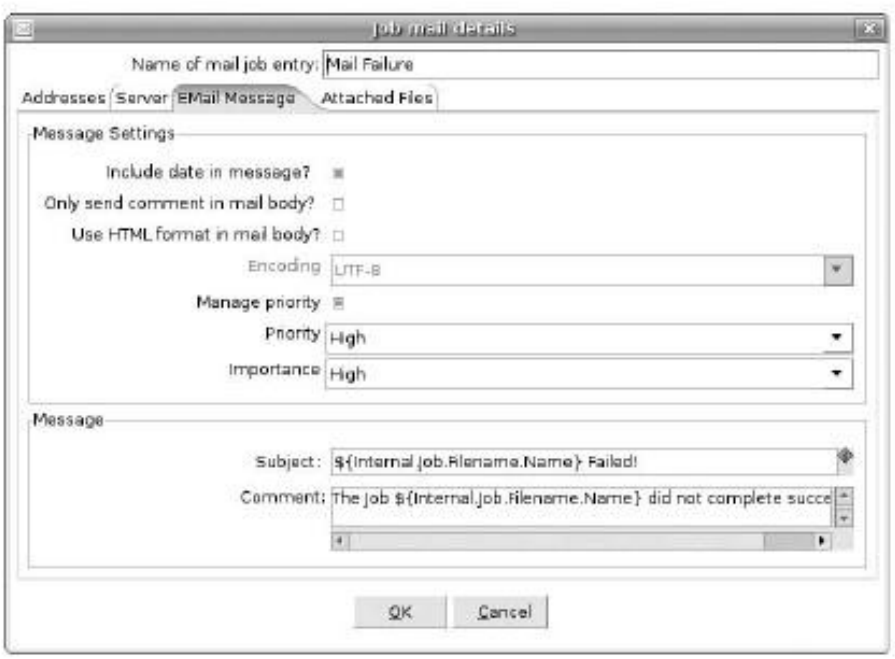


Figura 10-17: O Email Mensagem guia na janela de configuração da entrada de emprego Mail

O assunto da mensagem eo corpo estão especificados no assunto e comentar propriedades, respectivamente. Você pode usar livremente o texto e incluir referências variável
para essas propriedades. Por padrão, o PDI inclui um breve relatório da situação da transformação no corpo da mensagem, logo após o conteúdo fornecido no Comentário propriedade. Para evitar que esse relatório de status de ser incluída, selecione Apenas envie o comentário na caixa de correio do corpo. Opcionalmente, você pode selecionar
Use o formato HTML no corpo da mensagem a enviar em formato HTML e-mail. Alguns e-mail os clientes usam cabeçalhos prioridade da mensagem. Se quiser, você pode selecionar o Gerenciar
opção de prioridade para permitir isso. Quando esta estiver ativada, você pode definir a prioridade
Importância eo propriedades.

O extract_lookup_type e extract_lookup_value
Transformações

O objetivo da extract_lookup_type e extract_lookup_value trans-
formações é recuperar os dados do lookup_type e valor_procurado
tabelas do sistema de origem. O design de ambas essas transformações é extremamente simples: um passo de entrada tabela executa um SQL SELECT declaração sobre a
respectiva tabela. O fluxo de saída da etapa de entrada da tabela é liderada imediatamente a uma etapa da produção de texto de arquivo que grava os dados em arquivo. Figura 10-18 mostra a
concepção do extract_lookup_type transformação.

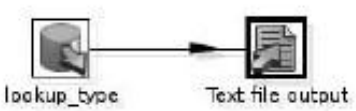


Figura 10-18: O projeto da transformação extract_lookup_type

O objetivo dessas transformações é extrair os dados tão rapidamente quanto possível a partir da base de dados mundial de filmes de classe. Embora seja possível tem a loja de transformação dos dados diretamente no wcm_staging banco de dados, nós preferimos descarregar os dados em um arquivo de texto, porque esta é mais rápido. Pode-se
argumentam que, para estas quantidades de dados, escrevendo diretamente para uma tabela no estadiamento
área deve ainda oferecer bom desempenho. Isso pode ser verdade, mas escrever desempenho não é a principal preocupação desta extração. A quantidade de tempo precisamos de ocupar o sistema de origem é a principal preocupação, que é por isso que nós optar pela solução mais rápida possível.

NOTA Neste projeto, os dados são puxados através da rede e armazenados em um arquivo no
o sistema de arquivos do host área de preparo. Isto pode não ser a melhor solução quando lidar com grandes volumes de dados. Se há um risco de congestionamento da rede, pode ser melhor para descarregar os dados para um arquivo no sistema do sistema de origem do arquivo e compactar
o arquivo antes de transferir para a área de preparo de host.

Este é um trade-off: A compressão aumenta a carga da CPU no sistema de origem. Em Além disso, este cenário aumenta a complexidade da implantação da integração de dados

solução, porque ela torna-se distribuída. No entanto, é bom saber que você pode fazê-lo. Ferramentas para distribuir soluções Pentaho Data Integration são discutidos em Capítulo 11.

A Transformação stage_lookup_data

O objetivo da stage_lookup_data transformação é para carregar a pesquisa dados em tabelas no wcm_staging banco de dados. A transformação é mostrado na Figura 10-19.

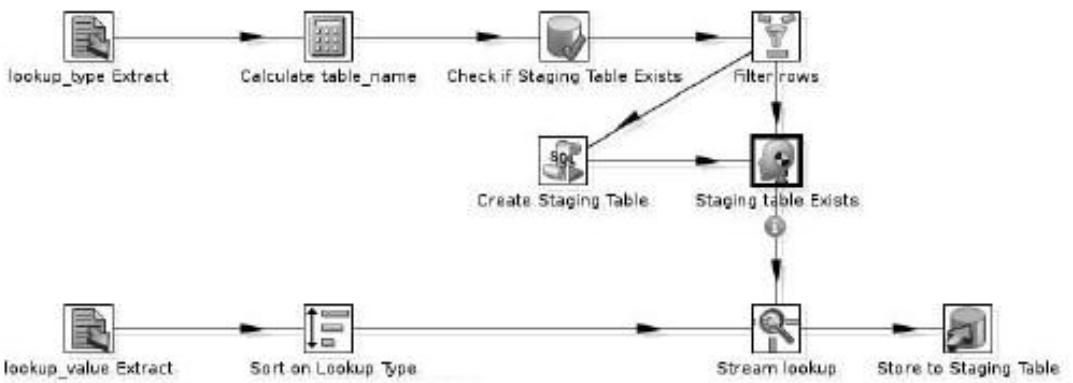


Figura 10-19: A transformação stage_lookup_data

Aqui está um resumo do que está acontecendo nessa transformação:

1. Extrato lookup_type e Extrato valor_procurado: Estes arquivos de entrada de texto etapas analisar os arquivos de texto criado pelo extract_lookup_type e transformações extract_lookup_value em um fluxo de registros.

2. Calcule table_name: Esta etapa calculadora utiliza os dados da Extrato lookup_type para gerar um nome de tabela para a tabela de teste. O nome da tabela é adicionada ao fluxo de saída no table_name de campo.

3. Verifique se o preparo tabela existe, as linhas de filtro, e as etapas existe tabela de preparação:
   Verifique se o preparo Tabela etapa existe usa o table_name campo para ver se a tabela de teste existe. O resultado da seleção é alimentado no filtro linhas etapa. Se a tabela de teste existe, a execução continua diretamente com o Encenação etapa existe tabela. Caso contrário, a transformação faz um desvio e continua com a etapa Criar tabela de preparação. Este é um Execute SQL passo que irá criar a tabela de teste de destino, e depois continua na Encenação etapa existe tabela.

4. Stream de Pesquisa e Ordenação de etapas de pesquisa Tipo: O Stream etapa de pesquisa
   junta-se, essencialmente, os fluxos de dados provenientes das duas etapas de entrada de arquivo de texto
   com base no lookup_type_id campo, que está presente em ambos os extratos. Isso adiciona o table_name campo para cada uma das linhas de entrada do Extrato valor_procurado. A Ordem de pesquisa tipo é necessário para o

operação correta da etapa de pesquisa de fluxo, que assume o principal fluxo de dados (o que vem a partir do extrato valor_procurado) é classificada valor de chave.

5. Guarde o estadiamento Table: Tabela insere saída etapa os valores de pesquisa na tabela de preparação adequado especificado pela table_name de campo.

Nós já discutimos a entrada de arquivo de texto, calculadora, e as etapas de saída de mesa das seções anteriormente neste capítulo e no anterior. Vamos agora brevemente discutir os passos que nós não encontramos antes.

Verificar se existe tabela de preparação: a tabela existe Etapa

Na Figura 10-19, a etapa rotulada Verificar se tabela existe Staging é da Mesa Existe tipo. Esta etapa é encontrado debaixo da categoria Pesquisa e seu ícone é um tambor com uma marca de seleção.

Como está implícito pelo seu nome, a Mesa etapa pode verificar se existe uma tabela é acessível sobre a conexão com o banco especificado. A janela de configuração da tabela Existe passo é mostrado na Figura 10-20.

Figura 10-20: A janela de configuração do passo existe tabela

A conexão de banco de dados é especificado pela propriedade Connection. O valor para a propriedade denominada campo TableName é o nome do campo na entrada fluxo que transmite o nome da tabela. Neste caso, o campo é chamado table_name e se origina na etapa table_name Calcular. A propriedade fieldname Resultado é usado para especificar o nome do campo que conterá o resultado da verificação. Este campo Boolean é adicionada ao fluxo de saída da etapa.

O Filtro de Passo linhas

O filtro de passo linhas é usado para escolher adequadamente entre duas alternativas caminhos de acordo com o resultado da etapa de Existe tabela. O Filtro de linhas é o passo encontradas na categoria Transformação e seu ícone é um funil (ver Figura 10-19).

O filtro de passo linhas fornece funcionalidade If-Then-Else base. A configuração da primeira etapa é mostrado na Figura 10-21.

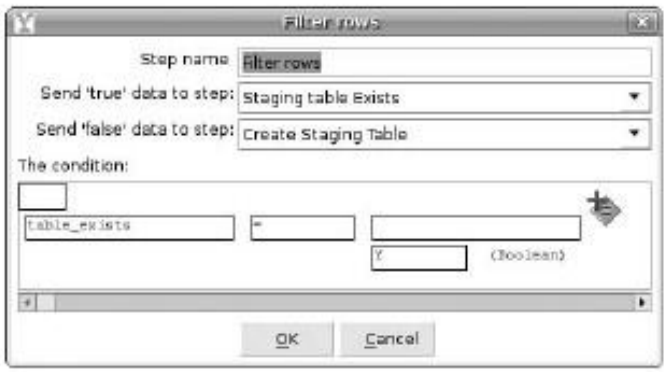


Figura 10-21: A configuração da etapa Filtrar linhas

Na etapa Filtrar linhas, você pode especificar uma condição, adicionando uma ou mais comparações. Na Figura 10-21, a apenas uma comparação é adicionado, que verifica se o valor da table_exists campo (que se originou a partir da tabela existe etapa) é igual à constante booleana Y.

NOTA Em Pentaho Data Integration, valores booleanos são representados usando o constantes da cadeia Ye Npara verdadeiro e falso, respectivamente.

Se necessário, comparações múltiplas pode ser combinada com a lógica habitual operadores. Para adicionar mais de uma comparação, clique no ícone de adição na parte superior direita da área do estado. Você pode clicar no recém-adicionada comparação com a editá-lo e escolher um operador lógico, como E ou OU combinar as comparações.

O filtro de passo deve ter duas linhas de saída de lúpulo. As etapas na outra extremidade da saída de lúpulo pode ser selecionado no drop-down list caixas para configurar ramificação. Enviar a 'verdadeira' de dados para a etapa será executada no caso, a condição avalia para TRUE, e Enviar "falso" dados para a etapa será executada em contrário. Neste exemplo, optamos por continuar com a etapa Criar tabela de teste no caso, o table_exists campo é FALSE.

Criar Staging Quadro: Execução de SQL dinâmico

A etapa de teste Criar tabela é do tipo Executar SQL. Nós descrevemos este tipo de passo, anteriormente neste capítulo, na seção"CREATE TABLE dim_date: Utilizando o Execute"Passo SQL Script.

Até agora, todas as transformações que temos discutido usado o SQL Execute passo para criar a tabela de destino, e essa transformação não é excepção. Não No entanto, é uma diferença importante na forma como essa etapa faz parte do transformação. Em todas as transformações anteriores, a etapa de execução do SQL apareceu
separado do restante da transformação, uma vez que não estava ligado a qualquer das outras etapas. Nessa transformação, a etapa de execução do SQL é parte do o fluxo de dados, e tem um salto de entrada e saída como as outras etapas.

Na transformação mostrado na Figura 10-18, o script SQL é dinâmico e parametrizado com dados de fluxo de entrada. Para ser mais específico, o SQL script é um modelo para um CREATE TABLE declaração que contém um espaço reservado para o nome da tabela. O espaço reservado é indicado com um ponto de interrogação. O SQL modelo de declaração é mostrado na Listagem 10-12.

Listagem 10-12: Um modelo de declaração de CREATE TABLE com espaço reservado para nome de tabela

```
CREATE TABLE? (
   lookup_value_idINT
, Lookup_textVARCHAR (50)
, Lookup_type_idINT
, Lookup_type_table_name VARCHAR (64)
, Lookup_type_column_name VARCHAR (64)
, PRIMARY KEY (lookup_value_id)
, UNIQUE (lookup_text)
)
```

A configuração da etapa Criar Tabela Staging é mostrado na Figura 10-22. Na janela de configuração, a execução de cada opção de linha está selecionado. Isso permite que o script SQL a ser executada para cada linha de chegada por meio de o fluxo de entrada e não apenas uma vez na fase de inicialização do transformação. Além disso, o table_name campo é especificado nos parâmetros grade. Ao manusear uma linha do fluxo de entrada, o espaço reservado (s) em o script SQL são substituídos com o valor do (s) domínio (s) especificado na grade. O resultado CREATE TABLE instrução é executada, a criação de um nova tabela.

O Passo do manequim

A encenação da tabela etapa existe é um passo Dummy. A etapa Dummy pode ser encontradas na categoria transformação e seu ícone (visível na Figura 10-19) é uma cabeça de manequim.

A etapa Dummy passa os dados da sua fluxos de entrada para sua saída fluxo (s) e não fazer nada. Apesar de não representar uma verdadeira ópera-ção, é útil para juntar e dividir recorde córregos. Isso merece alguma explicação.

Quando discutimos a transformação para o carregamento dos dim_demography dimensão, que explicou que a maioria das etapas pode ter várias entradas e registros de fluxos de saída do mesmo tipo, porque o tipo de operação por-formado pela etapa não é influenciada pela origem ou destino dos registros. Mas alguns passos são projetados para operar em múltiplos fluxos de entrada diferentes, e alguns passos de gerar múltiplos fluxos de saída diferente. Muitas vezes, isso implica

as correntes podem ter layouts registro diferente, mas a característica definidora é que o passo é projetado para conectar um significado diferente para diferentes correntes.

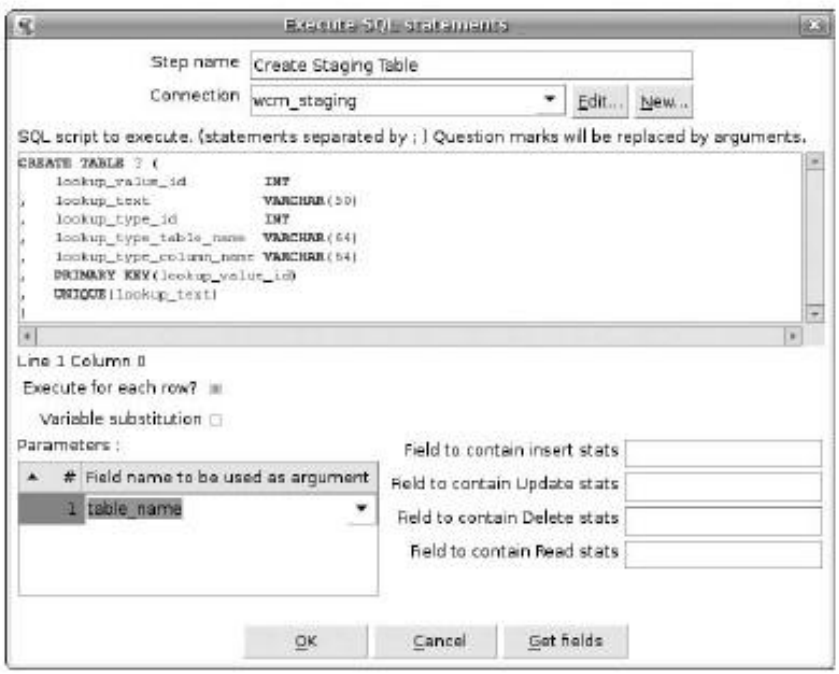


Figura 10-22: Configurando o SQL Execute passo para aceitar um parâmetro e executar uma vez para cada linha

Por exemplo, na transformação mostrada na Figura 10-19, o filtro passo linhas gera dois fluxos de saída com uma semântica diferente: o ramo 'verdadeiro' e o ramo 'falso'. Como você vai aprender em detalhes na próxima subseção, o Stream Busca aceita duas entradas diferentes correntes, ou seja, um fluxo de dados principal e uma pesquisa de fluxo. Olhe novamente a Figura 10-19. Observe as informações"pequena" (I) a meio caminho ícone abaixo do salto, saindo da etapa Dummy para o córrego etapa de pesquisa. Este ícone indica que este é realmente um fluxo de entrada especial, que é considerado a ter um papel distinto do fluxo de entrada principal.

Quando isso acontece, os dois ramos que saem das linhas de filtro passo juntos formam a fonte dos dados de pesquisa para a etapa de pesquisa Stream. Assim, a fim de permitir que o fluxo passo Pesquisa para usar todos os dados de ambos os ramos, que têm ser reunidos de alguma forma os dados podem ser tratadas como um único fluxo novamente. Este é o lugar onde o passo Dummy entra em jogo. Liderando as duas correntes para a etapa do manequim é o equivalente funcional de um SQL UNIÃO operação.

A Corrente Pesquisa Etapa

O Stream etapa de pesquisa encontra-se na categoria Pesquisa e seu ícone é um lupa.

O Stream etapa de pesquisa aceita duas entradas diferentes fluxos. Um fluxo é considerado o principal fluxo, eo fluxo de outro é a pesquisa de fluxo.
A etapa de obras, observando-se um registro da pesquisa de fluxo para cada linha no fluxo principal com base no resultado de uma comparação de valores de campo. Para o correcto funcionamento desta etapa, os registros da principal corrente deve ser classificadas de acordo com os campos que são usados para fazer a pesquisa. A configuração
para o fluxo de pesquisa utilizado na transformação para carregar os valores de pesquisa na área de teste é mostrado na Figura 10-23.

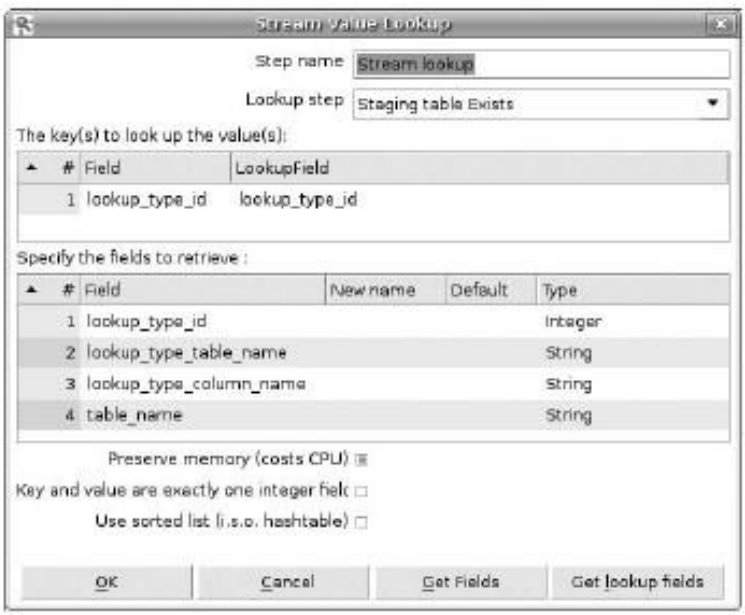


Figura 10-23: A configuração da etapa de pesquisa Stream

Usando a lista de Passo pesquisa drop-down, você pode selecionar quais passo fornece a pesquisa de fluxo de entrada. Ao clicar no botão Get Fields, você pode preencher o chave (s) para procurar o valor da (s) grelha. Na coluna Campo dessa rede, você pode especificar os campos do fluxo principal. Na coluna LOOKUPFIELD, você pode especificar os campos correspondentes da pesquisa de fluxo. Com o Get pesquisa botão de campos, você pode encher a especificar os campos a recuperar grade. Isso preenche a grelha
com todos os campos da pesquisa de fluxo. Campos nesta rede será adicionada à fluxo de entrada para gerar o fluxo de saída. As caixas abaixo desta rede podem ser utilizados para otimizar o algoritmo de pesquisa. A seleção Preservar a memória opção garante que as linhas do fluxo de pesquisa serão compactados. Isso reduz os requisitos de memória para a etapa, mas vai exigir mais ciclos de CPU para comprimir (e descompactar) os dados.

A chave e valor são exatamente uma caixa campo inteiro pode ser ativado no caso, a chave que liga o rio principal eo fluxo de pesquisa consiste em um campo Integer única e apenas um campo inteiro é obtido a partir da pesquisa

córrego. Esta caixa permite uma otimização que permite a pesquisa de dados para ser armazenados de forma mais eficiente, o que economiza memória e aumenta o desempenho. Se
o tamanho de registro dos registros de pesquisa é grande, você pode ativar o uso classificado caixa de lista para economizar memória.

Classificar em Lookup Type: o tipo Etapa Linhas

Na Figura 10-19, a Ordem de pesquisa do tipo degrau é do tipo passo Classificar linhas. Este passo é encontrada na categoria de Transformação. Você pode ver a correspondente ícone na Figura 10-19.

Esta etapa demora um fluxo de entrada não triados e cria um fluxo de saída ordenada baseado na especificação de um conjunto de campos. Opcionalmente, você pode especificar que
apenas único (com base em valores de chave) linhas devem ser passados para o fluxo de saída.
A janela de configuração é mostrado na Figura 10-24.

Figura 10-24: A janela de configuração do passo Classificar linhas

Você pode rapidamente preencher a grade de Campos utilizando o botão Get Campos. Se desejar,
você pode configurar o sentido de ordenação (ascendente ou descendente). Para campos do tipo String, você pode especificar se a honrar ou ignorar as diferenças de letras maiúsculas.

Pentaho Data Integration tenta realizar uma espécie de memória. Por padrão, 5.000 linhas podem ser ordenadas na memória. Além de 5.000 linhas, o algoritmo escreve resultados de classificação intermediária para arquivos temporários para economizar memória. Se você quiser
classificar mais de 5.000 linhas e você sabe que o jogo vai caber na memória, você pode definir o tamanho da propriedade Sort para o número máximo de linhas que você deseja classificar.

Você também pode usar o limite de memória livre (em%) propriedade para controlar uso de memória. Aqui você pode especificar a quantidade de memória que serão usados para classificar as linhas antes de armazená-classificados para um arquivo temporário. A quantidade de
memória está especificada como uma porcentagem da quantidade total de memória disponível para a máquina virtual Java.

Você pode configurar o local onde os arquivos temporários são escritos por especificar-
ção da propriedade do diretório de classificação. Você também pode especificar o prefixo TMP-arquivo, que
será usado para prefixar os nomes dos arquivos temporários escrito por esta etapa. Isso é útil principalmente para fins de depuração. Se você achar que o temporário arquivos estão crescendo muito grande, você poderia considerar selecionando o TMP Compress
Arquivos de caixa. Isso pode ajudar a economizar espaço em disco e, potencialmente, aumentar a per-
desempenho (devido à redução de I / O). Compressão virão a um custo de aumento Carga da CPU.

Guarde o estadiamento da tabela: Utilizar uma etapa de saída de mesa para carregar Várias tabelas

A etapa final da transformação mostrado na Figura 10-19 é uma mesa de saída etapa. Nós discutimos este tipo de passo em detalhes na subseção"Load dim_date: A etapa de saída de mesa"no início deste capítulo.

Em todas as transformações descritas anteriormente, foi utilizada a tabela para a etapa de saída
inserir dados em uma única tabela de destino. Neste caso, usamos o passo para carregar múltiplas tabelas de destino. Isso é configurado selecionando o É o nome da tabela definida em uma caixa de campo e especificando o table_name campo na Campo que contém o nome da propriedade da tabela.

A Dimensão Promoção

Esta seção demonstra como carregar o dim_promotion Dimensão da tabela. Este processo é relativamente simples para um número de razões:

- A tabela de dimensão é muito pequena: ele não contém muitas colunas, nem ela contém muitas linhas.
- Promoções são estáticos e não mudam ao longo do tempo. Portanto, nós não necessidade de acompanhar uma história de promoção. Isso simplifica o carregamento processo consideravelmente.
- Promoções estão em vigor por um período fixo de tempo. Isto torna mais fácil captar as variações dos dados.
- No caso de Classe Mundial Filmes, a promoção mapas tabela de dimensão somente a tabelas de origem poucos que estão disponíveis no mesmo sistema.

Ao desenvolver um plano para carregar uma tabela de dimensão, as seguintes conside-
ções entram em jogo:

- Mapeamento-In quais tabelas e colunas do sistema de origem que o dados de dimensão originou? Como essas tabelas relacionadas entre si? O mapeamento deve ser previamente conhecidos como um produto secundário da dados de projeto do armazém. No entanto, o desenvolvedor de integração de dados pode
não ter sido envolvido no processo de design de banco de dados, necessitando de um revisão do mapeamento pretendido.

- As alterações de dados, que tipo de mudanças que você espera ver na fonte tabelas, e que mudanças devem ser refletidas nos dados de dimensão? Quantas mudanças você espera estar lidando com, e quais são as Estima volumes de dados?

- Sincronização-How Com que freqüência você deseja sincronizar os dados em a tabela de dimensão com o sistema de origem? Como isso se relaciona com o disponibilidade real do sistema de origem e data warehouse?

Responder a estas perguntas não oferecer uma solução pronta, mas isso não ajudar você a ganhar uma compreensão do problema. Essa sub- em pé é um pré-requisito para o desenvolvimento de uma solução que possa satisfazer as requisitos do projeto de data warehouse.

Promoção de mapeamentos

Você pode encontrar detalhes sobre os mapeamentos para o dim_promotion tabela Capítulo 8. Para o dim_promotion tabela, você vê que existem três fontes de tabelas:

- promoção -Este quadro constitui a principal fonte de dados para o escurecer _promotion dimensão. Para cada linha da promoção tabela, haverá exatamente uma linha na dim_promotion tabela.

- site -No sistema de origem, o promoção tabela tem uma chave estrangeira para o site tabela para apontar o site que apresenta a promoção. No entanto, no armazém de dados, atributos descritivos do site tabela são dobrados no dim_promotion tabela, que é um exemplo de

  desnormalização.

- valor_procurado -Isso é muito similar ao site quadro: a promoção tabela tem uma chave estrangeira para a tabela para apontar o tipo de promoção, e os descrição textual do tipo é dobrado diretamente no de-normalizados dim_promotion tabela.

Dados Alterações Promoção

Agora que está claro qual a tabela que será o carregamento de dados, devemos analisar que tipo de alterações de dados (inclusões, alterações, remoções) será afeta a dimensão da promoção. Devemos também estimar as alterações quantas esperamos ver ao longo do tempo, e que os volumes de dados estão envolvidos.

- Para o promoção tabela, nós esperamos ver principalmente as adições. O volume de mudanças deve ser bastante baixa: mesmo se assumirmos uma nova promoção é iniciado em cada site Filmes Classe Mundial em uma base diária, nós ainda estar lidando com apenas 1.500 linhas por ano.

- Também devemos esperar algumas atualizações sobre promoções existentes. Não é improvável que uma promoção de sucesso pode ser estendida para durar mais

originalmente planejado. Mesmo se pode descartar essa possibilidade, ainda devemos levar em conta a possibilidade de que um erro entrou na fonte sistema seja corrigido posteriormente. Por exemplo, se uma data errada termina acidentalmente foi especificado na criação da promoção, é provável que o erro será corrigido mais tarde, actualizando a data de término. Outro possibilidade é que uma linha de divulgar errado pode ser substituída por uma correta um, ou até mesmo removido.

Apesar de todos esses cenários, nós esperamos ver um certo nível de estabilidade. Por exemplo, nós não esperamos uma promoção que já terminou a mudar ou ser removido. Além disso, parece razoável supor que o data de início das promoções que já começaram permanece fixo e que essas promoções não são subitamente removidas. É claro que, se nós pode fazer essas hipóteses é de que o negócio eo sistema de origem, mas estes são os pressupostos que fará para a Classe Mundial Filmes banco de dados.

- Para a tabela site, esperamos ver um número muito pequeno de linhas de todos os tempos. Durante um período de um ano ou assim, um novo site pode ser acrescentado.
  Modificações de linhas já existentes poderia ocorrer em uma base muito raras devido a eventuais alterações no título do site, e talvez a URI.
- Assumimos os dados da pesquisa de permanecer estático em todos os momentos. Nós esperamos muito
  algumas linhas aqui.

Sincronização de Freqüência

Aconteça o que acontecer, você deve sempre assegurar que as promoções que ocorreram em o passado, bem como promoções que estão atualmente ativos são carregados no Dimensão da tabela. Se você não fizer isso, você não será capaz de carregar as tabelas de fatos
como fact_customer e fact_order porque não pode ser capaz de pesquisa a chave para dim_promotion Dimensão da tabela.

Considerando que as promoções podem ser adicionados ou corrigidos durante cada dia de trabalho, parece sensato para garantir que as promoções são carregados diariamente base.

O load_dim_promotion Trabalho

Nós criamos um trabalho chamado load_dim_promotion para carregar a dimensão da promoção.
É mostrado na Figura 10-25.

O trabalho consiste de duas transformações principais:

- extract_promotion Isola- e extrai o conjunto de promoções a partir do tabela da promoção no sistema de origem que pode ter mudado desde a último carregamento.
- load_dim_promotion -Na verdade carrega os dados para o dim_promotion tabela de dimensão do data warehouse.

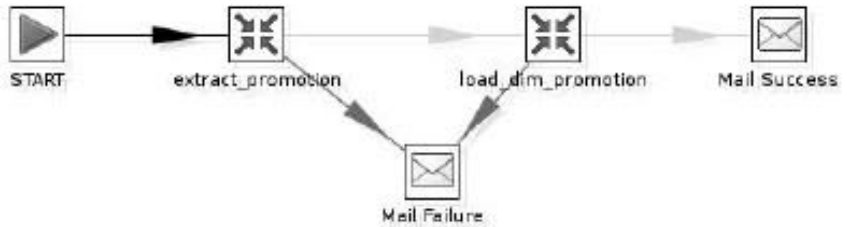


Figura 10-25: O trabalho load_dim_promotion

Como no trabalho de carregar a tabelas de pesquisa, usamos as entradas de emprego Mail para
fins de notificação.

A Transformação extract_promotion

A transformação extract_promotion é a primeira transformação na carga _dim_promotion trabalho. Sua finalidade é extrair essas linhas de promoção o sistema de origem que pode ter mudado desde a última vez que carreguei o dim_promotion dimensão. A transformação é mostrado extract_promotion na Figura 10-26.

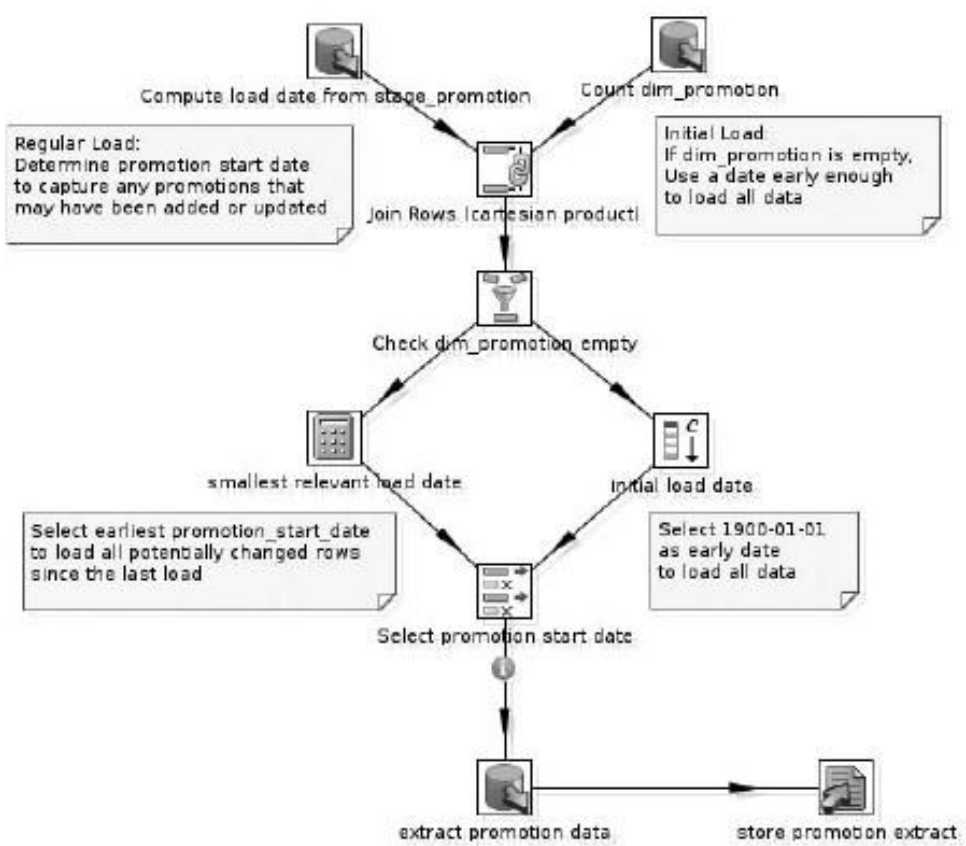


Figura 10-26: Extração de dados de promoção com a transformação extract_promotion

A transformação de obras como esta:

1. Contagem dim_promotion: Esta etapa de entrada de mesa rende uma única linha com um
   coluna que contém a contagem do número de linhas na dim_promotion
   tabela. Este é utilizado mais tarde para determinar se uma carga inicial deve ser
   realizada.
2. Calcule a data de carga de stage_promotion: Esta etapa de entrada de tabela calcula
   um valor para promotion_start_date que pode ser usado para selecionar todas as promoções
   linhas do sistema de origem que pode ter mudado desde a última vez
   nós sincronizados a tabela dim_promotion.
3. Junte-se a linhas: Isto é usado para criar um único registro da entrada de dados criado
   pelas duas etapas anteriores.

4. Verifique promotion_empty dim: Este filtro passo linhas utiliza a contagem de
   dim_promotion para determinar se é preciso fazer uma inicial ou um regular
   carga.

5. Menor data de carga correspondente ea data da carga inicial: Uma dessas datas é
   escolhidos para determinar o conjunto de linhas que serão obtidos a partir da fonte
   do sistema.

6. Selecione data de início da promoção: Descarta qualquer campos desnecessários do
   córrego, a fim de ser capaz de passar um único parâmetro para o extrato
   promoção etapa dados.

7. O extrato de promoção etapa dados realmente extrai os dados do
   tabela da promoção no sistema de origem.

8. A promoção da loja extrato etapa grava os dados extraídos promoção
   do sistema de origem para um arquivo para processamento posterior.

Determinar as alterações nos dados de Promoção

Para determinar quais linhas para extrair, raciocina do seguinte modo:

- Se o dim_promotion tabela está vazia, então estamos lidando com uma inicial
  carga, e temos de extrair todo o conjunto de promoções a partir da fonte
  do sistema.

- Se o dim_promotion tabela não estiver vazia, ela será preenchida por dois tipos de
  registros promoção:""promoções activas, ou seja, os registros para os quais a
  promotion_end_date reside no futuro e"acabado"promoções, os
  promoções para os quais a promotion_end_date está no passado.

Para manter o controle de alterar os registros de promoção, nós apresentamos uma tabela de preparação
chamada stage_promotion. Para executar cada transformação, que irá carregar as linhas
a partir do extrato e armazená-los aqui. Manter esses dados nas tabelas de preparo
permite rastrear todas as alterações na tabela de dimensão com a fonte,

que é uma grande ajuda na solução de problemas. A CREATE TABLE declaração para a stage_promotion tabela é mostrado na Listagem 10-13.

Listagem 10-13: A instrução CREATE TABLE para a tabela stage_promotion

```
CREATE TABLE wcm_staging.stage_promotion (
   promotion_idSMALLINTNOT NULL,
   website_idSMALLINTNOT NULL,
   promotion_titleVARCHAR (50) NOT NULL,
   promotion_type_lookup SMALLINTNOT NULL,
   promotion_start_date DATENOT NULL,
   promotion_end_dateDATENOT NULL,
   extract_filenameVARCHAR (255) NOT NULL,
   extract_linenumberINTEGERNOT NULL,
   load_timestampTIMESTAMPNOT NULL
                    DEFAULT CURRENT_TIMESTAMP
);
```

Observe que a estrutura desta tabela é muito parecida com a do original promoção tabela. A diferença é que esta tabela tem três colunas extra:

- Extrair nome de arquivo -O nome do arquivo usado para carregar os dados novamente.
- linenumber Extrato -Isto é usado para armazenar o linenumber no extrato. Isto torna mais fácil de carregar uma parte do extrato deve precisamos fazê-lo no futuro.
- Load_timestamp -Um timestamp gerado automaticamente.

Listagem 10-14 mostra o SQL usado para determinar a data de carga.

Listagem 10-14: Usando SQL para determinar promoções que podem ter sido alterados

```
SELECT MIN (promotion_start_date) data_inicial
FROMstage_promotion
ONDE promotion_end_date> load_timestamp
```

O coração da consulta mostrada na Listagem 14/10 é formado pelas ONDE condição. Todas as linhas que têm uma data maior que a data final de carga última pode mudaram entretanto, por isso devemos tratá-los.

Agora considere o que aconteceria se o stage_promotion tabela está vazia. Bem, ele iria retornar um valor NULL. Agora considere o que acontece no caso Rodamos o direito de consulta depois que terminar de carregar a dim_promotion e stage_promotion tabelas. Neste caso, estamos up-to-date para o momento e A consulta também retorna zero linhas.

Porque não podemos distinguir facilmente entre esses dois casos de zero linhas, nós explicitamente verificar se a tabela de dimensão está vazio. Isso é feito

usando um simples SELECT COUNT (*) . . . script. Se o resultado for zero, pode ser certeza de que nada foi carregado, e usamos 1900-01-01. Caso contrário, use a data determinado pelo script mostrado na Listagem de 10-14, e usá-lo como parâmetro para a etapa de extração atual. Figura 10-27 mostra como o campo de data é utilizado como um parâmetro.

Figura 10-27: Usando a data como parâmetro para uma entrada em degrau Tabela

Conforme mostrado na Figura 10-27, um ponto de interrogação é usado na instrução SQL onde o valor deve ser inserido. Além disso, os dados da lista Inserir etapa caixa deve ser usado para apontar exatamente quais dados passo será recebido.

Salvando o extrato e passando sobre o nome do arquivo

As linhas extraídos são salvos no disco usando um simples texto etapa arquivo de saída. Para obter um melhor controle sobre a extração e conseguir uma melhor manutenibilidade do trabalho como um todo, nós usamos alguns recursos PDI para trabalhar com arquivos. A configuração da etapa de produção de texto do arquivo é mostrado na Figura 10-28.

Note-se que foram selecionados a data Incluir no nome do arquivo e incluir o tempo em Nome caixas. Isto assegura que teremos nomes bastante originais para extratos. Também escolhemos os nomes Adicionar ao resultado checkbox. Isso permite que o nome do arquivo a ser passado entre dois ou mais transformações no trabalho.

Levantando o arquivo e carregar o Extrato

A transformação load_dim_promotion segue imediatamente o extract_pro- movimento de trabalho. O objetivo deste trabalho é ler o extrato na área de preparo e carregar a tabela dim_promotion. Figura 10-29 mostra a transformação inteira.

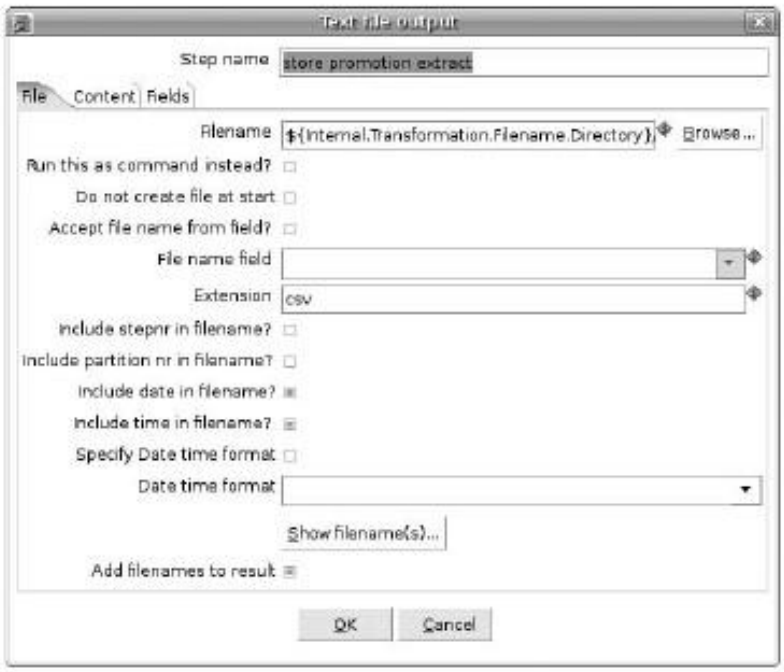


Figura 10-28: Configurando nomes de arquivo com data e hora e passar o arquivo para o resultado

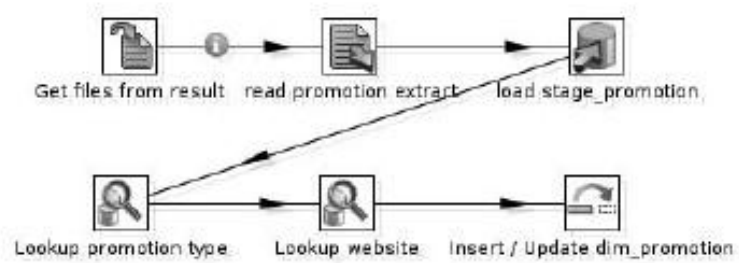


Figura 10-29: Carregando o extrato em dim_promotion

Aqui está um rápido resumo do que está acontecendo nessa transformação:

1. O Obter arquivos de etapa resultado: Esta etapa é obrigatória para pegar o arquivo que foi criado pela transformação extract_promotion.

2. leia promoção extracto: Esta é uma entrada de texto comum que nós configuramos para pegar o arquivo criado pela transformação anterior.

3. stage_promotion carga: Isso carrega o extrato no stage_promotion tabela.

4. Pesquisa tipo de promoção e do site de pesquisa: Estes passos são usados para encontrar e adicionar campos a partir do site relacionados e tabelas de referência de valor.
5. Inserir / dim_promotion Update: Esta etapa é utilizada para carregar o dim_promotion tabela. Se uma linha de promoção é novo, ele será adicionado. Se uma linha de promoção já existia, os seus valores são substituídos com a valores atuais.

Resumo

Neste capítulo, você expandiu suas habilidades PDI. Você aprendeu a:

- Executar SQL
- Gerar linhas
- Formatar números e datas
- Efetuar cálculos
- Junte-se a correntes
- Split córregos
- Criar postos de trabalho
- Passe os arquivos entre as transformações

Além das muitas coisas abordadas neste capítulo, há ainda mais coisas que não cubra. Por favor, consulte o site do livro para download de todos os transformações e empregos para carregar o World Class armazém de dados de filmes.

CAPÍTULO

11

Implantando Dados Pentaho Soluções de Integração

Como você aprendeu nos capítulos anteriores, durante a concepção eo desenvolvimento fase, a solução Pentaho Data Integration (PDI) é executada utilizando principalmente Spoon. Após a fase de desenvolvimento, a solução de integração de dados é normalmente mudou-se para um ambiente de servidor, seja para testes ou de produção finalidades.

No servidor, transformações e empregos não são geralmente lançados utilizando uma ferramenta gráfica do usuário, como Spoon. Em vez disso, as medidas são tomadas para assegurar a execução automatizada. Dependendo dos requisitos da integração de dados solução, a execução pode ser agendada ou conduzido de maneira contínua por algum processo em segundo plano. De qualquer forma, o administrador deve agir no sentido de tomar as solução fora do ambiente de desenvolvimento para colocá-lo para trabalhar em seu alvo plataforma. Nós nos referimos a este processo como implantação.

Há mais a implantação do que instalar o software e configuração de execução automática. As medidas devem ser postas em prática para permitir que o sistema administradores para verificar rapidamente, e se diagnosticar, necessário e reparação, o solução de integração de dados. Por exemplo, deve haver alguma forma de notificação para confirmar se a execução automática do ocorrido. Além disso, os dados devem ser reunidos e analisados para medir o quão bem os processos são executados. Nós nos referimos a essas atividades como monitoramento.

Neste capítulo, apresentamos uma visão geral dos recursos e ferramentas que você pode usar para organizar a implantação e acompanhar sua solução PDI. O objetivo do presente capítulo não é para ditar como você deve administrar as suas soluções de PDI. Em vez disso, Neste capítulo deve fornecer informações suficientes para tomar decisões informadas decisões sobre quais ferramentas e estratégias de melhor atender às suas necessidades.

Configuration Management

Empregos e transformações dependem de recursos como diretórios do sistema de arquivos, arquivos e servidores de banco de dados. Durante o desenvolvimento, esses recursos são normalmente
reservados para fins de desenvolvimento. Por exemplo, em vez de apontar um transformação dos sistemas de fonte real, de desenvolvimento ou versão de teste do o sistema de origem é usado. Da mesma forma, a saída da transformação é dirigido para uma versão de desenvolvimento ou teste do sistema de destino.

Para a implantação, todos os elementos que são específicos para o ambiente de desenvolvimento
mento também devem funcionar correspondente no sistema de destino. Em alguns casos, pode ser possível fazer isso sem alterar a solução. Por exemplo, em
no caso de recursos de arquivo, pode ser possível usar consistentemente caminhos relativos, que são resolvidos dinamicamente em tempo de execução. Em outros casos, algo deve acontecer para apontar os componentes da solução de integração de dados para o recursos adequados para o ambiente de destino.

Uma maneira de oferecer os recursos certos para a plataforma alvo é modificar a transformação ou o trabalho em conformidade. No entanto, este não é um muito bom solução. Pode ser simplesmente muito trabalho para substituir todas as referências ao banco de dados
servidores e coisas do género em toda a solução. Mais importante, há uma visível possibilidade de modificação de um emprego ou transformação irá introduzir erros.

A melhor maneira de prever os recursos adequados é de alguma forma parametrizar todos componentes que são dependentes do ambiente. Desta forma, estes dados pode ser fornecido em tempo de execução, sem alterar o trabalho ou a transformação em si.

Nós nos referimos a todos os dados dependentes de ambiente como o configuração. O processo de
responsável por manter esses dados e fornecendo-lhe a solução é chamada gerenciamento de configuração. Existem algumas construções que permitem a configuração gestão:

- Variáveis de configuração empregos e transformações
- conexões JNDI para o gerenciamento de conexões de banco de dados
- Um repositório para gerenciar vários desenvolvedores trabalhando nos mesmos dados solução de integração, e também para gerenciar as conexões de banco de dados

Esses tópicos são discutidos em detalhe no restante desta seção.

Usando variáveis

A maioria das propriedades de etapas, as entradas de emprego e conexões de banco de dados pode ser param-
eterized usando variáveis. As variáveis são espaços reservados para os valores. Durante a execução,
os valores reais para essas variáveis se torna conhecido, ea ocorrência de cada variável é substituído pelo seu valor de tempo de execução. Para ser mais preciso, o real substituição da variável com o valor ocorre na inicialização da etapa

fase, e permanece fixa durante a fase de execução (a inicialização e
fases de execução são explicados no Capítulo 10 na seção"CREATE TABLE
dim_date: Utilizando o Execute SQL"Passo Script).

NOTA etapas de transformação, muitas vezes permitir que as propriedades de ser parametrizada por
áreas específicas do fluxo de entrada. Embora este dispositivo pode às vezes ser usados
para a gestão de configuração, muitas vezes é melhor manter uma clara separação
entre a transformação de dados reais (que é o domínio dos campos) eo
configuração do ambiente específico (o que é melhor feito com variáveis).

Variáveis em propriedades de configuração

Nos diálogos de configuração, um ícone de sinal vermelho minúsculo dólar aparece no lado direito
das propriedades onde as variáveis podem ser inscritas. Este ícone é mostrado na
Figura 11-1.

Figura 11-1: O ícone variável

Para se referir a uma variável, você pode usar uma sintaxe semelhante ao UNIX ou
variáveis de ambiente Windows. O código a seguir mostra duas alternativas
maneiras de se referir a uma variável chamada foo:

```
$ {Foo}
%%%% Foo
```

Isso é perfeitamente válido para incorporar referências variáveis dentro de uma propriedade literal
valor. Por exemplo, a Figura 11-2 mostra a janela de configuração de um arquivo de texto
etapa de saída que usa a variável os_user para parametrizar parte do arquivo
caminho do sistema para o arquivo de saída.

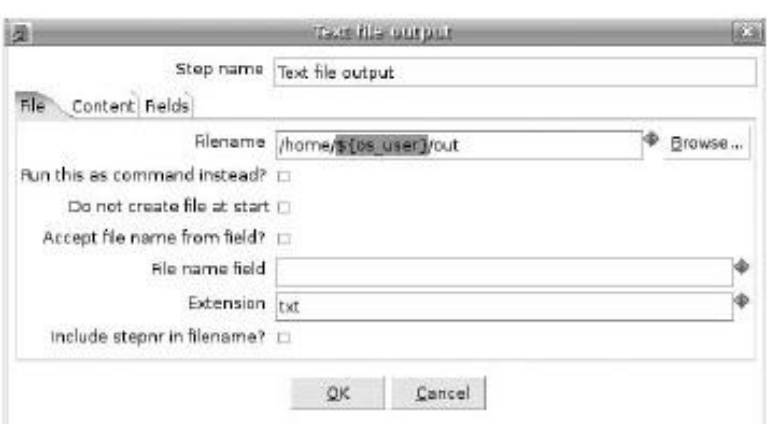


Figura 11-2: Incorporar referências variáveis nos valores dos imóveis

NOTA Se você não quer o tipo da variável, você também pode usar o teclado atalho Ctrl + Barra de espaço, que traz uma lista onde você pode selecionar o desejado variável. Por exemplo, a Figura 11-3 mostra a lista de opções de todas as variáveis disponíveis. Observe a dica exibir o valor atual da variável.

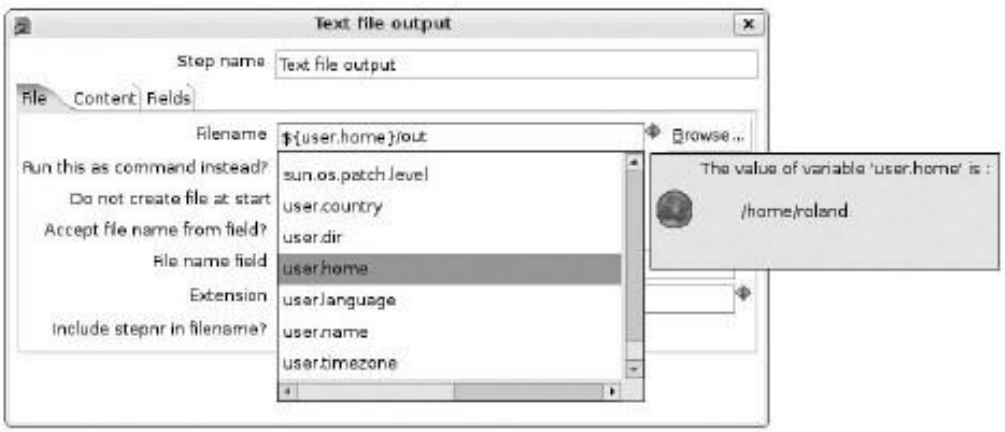


Figura 11-3: Escolher as variáveis de uma lista

Variáveis de Usuário

Como o nome implica, variáveis definidas pelo usuário são criadas pelo usuário. A os_user variável mostrado na Figura 11-2 é um exemplo típico de tal definidos pelo usuário variável. Variáveis definidas pelo usuário obter o seu valor em uma das seguintes maneiras:

- As transformações podem criar e / ou alterar o valor das variáveis usando um Definir Variáveis etapa.
- Um conjunto de variáveis de entrada de emprego estão disponíveis para ajustar as variáveis a partir de um trabalho.
- Em Spoon, você pode definir valores padrão das variáveis por posto de trabalho e por transfor-
  mações na grade que aparece no canto inferior direito do Executar uma transformação e executar um trabalho de diálogos. Na fase de desenvolvimento, esta permite a utilização de variáveis, mesmo quando a execução do trabalho ou de transformação
  autônomo dentro Spoon.
- Em Spoon, você pode criar e atribuir variáveis globais no Envi Set-
  variáveis ente diálogo. Você pode chamar isso de diálogo através do menu Editar Ambiente valores definidos. Usando Ambiente MenuEditShow
  Valores, você pode inspecionar todas as variáveis disponíveis e seus valores. Ambos desses diálogos são úteis para o desenvolvimento ou testes múltiplos empregos ou transformações que dependem de um conjunto comum de variáveis.
- Variáveis podem ser definidos como <NAME> = Valor par na kettle.properties arquivo. Este arquivo reside no . Chaleira Diretório sob o diretório home do usuário.

NOTA Para sistemas baseados em UNIX, a localização da kettle.properties é

/ Chaleira home / <user> /..

Para o Windows, esta é geralmente C: \ Documents and Settings \ <usuário> \ chaleira..

variáveis definidas pelo usuário que são definidas no kettle.properties arquivo tem escopo global. Isso significa que eles sejam acessíveis em todos os postos de trabalho e transformações. Para
Defina a entrada de passo variáveis e trabalho, o escopo pode ser definido na variável coluna âmbito tipo, conforme mostrado na Figura 11-4.

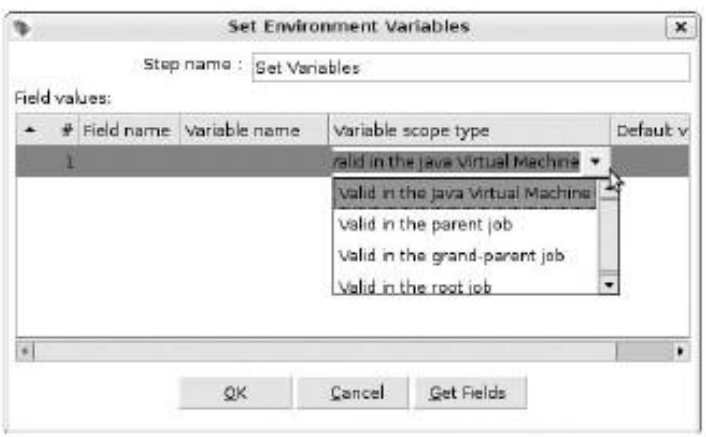


Figura 11-4: Escolhendo uma variável do escopo de uma etapa Definir Variáveis

Os escopos disponíveis são:

- Válido no trabalho de pai-Este o espaço está disponível tanto para o conjunto de variáveis
  etapa de transformação, bem como a entrada de emprego. Variáveis com este escopo estão disponíveis no trabalho que contém a entrada de transformação ou de emprego , respectivamente. Outras transformações constantes que o trabalho também pode se referir a
  da variável.
- Válido no emprego atual-Este o espaço está disponível para o conjunto de variáveis Emprego entrada. Variáveis com este escopo estão disponíveis no trabalho que contenham
  e as transformações e os trabalhos nela. Funcionalmente neste âmbito é equivalente ao Válido no âmbito de trabalho dos pais disponíveis no conjunto Variáveis etapa de transformação.
- Válido no pai-avô de emprego Este o espaço está disponível para o conjunto Variáveis etapa de transformação. Variáveis com este escopo estão disponíveis no trabalho que contém essa transformação (o pai) e também no
  trabalho que contém esse trabalho (pai-avô). Todos os trabalhos e as transformações contido pelo trabalho de avós pode fazer referência ao valor, também.
- Válido na raiz de emprego Variáveis com este escopo estão disponíveis no alto nível de emprego e todos os trabalhos e as transformações directa ou indirectamente

continha referências nele pode variável. Este espaço é apoiada por tanto passo da transformação e da entrada de emprego.

- Válido na Máquina Virtual Java (JVM)-O variável é um verdadeiro global variável e é visível a todos os trabalhos e as transformações que acontecem a rodar na mesma instância da máquina virtual Java. Este espaço é suportada tanto pela etapa de transformação e da entrada de emprego.

ATENÇÃO Use variáveis com o escopo da máquina virtual com cautela. Em cenários onde uma máquina virtual Java é utilizado para executar várias instâncias do PDI motor, como durante a execução dentro do servidor de BI Pentaho, todos os trabalhos em execução referem-se para a mesma instância da variável. Isso, certamente exclui qualquer utilização caso em que o valor da variável que precisa ser mudado. Se o valor for mudou, todos os trabalhos que fazem referência a variável quase que imediatamente ver o novo valor da variável.

Variáveis internas

Built-in variáveis refletem propriedades das coisas que têm a ver com a execução ambiente de tempo e Pentaho Data Integration si. Essas variáveis são pré-definidos, e os valores são automaticamente preenchido pelo motor.

Built-in variáveis incluem o nome da etapa de transformação de corrente, o Local da transformação em curso, bem como o nome do sistema operacional.

Variáveis Exemplo: Conexões de banco de dados dinâmico

Para ilustrar como realmente usar variáveis, vejamos um exemplo prático que ilustra como você pode usar variáveis para gerenciar a conexão de banco de dados configuração. No exemplo, as variáveis são utilizadas para configurar o MySQL conexão, mas os mesmos princípios se aplicam a qualquer conexão com o banco.

1. Criar uma nova transformação e armazená-lo como set_variables em um diretório no sistema de arquivos. Adicione um passo Gerar linhas da categoria de entrada. Configure o passo, adicionando o HOST,DB,PORT,USUÁRIO,SENHAE TABELA variáveis, como mostrado na Figura 11-5. Observe os valores inscritos no Valor padrão da coluna da grade e da definição do tipo variável alcance. Executar a transformação uma vez, isso irá criar as variáveis do Spoon ambiente. Isso torna mais conveniente para se referir a essas variáveis a partir de outras transformações e trabalhos que você pode estar projetando.

NOTA Observe que no set_variables etapa, a senha é inserida na planície texto. Isto constitui um risco de segurança, como qualquer um que pode ler o arquivo de transformação também pode ler a senha. Este risco de segurança podem ser mitigados usando senhas ofuscado. Ofuscado senhas são discutidos em mais detalhes na seção"Usando senhas de banco de dados ofuscado"mais adiante neste capítulo.

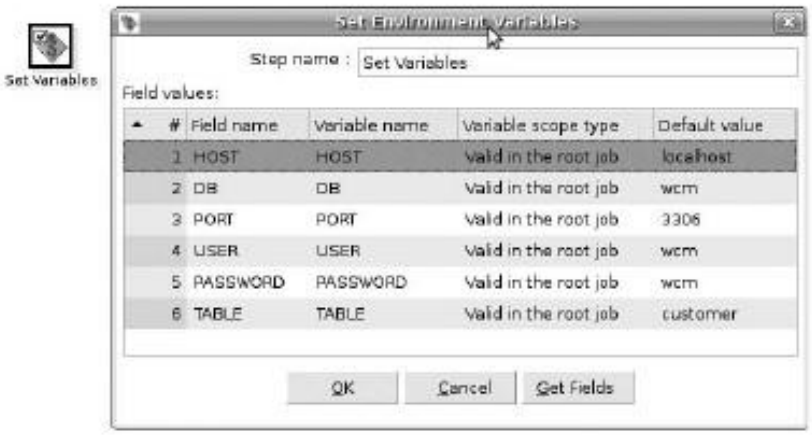


Figura 11-5: Definindo variáveis

2. Criar uma outra transformação chamada count_rows e armazená-lo no mesmo localização como set_variables transformação. Adicionar um banco de dados MySQL conexão chamada Fonte e usar referências variáveis para a conexão propriedades, como mostrado na Figura 11-6.

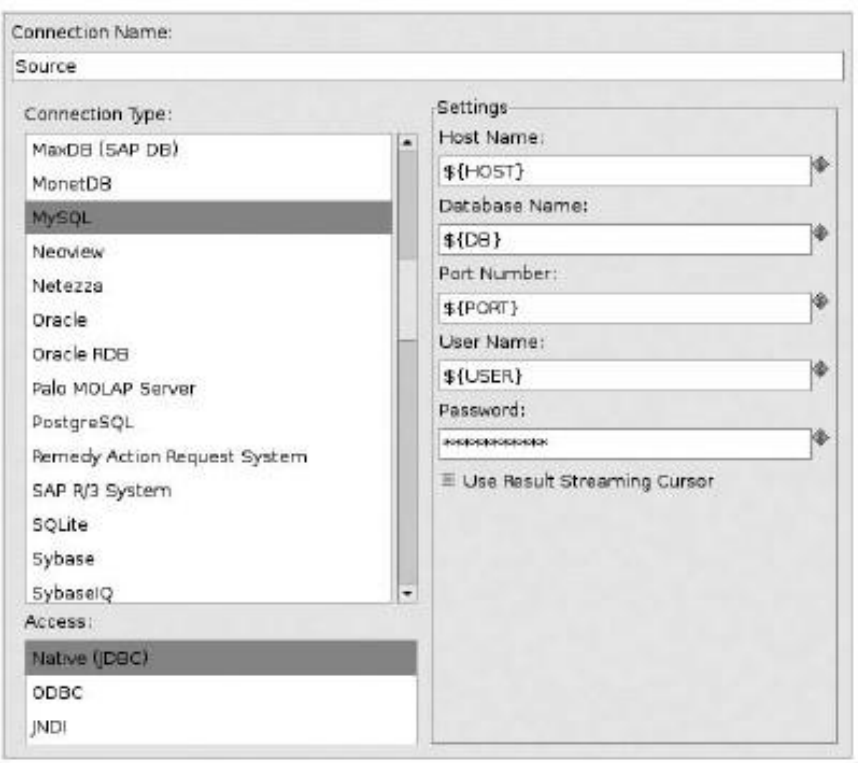


Figura 11-6: Uma conexão de banco de dados variáveis

Note-se que todas as propriedades no quadro Configurações são parametrizados usando
referências de variáveis. (A Senha valor da propriedade é mostrado como uma série

de asteriscos, embora o seu valor é realmente $ {PASSWORD}.) Teste o conexão. Em caso de falha, a causa mais provável é que você se esqueceu de executar o set_variables transformação (ou talvez um erro de digitação ou mais dos valores).

NOTA Embora as propriedades da conexão pode ser parametrizado com variáveis, você não pode usar variáveis para definir dinamicamente o tipo de conexão com o banco. No entanto, existe uma solução para essa limitação que possa atender às suas necessidades. Você pode criar uma conexão de banco de dados genéricos, e usar variáveis para parametrizar o Custom URL de conexão e Custom Class Driver propriedades de nome. Este permite que você alterne entre, digamos, um banco de dados MySQL para um banco de dados Oracle dependendo dos valores das variáveis.

Esta solução pode não ser adequado em todos os casos. Em uma conexão JDBC específico, o tipo de configuração de banco de dados SQL influências que dialeto é usado pelo PDI para comunicar ao banco de dados. Alguns passos se comportam de maneira diferente dependendo do recursos do banco de dados, e por essa razão alguma funcionalidade não está disponível para conexões de banco de dados genéricos.

3. Dê um passo de entrada de tabela e configurá-lo para usar o Fonte conexão. Digite a seguinte instrução SQL:

```
'$ {Tabela}' SELECT table_name como
, COUNT (*) AS row_count
FROM $ {tabela}
```

A variável TABELA aparece duas vezes na instrução SQL: uma vez como é no DA cláusula, quando o seu valor será utilizado como identificador de tabela, e uma vez entre aspas simples no SELECT lista, onde é usado como uma string literal. Certifique-se de verificar as variáveis substituir no script? caixa de forma que estes instâncias de uma variável são substituídos com o valor da TABELA variável em tempo de execução.

ATENÇÃO Na instrução SQL anterior, as variáveis são substituídas por suas valor quando a transformação é inicializado. Devido a substituição de variáveis é basicamente, uma ação de substituição de texto simples, pode levar a resultados inesperados quando usá-lo para gerar SQL (ou script em geral). Se o valor da TABELA variável não é um identificador simples, mas contém caracteres que têm um significado de suas próprias na linguagem SQL, um erro de sintaxe poderia ocorrer. No pior caso, a instrução SQL resultante pode mesmo ser prejudicial e entregar o resultado errado ou apagar acidentalmente dados.

Por esta razão, deve-se evitar o uso de variáveis no script, se possível. Para a Tabela passo de entrada, muitas vezes você pode usar parâmetros ao invés de variáveis. Os parâmetros são espaços reservados para o valor de expressões, e são indicados com um ponto de interrogação. Eles

pode aparecer em qualquer lugar onde você normalmente pode ter um valor de expressão como um literal ou uma referência de coluna. No caso do exemplo anterior, a variável em o SELECT lista podem ter sido omitidas, usando um parâmetro para que a primeira parte da declaração seria:

```
SELECT? AS table_name
```

4. Adicionar um texto passo a saída do arquivo, e adicionar um salto de entrada da tabela entrada em degrau. Configure o passo e editar o Nome do arquivo propriedade assim que lê

```
Internal.Transformation.Filename.Directory $ {} / $ {tabela}
```

como mostrado na Figura 11-7. Observe as duas variáveis separadas por uma barra. A primeira é uma variável interna que lhe permite construir caminhos de sistema de arquivo relativo
para o local de transformação. A segunda é uma variável definida pelo usuário que foi criado quando você executou o set_variables transformação.

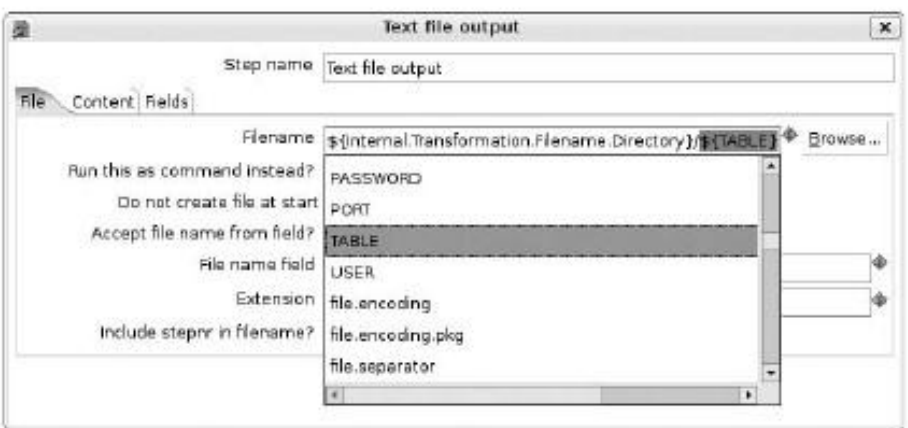


Figura 11-7: arquivo de saída variável

DICA Você pode usar o atalho Ctrl + Barra de espaço do teclado para abrir uma lista dos actuais variáveis e selecioná-los. Conclua a configuração, adicionando o table_name e row_count campos, pressionando o botão Get Campos na guia Campos.

5. Criar um emprego e guarde-o no mesmo local que as transformações. Adicionar uma entrada de trabalho START e duas entradas de emprego Transformação. Configurar um
das etapas de transformação. Renomeá-lo para set_variables e modificar o transformação propriedade filename para que ele lê

```
$ {} Internal.Job.Filename.Directory set_variables.ktr /
```

Note que este é semelhante ao Nome do arquivo configuração para o arquivo de texto-para fora
colocado na etapa count_rows transformação: esta variável tem o efeito de fornecendo um caminho de sistema de arquivo relativo ao do trabalho atual. Criar um hop
a partir da entrada START para o set_variables Emprego entrada. Renomeie o outro entrada de transformação do emprego para count_rows e aponte para o count_rows

transformação. Dê um salto que vai do set_variables entrada para o trabalho count_rows Emprego entrada. Seu trabalho agora deve ser semelhante à Figura 11-8. (Note que na figura, o conteúdo das entradas do trabalho de transformação Também são mostrados a seguir o trabalho real. Isso é adicionado para o esclarecimento sobre
tela, o trabalho se parece com a metade superior da Figura 11-8.) Execute o trabalho.

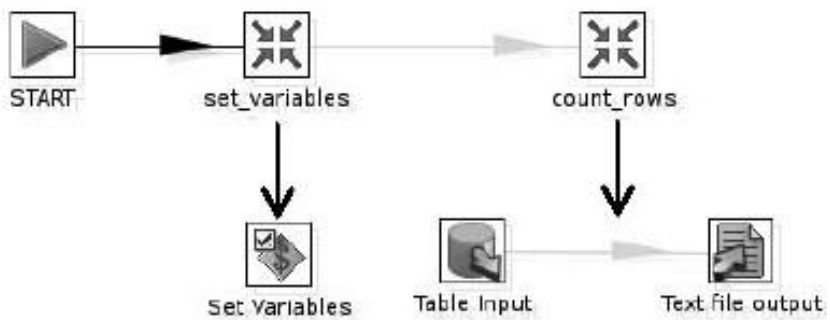


Figura 11-8: Um trabalho com uma conexão de dados variáveis

Neste exemplo, você criou um trabalho que chama duas transformações. O primeiro transformação utiliza o conjunto de variáveis passo para criar um número de definido pelo usuário
variáveis. Essas variáveis, em seguida, tornou-se disponível a nível do trabalho de chamada. A transformação subsequente pegou as variáveis utilizadas e seus valores
para configurar uma conexão de banco de dados. Além disso, os caminhos de arquivo foram criados em relação
para o local do trabalho e arquivos de transformação usando as variáveis.

Mais sobre a etapa Definir Variáveis

Do ponto de vista do gerenciamento de configuração, é claro que o uso de variáveis internas contribui para a facilidade de gerenciamento de configuração. Relativa eliminar caminhos hard-wired caminhos do emprego e transformações, tornando-
mais fácil a transferência de arquivos a partir do ambiente de desenvolvimento para a implantação
ambiente sem quebrar a solução de integração de dados.

Pode ser menos claro como as variáveis definidas pelo usuário criadas pelo conjunto etapa Variáveis melhorar a gestão de configuração. No exemplo, parece
as propriedades de configuração para a conexão do banco de dados ainda são codificados dentro de uma transformação, eles simplesmente se mudaram de uma transformação para o outro. No entanto, ainda é uma melhoria se você considerar que transformações mais como row_count podem ser adicionados ao trabalho. Estes outras transformações podem usar as mesmas variáveis para definir o seu banco de dados conexões no exatamente da mesma maneira, permitindo que todos os parâmetros de conexão para
todo o trabalho a ser alterada pela transformação de edição de um único.

Assim, o conjunto de variáveis passo pode ser usado como um único ponto de definição para
parametrizar recursos como conexões de banco de dados para um trabalho inteiro, ou nomes de arquivos e diretórios. Embora esta seja uma boa coisa, ainda é indesejável

que o set_variables transformação teria de ser editado para ajustar o
transformação para trabalhar em outro ambiente.

Felizmente, o conjunto de passo variáveis não se limita aos valores padrão usados no exemplo. Pelo contrário, o modo normal de operação é a utilização do fluxo de entrada Variáveis da etapa Definir como preencher os valores. Desta forma, os dados de configuração pode
ser armazenados em um recurso externo para qualquer trabalho ou transformação. Isso permite que o
configuração a ser gerenciado usando arquivos ou tabelas de dados ou quaisquer dados fonte pode ser lido por Pentaho Data Integration.

Defina variáveis Gotchas Etapa

Um ponto importante a lembrar é que o conjunto etapa de transformação de variáveis existe para que as transformações podem transmitir dados de configuração para os empregos e
outras transformações. Você não deve tentar usar o conjunto de variáveis para o passo transmitir informações dentro de uma mesma tarefa, e / ou transformação.

Você não pode de forma confiável as duas variáveis definir e utilizar dentro de um mesmo trans-
formação, porque a substituição de variável ocorre na fase de inicialização.
Na fase de execução, o valor permanece fixo até o trabalho ou a transformação está terminado. Se você achar que você gostaria de definir as variáveis usando dados de um transformação para configurar as etapas subseqüentes da mesma transformação, a solução é criar um novo trabalho com duas transformações. Dessa forma, você pode definir as variáveis na primeira transformação, e usá-los na outra.

Lembre-se também que o conjunto de variáveis etapa exige exatamente uma linha de entrada.
Você pode achar que é contra-intuitivo que o conjunto de variáveis etapa não suporta várias linhas de entrada, mas tente imaginar como seria trabalhar nos casos em que a transformação é parte de um trabalho. O único trabalho que considera a variável como é depois
a transformação inteira é executado. O trabalho não pode ver qualquer um dos valores da variáveis poderiam ter tido durante a transformação, ela só pode ver o que
valor foi modificada pela última atribuído à variável.

Usando conexões JNDI

Na seção anterior, você aprendeu a usar as variáveis para a configuração
de gestão. Embora as variáveis, a solução mais flexível para lidar
com múltiplas configurações, eles têm a desvantagem de que os postos de trabalho e transformações precisam ser projetado especialmente para se beneficiar deles.

Existe uma forma alternativa para gerenciar as configurações de conexão do banco de dados.
Você pode considerar usar JNDI em vez de conexões JDBC puro.

O que é o JNDI?

JNDI (""Gindy pronunciado) é um acrônimo para Java Naming and Directory Interface. JNDI é uma forma geral, para atribuir um nome para se referir a um recurso muito muito qualquer tipo, tais como conexões de banco de dados, URLs, arquivos e classes.

NOTA JNDI é parte da plataforma Java. É muito usado em muitos Java aplicações para resolver problemas de configuração. Por exemplo, uma parte considerável da arquivos de configuração do servidor de BI Pentaho usar JNDI para gerenciar os componentes. Em Aliás, se você deseja executar suas transformações e emprego no interior do Pentaho BI Server, JNDI deve ser a sua forma preferida de criar conexões PDI, como você pode gancho nas conexões JNDI configurada no nível do servidor.

Uma discussão completa sobre JNDI está fora do escopo deste livro. Se você gosta, você pode descobrir mais sobre JNDI no site da Sun: http://java.sun.com/products/jndi/.

No âmbito do PDI, uma conexão JNDI é simplesmente uma chamada de conexão JDBC para que os detalhes da conexão exata são armazenados fora das transformações e do emprego. Sempre que as transformações e os trabalhos devem se referir ao JNDI con- interligação, fazem isso apenas usando o nome da conexão. Sempre que o real conexão com o banco de dados precisa ser criada, o nome é resolvido para o dados de configuração, que são realmente utilizados para instanciar a conexão.

Criando uma conexão JNDI

Para criar uma conexão JNDI, usar um editor de texto para abrir o arquivo jdbc.properties localizado no simples jndi diretório sob o diretório home PDI.

NOTA A jdbc.properties arquivo é simplesmente parte de um JNDI particular implementação fornecida pelo PDI. Outras aplicações, tais como o BI Pentaho servidor, têm implementações mais avançadas JNDI, que exigem uma diferente configuração do procedimento.

Muito provavelmente você vai descobrir que o arquivo contém um número de linhas já. Você necessidade de adicionar linhas para configurar o nome da classe do driver, JDBC seqüência de conexão,
nome de usuário e senha, e uma linha genérica para identificar que este recurso JNDI é uma conexão de dados. A sintaxe geral é mostrada na Lista 11-1.

Listando 11-1: Sintaxe para conexões JNDI em jdbc.properties

```
<jndi-name> / tipo = javax.sql.DataSource
<jndi-name> motorista / = totalmente qualificado da classe JDBC driver name>
<jndi-name> / url = <Driver e conexão connectstring> específicas
<jndi-name> / user = usuário do banco de dados>
<jndi-name> user / password = banco de dados>
```

DICA Se você tiver dificuldade para descobrir os valores exatos para o nome da classe do driver e / ou a seqdeconexao, você pode achar esta dica útil. Primeiro, crie um ordinário JDBC de conexão na caixa de diálogo Database Connection. (Este processo é descrito em detalhes no Capítulo 9.) Em seguida, pressione o botão Lista de recursos na parte inferior do

Database de diálogo Connection. Isto traz uma lista de propriedades para o banco de dados conexão. Olhe para as propriedades nomeadas Classe Driver e URL. Figura 11-9 mostra um exemplo de um recurso de lista.

Figura 11-9: Determinação da classe do driver JDBC e URL com a lista de recursos

Listando 11-2 mostra que cinco linhas como pode parecer para criar um JNDI conexão chamada wcm para se conectar ao wcm banco de dados.

Listagem 11-2: Um exemplo de conexão JNDI para o banco de dados WCM

```
#
# Uma conexão JNDI para o banco de dados chamado WCM wcm
#
wcm / tipo javax.sql.DataSource = # define uma conexão JNDI DB
wcm / driver com.mysql.jdbc.Driver = # nome do driver de classe Java
url wcm / = jdbc: mysql: / / localhost / mca # JDBC URL (seqdeconexao)
wcm / user = # wcm nome de usuário do banco de dados
senha wcm / mca = # senha para o usuário do banco de dados
```

ATENÇÃO Esteja ciente de que a senha do banco de dados é armazenada em texto simples jdbc.properties.

Depois de adicionar os detalhes necessários para a jdbc.properties arquivo, você ainda precisa para criar a conexão no nível de transformação ou de trabalho. Neste caso, você necessidade de selecionar o tipo apropriado de conexão ao MySQL (que esta informação não faz parte da configuração do JNDI em jdbc.properties), Defina o tipo de acesso a JNDI, em seguida, digite o nome do JNDI.

Conexões JNDI e Implantação

Se você sempre usam conexões JNDI, você não precisará alterar o seu emprego e transformações, nem é preciso passar o parâmetro adequado valores durante a execução. Tudo que você precisa fazer é garantir que tanto o desenvolvimento ambiente e do ambiente de implantação estão configurados para suportar todos os utilizados nomes JNDI.

Por exemplo, se você deseja executar tarefas e transformações no ambiente PDI-
mento da plataforma de implantação, você precisará modificar o jdb.properties
arquivo lá em conformidade. Se a sua plataforma de distribuição consiste no Pentaho
servidor de BI, você vai precisar usar o console de administração para configurar o
adequadas nomes JNDI em que acabar também.

Trabalho com o Repositório PDI

Virtualmente todos os componentes em Pentaho Data Integration pode interagir com um
repositório. O repositório é um banco de dados relacional que é utilizado para armazenar
empregos, transformações e objetos, como conexões de banco de dados.

Usando um repositório pode ajudá-lo a organizar o desenvolvimento, bem como implantar-
mento de soluções de integração de dados. O repositório pode ser utilizado por múltiplos
desenvolvedores simultaneamente, formando assim um local de armazenamento centralizado.
Porque
o repositório é apenas um banco de dados, pode ser feito como um, e os relatórios podem ser
correr contra ele.

Usando um repositório também simplifica o gerenciamento de configuração de banco de
dados
conexões. No repositório, conexões de banco de dados são armazenados como separar
objetos que podem ser reutilizados em vários trabalhos e transformações.

O restante desta seção descreve como usar o repositório quando
trabalhar com ferramentas Pentaho Data Integration, como colher de cozinha, e
Pan. A seção"rodando dentro do Pentaho BI Server"mais adiante neste capítulo,
descreve como usar um repositório de PDI para a execução de tarefas e transformações como
seqüência de componentes de ação.

Criando um Repositório PDI

Você pode criar um novo repositório utilizando a caixa de diálogo Selecione um repositório
Spoon.
Este diálogo é aberta por padrão quando você inicia o Spoon, a menos que você modificou o
opções para evitar esse comportamento. Você pode controlar isso usando o Show repositório
na caixa de diálogo de inicialização na janela de opções. Você também pode invocar o
diálogo a partir do repositório Spoon. Isto é feito através dos seus principais
RepositoryConnect ao repositório ou usando themenu usando Menu
atalho do teclado Ctrl + R. O repositório de diálogo é mostrada na Figura 11-10.

Você pode criar um novo repositório, premindo o botão Novo no repositório
diálogo. Uma caixa de diálogo informações do repositório intitulado abre. Neste diálogo, você
deve
especificar a conexão do banco de dados para se conectar ao repositório.

conexões de banco de dados para o repositório não são diferentes de Pen-ordinária
Taho banco de dados conexões Integração. Não há requisitos especiais
para um determinado RDBMS, você deve ser capaz de usar todos os produtos populares,
como o Microsoft SQL Server, MySQL, Oracle, PostgreSQL e sem nenhum problema. Veja
Capítulo 9 para obter informações sobre como criar conexões de banco de dados.

Figura 11-10: O diálogo repositório

NOTA No momento da redação deste artigo, não é possível usar o SQLite como o sistema de banco de dados subjacente. Se você está procurando uma solução de banco de dados integrado para o repositório, você pode tentar H2 ou Derby.

Quando você seleciona a conexão de banco de dados, clique no botão Criar para criar um novo repositório. Ou, para atualizar um repositório existente, clique no Botão Upgrade. No entanto, há uma série de coisas a considerar Ao atualizar um repositório existente. Portanto, a atualização é coberto por um subseção mais adiante neste capítulo.

NOTA Criando o repositório envolve a criação de tabelas de banco de dados. Você deve assegurar a conta especificada na conexão com o banco tem privilégios suficientes para esta tarefa.

Depois de confirmar que você realmente deseja modificar o banco de dados, o repositório as tabelas são criadas no banco de dados especificado. Note que isto não irá criar um novo esquema. As tabelas do repositório são criados no esquema definido no nível conexão.

Conectando-se ao repositório

Para utilizar um repositório para ler e guardar soluções PDI, primeiro você deve estabelecer uma conexão a um. Isso é feito especificando as credenciais de um repositório do usuário na caixa de diálogo Repository (ver Figura 11-10). Especifique o nome

do repositório do usuário no campo Login ea senha nas
de campo. Em seguida, clique em OK.

NOTA Credenciais para o repositório do usuário são diferentes dos dados da conta associado com a conexão do banco de dados que é usado para o repositório.

Um recém-criado repositório vem com dois usuários pré-definidos:

- A admin usuário (a senha padrão: admin) Tem privilégios totais para o repositório, e deveriam ser usados principalmente para criar novos usuários do repositório de
  para os desenvolvedores ETL.
- A convidado usuário (a senha padrão: convidado) Só pode ler a partir do reposicionamento-história, e deve ser usado para explorar o repositório.

Normalmente, você poderia criar outros usuários e configurar os para executar tarefas específicas com o repositório. Repositório de gerenciamento de usuários é descrito em detalhes na seção"Repositório Administrando Contas de Usuário"mais adiante neste capítulo.

Depois que uma conexão a um repositório é criado, automaticamente será Spoon usá-lo. Por exemplo, quando conectado a um repositório, as ações Salvar Arquivo Abrir Arquivo e correspondem a armazenar e carregar a partir do repositório. Quando não há nenhuma conexão ativa dessas ações padrão para armazenamento e carregamento
arquivos do sistema de arquivos.

Você sempre pode desconectar-se do repositório, escolhendo o Disconnect repositório opção do menu Repositório, ou usando o atalho do teclado Ctrl + D.

NOTA Dentro Spoon, às vezes pode ser confuso quando você tem que trabalhar com múltiplos repositórios. Para descobrir rapidamente se você estiver conectado e ver que repositório que você está conectado, olhar para o título da janela do aplicativo principal janela. Se você estiver conectado a um repositório, o título diz:

Spoon - [<Repositório > Nome] [<repository user>]

onde <Repositório > Nome representa o nome do repositório, e <Repositório > Usuário representa o repositório do usuário.

Conectando-se automaticamente um repositório padrão

Você pode achar que é inconveniente para explicitamente logon no repositório. Se você geralmente trabalham com apenas um único repositório, você pode configurar Pentaho Integração de Dados para se conectar automaticamente a um padrão específico repositório. Este
configuração para o repositório padrão será usado por Spoon, mas também a outros Pentaho Data Integration ferramentas.

Para configurar o login automático para o repositório padrão, use um editor de texto para abrir a kettle.properties arquivo localizado na . Chaleira diretório abaixo seu diretório home. Em seguida, adicione as linhas para as variáveis KETTLE_REPOSITORY,

KETTLE_USERE KETTLE_PASSWORDE atribuir valores para o nome do repositório, o nome do repositório do usuário ea senha do usuário. Veja Listagem 11-3 para um exemplo.

Listagem 11-3: kettle.properties Modificar para se conectar automaticamente a um repositório

```
# Este é .kettle <user-home> / kettle.properties
#
# Automaticamente login como admin / admin no PDI_REPO Repositório
#
KETTLE_REPOSITORY = PDI_REPO
KETTLE_USER = admin
KETTLE_PASSWORD = admin
```

Note que o valor de KETTLE_REPOSITORY deve ser o nome de um repositório encontrado na repositories.xml arquivo, e os valores para KETTLE_USER e KETTLE_PASSWORD deve corresponder a um repositório do usuário do referido depósito.

ATENÇÃO A kettle.properties arquivo é em texto puro. Embora resida no diretório oculto . Chaleira, ele ainda pode ser lido por qualquer pessoa que pode acessar o arquivo. Como as credenciais repositório são armazenados sem criptografia, você deve estar ciente que automaticamente se conectar ao repositório desta forma constitui um risco de segurança.

O Explorer Repositório

Se você estiver conectado ao repositório, você pode abrir o Repository Explorer para analisar e gerenciar seu conteúdo. Você pode chamar o Repository Explorer por escolher a opção Explorar repositório no menu Repositório, ou usando o atalho Ctrl + E. O repositório do Windows Explorer é mostrado na Figura 11-11.

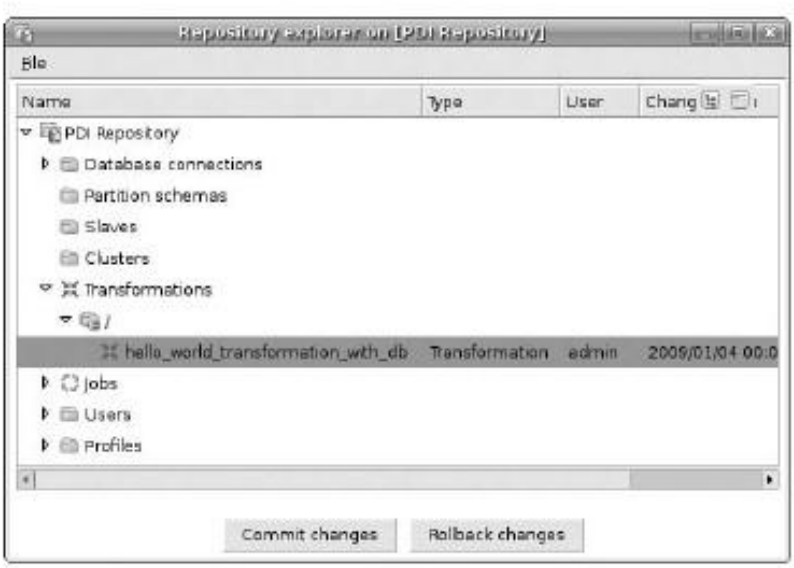


Figura 11-11: O Explorer Repositório

O Repository Explorer exibe o repositório em uma árvore. Para trabalhar com um item de repositório, clique com o botão direito do mouse para abrir um menu de contexto. A partir daí, escolher as medidas adequadas. Desta forma, você pode:

- Gerenciar conexões de banco de dados ligações são armazenadas em um nível mundial no repositório e podem ser compartilhados por todas as transformações e postos de trabalho no mesmo repositório. O menu de contexto das conexões de banco de dados nó lhe permite criar novas conexões. O menu de contexto para o indivíduo conexões permite que você editar ou excluir a conexão correspondente.
- Gerenciar diretórios do repositório-In repositório, transformações e os trabalhos são sempre armazenados em algum diretório, bem como os arquivos são armazenados em um diretório específico no sistema de arquivos. Isso é útil para estruturar os dados solução de integração, mantendo elementos relacionados. Há uma sepa- taxa de árvore de diretórios de ambos os trabalhos e transformações. Em ambos os casos, o diretório raiz é built-in e nomeado /(Barra do personagem). Você pode use o menu de contexto de diretório para criar, renomear e excluir diretórios.
- Exportação de empregos e transformações para arquivo Escolher Exportação ou empregos transformações salvar cada trabalho individual ou transformação de um set- devem ficar separados . KJB ou . KTR arquivo, respectivamente. Essas ações também recursivamente exportação a estrutura de diretórios ao conteúdo a jusante do diretório selecionado.
- Despejo de um diretório e todo o seu conteúdo em um único arquivo XML-Você pode fazer isso escolhendo a exportar todos os objetos de uma opção de arquivo XML. Isto é muito conveniente no caso de necessidade de implantar uma filial de seu PDI soluções para outro repositório. Você também pode copiar a todo o repositório para tal um arquivo XML usando o menu Arquivo do Explorer Repositório janela. Neste menu, você também vai encontrar uma opção para importar o conteúdo de tal arquivo de despejo de um XML.
- Gerenciar contas de usuários-Este tema é discutido em mais detalhe no subseção seguinte.

DICA Você também pode exportar postos de trabalho e as transformações do arquivo principal. Seleção Menu Arquivo Exportar todos os recursos vinculados a XML é geralmente a mais conveniente opção, uma vez que a exportação não somente o trabalho atual / transformação, mas também todos transformações e empregos em que ele é dependente. Ao escolher esta opção, você será solicitado a digitar a localização de um Zip. arquivo. Depois de armazenar todos os trabalhos ligados e transformações no Zip. arquivo, uma caixa de diálogo aparece para informar de como abrir trabalhos individuais ou transformações armazenados no Zip. arquivo (sem descompactar primeiro o arquivo). Este diálogo é mostrado na Figura 11-12.

Administrando Contas de Usuário do Repositório

Já mencionamos que, por padrão, um novo repositório fornece duas usuário contas: uma admin conta para a administração do repositório, e uma convidado conta

repositório para a introspecção. Para beneficiar do repositório ao desen-
o uso das soluções de integração de dados, você precisa criar contas de usuário de dados desenvolvedores de integração para que eles possam usar o repositório para armazenar e recuperar
seu trabalho.

Figura 11-12: Diálogo informando sobre o trabalho / exportação de transformação

Antes de explicar em detalhes como criar novos usuários, é necessário considerar algumas coisas sobre o gerenciamento de conta de usuário em geral. Em praticamente
todos os assuntos de gestão de conta de usuário, há duas coisas a considerar: identificação e autorização. Identificação se preocupa com verificação de que um usuário do sistema corresponde ao usuário no mundo real. Autorização tem que fazer com que determina as ações que um usuário do sistema tem permissão para executar.

Para a identificação, todos os usuários do mundo real PDI Repositório precisa de um nome de usuário
e senha, que servem como credenciais. O usuário do mundo real é esperado manter a senha secreta para que o sistema pode-se supor que um pedido de login composto por uma combinação específica de usuário e senha, de fato, identificar o usuário do mundo real.

Para autorização, os usuários têm um associado perfil. Um perfil é um chamado recolha de permissões que determinam quais funcionalidades o utilizador pode aceder. Em um recém-criado repositório, três perfis como já estão presentes:

- Este administrador é o perfil padrão para o built-in admin usuário. Ele permite ao usuário usar todas as funcionalidades do PDI, incluindo a conta do usuário
de gestão.
- Read-only-Este é o perfil padrão para o built-in convidado usuário.
- Usuário Este perfil é adequado para regular os desenvolvedores de integração de dados.

Para criar um usuário novo repositório, abra o Repositório Explorer e clique com o botão direito
no nó de usuário (ou um nó de um usuário em particular) para abrir seu menu de contexto. Escolha Novo usuário. Um pouco de diálogo intitulada Informações do Usuário é exibida. Na Usuário de diálogo Informações, você deve especificar o nome de usuário no campo Login,

ea senha no campo Senha. Além disso, você pode usar o perfil
caixa de lista para atribuir um perfil para o novo usuário. Figura 11-13 mostra o Repositório Explorer, o usuário menu de contexto do nó, eo Usuário de diálogo Informações.

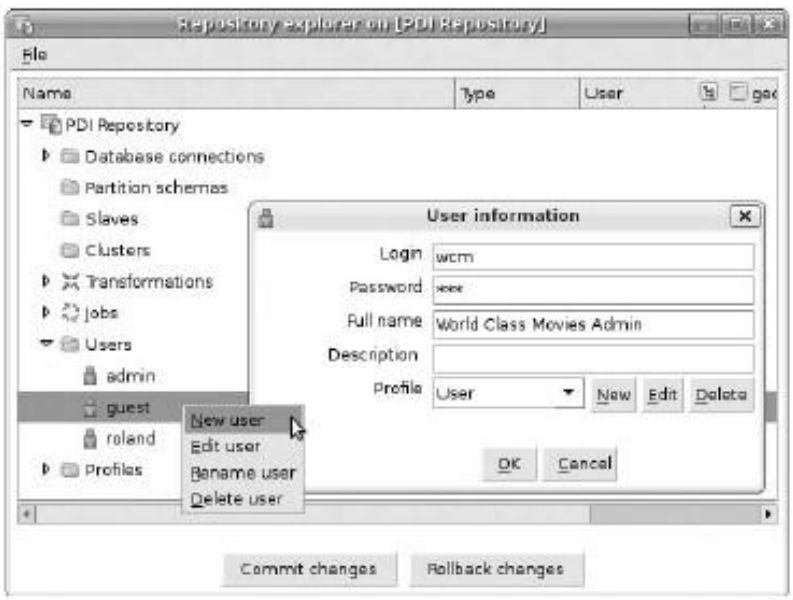


Figura 11-13: Criando um novo usuário repositório

A utilização de diálogo Informações também permite que você crie novos perfis agrupamento de uma ou mais permissões individuais. No entanto, encontramos o built-in Lista de perfis a ser suficiente para a maioria, senão todos, os casos de uso. O número real de permissões é bastante pequena, e as permissões individuais não são refinadas suficiente para construir perfis muito mais significativa.

Cada usuário que está conectado ao repositório pode usar o Edit usuário atual opção no menu Repositório de invocar as informações do usuário de diálogo para modificar sua própria conta. Alternativamente, você pode usar o atalho Ctrl + U.

Como PDI se mantém informado dos Repositórios

Quando as ferramentas de PDI devem se conectar a um repositório particular, eles olham na repositories.xml arquivo. Este arquivo reside no . Chaleira diretório abaixo diretório home do usuário.

NOTA Para sistemas baseados no Windows, a localização mais provável é:

C: \ Documents and Settings \ <usuário> \ chaleira.

Para sistemas baseados em UNIX, o local é:

/ Home <user> /. Chaleira

Listagem 11-4 mostra o que o conteúdo (parcial) da repositories.xml arquivo poderia parecer.

Listagem 11-4: Conteúdo do arquivo repositories.xml

```
<? Xml version = "1.0" encoding = "UTF-8"?>
<repositories>
  <connection>
    <name> Repo Connection </ name>
    <servidor> localhost </ servidor>
    <tipo> MYSQL </ tipo>
    <access> Native </ acesso>
    <database> pdi_repository </ database>
    <port> 3306 </ porto>
    <username> root </ username>
    <senha> criptografados 2be98afc86aa7f2e4cb79ce7dc781bed6 </ senha>
  </ Conexão>
  <repository>
    <name> PDI REPO </ name>
    <description/>
    <connection> PDI Repository2 </ conexão>
  </ Repository>
</ Repositórios>
```

Mencionamos o repositories.xml arquivo, pois você pode querer copiá-lo para outro diretório home do usuário, ou para outra máquina, por exemplo, o implantação ambiente. Você pode até precisar editá-lo manualmente para ajustar o banco de dados de parâmetros de conexão com o ambiente de implantação.

Como você pode ver na listagem 11-4, o formato é bastante auto-explicativo, e você não deve ter nenhum problema manualmente editando este.

Atualizando um repositório existente

Se você planeja usar o Pentaho Data Integration com um repositório que foi criado por uma versão anterior do Spoon, você deve atualizar o repositório. O processo para atualização de um repositório existente é idêntico ao que, para criar um novo repositório: clique no botão Upgrade para atualizar um repositório existente ou a criar botão para criar um novo se ainda não existe.

ATENÇÃO Se você pretende atualizar o repositório, aconselhamos que você backup do repositório de idade. Você pode usar as suas ferramentas de banco de dados para criar um ordinário banco de dados de backup.

Além de fazer um backup do seu banco de dados, são fortemente aconselhados para exportar todos os objetos do repositório para um arquivo XML. Você pode fazer isso usando Explorer Repositório descrito anteriormente. Outra possibilidade é usar o Pan ferramenta de linha de comando, descritos mais adiante neste capítulo. O benefício da exportação de todos os

objetos em um arquivo é que ele permite importar rapidamente a solução para um recém-criado vazio repositório. Isto torna mais fácil superar qualquer problema que você pode encontrar a atualização do repositório.

Em execução no ambiente de implantação

Você já viu vários exemplos de trabalhos em execução e as transformações Usando a colher. No entanto, Spoon não é um ambiente de execução típica de fins de implantação. Spoon é completamente dependente da interação do usuário, e exige uma interface gráfica, os quais não são susceptíveis de ser disponíveis em um ambiente de implantação.

Em ambientes típicos de implementação e execução de trabalho de transformação é automatizados e muitas vezes desencadeados por algum tipo de programação programada. Nesta seção, vamos dar uma olhada nas ferramentas que podem ser utilizados para esses fins.

Correndo na linha de comando

Empregos e as transformações podem ser iniciadas usando as ferramentas de linha de comando Cozinha e Pan, respectivamente. Pan e cozinha são invólucros leve em torno do mecanismo de integração de dados. Eles fazem pouco mais do que interpretar linha de comando parâmetros e chamar o motor para lançar uma transformação ou trabalho. Essas ferramentas são úteis principalmente para a integração Pentaho Data Integration soluções com scripts nível de sistema operacional e as soluções de agendamento.

Cozinha e Pan são iniciados usando shell scripts, que residem no Pentaho Dados diretório de instalação do Integration. Para o Windows, os scripts são chamados Kitchen.bat e Pan.bat , respectivamente. Para sistemas baseados em UNIX, os scripts são chamados kitchen.sh e pan.sh.

NOTA Os scripts para sistemas operacionais baseados em UNIX não são executáveis pelo padrão. Elas devem ser feitas usando o executável chmod comando.

Parâmetros de linha de comando

A cozinha eo Pan interface do usuário consiste em um número de linha de comando parâmetros. Correr Cozinha e Pan sem nenhum parâmetro resulta em uma lista de todos os parâmetros disponíveis. Basicamente, a sintaxe para especificar os parâmetros consiste de uma barra (/) Ou traço (- personagem), imediatamente seguido por o nome do parâmetro:

```
[/ -] Nome de valor] [[:=]
```

A maioria dos parâmetros aceitar um valor. O valor do parâmetro é especificado diretamente após o nome do parâmetro ou dois-pontos (:) Ou igual a um personagem (=), seguido do valor real. O valor pode, opcionalmente, ser colocada em um único (») Ou dupla (") Aspas. Isso é obrigatório no caso de o parâmetro valor em si contém caracteres espaço em branco.

ATENÇÃO Usando o travessão e iguais caracteres para especificar os parâmetros podem levar a problemas em plataformas Windows. Atenha-se a barra e dois pontos para evitar problemas.

Apesar de postos de trabalho e as transformações são funcionalmente muito diferentes, há praticamente nenhuma diferença no lançamento a partir da linha de comando. Por isso, Cozinha e Pan partes mais dos seus parâmetros de linha de comando. O genérico parâmetros de linha de comando podem ser categorizadas como segue:

- Especifique um emprego ou de transformação
- Controlar o registro
- Especificar um repositório
- Lista de repositórios disponíveis e seus conteúdos

Os parâmetros comuns de linha de comando para ambos Pan e da cozinha estão listados na Tabela 11-1.

Tabela 11-1: Os parâmetros genéricos de linha de comando para a cozinha eo Pan

NOME-	VALOR	OBJETIVO
NOREP	Y	Não se conectar a um repositório. Útil para ignorar o login automático.
Rep	Nome do repositório	Conecte-se com o repositório nome especificado.
Usuário	username Repositório	Conecte-se com o repositório especificado nome de usuário.
Pass	Repositório senha do usuário	Conecte-se com o repositório especificados senha.
Listrep	Y	Mostra uma lista de repositórios disponíveis.
Dir	Caminho	Especifique o diretório do repositório.
Listdir	Y	Lista do repositório disponível Emprego / diretórios do repositório.
Arquivo	Nome do arquivo	Especifique um emprego ou de transformação armazenadas em um arquivo.
Nível	Erro |Nada |Basic | Detalhada |Depurar |Rowlevel	Especifique a quantidade de informação devem ser registrados.
Logfile	Nome do arquivo para log	Especificar a qual arquivo que você deseja log. Por padrão, as ferramentas de log para o saída padrão.
Versão		Visualizar a versão, número de revisão e data de criação da ferramenta.

Embora os nomes dos parâmetros são comuns a ambos Cozinha e Pan, o semântica da dir e listdir parâmetros são dependentes da ferramenta.
Para Cozinha, esses parâmetros se referem aos diretórios nos repositórios 'trabalho. Para Pan,
estes parâmetros se referem aos diretórios transformação.

Executar trabalhos com Cozinha

Além dos parâmetros genéricos de linha de comando, cozinha suporta um par de outros parâmetros, mostrados na Tabela 11-2.

Tabela 11-2: parâmetros de linha de comando específico para a cozinha

NOME	VALOR	OBJETIVO
Trabalho	Nome do trabalho	Especifique o nome de um trabalho armazenado no repositório.
listjobs	Y	Lista de postos de trabalho disponíveis no diretório do repositório especificados pelo dir parâmetro.

Listagem 11-5 fornece alguns exemplos de linhas de comando típica cozinha.

Listagem 11-5: linhas de comando típica cozinha

```
#
# Lista todos os parâmetros disponíveis
#
home PDI->. / kitchen.sh


#
# Executar o trabalho armazenado em / home / foo / daily_load.kjb
#
home PDI-> arquivo / kitchen.sh:. / home / foo / daily_load.kjb


#
# Executar o trabalho daily_load do pdirepo repositório chamado
#
Home> PDI / kitchen.sh / rep:. Pdirepo / user: admin / pass: admin \
      > Dir: / trabalho /: daily_load.kjb
```

Correndo com Transformações Pan

Os parâmetros Pan-específicas da linha de comando são completamente equivalente ao os Cozinha específicos. Eles são mostrados na Tabela 11-3.

Tabela 11-3: parâmetros de linha de comando específico para a cozinha

NOME	VALOR	OBJETIVO
Trans	Nome do trabalho	Especifique o nome de um trabalho armazenado no repositório.
Listtrans	Y	Lista de postos de trabalho disponíveis no diretório do repositório especificados pelo dir parâmetro.

Usando parâmetros personalizados de linha de comando

Ao utilizar ferramentas de linha de comando para executar tarefas e transformações, pode ser útil usar parâmetros de linha de comando para transmitir dados de configuração. Para a linha de comando da cozinha ferramentas e Pan, você pode usar o Java Virtual Machine propriedades para obter o efeito de parâmetros personalizados de linha de comando. A sintaxe para passar esses parâmetros""é costume:

```
-D = valor <name>
```

A linha de código a seguir ilustra como tal parâmetro pode aparecer em uma linha de comando da cozinha:

```
home PDI-kitchen.sh> / arquivo:-Dlanguage = en
```

Em transformações, você pode usar o Get System Info passo para obter o valor dos parâmetros de linha de comando. Você pode encontrar este passo na entrada categoria. O Get System Info etapa gera uma linha de saída com um ou mais campos com um valor gerado pelo sistema. O Get System Info passo é configurado através da criação de campos e escolher um determinado tipo de sistema de valor a partir de uma lista pré-definida (ver Figura 11-14).

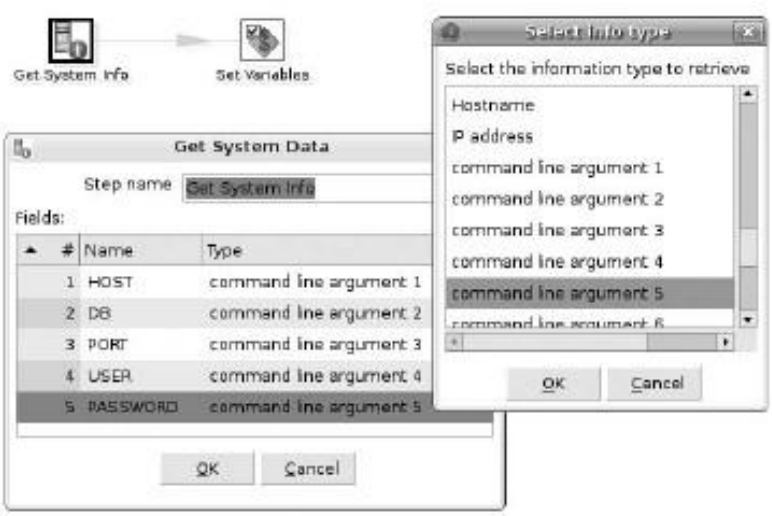


Figura 11-14: Capturando parâmetros de linha de comando com o Get System Info etapa

Entre os valores do sistema pré-definidos, você vai encontrar um argumento de linha de comando através argumento de linha de comando 10. Os campos com um desses tipos automaticamente ter o valor da correspondente D = valor <name> linha de comando parâmetro.

Usando os valores dos parâmetros de linha de comando como entrada para um conjunto de variáveis etapa, você pode fazer os parâmetros de linha de comando acessível a outros trabalhos e transformações. Isso também é mostrado na Figura 11-14.

Usando senhas de banco de dados Obfuscated

Para mitigar o risco à segurança de passar senhas em texto puro no comando
linha de ferramentas como cozinha e Pan, você pode usar ofuscado senhas de banco de dados.
Você pode usar esta técnica para especificar a senha para o repositório con-
fio Bluetooth para as ferramentas de linha de comando. Você pode usar a mesma técnica para qualquer
conexões de banco de dados que você configurou o uso de variáveis, como discutido anteriormente
neste capítulo.

Para usar senhas ofuscado, primeiro você precisa para executar o Encr.bat script (Para sistemas Windows) ou encr.sh script (para sistemas UNIX-like), passando a senha em texto puro. Aqui está um exemplo de comando para ofuscar a passagem palavra wcm:

```
shell wcm encr.sh chaleira->
```

Este comando retorna o seguinte resultado:

```
Criptografado 2be98afc86aa7f2e4cb79ce10be85acd7
```

A saída de texto é a senha real ofuscado. Agora você pode usar este
no lugar da senha em texto puro. Certifique-se de incluir a completa ofuscado
senha, incluindo a parte legível, Criptografado. Isso permite identificar PDI
a senha, ofuscado.

Rodando dentro do Pentaho BI Server

A Plataforma Pentaho fornece componentes para integrar Pentaho Data Inte-
trabalhos de integração e transformações nas seqüências de ação. Isso permite que os postos de trabalho e
transformações a ser agendada com o agendador de quartzo que é construído em
o servidor. Outro benefício da incorporação de soluções PDI em seqüências de ação
é que ele permite a interação humana. Por exemplo, a seqüência de ação
pode levar o usuário para a entrada, que pode então ser usado para parametrizar uma transformação ou de emprego.

A integração do PDI para a plataforma Pentaho não pára com a mera
execução de tarefas e transformações. transformações PDI pode transferir um
conjunto de resultados para a seqüência de ação de chamada, permitindo relatórios para desenhar os dados diretamente
do motor PDI. A seqüência de ação pode fazer praticamente tudo com
o conjunto de resultados: mostrar os dados em um relatório ou em um gráfico, percorrer as linhas
e chamar um trabalho de impressão, envie uma mensagem de e-mail, use os dados da linha como parâmetros
para um relatório, e assim por diante e assim por diante.

Transformações em seqüências de ação

As transformações podem ser incorporadas em seqüências de ação utilizando o Pentaho
Integração de Dados ação do processo. As instruções a seguir descrevem como
utilizar uma transformação em uma seqüência de ação.

1. Abra a seqüência de ação em Pentaho Design Studio. No processo
seção Ações, clique no botão Adicionar ou invocar o menu de contexto para adicionar
uma ação de processo do tipo Pentaho Data Integration. Você pode encontrar este
no menu Ação processo, selecionando Adicionar obter dados de Pentaho
Integração de Dados.

2. Configure o Pentaho Data Integration processo de ação de apontar para um
transformação. Para uma transformação armazenadas em um arquivo, use o botão Procurar
botão para selecionar o apropriado . KTR arquivo. Para uma transformação armazenados em
o repositório, marque a caixa de seleção Use Chaleira Repository, e selecione o
diretório desejado e transformação por meio do Diretório e Transfor-
mações listas drop-down.

NOTA Você deve configurar o servidor de BI para dizer-lhe que PDI repositório de usar. A procedimento para fazer isso é descrito mais tarde nesta seção.

3. Se a transformação tem um passo Get System Info para capturar linha de comando
parâmetros, você pode usar a transformação seção Entradas para especificar
quais as entradas a partir da seqüência de ação deve ser passado para o trans-
formação. O Get System Info passo é descrito em detalhes anteriormente neste
capítulo, na seção''Usando parâmetros personalizados de linha de comando''.

4. Se você quer usar o fluxo de saída de uma das etapas de transformação
na seqüência de ação, especifique o nome da etapa de transformação em
a propriedade Step Transformação. Note que embora a propriedade oferece
uma caixa de listagem, você ainda deve digitar manualmente o nome da transformação
etapa. Você também deve fornecer um nome na propriedade Nome de saída ao mapa
o conjunto de resultados para uma saída seqüência de ação. Esse nome pode ser usado
por ações processo subseqüente da seqüência de ação.

Figura 11-15 mostra uma seqüência de ação simples contendo três processos ações: uma ação Secure Descrição Filtro para levar o usuário para uma data, uma Pentaho Data Integration ação para executar uma transformação e, finalmente, Pentaho Relatório de ação para executar um relatório com base no conjunto de resultados retornado pela Pentaho
Integração de Dados ação.

Empregos em seqüências de ação

O procedimento para executar um trabalho dentro de uma seqüência de ação é muito semelhante ao
que, para transformações. Os trabalhos são executados pela Pentaho Data Integration Job ação do processo. Você pode encontrar isso no menu Ação processo em Adicionar Execute Pentaho Data Integration Job.

Tal como acontece com as transformações, você pode se referir tanto a um arquivo ou um item na
repositório. Você também pode especificar as entradas para um trabalho, assim como é possível para um
transformação. No entanto, um trabalho não pode retornar um conjunto de resultados para a seqüência de ação.

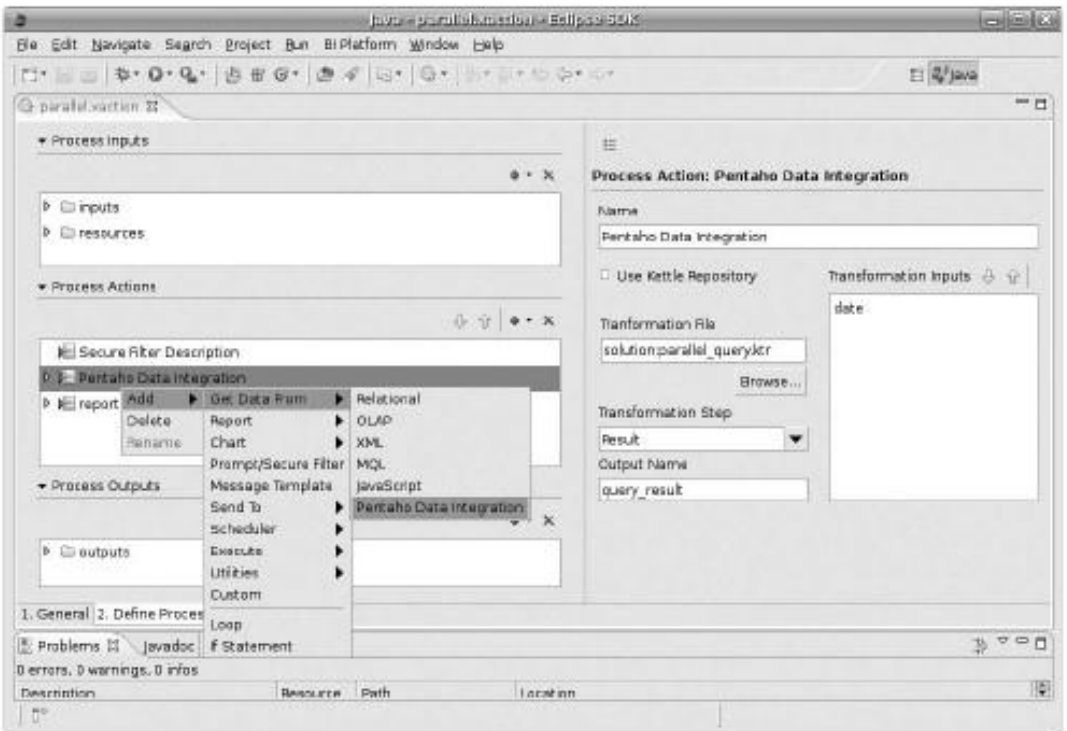


Figura 11-15: Usando uma transformação em uma seqüência de ação

O servidor Pentaho BI e do PDI Repositório

O servidor Pentaho BI tem um arquivo de configuração separado que especifica onde procurar emprego e transformações. Por padrão, o servidor está configurado para funcionar com arquivos de trabalho e transformação. Se você quiser usar postos de trabalho e as transformações
armazenadas em um repositório PDI, você precisa configurar o servidor de acordo. Para exemplo, se você está recebendo mensagens de erro como o seguinte, você sabe
Você terá que configurar isso.

```
Kettle.ERROR_0019 - RDBMS acesso a transformação não é permitido quando
tipo de repositório é 'arquivos'
```

Para configurar o servidor para usar o repositório do PDI, use um editor de texto abrir o settings.xml arquivo localizado na pentaho-solutions/system/kettle
diretório. O conteúdo desse arquivo deve ser algo parecido Listagem 11-6.

Listagem 11-6: Conteúdo do arquivo settings.xml padrão

```
<kettle-repository>
  <! - A localização do arquivo Chaleira repositories.xml
      deixar em branco quando padrão (HOME .kettle / repositories.xml) ->
  <repositories.xml.file> </ repositories.xml.file>
  <-! Arquivos ou RDBMS ->
```

```
  <repository.type> arquivos </ repository.type>
  <! - O nome do repositório para usar ->
  <repository.name> </ repository.name>
  <! - O nome do repositório do usuário ->
  <repository.userid> </ repository.userid>
  <! - A senha ->
  <repository.password> </ repository.password>
</ Chaleira repositório>
```

Os comentários no settings.xml arquivo é auto-explicativo:

- Definir o conteúdo do repository.type elemento para rdbms se você quiser usar um repositório PDI.
- Coloque o nome do repositório que você deseja usar na repository.name elemento. Este deve ser o nome de um repositório existente definido no PDI repositories.xml arquivo.
- Definir o repository.name elemento para o nome do usuário do repositório que é ligar para o repositório.
- Definir o repository.password elemento para a senha do repositório usuário.

Ao contrário do que a sugestão nos comentários que aparecem no settings.xml arquivo, não é possível especificar um determinado repositories.xml arquivo usando o repositories.xml.file elemento. Se você tentar configurá-la, uma mensagem de erro como a seguinte indica que ele ainda não está implementado na sua plataforma:

```
Kettle.ERROR_0017 - Desculpe, especificando o arquivo repositories.xml é para
utilização futura, tendo o arquivo padrão para agora.
```

Execução remota com Carte

Carte é um servidor que executa trabalhos leves ou transformações em resposta a uma solicitação HTTP. Um servidor é chamado de Carte servidor escravo, porque ele responde aos pedidos de outros lugares, e executa o trabalho associado, em nome da da parte requerente. Vários servidores escravos que são usados para executar um conjunto único trabalho ou transformação são referidos como um cluster.

Execução remota e aglomeração são características muito poderosas, estabelecendo o PDI além de um produto de integração de dados que é adequado para a computação em nuvem ambientes como o Amazon EC2. Isso significa que o processo de integração de dados pode de maneira rentável ser escalado dinamicamente, de acordo com a demanda.

Neste livro, não é possível descrever todos os casos de uso e as vantagens da esses recursos. Em vez disso, vai descrever a sua arquitectura e componentes. Este deve servir como uma boa introdução para os leitores que queiram utilizar esses recursos.

Por execução remota?

De uma forma ou de outra, todas as razões para exigir a execução remota ou clustering estão relacionados a problemas de desempenho de superação. A lista a seguir resume mais importantes problemas de desempenho que podem ser resolvidos com controle remoto execução e / ou agrupamento:

- Escalabilidade
- Disponibilidade
- Redução de tráfego de rede
- Redução da latência

Vamos agora explicar esses conceitos com mais detalhes.

Escalabilidade

Cada computador tem um potencial de transformação que é, em última análise limitada pelo hard-
ware: CPU, memória, disco e rede. Contanto que nenhum desses componentes está esgotado pelo trabalho, não há problema. No entanto, uma crescente carga de trabalho, mais cedo ou mais tarde porque um dos recursos a serem maxed para fora, limitando assim o desempenho do sistema como um todo. O termo escalabilidade refere-se a quão bem o sistema pode ser cultivada para acompanhar uma crescente carga de trabalho.

Há duas maneiras fundamentais para o crescimento do sistema:

- Atualizar os componentes de hardware com outros que oferecem melhor desempenho. Por exemplo, você pode substituir um processador de 1 GHz com um que trabalha a 2 GHz.
  Esta estratégia é conhecida como scale-up.
- Adicionar mais computadores com semelhantes ou comparáveis potencial de transformação.
  Esta estratégia é conhecida como scale-out.

Com Carte, recursos remotos PDI é permitir a execução de agrupamento, e este permite a integração de dados de carga de trabalho a serem distribuídos entre um número de computadores que funcionam em paralelo para fazer o trabalho. Paralelização permite mais trabalho a ser feito no mesmo espaço de tempo. Este é um exemplo típico de um arquitetura scale-out.

Disponibilidade

Disponibilidade do serviço normalmente não é o objetivo principal para o agrupamento de dados em um
integração do ambiente. No entanto, ao agrupamento para fins de expansão, maior disponibilidade vem como um bônus extra. Isto merece uma explicação.

Quando um processo depende de um único computador, esse computador se torna um ponto único de falha. Se o computador torna-se indisponível para qualquer motivo, o serviço que fornece normalmente fica indisponível também.

Um cluster não tem tal ponto único de falha. Com capaci Carte de clustering-
dade, no caso em que um computador se torna indisponível, o cluster restantes
ainda podem ser usados para processar trabalhos e transformações. O restante pode agrupar
demorar um pouco mais para realizar a tarefa, porque há menos laboriosos de
distribuir a carga, mas a verdade é que o próprio serviço ainda está disponível.

Redução de tráfego de rede

Se o seu problema de integração de dados envolve grandes volumes de dados em máquinas remotas,
ou várias fontes de dados em máquinas remotas, puxando os dados através da rede
a um computador para executar uma transformação ou de trabalho podem facilmente sobrecarregar o
rede. Isso pode resultar em mais tempo de processamento, ou pior, a deterioração
de outras aplicações que dependem da rede.

Ao empurrar todas as transformações agregar e filtrar o mais próximo possível
para o local onde os dados de origem é armazenado, uma redução considerável na
o volume de dados que deve ser puxada por toda a rede pode ser alcançado, assim,
descarga na rede.

Por exemplo, suponha que você queira fazer uma análise básica do logs de acesso de um cluster de servidores web. Digamos que você deseja contar o número de visitas a cada página. Normalmente, o número de páginas exclusivas será muito pequeno em comparação ao número total de pedidos. Fazendo a análise localmente em um computador
implicaria copiar os logs de acesso, causando uma carga significativa no
rede. Remotamente a execução da análise sobre o servidor onde o log é armazenado
usando Carte vai aumentar a carga sobre os servidores remotos, mas significativamente
diminuir a carga sobre a rede, pois somente o resultado relativamente pequeno de
a análise terá que viajar para a rede.

Redução da latência

operações de pesquisa sobre fontes de dados remotas ter tempo para viajar por todo o
rede. Se a operação de pesquisa em si é bastante rápido, o tempo será perdido à espera
para os circuitos de rede. Usando a capacidade de execução remota de Carte, você pode
minimizar a latência, realizando a pesquisa mais próxima da fonte de pesquisa
de dados.

Note que esta estratégia para reduzir a latência irá funcionar apenas para casos específicos. Se você precisa olhar para os mesmos dados, muitas vezes, o cache local é susceptível de ser
uma melhor solução para reduzir a latência. Por outro lado, se quase todas as pesquisas
são originais, ou a quantidade de dados de pesquisa é muito grande para cache, realizando a
pesquisa à distância pode ajudar.

Correndo Carte

Carte é fácil de configurar. Assim como as outras ferramentas PDI é iniciado executando um shell
script que reside no diretório home do PDI. Em sistemas Windows, este script
é chamado carte.bat. Para sistemas baseados em UNIX, o script é chamado carte.sh.

Correndo Carte sem nenhum parâmetro revela a disposição de linha de comando parâmetros e alguns exemplos de uso (ver listagem 11-7).

Listagem 11-7: Correndo carte sem parâmetros

```
Shell>. / Carte.sh
Uso: Carte [endereço Interface] [Porto]


Exemplo: 127.0.0.1 Carte 8080
Exemplo: 192.168.1.221 Carte 8081
```

A Endereço de interface parâmetro é usado para especificar o endereço IP do adaptador de rede que está a ser utilizado pelo servidor Carte. A Porto parâmetro é usado para especificar em qual porta o servidor deverá atender aos pedidos. Ambos parâmetros são obrigatórios.

NOTA Certifique-se de escolher uma porta que não esteja sendo usado por outro servidor. Na primeiro exemplo fornecido pela saída Carte mostrado na Lista 11-7, porto 8080 é utilizado. A porta 8080 também é a porta padrão do pré-servidor de BI Pentaho. Você pode receber uma mensagem de erro como o seguinte se a porta já está tomada:

```
2009/01/23 11:37:39.759:: WARN: não SocketConnector@127.0.0.1: 8080
java.net.BindException: Endereço já em uso: JVM_Bind
```

Criando Servidor Slave

Antes que você pode usar um servidor remoto para executar Carte sua transformação ou trabalho, você deve criar um objeto do servidor escravo. Neste contexto, um servidor escravo é como uma conexão de banco de dados: um descritor de chamada para localizar um servidor remoto.

Para criar um servidor escravo, habilitar o modo de visualização no painel lateral. Expanda a transformação em curso ou trabalho, e botão direito do mouse na pasta do servidor escravo no lado do painel. No menu de contexto, escolha Novo e, a caixa de diálogo Server escravo aparecer. Na página da guia de serviço deste diálogo, você pode definir as propriedades de o servidor Carte, como mostrado na Figura 11-16.

Especifique um valor para o nome do servidor, nome ou endereço IP, Port, Username, e campos de senha. O nome do servidor será usado para se referir a este particular Carte serviço no PDI. O nome do host / endereço IP ea porta identificar os Carte servidor na rede.

O nome de usuário e senha de imóveis são obrigados a fornecer um nível básico de segurança. Por padrão, você pode usar admin para ambos. Isto é suficiente quando executado em um ambiente onde a rede está protegida e inacessível aos usuários não autorizados. Se você precisa executar Carte em uma rede de mais público, você pode ligar em mais serviços de autenticação avançada para oferecer uma melhor segurança.

NOTA Você pode usar variáveis para parametrizar as propriedades do servidor escravo. Um caso de uso para eles é clara quando o ambiente de computação em nuvem não apoio endereços IP fixos. Este é o caso com o EC2 da Amazon.

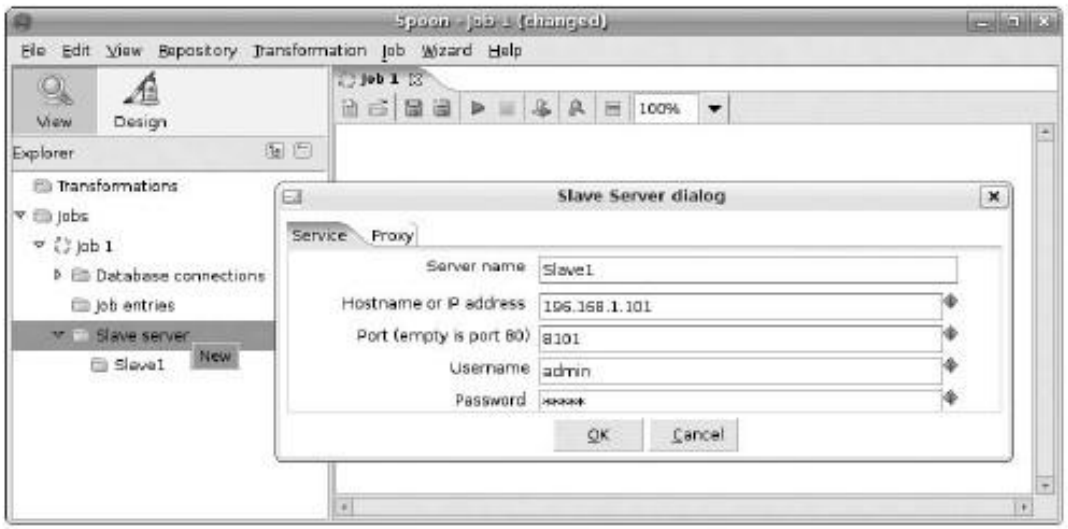


Figura 11-16: Definição de um novo servidor escravo

Remotamente Executando uma transformação ou de trabalho

Você pode usar os botões na parte de execução local ou remota na topo da executar uma tarefa ou executar um diálogo Transformação de escolher se você deseja executar o trabalho / transformação na máquina local ou em um de seus servidores escravos. Para executar remotamente, verifique o botão Executar remotamente, e usar o combobox host remoto para selecionar o servidor escravo que deseja usar. Isso é mostrado na Figura 11-17.

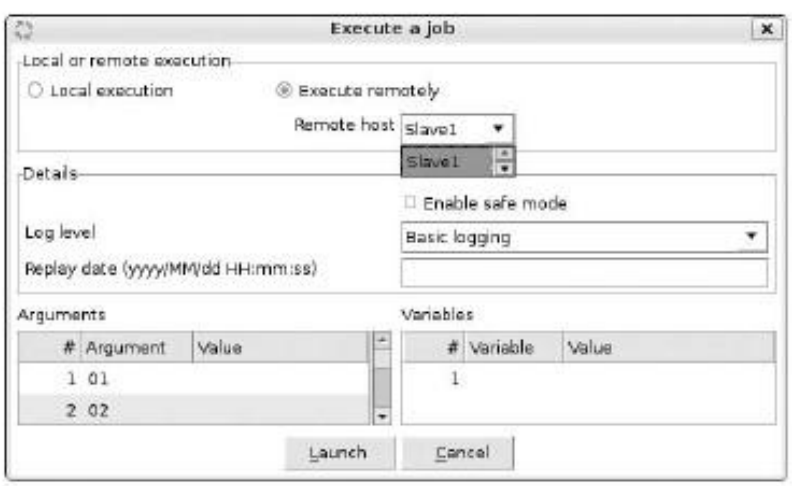


Figura 11-17: Executando um trabalho remotamente

Clustering

Você pode executar transformações em um cluster de servidores escravos. Para usar tal cluster, é necessário definir uma esquema de cluster. Um esquema de cluster é simplesmente um nome

coleção de servidores escravos. Após a definição do esquema, você pode atribuir um cluster para as etapas de sua transformação, fazendo com que passo a ser processado no escravos desse cluster.

Para começar com cluster, primeiro é necessário definir um número de escravos servidores, como discutido na seção anterior. Depois, você pode definir uma aglomeração esquema. Ao lado do painel, está no modo de visualização, você pode ver um aglomerado Chaleira
esquemas nó na parte inferior da árvore. Se você clicar o botão direito do nó, um menu de contexto será exibido, a partir daí você pode criar um novo item. Isso traz à tona o agrupamento de diálogo de esquema, mostrado na Figura 11-18.

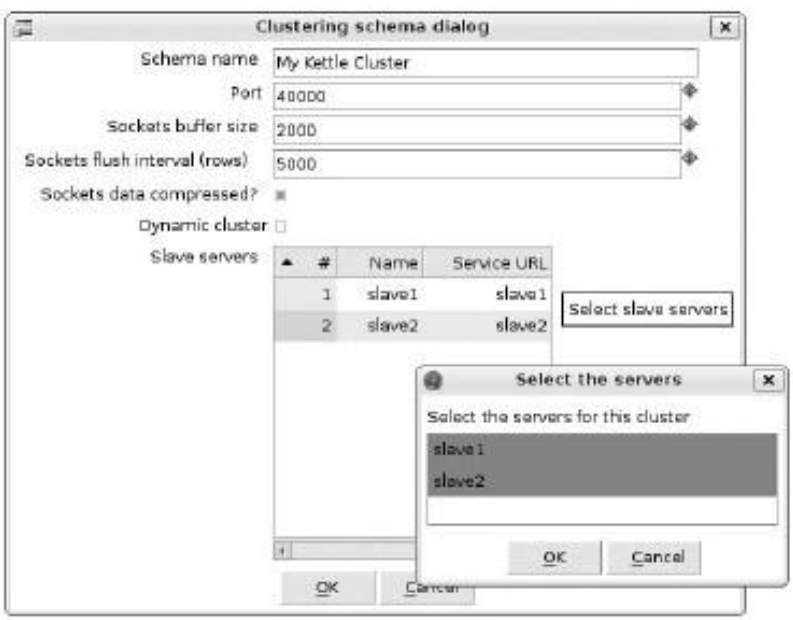


Figura 11-18: O esquema de cluster de diálogo

Na caixa de diálogo de esquema de cluster, você deve digitar um nome de esquema. O esquema
nome deve ser exclusivo para a transformação. Você pode atribuir servidores escravos com o esquema, pressionando o botão Select escravo servidores. Isso traz a Selecione a caixa de diálogo servidores, listando todos os servidores escravos disponíveis. Neste diálogo,
você deve selecionar todos os servidores escravos que precisam participar do cluster. (Você pode selecionar vários itens mantendo pressionada a tecla Ctrl e clicando em a entrada desejada.) Quando você terminar de escolher os servidores, clique no botão OK para
designar os servidores escravos. Note que isso irá apagar qualquer atribuição prévia de servidores escravos ao cluster.

Depois de definir o esquema de cluster, você pode atribuir o cluster para o seu etapas de transformação. Basta selecionar a etapa e clique com o botão direito para abrir o menu de contexto. Escolha o Clustering . . . opção. Isso traz uma caixa de diálogo de onde você pode escolher um dos esquemas disponíveis clustering. Clique em OK para confirmar o diálogo eo cluster é atribuído à etapa. Você pode limpar o associação com um grupo, trazendo o diálogo novamente e pressionar Cancelar.

Resumo

Neste capítulo, você aprendeu técnicas diferentes que você pode usar para executar e implementar
Pentaho Data Integration postos de trabalho e as transformações nos sistemas de produção. Você
aprendi:

- Como usar variáveis para gerenciar recursos como conexões de banco de dados, nomes de arquivos e diretórios
- Como usar conexões JNDI
- Como trabalhar com repositórios PDI
- Como usar a cozinha e Pan para executar as transformações e as tarefas da linha de comando
- Como executar tarefas dentro do servidor de BI Pentaho usando uma seqüência de ação
- Como executar um trabalho ou transformação remotamente usando Carte
- O agrupamento é, e quais os benefícios que tem
- Como servidores de grupos de vários escravos em um cluster

Parte

IV

Business Intelligence Aplicações

Nesta parte

CAPÍTULO

12

A camada de metadados

Muitos dos temas relacionados com a inteligência de negócio, tais como a integração de dados
e armazenamento de dados, podem ser entendidos como soluções para os problemas relativos
abstração, a acessibilidade ea transmissão dos dados.

Nos capítulos anteriores, você aprendeu que o armazém de dados fornece um
ofertas substanciais de abstração a partir dos dados acumulados em vários
fontes de dados. Central para que a realização é a organização de dados em
orientadas para um assunto esquemas estrela, reduzindo consideravelmente a complexidade do trans-
lando perguntas a partir do final de negócios para consultas de banco de dados.

Apesar de estabelecer um data warehouse resolve alguns dos abstração de dados
e as questões de acessibilidade, ainda não é ideal para o fornecimento de dados para ferramentas de relatórios.
Os utilizadores empresariais tentar obter dados a partir do armazém podem ter dificuldades para obter o
informação que eles querem em um formato que possa compreender, ou o sistema pode precisar
para ser refinado para se certificar de que os dados podem ser acessados de forma útil. Neste

Metadados Resumo

Capítulo você aprender como a adição de uma camada de metadados pode ajudar neste aspecto.

Nesta primeira seção, vamos explicar brevemente o que tipos de coisas que nós estamos falando
sobre quando usamos o termo metadados'','e qual problema ela resolve. Mais tarde
Neste capítulo, vamos dar uma olhada no uso de metadados Pentaho.

O que são metadados?

O termo metadados é um pouco em demasia. Em um sentido geral, significa que os dados sobre''
de dados.''Dependendo do contexto, há um monte de coisas diferentes para dizer

"Dados sobre", e tecnicamente tudo isso se qualifica como metadados. Por exemplo, a maioria
RDBMSs apoio listando todas as bases de dados disponíveis e objetos de esquema. Esta é uma
exemplo típico de metadados, que descreve os tipos disponíveis e as formas de
dados armazenados no banco de dados.

A plataforma Pentaho oferece as suas instalações para armazenar e acessar metadados. No contexto deste capítulo, usamos o termo metadados Pentaho para denotar a facilidade de metadados que é parte da plataforma Pentaho.

As vantagens da Camada de Metadados

Como mencionado anteriormente, o data warehouse não resolve todos os problemas em entregar os dados para ferramentas de relatórios. Nesta seção, nós tomamos um olhar mais atento
estes problemas não resolvidos e mostrar como uma camada de metadados podem ajudar a resolver
elas.

Utilizando Metadados para fazer uma interface mais amigável

Do ponto de vista das ferramentas de comunicação e de visualização, o data warehouse é"apenas"um banco de dados relacional. Usando ainda requer considerável conhecimento e experiência com a linguagem de consulta de banco de dados (que normalmente é alguma dialeto da Structured Query Language, SQL). Na maioria dos casos, isso causa design de relatório para estar fora do alcance do usuário típico de negócios. O Pentaho camada de metadados podem aliviar este problema até certo ponto.

Você pode usar a camada de metadados para descrever as tabelas e suas colunas e relacionamentos. Uma vez que este é descrito, coleções de colunas pode ser definida que são susceptíveis de aparecer juntos em um relatório. Estes podem então ser apresentado à
o usuário, utilizando uma interface de assistente. Isso permite que o usuário final a criação de relatórios
on-the-fly, basta escolher as colunas de interesse e colocá-los numa relatório. Devido à metadados definido"nos bastidores", o mecanismo de relatório sabe como gerar a consulta de dados adequados para entregar o especificado os resultados.

Adicionando Independência Flexibilidade e esquema

Suponha que você acabou de construir cerca de 50 ou mais relatórios diretamente em seus dados
armazém. De repente, os dados da equipe de projeto do armazém decide que, no interesse de desempenho da consulta, faz sentido separar uma mini-dimensão da tabela de dimensão do produto. Este tem potencialmente um impacto sobre todos os relatórios que você construiu que são dependentes da dimensão do produto. Na verdade, você
pode precisar exatamente o que relata a investigação será afetado.

A camada de metadados podem ajudar a limitar o impacto do esquema do banco alterações. Como a camada de metadados permite que você especifique os relatórios em um coleção predefinida de tabelas usando conhecida juntar os caminhos, as mudanças de esquema pode

ser resolvidos através do mapeamento do esquema novo banco de dados para o''esquema''resumo
apresentados os relatórios da camada de metadados.

Esse tipo de mudança pode ir muito além da simples renomeação de tabelas e colunas. Por exemplo, se a equipe do data warehouse decide mudar-se de um esquema em estrela para um floco de neve (ou o contrário), o impacto resultante pode ser completamente amortecido por uma camada de metadados. Da mesma forma, as soluções que
reportará diretamente sobre (uma cópia) do sistema de origem pode ser gradualmente transferido para
um ambiente de pleno direito do armazém de dados. Estas alterações envolvem uma ampla modificação da camada de metadados, mas esta seria uma one-shot operação, evitando o caminho potencialmente mais tempo e esforço, consumo de mudando todos os relatórios.

Privilégios de acesso do Refino

Outro aspecto da entrega de dados que não está completamente resolvido pelos dados armazém é privilégios de acesso. Embora a maioria dos RDBMSs proporcionar um acesso camada que envolve a autenticação e autorização, este não é freqüentemente bastante finos para aplicações (incluindo aplicações de BI). O nativo
RDMBS camada de acesso é tipicamente implementado pelos utilizadores concessão do privilégio
para um uso particular (leitura, escrita, ou alteração) da base de dados designada objetos (como tabelas, exibições e / ou procedimentos armazenados). O RDBMS nativas camada de acesso geralmente não oferecem a possibilidade de controle de acesso a dados sobre o
linha de nível.

A camada de metadados permite que a autorização para ser definida em vários níveis. A autorização pode ser controlada de forma flexível a nível de usuário ou papel, e pode ser orientada a objetos em sua totalidade, ou linhas individuais. Isso permite que refinadas políticas de acesso.

Manipulação de localização

saída do relatório contém os dados, bem como metadados. Por exemplo, a saída do de um relatório de vendas pode mostrar os números reais de vendas junto com dados, tais como
o país eo estado nomes e datas de vendas. Mas, além disso a saída do relatório, normalmente
contém os rótulos de texto que descrevem os dados que está sendo mostrado. Por exemplo, uma venda
relatório com um layout de colunas podem ter um título que se lê''Country''direito acima da coluna que contém os nomes de países.

O problema com esses rótulos é que eles não são de idioma neutro. Este pode ser um problema em organizações multinacionais ou multilingues. Por exemplo, World Class Filmes tem gerentes Inglês e de língua francesa. Ambos precisam para ver os dados mesmo relatório, mas os rótulos relatório deve ser localizada no linguagem apropriada, consoante os pedidos em que o gerente do relatório.

A camada de metadados Pentaho suporte a várias localidades. Descritiva adequada-laços, como etiquetas e descrições de objetos de dados, como tabelas e colunas

pode ser associado com textos dependente de localidade. Isso permite que os relatórios devem ser
adaptados para cada idioma.

Cumprimento de formatação consistente e Comportamento

Atributos das tabelas de dimensão, por vezes, manter os dados pré-formatados. Para exemplo, tabelas de dimensão de data geralmente contêm muitas colunas para armazenar diferentes
representações de texto a partir da data do calendário. Isto é completamente diferente de métricas
armazenadas nas tabelas de fatos. Métricas são tipicamente numéricos, e, muitas vezes, os relatórios
mostram os dados agregados da métrica, em vez de valores brutos estocados em linhas de fato.

Para alguns tipos de dados de métricas, formatação especial dos dados pode ser desejável. Por exemplo, a métrica monetária, tais como custo, lucro e volume de negócios deve aparecer em relatórios com o símbolo da moeda adequada e separadores para decimais e milhares. Alguns atributos de dimensão ainda podem requerer formatação adicional, especialmente se a formatação desejada não pode ser alcançado simplesmente armazenar um valor. Por exemplo, o texto representa URLs podem precisar a ser processado de uma forma que os distingue de outros dados. Para alguns tipos de saída do relatório, tais como documentos PDF ou páginas web, ele mesmo pode desejável anexar um comportamento específico para os dados da URL, como a abertura do site adequado quando o rótulo é clicado.

A camada de metadados Pentaho permite que grupos de visual e comportamental adequada-
chamados laços conceitos para ser anexado aos objetos de dados, como tabelas e colunas. Conceitos podem ser baseadas em conceitos existentes a partir do qual herdam seus propriedades. Se desejar, um conceito pode substituir as propriedades de seu pai con-conceito. Isso permite que você crie uma hierarquia de conceitos que podem ser usados consistentemente aplicar as propriedades visuais e comportamentais para itens de dados.

Âmbito de aplicação e uso da Camada de Metadados

A lista a seguir oferece um breve panorama de como Pentaho usa os metadados camada na prática. Estes pontos estão ilustrados na Figura 12-1.

- Metadados de entrada do banco de dados, bem como metadados definidos pelo usuário, é definido usando o Pentaho Metadata Editor (PME) e armazenados no repositório de metadados.

- Os metadados podem ser exportados a partir do repositório e armazenados na forma de . Xmi arquivos, ou em um banco de dados. Os metadados são associados a uma Pentaho solução Pentaho no servidor, onde ele pode ser usado como um recurso para metadados baseados em serviços de informação.

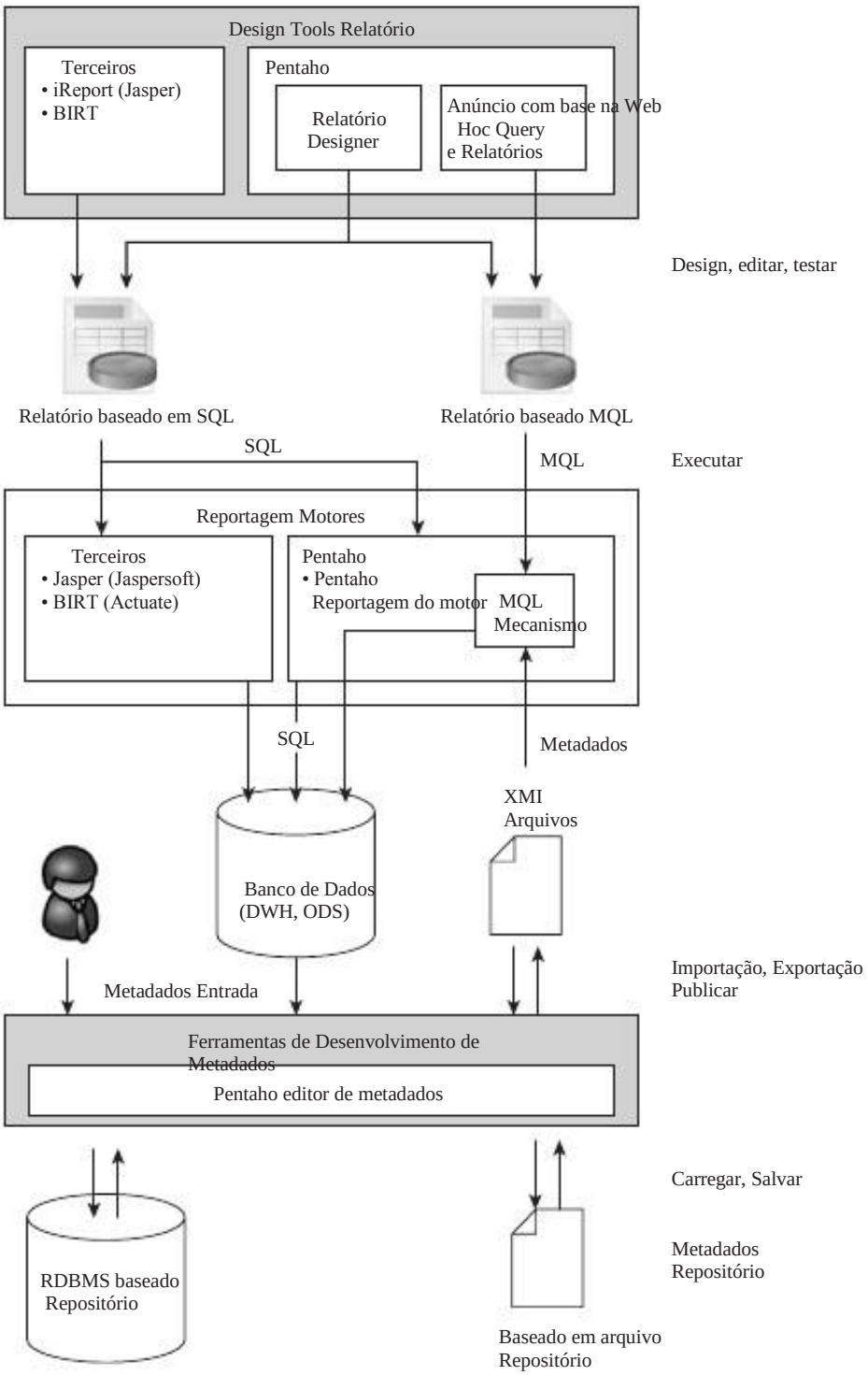


Figura 12-1: Resumo de alto nível do escopo e do uso de Pentaho Metadata

- Usando as ferramentas de projeto Pentaho relatório, os usuários finais podem criar relatórios sobre
os metadados. Isso permite que os relatórios a serem construídas sem o conhecimento da
detalhes físicos do banco de dados subjacente, e sem nenhum conhecimento de SQL. Em vez disso, o relatório contém uma especificação de alto nível da consulta resultado, que é definida através de uma interface gráfica de usuário.
- Ao executar relatórios baseados em Pentaho metadados, o mecanismo de relatório interpreta o relatório. Consulta às especificações são armazenadas no relatório um formato chamado Metadados Query Language (MQL), que é resolvido contra os metadados. Neste ponto, o SQL correspondente é gerado e enviado para o banco de dados. Além deste ponto, o processamento do relatório é bastante
semelhante ao"normal"relatórios baseados em SQL. O banco de dados responde ao consulta através do envio de um resultado de dados, que é processado como saída do relatório.

Atualmente, o uso da camada de metadados Pentaho é limitado a relatar. No futuro, o apoio de metadados para outros componentes da plataforma, como análise, mineração de dados e integração de dados podem ser adicionados.

NOTA metadados Pentaho é atualmente baseada na Common Warehouse Metamodelo (CWM) especificação criada e mantida pelo Object Management Group (OMG). O CWM é uma plataforma aberta e independente de fornecedor norma que especifica a troca e representação de business intelligence metadados.

Para obter mais informações sobre o CWM, consulte o catálogo da OMG Modelagem e especificações de metadados. Você pode encontrá-lo em http://www.omg.org/technology/cwm/.

Metadados Características Pentaho

Nesta seção, descrevemos brevemente as características fundamentais dos metadados Pentaho
camada.

Banco de Dados e Abstração de consulta

A camada de metadados Pentaho pode conter muitos tipos distintos de estrutura componentes, e é fácil perder de vista o quadro geral. Portanto, vamos examinar primeiro a camada de metadados em um alto nível antes de mergulhar nos detalhes.

Relatório de Definição: Ponto do usuário de negócios de visão

Considere os requisitos de uma empresa típica do usuário, digamos, o gerente de Vendas e Aluguéis no Filmes Classe Mundial. Para cada site, o gerente

gostaria de ver o número de encomendas feitas por cada mês em 2008 por
clientes dos Estados Unidos. Mesmo sem nenhum conhecimento da sub-
mentindo banco de dados, a maioria dos usuários de negócios são perfeitamente capazes de compreender o
estrutura e elementos de saída do relatório. Um relatório como este terá:

- Uma seção encabeçada pelo título website ou URI
- Em cada seção, de 12 linhas com uma etiqueta indicando o mês
- Para cada mês, o tema do relatório propriamente dito, isto é, o número de ordens

Além destes itens visíveis, o relatório tem dois itens que são invisíveis
usado para fazer a escolha adequada:

- Um item de ano para selecionar apenas os pedidos feitos no ano de 2008
- Um item país que serve para selecionar apenas os clientes dos Estados Unidos

Agora, nós descrevemos a estrutura do relatório, em termos de seções e seção
conteúdo, mas é importante perceber que esta é uma questão de apresentação:
A informação foi transmitida pelo relatório permanecerá o mesmo independentemente da
se ele contém uma seção por mês com linhas para cada site ou
uma seção por site que contém linhas para cada mês. Se nós esquecemos sobre
a ordem do agrupamento por um momento, acabamos com uma coleção de linhas
consistindo de um site, um mês, eo número de encomendas. Em outras palavras,
os dados do relatório é tabular e contém os itens website, número e série
das encomendas.

Relatório de Execução: o desenvolvedor de SQL View

Suponha que você deseja recuperar os dados do mesmo relatório diretamente do Mundo
Classe armazém de dados de filmes usando SQL. Se você tem as habilidades de SQL, que certamente
não é difícil (embora possa ser um pouco entediante). Mesmo assim, esta secção anda
lo através do processo passo a passo para ilustrar alguns conceitos sobre Pentaho
metadados.

Uma maneira comum de atacar este problema no SQL é começar com a tabela
mais centrais para a questão. Neste caso, o relatório é sobre o número
de encomendas, assim que você começar com o fact_orders tabela. Você vê que contém o
customer_order_id. Se você aplicar o COUNT função agregada em combinação
com o DISTINCT modificador, você pode usar isso para contar o número de indivíduos
ordens.

A partir da fact_order tabela, use uma junção de"procurar"o cliente que
colocado na ordem do dim_customer tabela. Você precisa disto para que você possa usar o
country_name coluna para restringir os resultados para os clientes dos Estados Unidos
apenas. Você pode usar outro junção entre fact_order e dim_website onde

encontrar o website_title. Finalmente, você usa outro junção entre fact_order e dim_date para obter a data do pedido. Você precisa disto para que você possa restringir o resultado de encomendas feitas no ano de 2008 e para produzir etiquetas para cada mês item de relatório.

Conceitualmente, as mesas se juntou a servir para"alargar"o fact_order mesa, estendê-lo com atributos de tabelas associadas. Agora você pode aplicar o critérios para o ano (deve ser igual a 2008) eo país (deve ser igual grupo para os Estados Unidos) e depois pelo título do Web site e um mês a contar da número de ordens distintas, o que lhe dá o resultado final. Figura 12-2 mostra uma representação gráfica desses caminhos e juntar os itens de relatório.

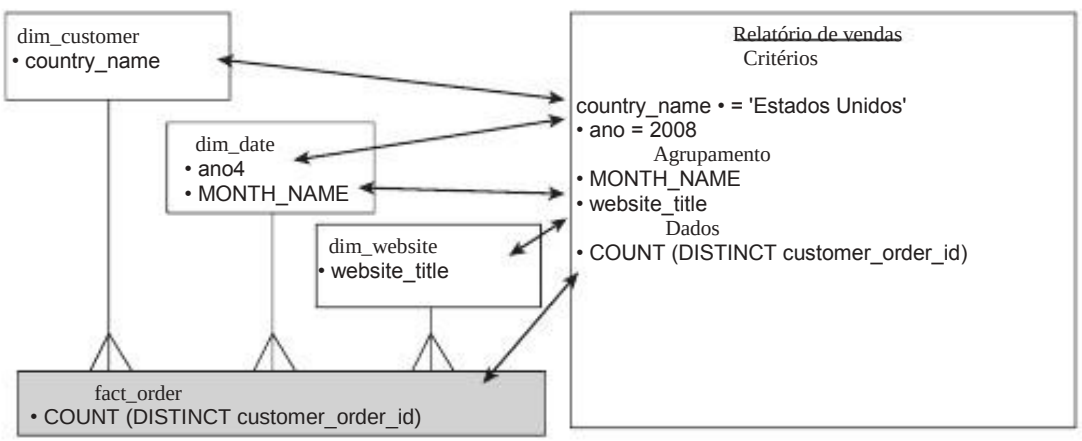


Figura 12-2: Derivando itens de relatório de tabelas associadas

Para completar, a instrução SQL é mostrado na Lista 12-1.

Listagem 12-1: A instrução SQL para recuperar o número de pedidos, agrupados por site título e no mês

```
SELECT      dim_website.website_title
,           dim_date.month_name
,           COUNT (DISTINCT fact_order.customer_order_id) AS count_orders
DA          fact_order
INNER JOIN  dim_customer
ON          dim_customer.customer_key fact_order.customer_key =
INNER JOIN  dim_website
ON          dim_website.website_key fact_order.website_key =
INNER JOIN  dim_date
ON          dim_date.date_key fact_order.local_order_date_key =
ONDE        dim_date.year4 = 2008
E           dim_customer.country_name = 'Estados Unidos'
GROUP BY    dim_website.website_title
,           dim_date.month_name
```

Mecânicos de Abstração: A camada de metadados

É provável que os passos que você acabou de passar por está além das habilidades técnicas da maioria dos usuários de negócios e, certamente, além de descrições de seu trabalho. Mas o que se
os detalhes de juntar as mesas tinham sido tomado cuidado de antes? E se você
usuários apresentam apenas um conjunto de itens a partir do qual pode escolher o que quer acontecer de achar interessante? E se o customer_order_id item foi
concebido para representar directamente o COUNT DISTINCT operação?

Esta é, de fato, exatamente como funciona Pentaho abstração de metadados. Negócios desenvolvedores de inteligência que têm as habilidades eo conhecimento de fundo do banco de dados de relatórios ou armazém de dados pode definir a existência de tabelas reais e colunas em que é chamado de camada física. Os objetos da física
camada são os blocos de construção para a camada de lógica.

Na camada lógica, os quadros da camada física são redefinidas e
às vezes enriquecida com colunas extras que são derivados de colunas em
a camada física através da aplicação de funções e outras expressões para eles. Outro tipo de objeto encontrado na camada lógica é a relacionamento. Um relacionamento define essencialmente como dois quadros na camada lógica pode ser juntou. Às vezes, existem múltiplas relações entre as duas tabelas, que
É por isso que você pode basear várias tabelas na camada lógica em uma única e mesma tabela na camada física. Isto vem a calhar quando se trata de role-playing
tabelas de dimensão.

A camada de entrega é onde as seleções de colunas da camada lógica são
agrupados em unidades que façam sentido para o usuário corporativo. Este é o único parte da camada de metadados que será visível para os usuários empresariais. A partir daqui, eles podem pegar os itens de interesse para a construção de relatórios. Esta seleção é indicado em um
formato XML especial chamado MQL (metadados Query Language).

O MQL""'consulta pode ser usada para gerar uma consulta SQL (ver Figura 12-1). A origem dos itens MQL pode ser rastreado para a camada de lógica e de
lá para a camada física. Com base nessas informações, pode-se derivar as tabelas deve ser juntado, e as colunas que devem ser agrupados por diante.

Propriedades, Conceitos e Herança na Metadados Camada

Nesta seção, discutimos conceitos e propriedades, que são fundamentais
blocos de construção da Pentaho Metadata Layer. Além disso, descrevemos como conceitos podem herdar as propriedades de um outro.

Propriedades

Os objetos na camada de metadados pode ter um número de propriedades. As propriedades são
itens nomeados que são usados para associar diferentes tipos de informação com

os objetos na camada de metadados. As propriedades podem ser divididos em um número de categorias:

- As propriedades gerais, tais como nome e descrição
- propriedades visuais, tais como fonte, cor, e se o objeto é visível em todos os utilizadores finais
- Descritores de Modelo, tais como a expressão subjacente, tipo de dados e agre- Estado inquérito

objetos de metadados pode ter uma coleção de propriedades. Dependendo do tipo do objeto de metadados, algumas propriedades são obrigatórias e sempre presente.

Conceitos

No contexto da Pentaho metadados, um conceito é uma coleção de propriedades que pode ser aplicado como um todo a um objeto de metadados. No máximo, um conceito pode ser anexado a um objeto de metadados dessa maneira.

Um exemplo de um conceito seria quantia em dólar. Ao adicionar propriedades que fornecem a formatação visual correta de valores em dólares, e especificando uma tipo de dados comuns (por exemplo, um tipo decimal com pelo menos duas posições depois da vírgula) ea regra de agregação (por somatório, por exemplo), você pode formatar rapidamente todos os objetos de coluna que mantêm os valores que representam dólar montantes. Ao aplicar o conceito ao invés de localmente modificar individuais propriedades no nível do objeto, é assegurar que os metadados é consistente e facilmente sustentável.

Os conceitos são construídos em cima de conceitos já existentes. Isto é explicado em detalhes no próxima seção,"herança."

Herança

As propriedades podem ser gerenciados usando um recurso chamado herança. Herança ocorre baseando um objeto, o criança objeto, em outro objeto, a pai objeto. Em caso de herança, propriedades da criança e seus valores são obtidos por referência para as propriedades e valores de propriedade do objeto pai. Em um nível mais elevado, o objeto pai pode-se herdar de seu objeto próprio pai, que institui um corrente ou hierarquia de herança. Isso permite que mudanças nas propriedades em cascata jusante da cadeia de herança.

Objetos na cadeia de herança não são obrigados a herdar todas as propriedades de seu objeto pai. Em vez disso, eles podem mudar algumas ou todas as suas herdada propriedades e fornecer um valor local que se desvie o valor do pai objeto. Quando um objeto fornece um valor local de uma propriedade, ou define um propriedade que não está presente no objeto pai, a cadeia de herança é discriminados em relação a essa propriedade, e o objeto filho é dito substituir o propriedades de seu objeto pai.

A camada de metadados tem dois níveis de herança:

- Metadados objetos herdam as propriedades de seus objetos de metadados ancestral.
- Conceitos de herdar propriedades de seus conceitos ancestrais.

Metadados objetos herdam as propriedades de seu objeto pai. Por exemplo, tabelas lógicas e suas colunas herdar as propriedades de seus respectivos tabelas físicas e colunas. Isto é muito útil, pois permite um único ponto de definição para aquelas propriedades que geralmente precisam ser os mesmos. Por exemplo, o tipo de dados, formato de dados, e talvez a descrição definida para um físico coluna provavelmente pode ser reutilizado por colunas descendente em ambas as lógicas camada ea camada de entrega. A herança garante que, no caso do físico modelo é alterado, a alteração é imediatamente captado pelo objetos derivados.

Conceitos são baseados em conceitos já existentes, e herdar as propriedades de seu conceito pai. Na raiz da hierarquia é um conceito especial embutido conceito, o Base conceito.

hierarquias de conceitos permitem uma gestão adequada de propriedades relacionadas. Para exemplo, suponha que você queira aplicar formatação consistente dos números. Você poderia começar pela criação de um conceito genérico Número que herda da Base conceito. O conceito de número seria substituir apenas uma ou algumas propriedades de o conceito de base que são comuns a todos os itens numéricos. Por exemplo, poderia substituir o texto de propriedade de alinhamento e configurá-lo para a direita em vez da esquerda.

Localização de Imóveis

propriedades gerais, como o nome ea descrição pode ser localizada assim que pode ser exibido em vários idiomas. Isto é feito criando primeiro todos locais apropriados e em seguida, especificando o texto adequado para cada localidade.

Criação e manutenção de metadados

Esta secção explica brevemente os componentes que compõem a camada de metadados bem como as relações que os conectam. No restante deste capítulo, descrevemos esses componentes em mais detalhes, e explicar como criá-los utilizando o Pentaho Metadata Editor.

O editor de metadados em Pentaho

Pentaho oferece a Pentaho Metadata Editor para criar e editar metadados. Você Pode transferir esta ferramenta a partir da página do projeto Pentaho em sourceforge.net.

O Pentaho Metadata Editor é distribuído como um único Zip. (Para o Windows plataformas) ou . Tar.gz (Para plataformas baseadas em UNIX) arquivo. Descompactando

o arquivo gera um diretório que contém o software. Após descompactar o arquivo, você pode iniciar o editor, executando o MetaEditor.bat (Em plataformas Windows) ou metaeditor.sh (Para plataformas baseadas em UNIX) script.

O Repositório de Metadados

metadados Pentaho é armazenado em seu próprio repositório, que é distinto de ambos o Pentaho solução repositório e Pentaho integração de dados repositório. Atualmente, o Pentaho Metadata Editor é a única aplicação que se destina para editar o conteúdo do repositório de metadados.

Por padrão, a PMA utiliza arquivos binários para armazenar metadados. Estes arquivos, chamados mdr.btx e mdr.btd, São encontrados no diretório home do editor de metadados.

Você pode alternar de um arquivo baseado em repositório de armazenamento para um banco de dados baseado em repositório com bastante facilidade. Ao lidar com uma grande camada de metadados, desempenho do repositório baseado em arquivo pode diminuir significativamente. Neste caso, utilizando um repositório de banco de dados baseados na Internet podem aumentar o desempenho. Além disso, uma banco de dados baseado em repositório é mais adequado no caso de múltiplos desenvolvedores estão edição da camada de metadados simultaneamente.

O procedimento, descrito no README.txt arquivo encontrado no jdbc Diretório sob o diretório home do editor de metadados, é o seguinte:

1. Faça um backup do repository.properties arquivo localizado na jdbc diretório. Você pode mantê-lo no mesmo diretório, ou movê-lo para a segurança em outro lugar. O backup permite que você restaure o arquivo original com base em repositório de configuração.

2. A jdbc contém um número específico de RDBMS propriedades. arquivos. Substituir o original repository.properties arquivo com uma cópia do o RDBMS específico propriedades. ficheiro de escolha. Por exemplo, para armazenar o repositório em um banco de dados MySQL, faça uma cópia do MySQL.properties e renomeá-lo para repository.properties .

3. Abra o modificado repository.properties arquivo e editá-lo para apontar para seu banco de dados. Você deve fornecer valores para um número de propriedades. Os nomes dessas propriedades todos começam com MDRStorageProperty.org. netbeans.mdr.persistence.jdbcimpl. Esse prefixo é seguido por um ponto e um nome que configura uma propriedade de uma conexão JDBC. Típica nomes de propriedade (sem o prefixo) são:

   - driverClassName: O nome da classe Java do driver
   - url: A seqüência de conexão JDBC
   - userName: O nome do usuário do banco
   - senha: A senha do usuário do banco

NOTA Descobrimos que a carga eo desempenho salvar é bastante reduzida quando se utiliza o repositório de dados, em oposição ao repositório baseado em arquivo. Se você está considerando usar o repositório do banco de dados de base, você deve sempre levar algum tempo para medir o impacto no desempenho de sua situação específica. É difícil fornecer uma estimativa aqui, como o efeito real depende de uma série de fatores, tais como seu hardware, o RDBMS eo tamanho de sua camada de metadados.

Metadados Domínios

A camada de metadados Pentaho como um todo está organizado em uma ou mais metadados domínios. Um domínio de metadados é um recipiente para uma coleção de objetos de metadados
que podem ser usados juntos, como uma fonte de metadados para uma solução Pentaho. (Neste contexto, usamos o termo solução Pentaho""como definido no capítulo 4: uma coleção de recursos, tais como relatórios e seqüências de ação que residem em um única pasta no pentaho soluções diretório.)

Você pode criar um arquivo novo domínio, escolha Arquivo Novo Domínio Arquivo a partir do menu principal. Você pode excluir domínios escolhendo Excluir Domínio a partir do menu principal. Isso abrirá uma janela onde você pode escolher o domínio que você deseja remover a partir do repositório.

As subcamadas da Camada de Metadados

As seções seguintes descrevem os componentes do ambiente físico, lógico e camadas de entrega que estão incluídos dentro da camada de metadados.

A Camada Física

Os objetos que residem na camada física de um domínio de metadados são des- indicadores que correspondem mais ou menos um-para-um com objetos de banco de dados. A
seguintes objetos residem na camada física:

- Conexões de banco de dados descritores de conexão
- Física quadros-Descritores de tabelas de dados e pontos de vista
- Física Tabela Colunas- definições de uma tabela física

As subseções seguintes abordam cada um desses objetos.

Conexões

Aobjeto de conexão representa uma conexão de banco de dados. É uma descrição da conexão
tor, bem como os utilizados por Pentaho Data Integration.

Para criar uma nova conexão no editor de metadados, você pode usar o menu principal e escolher Ficheiro New Connection, ou você pode botão direito do mouse no Conexões nó na árvore do lado esquerdo do editor de metadados e

escolha Nova conexão no menu de contexto. Isso traz um diálogo que é, para todos os efeitos, idêntico ao Pentaho Data Integration Database Conexão de diálogo. (Consulte o Capítulo 9 para obter mais detalhes sobre como usar esta janela para criar uma conexão de banco de dados.)

Imediatamente após a conexão é criada, uma caixa de diálogo aparece. A caixa de diálogo apresenta as tabelas base que residem no banco de dados especificado, oferecendo a importar elas. Cancelar a caixa de diálogo para agora. Importando tabelas serão discutidos extensivamente na seção seguinte.

Depois que uma conexão é criada, a conexão aparece como um nó de árvore no exibição em árvore no painel esquerdo. Você pode botão direito do mouse no nó de conexão para trazer o seu menu de contexto. As opções oferecidas há de novo, bastante semelhantes aos você viu no Pentaho Data Integration. Você pode:

- Explore a conexão do banco de dados usando o Database Explorer.
- Use o Database Explorer para importar as tabelas (ou visões).
- Abra o editor de SQL para executar instruções SQL arbitrários.
- Duplicar a conexão.
- Excluir a conexão.

Se você gosta, você pode configurar conexões JNDI para o editor de metadados. O processo é idêntico ao de adicionar conexões JNDI para Pentaho Data Integração. Para usar JNDI, primeiro você precisa adicionar os descritores de conexão ao jdbc.properties arquivo localizado na simples jndi diretamente abaixo do diretório o diretório home do editor de metadados.

Tabelas e colunas Física Física

A camada de metadados podem descrever tabelas base e pontos de vista em um banco de dados usando chamada Os objetos físicos tabela. Os objetos físicos Tabela estão construindo de baixo nível blocos da camada de metadados. Eles oferecem uma abstração do banco de dados real tabelas.

Os objetos físicos tabela são objetos filho direto de conexões (discutido no subseção anterior). Ou seja, um objeto de tabela física é diretamente dependente em cima de um objeto de conexão existente.

Você pode importar tabelas físicas na camada de metadados clicando em uma conexão e escolher tanto a opção Importar tabelas ou importação de Explorer a opção no menu de contexto. A opção Importar tabelas permite que você apenas para importar tabelas base. A opção Importar do Explorer abre um banco de dados Explorer como você viu no Pentaho Data Integration. A partir daqui, você pode importação tanto a base tabelas e exibições.

Física Colunas é um filho direto de objetos físicos tabela. A Física objeto coluna representa uma coluna de banco de dados real, como uma tabela física representa uma tabela real no banco de dados. Colunas Física são normalmente adicionados para a camada de metadados automaticamente ao importar tabelas físicas.

Para editar uma tabela (ou suas colunas), botão direito do mouse e selecione Editar a partir do contexto
menu. A Física Quadro de diálogo Propriedades abre, como mostrado na Figura 12-3. Em no lado esquerdo da janela, você pode selecionar a tabela ou uma de suas colunas de uma árvore. Selecionando um item na exibição de árvore carrega a propriedade apropriada da página no lado direito da janela. Você pode navegar a uma propriedade particular rapidamente selecionando-o na vista de árvore propriedades esquerdo da página de propriedades.

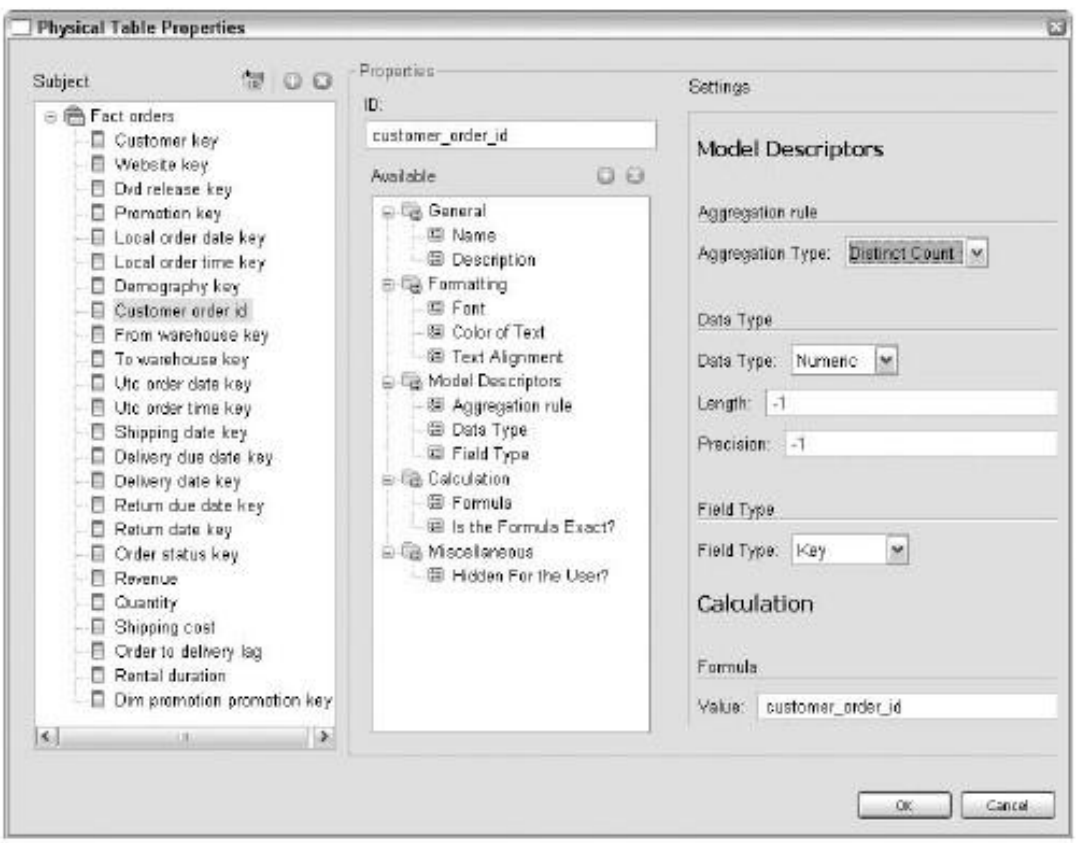


Figura 12-3: O quadro de diálogo Propriedades Físicas

Na caixa de diálogo, você também pode criar novas colunas personalizadas. Isso é útil se você deseja criar uma coluna calculada. Por exemplo, o COUNT (DISTINCT customer_order_id) que foi usado no exemplo de relatório pode ser adicionado desta forma.

Para criar uma nova coluna, clique no pequeno botão com o sinal de adição que aparece na parte superior da janela. Em seguida, editar as propriedades. Você deve pelo menos revisão
todas as propriedades no modelo de descritores e categorias de cálculo:

1. Tipo de Agregação-se aplicável, especificar a função de agregação. Para exemplo, o _order fato tabela no nosso exemplo Relatório de Vendas teria distinto agregador de Contagem.

2. O tipo de dados tipo de dados da expressão.

3. Tipo do campo-Este pode ser usado para especificar se a coluna é uma chave coluna, ou uma métrica ou um atributo de dimensão.

4. Este cálculo deve conter a expressão SQL que define esta coluna. Normalmente, isso é simplesmente o nome da coluna.

Se você precisa de definições de coluna personalizadas, note que você também pode definir na camada lógica. Na seção seguinte, vamos ilustrar a para o COUNT (DISTINCT customer_order_id) item.

Não há nenhuma regra dura e rápida que lhe diz onde adicionar estes personalizada colunas. Em alguns casos, pode ser necessário a coluna personalizada em vários locais, caso em que é provavelmente melhor para adicioná-lo ao nível de uma tabela física. Se a coluna personalizada é específico para um uso particular da tabela, é provavelmente melhor para incluí-lo na camada lógica ao nível dos quadros de Negócios.

A camada lógica

A camada lógica, literalmente, fica entre a camada física ea apresentação camada. A finalidade da camada lógica é descrever como os objetos do camada física se relacionam com o negócio. Os utilizadores empresariais só interagem com esses objetos de negócio ea camada de lógica, portanto, isola-los do técnico implementação a nível físico. Até certo ponto, isso permite que um certo grau de independência de banco de dados de esquema de relatórios.

Modelos de Negócios

A camada lógica é organizado em Modelos de Negócios. Funcionalmente, você pode pensar de um modelo de negócio como um data mart, ou seja, um subconjunto do data warehouse focado em um assunto particular do negócio.

Modelos de Negócios contém tabelas de negócios, relacionamentos e negócios Vistas. Tabelas de negócios e relacionamentos formam o back-end o Modelo de Negócio. Tabelas de negócios são os modelos de tabelas físicas e Relacionamentos definem como mesas de negócios podem ser combinados (entrou).

Vistas Business formam o front-end do modelo de negócios, e servem para apresentar o conteúdo do modelo para o usuário final. Uma visão de negócios é um recipiente para uma ou mais categorias de negócios e, como você verá mais tarde, um Categoria Negócios é funcionalmente similar a um esquema em estrela em um data mart.

Você pode criar um novo modelo de negócio clicando os modelos de negócios nó e escolha a opção Novo modelo de negócio. Na janela de propriedades, use a caixa de lista que aparece no canto superior direito da caixa de diálogo para especificar o banco de dados conexão. Atualmente, apenas uma conexão é suportada por Modelo de Negócios.

Tabelas e colunas de negócios Business

Tabelas de negócios residem em um Modelo de Negócio e decorrem directamente da Tabelas físicas. Da mesma forma, Negócios colunas são directamente derivados Colunas Física.

Em certa medida, mesas de negócios reproduzir fielmente a estrutura do seu
físico correspondente da tabela. No entanto, há uma diferença importante
Tabelas de Física: Uma mesa de negócios não representa a tabela real;
ao contrário, representa um uso particular de uma tabela física. Isso merece alguma
explicação.

A dim_date tabela é uma tabela de dimensões conformadas. É utilizado em muitos
diferentes papéis ao longo do data warehouse de Classe Mundial Filmes. Em um Busi-
ness modelo, essas tabelas de role-playing dimensão que cada um ser representado
por seus próprios negócios tabela. Por exemplo, num modelo de negócio para o Cliente
Ordens do data warehouse de Classe Mundial Filmes, poderíamos ter separado
Tabelas de negócios para a data do pedido, a data de transporte, data de entrega, eo retorno
data de vencimento.

Você pode criar tabelas lógicas rapidamente, arrastando uma tabela física para o
Tabelas de negócios nó. As Colunas física serão automaticamente importados,
também, e será representada por colunas de negócios.

Tal como acontece com as tabelas físicas, você também pode adicionar colunas
personalizadas para Empresas
Tabelas. Figura 12-4 mostra o negócio de diálogo Tabela. Na caixa de diálogo, você pode ver
um cliente Ordem coluna Contagem de ser definido. Seu Tipo de agregação de propriedade
será substituído e definido como o valor Distinct Count. Além disso, o tipo de campo é
substituído e definido como Fato. Essas modificações permitirão ao usuário de negócios
simplesmente escolher o item Contagem de Ordem, em vez de especificar explicitamente o
COUNT
função e os DISTINCT modificador na customer_order_id item.

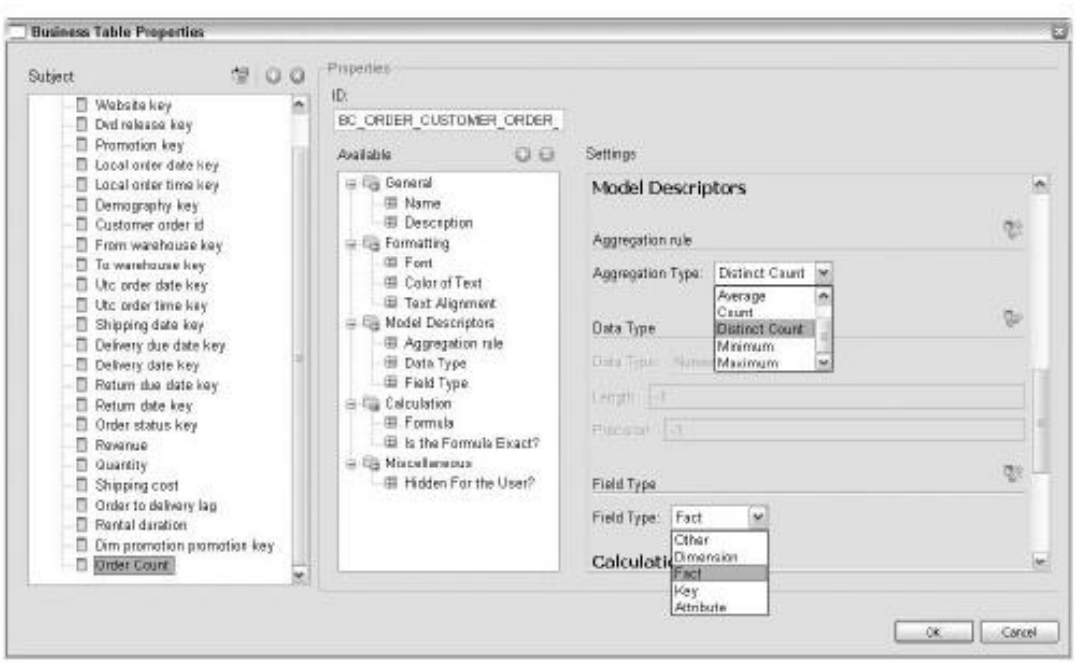


Figura 12-4: O negócio Quadro de diálogo Propriedades

Relacionamentos

Relacionamentos definir um caminho de junção entre duas tabelas de Negócios. Geralmente falam-
ção, cada mesa de negócios que aparece em um Modelo de Negócio deve ser relacionado com
pelo menos um outro negócio Quadro no mesmo modelo de negócio.

Não há nenhuma exigência lógica que dita que cada tabela de negócios deve estar relacionada a pelo menos uma outra tabela. No entanto, faz sentido fazê-lo de qualquer maneira. A razão é que, se um modelo de negócio é de fato um subconjunto próprio de
o armazém de dados focada em um aspecto particular do negócio, então todas as suas Tabelas de negócios deve de alguma forma contribuir para isso. Uma Mesa de Negócios que não é
relacionados com qualquer outro negócio de mesa tem aparentemente nada a ver com qualquer
das outras tabelas. Se esse for realmente o caso, ele provavelmente não deve ser parte de Neste particular Modelo de Negócios.

Você pode criar um novo relacionamento com o botão direito do mouse sobre o Relacionamento
nó e em seguida, escolhendo Nova Relação. Na caixa de diálogo que aparece, você pode selecionar as tabelas que estão relacionados, e especificar quais colunas devem ser comparados. Figura 12-5 mostra o diálogo de Relacionamento.

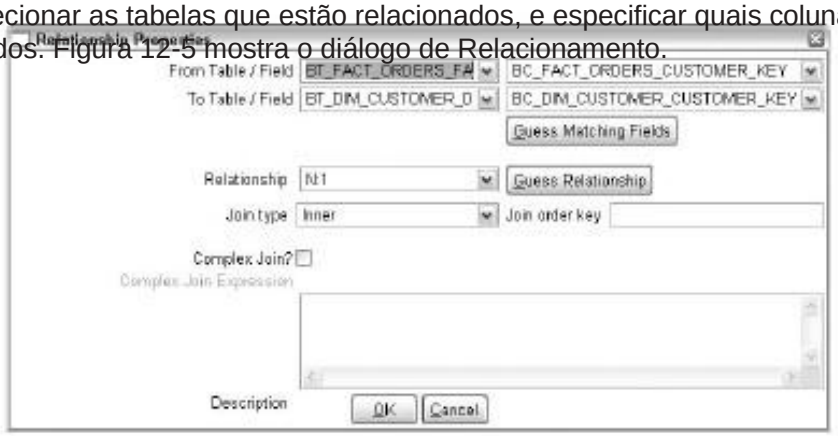


Figura 12-5: A relação de diálogo Propriedades

Uma coisa para manter em mente sobre os relacionamentos metadados Pentaho é que eles não são chaves estrangeiras. As tabelas em um relacionamento têm papéis diferentes: Existe uma"tabela"e de"Para uma tabela", mas não há sentido implícito na
sentido de que a partir da tabela deve ser o pai""Para a tabela e tabela deve
ser a criança"tabela", ou vice-versa. Em vez disso, a direção do relacionamento deve ser definida explicitamente usando a caixa de lista de Relacionamento. Na Figura 12-5 está definido para N: 1,
o que significa que pode haver várias linhas na tabela a partir de uma única linha na Para a mesa.

Por uma questão de sustentabilidade, é geralmente uma boa idéia de estabelecer um convenção para determinar se a criança ou a tabela pai deve ser
inserido como o da tabela (e, inversamente, se a mãe ou a criança
tabela deve ser inserida como a tabela). Na Figura 12-5, foi escolhido o

Pedidos tabela de negócios (que é a criança na relação, e mapas para o fact_orders tabela de fatos no armazém de dados) como o de mesa eo Tabela de negócios do cliente (que é a tabela pai na relação, e mapas para o dim_customer tabela de dimensão no armazém de dados) como o Para tabela.

A razão para a convenção usada na Figura 12-5 é um simples: tentar imitar o que seria o caso se seria a definição de uma chave estrangeira restrição. Nesse caso, também, a tabela filho""possui a chave estrangeira, que é evidenciando a partir da tabela filho para a tabela pai. Entretanto, se você sentir isso não é conveniente por alguma razão, você está livre para usar outra convenção, como Contanto que você tenha em mente que a propriedade Relações reflete corretamente o direção da relação.

A camada de entrega

A entrega camada contém objetos de metadados que são visíveis ao fim usuário, como Exibições de Negócios e categorias empresariais.

Vistas Negócios

Uma visão de negócios é uma coleção de chamadas categorias de negócios. Você pode pensar de uma visão empresarial como um data mart. Um data mart é uma coleção de funcionalmente relacionados com esquemas em estrela. Da mesma forma, uma visão empresarial é uma coleção de funcionalmente Categorias de negócios relacionados.

Você não precisa criar explicitamente uma visão empresarial. Não é simplesmente uma Visão empresarial em cada modelo de negócio.

Categorias de negócios

Uma categoria de negócios é um conjunto coerente de negócios relacionados com colunas. Funcionalmente, uma categoria de negócios pode ser pensado como um esquema em estrela. Como tal, uma categoria de negócios normalmente irá conter todos os itens que podem ser utilizados para comunicar sobre uma única tabela de fatos.

Dito isto, Categorias de negócios não parecem ter qualquer estrutura interna na medida em que o usuário final está em causa. A categoria Business simplesmente forma uma coleção de itens que podem ser usados juntos em um relatório, muito parecido com o Vendas exemplo Relatório apresentado anteriormente neste capítulo.

Você pode criar uma categoria de negócios com o botão direito do mouse sobre o Vista Business nó e escolha Nova categoria no menu de contexto. Para preencher a categoria com colunas, basta arrastar todas as colunas de interesse das tabelas de Negócios e soltá-los dentro da categoria.

Figura 12-6 mostra uma tela do editor de metadados mostrando uma simples Modelo de negócios ao longo das linhas do exemplo de relatório no início da Ordem o capítulo.

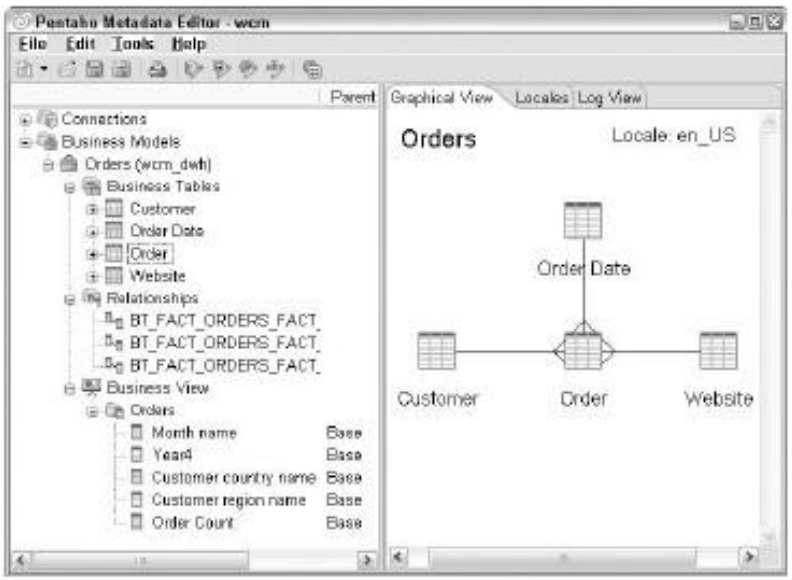


Figura 12-6: Um simples modelo de negócio

Como você pode ver, a figura contém as tabelas de negócios que correspondem a as tabelas reais mostrado na Lista 12-1. Da mesma forma, as relações entre as tabelas correspondem às JOIN cláusulas da Listagem 12-1. Uma vez implantado, os usuários podem criar todos os tipos de relatórios que envolvam ordens, as datas de pedido, clientes e sites sem precisar escrever SQL-se, em vez disso, o Pentaho camada de metadados é usado para interpretar as solicitações dos usuários para os itens do modelo (MQL) e gerar os comandos de banco de dados apropriado (SQL) para produzir os dados do relatório.

Implantação e uso de metadados

Depois de criar o Modelo de Negócio (s), você deve implantar a camada de dados antes você pode usá-lo para criar relatórios. Nesta seção, descrevemos como publicar os metadados. No próximo capítulo, você vai aprender como você pode realmente construir relatórios sobre uma camada de metadados já implantados.

Exportação e importação de arquivos XMI

Você pode criar relatórios sobre metadados usando a fonte de dados de metadados. Esta é explicadas em detalhe no Capítulo 13. Para criar um relatório baseado em metadados, você Deve informar o Report Designer, onde os metadados.

O Report Designer consome metadados no XML Metadata Interchange (XMI) formato. Para criar um arquivo XMI para o seu metadados, use o menu principal e escolha Arquivo Exportar para arquivo XMI. Da mesma forma que você pode usar o arquivo de importação de XMI opção Arquivo para carregar a camada de metadados com os metadados existentes.

Publicação de metadados para o servidor

Se os relatórios devem ser executados no servidor, os metadados devem estar disponíveis para o
servidor. Os metadados são armazenados no servidor como arquivos XMI. Você pode ter um XMI arquivo por Pentaho. Este arquivo deve ser chamado metadata.xmi.

Você pode simplesmente exportar os metadados para um arquivo XMI e, em seguida, basta copiar o
XMI arquivo para o diretório solução adequada no servidor. No entanto, para um servidor de produção, não é provável que todos os desenvolvedores de BI tem acesso directo à
sistema de arquivos do servidor. Portanto, o Pentaho BI Server oferece um serviço de que permite que você publique os metadados a partir do editor de metadados.

Você pode publicar metadados para o servidor de BI Pentaho no menu principal por escolha Arquivo Publicar servidor. Este aparece o diálogo Publicar Server, mostrado na Figura 12-7.

Figura 12-7: A Publicar diálogo Servidor

Para publicar, você deve configurar a configuração de publicação. A Local de publicação deve ser o nome de um diretório existente que residem sob o pentaho soluções diretório. A URL Publicar Web deve ser apontado para o Pentaho BI Server. Para obter a senha de publicação, você deve usar a senha que foi definido no publisher_config.xml arquivo. Essa configuração é coberto em Capítulo 3. Finalmente, você deve usar o ID de usuário e senha de um usuário que tem a função de Administrador (''Joe''e''senha''para uma instalação padrão).

Atualizando os Metadados

Depois de publicar ou copiar o arquivo XMI para o servidor, você deve dizer ao servidor para recarregar os metadados. Isso pode ser feito a partir do usuário através do console
No menu, escolha Ferramentas Refresh Reporting Metadados como mostrado na Figura 12-8.

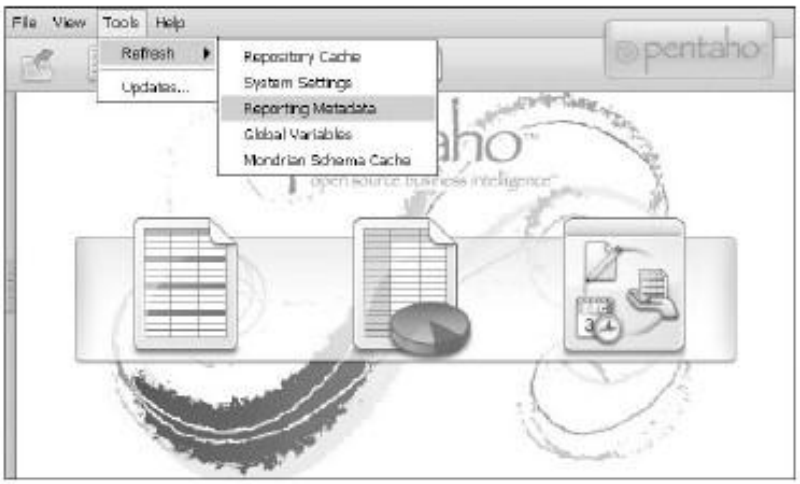


Figura 12-8: Atualizando os metadados com o usuário do console

Alternativamente, você pode atualizar os metadados usando o Servidor de Administração Console. Para atualizar os metadados do Server Administration Console, pressione o botão Modelos de Metadados no painel Server Refresh BI no Administração página da guia, mostrado na Figura 12-9.

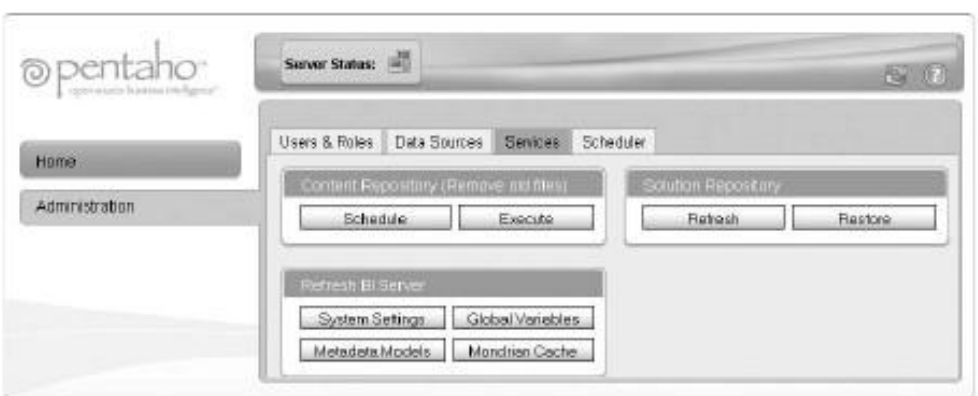


Figura 12-9: Atualizando os metadados com a Administração Server Console

Resumo

Este capítulo apresenta a camada de metadados Pentaho. Os metadados Pentaho camada permite que você apresente o seu banco de dados ou armazém de dados de uma forma que
é mais compreensível para usuários empresariais. Isto permite-lhes apresentar relatórios sem diretamente escrevendo em SQL. O capítulo seguinte descreve como você pode realmente usar a camada de metadados para criar relatórios.

Neste capítulo, você aprendeu:

- Que metadados é
- As vantagens de uma camada de metadados
- As características da camada de metadados Pentaho
- Como os metadados Pentaho é organizado em conceitos e propriedades e como A herança é usada para gerenciá-los
- Como usar o editor de metadados Pentaho (PME)
- A divisão da camada de metadados Pentaho em física, lógica e camadas de apresentação
- Como a camada física é organizada em conexões de banco de dados, tabelas e colunas
- Como a camada lógica é organizado em modelos de negócios, mesas, colunas, e relacionamentos
- Como a camada de apresentação é organizada em visões de negócios e categorias
- Como publicar a camada de metadados para o servidor Pentaho

CAPÍTULO

13

Usando o Pentaho Ferramentas de Relatórios

A forma mais comum a publicação de informações para os usuários finais é a criação de relatórios.
Na verdade, quando você olha para um típico ambiente de Business Intelligence (BI), cerca de 75 a 80 por cento do uso e conteúdo distribuído é composto de comunicação. Outra cento 15-20 utiliza ferramentas analíticas para OLAP, e apenas um número limitado número de pessoas (de 0 a 5 por cento) trabalhar com as ferramentas de mineração de dados. O mesmo
0-5 por cento está sendo tradicionalmente utilizada para indicar o tamanho do usuário população que utiliza um painel de gestão, mas isso está mudando rapidamente.
De fato, em uma solução Pentaho, a maioria dos usuários de acesso provável primeiro um painel
que exibe o conteúdo de BI sob medida para suas necessidades. Novamente, uma grande percentagem
deste conteúdo painel será composto de relatórios, portanto, de informação é um elemento-chave
de qualquer solução de BI. Este capítulo apresenta os dois relatórios Pentaho ferramentas, o Pentaho Web-based Ad Hoc de Consulta e Reporting Tool e os mais avançados Pentaho Report Designer. Temos tido muito prático abordagem, oferecendo muitas mãos sobre exercícios para que você possa acompanhar ao explorar as diferentes ferramentas. Supõe-se que você tem acesso a ambos os do Pentaho BI Server e Report Designer Pentaho.

Reporting Arquitetura

Todas as soluções de comunicação moderna tem uma arquitetura semelhante, como mostra Figura 13-1. A figura mostra as diferentes componentes de um relatório arquitetura:

- Um gerador de relatório para definir a especificação do relatório
- A especificação do relatório em um formato XML aberto

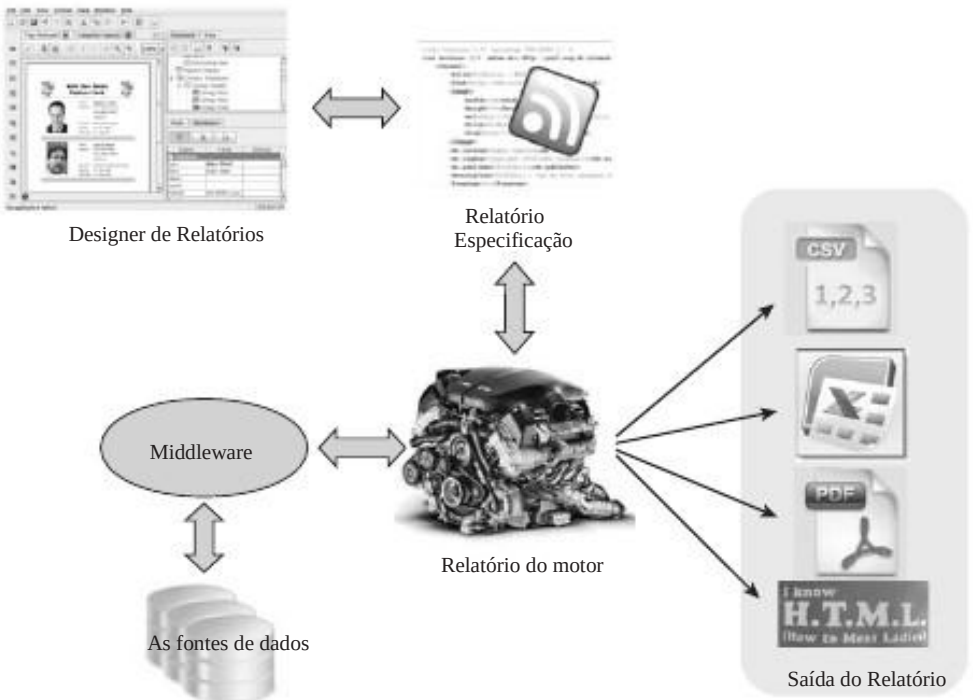


Figura 13-1: arquitetura Reporting

- Um mecanismo de relatório para executar o relatório de acordo com a especificação e processar a saída em formatos diferentes

- Definição de conexão de banco de dados que pode utilizar o middleware do padrão, como
  JDBC para se conectar a diferentes fontes de dados. Na versão mais recente do Pentaho
  Reporting, as consultas são executados diretamente a partir do mecanismo de relatório.

O modelo na Figura 13-1 é muito comum para um programa open source de relatórios solução. Pentaho não contém apenas os recursos para executar relatórios Pentaho mas também inclui as bibliotecas JasperReports para distribuir ou relatórios BIRT. A Pentaho mecanismo de relatório era conhecido anteriormente como JFreeReports, o designer é
uma versão completamente re-projetado do designer JFree relatório, que é agora chamado o Pentaho Report Designer, ou PRD. Embora a funcionalidade de outras soluções de comunicação, principalmente o projeto BIRT, já ultrapassou que da ferramenta Pentaho, sempre houve algumas vantagens importantes quando trabalham com a suíte de BI Pentaho que justificar o favorecimento do PRD e do JFree motor:

- Os relatórios criados com o PRD pode ser publicado diretamente no BI Pentaho Server a partir do menu PRD. Isso faz com que a implantação de novos relatórios ou existentes
  tão fácil quanto salvar uma planilha.

- PRD pode usar a camada de metadados Pentaho como fonte de dados, tornando-se um ferramenta ideal para usuários de poder sem conhecimento de SQL para criar as suas próprias
relatórios avançados.
- Os relatórios criados por usuários finais utilizando o Pentaho Web-based Ad Hoc Consulta e ferramenta de relatório pode ser estendida com PRD (embora depois de modificá-los com PRD, não pode ser editado com a web interface mais).

- PRD é muito fácil de usar depois de fazer um esforço para se familiarizar com as opções disponíveis, este capítulo pretende dar ao usuário inexperiente um avanço no uso do PRD para criar relatórios sofisticados.

Durante o primeiro semestre de 2009, o PRD tem uma revisão completa e está agora funcionalmente a par com a principal fonte de informação aberta outras soluções, e em alguns aspectos, o conjunto de recursos mesmo superior ao das ferramentas concorrentes.

Relatórios baseados na Web

O portal web Pentaho não é apenas para visualização e análise de conteúdo, mas também oferece recursos de relatórios ad hoc. A ferramenta de relatórios ad hoc trabalha em uma forma intuitiva, orientando o usuário através de um assistente de quatro etapas. Os relatórios
que podem ser criados usando o construtor do relatório baseado na web se limitam a agrupar listas, sem gráficos, quadros ou tabelas de referência cruzada. Isso limita a usabilidade do ferramenta baseada na Web para criar relatórios perspicazes BI-tipo, mas ele ainda pode ser usado por
usuários finais para obter rapidamente uma visão detalhada sobre questões específicas. O nome oficial
da ferramenta web é Web consulta ad hoc e cliente de relatório, ou WAQR para breve. O WAQR pode trabalhar apenas com modelos de metadados, que têm de ser criadas e publicado no primeiro servidor. Esse processo é explicado no Capítulo 12.

Criando um relatório é um processo simples. Você pode começar em um dos três maneiras: clique no botão Novo Relatório sobre a tela de boas vindas, clique no Relatório ícone de opção na barra principal, ou selecione Arquivo novo relatório a partir do menu. Todos estas opções de iniciar o Assistente de Relatório Ad Hoc, onde você começa escolhendo um modelo de negócio e um modelo. O modelo de negócios determina os dados que serão utilizados no relatório, o modelo determina que o relatório será
aparência.

NOTA Tenha cuidado quando você seleciona um modelo diferente depois de modificar o relatório configurações. Seleção de outra modelo redefine layout de página, tamanho do papel, bem como o relatório
cabeçalho para os valores padrão.

Depois de selecionar o modelo de negócio e instalada em seguida, leva Pentaho você ao fazer seleções parte do assistente. O lado esquerdo da tela
exibe todos os elementos disponíveis dados agrupados por visão empresarial, o direito

lado contém os grupos, os detalhes, e as caixas de filtro, onde os campos selecionados podem
ser colocado. Grupos permite que você adicione (sub) cabeçalhos e (sub) totais e pode
tem cinco níveis de aninhamento. Campos colocados na caixa de detalhes será a
parte interna do menor nível de grupo. A última caixa é para elementos filtrantes, estes
campos não farão parte da saída do relatório, mas pode ser usado para filtrar a
de dados. Os filtros também podem ser colocados em campos de grupo e de pormenor, mas o conteúdo do
esses campos serão sempre visíveis no relatório.

O relatório mais simples imaginável é a adição de apenas um único campo para os detalhes caixa, que também é o requisito mínimo para o repórter ad hoc para o trabalho.
Os relatórios podem ser visualizados em HTML (a opção padrão), PDF, CSV ou XLS.
Com estas duas últimas opções, Pentaho adiciona uma maneira conveniente de extração de dados
a partir de um data warehouse para a análise de qualquer ferramenta que o usuário está acostumado.
Quando todo o grupo, detalhes e campos de filtro foram adicionados, você pode clicar em Avançar,
que traz uma tela quase vazio, para além do mesmo grupo, detalhe,
e caixas de filtro que são posicionados do lado esquerdo. Este Customize Seleções
tela é onde o verdadeiro trabalho pode ser feito e contém muitas opções que
não são imediatamente óbvios:

- Classificando-Informação podem ser classificados em campos de grupo e detalhes. WAQR
  adiciona automaticamente os campos de grupo para seleção de classificação. Não é possível
  para remover essa triagem, a única opção é mudar a ordem de classificação. Detalhe
  campos não são adicionados automaticamente. Quando você clica em um campo de detalhes, o tipo
  tela aparece no lado direito da tela, onde o campo pode ser adicionado
  ea ordem de classificação pode ser alterada.

- Filtragem de Qualquer campo pode ser usado para filtrar, e várias condições podem ser combinadas usando os operadores E e OR. As condições disponíveis
  dependem do tipo de campo utilizado no filtro. Se for um campo de caractere,
  condições, tais como começa com ou contém Estão disponíveis, para uma data, a
  condições em,antesE depois podem ser utilizados, e para valores numéricos,
  operadores, tais como =,> =E <estão disponíveis. A opção de seleção está disponível
  onde os valores podem ser escolhidos de uma lista de valores. Isso é implementado
  como uma tela de pesquisa onde você pode usar o *personagem como um curinga. Se
  você deseja exibir todos os valores de uma determinada coluna, insira *e pressione Pesquisar.

- Agregando e formatação Vários funções de agregação e de campo
  formatos estão disponíveis para os campos de detalhe. valores não-numéricos só pode
  ser contados, mas para valores numéricos, o padrão funções de cálculo
  média,contagem,soma,minE max estão disponíveis. Estes resumos são
  colocado dentro de cada grupo ou subgrupo. Basta clicar em um campo de detalhes e
  as opções de formatação se tornará visível. Cada campo pode ser formatado
  individualmente.

- Agrupamento e paginação Cada grupo pode ser usado para a criação de uma página quebrar logo após ou antes de um novo grupo começa. Você também pode escolher se um total de grupo devem ser adicionados e se cabeçalhos de grupo devem ser repetidos em cada página. Para obter essas configurações, você precisará selecionar o nível correspondente (Nível 1 a Nível 5), que irá exibir o agrupamento disponíveis e opções de paginação.

A tela final com as configurações do relatório contém a orientação da página e tamanho e pode ser usado para inserir cabeçalhos e rodapés de relatório. Imprimir a data ea página números são adicionados automaticamente. Você verá que o botão Avançar na inferior direito da tela é cinza agora. Este é o comportamento correto: salvar o relatório não faz parte do assistente, mas deve ser feito usando o menu ou os botões de atalho na tela principal Pentaho. O relatório é salvo em o formato que era o formato de visualização ativa no momento de salvar, então se você selecionados PDF como a opção de visualização, o relatório salvo abrirá como um arquivo PDF.

Usos Práticos da WAQR

A opção WAQR é uma ótima maneira de começar a construir os seus relatórios com o primeiro Pentaho BI Suite, mas tem várias limitações que tornam improvável que WAQR será sua principal ferramenta de comunicação. Como já mencionado, gráficos e tabelas não estão disponíveis, e as opções de formatação são muito limitados. Por exemplo, ele não é possível modificar o tipo de fonte ou a cor dos valores exibidos na relatório a menos que você modifique as configurações da camada de metadados. A maneira como olhamos para ele, WAQR pode ser uma boa ferramenta em dois casos a seguir:

- Exportar dados-Seleção e exportação de dados para uma planilha ou um arquivo CSV é provavelmente a opção mais utilizada de WAQR. Há, naturalmente, muitas outras maneiras de obter dados de um armazém de dados em uma planilha, mas a velocidade ea facilidade de uso do WAQR para este efeito é difícil de bater.

- Quickstart relatório Relatórios criado com WAQR e salvos no servidor pode ser aberto a partir do Report Designer Pentaho (PRD) de novas modificação. Como criar um relatório básico de WAQR é geralmente muito mais rápido do que com o designer de relatório, isso pode poupar-lhe um considerável quantidade de tempo. Uma ressalva, porém: você terá direitos de acesso para o pasta onde o servidor de relatórios são salvos.

DICA Quase nada na plataforma Pentaho pode ser alterada ao seu gosto, incluindo o relatório de modelos. Os modelos estão armazenados na pasta do servidor de BI pentaho-solutions/system/waqr/templates. Cada modelo é armazenado em sua pasta própria, por isso a maneira mais fácil de adicionar o seu próprio modelo é uma cópia do existentes pastas e renomeá-lo. PRD pode ser usado para criar e modificar templates,

e as informações detalhadas sobre a modificação manual de modelos pode ser encontrado no Wiki em Pentaho http://wiki.pentaho.com/display/ServerDoc1x/Adhoc Reporting + + Modelos.

Pentaho Report Designer

A Pentaho Report Designer (PRD) é o front-end gráfico para criar, edição e publicação de relatórios para a plataforma Pentaho BI.

Uma das principais vantagens da utilização do PRD mais construtores outro relatório é o capacidade de utilizar modelos Pentaho metadados como fontes de dados. Os relatórios também podem ser publicado diretamente para o servidor Pentaho do designer para o uso no Pentaho Portal do Usuário. O novo . PRPT formato do arquivo é automaticamente reconhecido pelo a aplicação de servidor Pentaho assim um relatório PRD podem ser executados no portal sem a necessidade de adição de invólucros extra em torno dele.

Basicamente, existem dois tipos de escritores relatório: anilhadas e fluxo orientado ferramentas. ferramentas Banded dividir um relatório em um ou mais grupos de dados de relatório onde elementos podem ser colocados, enquanto as ferramentas baseado em fluxo permitem um formato mais livre colocação de elementos em uma página. PRD é um editor de relatório em faixas, assim como o bem conhecido e amplamente utilizado Crystal Reports. Embora autores do relatório em faixas são mais rigorosas na forma como os elementos de relatórios diferentes podem ser usados em um relatório, PRD permite a utilização de sub-relatórios, que em muito melhorar a flexibilidade e disposição opções.

RELATÓRIO DE ARQUIVOS PRD

Um relatório do PRD é armazenado como um . PRPT ficheiro de pacote. Este pacote contém um colecção de arquivos XML que definem o relatório. A layout.xml arquivo contém todas as informações de layout, enquanto o ds.xml *- arquivos contêm as definições de consulta.

Tenha em atenção que quando uma conexão JDBC puro é usado, as senhas são armazenadas como texto simples. É melhor usar as conexões JNDI e deixar o servidor de lidar com a definições de segurança.

Além da orientação da página das ferramentas de comunicação diferentes é outra distinção importante: WYSIWYG versus visão da estrutura. WYSIWYG (What You See Is What You Get) designers de relatórios permitem que você trabalhe em uma lona e os resultado final é imediatamente visível para o designer do relatório. PRD não é uma completa designer WYSIWYG, para que na maior parte trabalha com uma tela de design que mostra a estrutura do relatório, não o conteúdo e layout final. Uma opção de visualização é disponíveis para ver como o relatório vai olhar para um usuário final. Qualquer relatório pode também ser visualizada em diferentes formatos de saída disponíveis: PDF, HTML, XLS, RTF, e CSV.

NOTA Embora PRD não é um editor WYSIWYG completo, você pode alterar a maioria das opções de formatação diretamente no painel de propriedades, quando em modo de visualização.

As próximas seções explicam como PRD pode ser usado para criar relatórios perspicazes. A fim de ilustrar as diferentes partes do designer de relatório, nós usamos muito exemplo simples de um relatório com os anos, trimestres e meses a partir de uma dimensão de data WCM. Por fim, mostramos como construir uma venda mensal relatório usando o conjunto de ferramentas completo.

A tela do PRD

Quando você começa PRD, pela primeira vez, o aplicativo apresenta um Welcome tela, como mostrado na Figura 13-2.

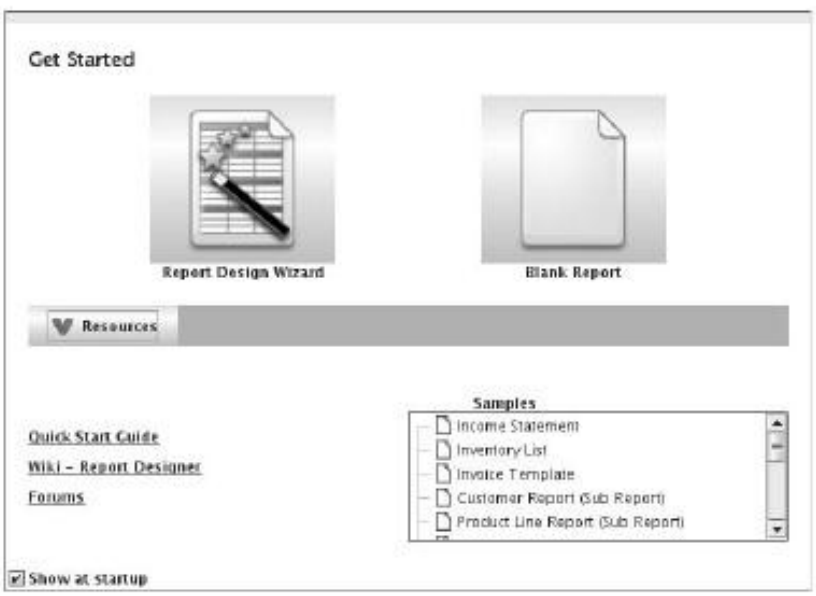


Figura 13-2: Pentaho Report Designer tela de boas vindas

Figura 13-2 mostra a tela de boas vindas com a guia Recursos aberta. Por padrão nesta parte da tela não será visível. Se você fechar o Welcome tela ou você tem a opção Mostrar na inicialização desligada, você pode reabrir tela clicando em Ajuda Bem-vindo. A tela de boas-vindas realmente vem em acessível, pois contém todos os atalhos para você começar rapidamente. Clicar a opção exibe Recursos links para o Guia de Início Rápido e Pentaho Wiki com a documentação do produto disponível. O link leva você Fóruns ao Fórum Pentaho onde você pode postar e responder perguntas, ou encontrar respostas de outros usuários que não podem ser encontrados na documentação. A Recursos de parte da tela também contém uma lista de exemplos para ajudá-lo iniciado rapidamente.

O layout da tela do PRD é bastante simples e se parece muito com qualquer outra ferramenta de relatório sobre o mercado. Há uma barra de menu no topo, a galeria com blocos relatório edifício do lado esquerdo, ea tela de propriedades do lado direito. A parte central da tela está reservado para a tela de design próprio onde você pode construir os relatórios e sub-relatórios. PRD oferece duas maneiras de construir um novo relatório:

- A nova opção (ou em branco de relatório no ecrã Welcome) cria um novo relatório vazio para você.
- O Assistente de Relatório leva você através das quatro etapas necessárias para criar uma relatório completo.

O Assistente de Relatório funciona de forma semelhante ao WAQR conforme descrito no seção"Usos Práticos WAQR", mas suas opções são um pouco diferentes:

1. Selecione um modelo para o layout do relatório exigido.
2. Definir os dados necessários, incluindo filtros e tipos.
3. Definir e agrupando os itens disponíveis a partir da consulta.
4. formatos campo Definir e funções de agregação.

Vamos pular o assistente no restante deste capítulo e se concentrar em configurar manualmente um relatório. É melhor que constrói a sua compreensão do ferramenta, o assistente é apenas um atalho.

Estrutura do relatório

Um relatório do PRD é dividido em várias seções de diferentes tipos. Algumas destas são padrão, como cabeçalhos de página e relatório, outros são flexíveis e podem ser adicionadas ou removidas por um designer de relatório. Quando você inicia um novo relatório em branco, a estrutura básica é imediatamente visível, como é mostrado na Figura 13-3.

Para criar Figura 13-3, a guia Estrutura no canto superior direito foi ativado e os grupos de árvores e detalhes do corpo foram ampliadas, clicando sobre eles. Quando você clica em um elemento, como o relatório mestre na tela, o editor de propriedades aparece no canto inferior direito. Qualquer estrutura ou conteúdo elemento tem propriedades que podem ser alteradas programaticamente ou usando o designer. As propriedades são divididas em estilos e atributos, onde o estilo propriedades são usadas para determinar o que um elemento parece, e os atributos determinar o conteúdo eo comportamento de um elemento.

DICA Para evitar que a tela fique demasiado cheio ao projetar seu relatório, os grupos e os cabeçalhos podem ser escondidos, clicando neles na Estrutura navegador e selecionar o atributo comum esconder sobre tela, que é uma simples checkbox. Marcando ou desmarcando esta opção não tem efeito sobre o relatório saída é apenas um projeto de ajuda.

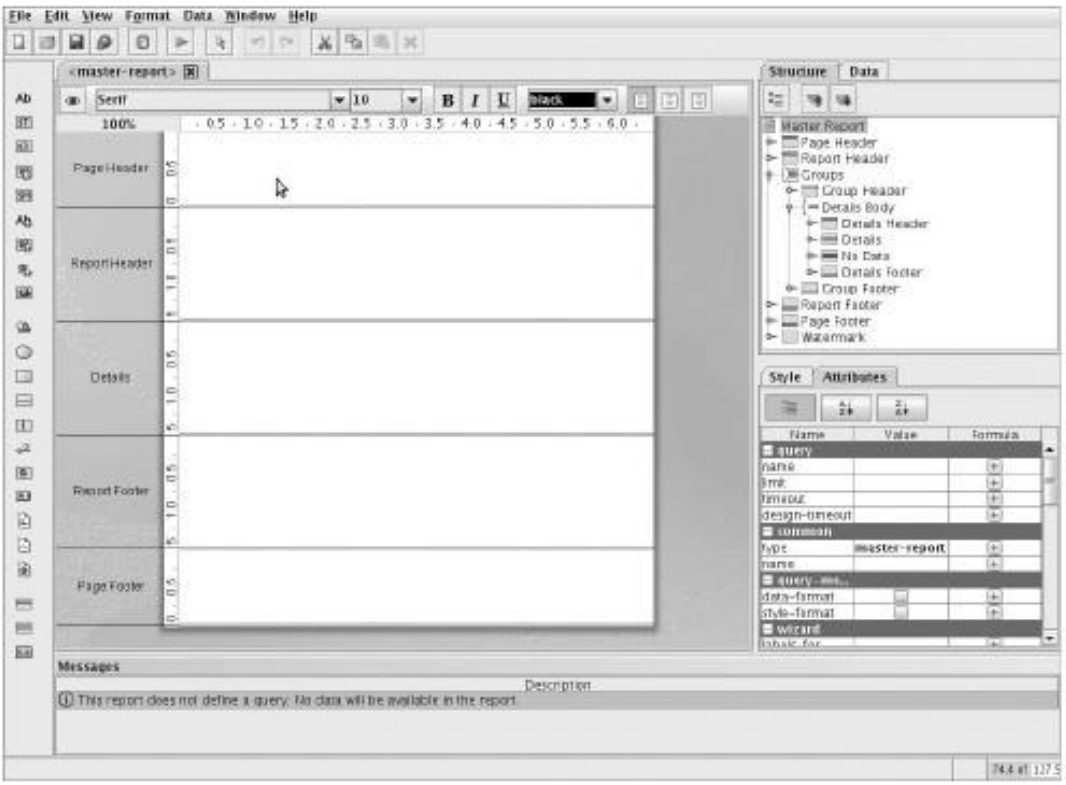


Figura 13-3: Estrutura do relatório

Os pontos básicos que compõem um relatório são:

- Cabeçalho / Rodapé de Página-Qualquer conteúdo colocado aqui será adicionado a cada
página do relatório. Exemplos de conteúdo são os números de página, data de impressão,
e os logotipos da empresa. A página de propriedades de estilo de comportamento pode ser usado para
indicar se o cabeçalho ou rodapé de página deve ser impresso na primeira ou a última página ou não.
- Relatório / Cabeçalho Rodapé-Qualquer conteúdo colocado aqui só será exibido uma vez. O uso típico de um cabeçalho do relatório é uma página com todos os relatórios
parâmetros, um resumo breve relatório, bem como o título do relatório. O rodapé é freqüentemente usado para exibir totais de relatório.
- Cabeçalho do Grupo / rodapé-A relatório tem pelo menos um grupo para organizar conteúdo. Cada grupo tem um cabeçalho e rodapé para colocar em nível de grupo rótulos ou subtotais. Os grupos podem ser aninhados em outros grupos, criando um estrutura hierárquica relatório.

- Detalhes Corpo Só o grupo mais interno contém o Corpo de detalhes,
que contém a banda Detalhes, onde as linhas individuais de uma consulta conjunto de resultados pode ser colocado. O cabeçalho e rodapé detalhes também fazem parte da
Corpo detalhes.

- Dados não-Este é um tipo especial de banda que pode ser usado para mostrar informações quando o conjunto de resultados da consulta está vazia. É parte dos detalhes
  Grupo Corpo.
- Watermark-Este banda especial pode ser usado para adicionar uma marca d'água que ser impresso como um fundo em cada página. uso comum é para mostrar um texto, como o''Projeto''ou''confidencial''ou um logotipo da empresa em cada página.

Relatório Elementos

O lado esquerdo da tela do design na Figura 13-3 contém a paleta de base
blocos de construção para exibição de conteúdo em um relatório. Tabela 13-1 contém uma breve
descrição desses elementos. Todas as opções são visualizados como ícones com o texto a ser exibido quando o mouse passa sobre o ícone.

Tabela 13-1: elementos Reporting

ELEMENTO	DESCRIÇÃO
etiqueta	A maioria dos elementos básicos para adicionar texto estático e rótulos de coluna a um relatório. Contém um rótulo para Assistente para criar facilmente cabeçalhos de coluna.
text-campo	Exibe valores de texto a partir de um conjunto de dados.
campo de número	Exibe os valores numéricos de um conjunto de dados. Contém formato atributos específicos para campos numéricos.
data-campo	Exibe valores de data a partir de um conjunto de dados. Contém formato atributos específicos para campos de data.
Mensagem de-campo	A maioria dos campos de dados dinâmico avançado. Pode conter combinado texto, o campo de referências e funções em uma única célula e permite para formatar os dados ao mesmo tempo. Por exemplo: Cliente: $ (nome) $ membro (apelido) desde: $ (data, date_registered, MM / dd / aaaa)
recursos rótulo	Baseado em um arquivo de recurso, PRD pode traduzir textos na etiqueta outras línguas.
recursos campo	Baseado em um arquivo de recurso, PRD pode traduzir no conteúdo do campo outras línguas.
recursos mensagem	Baseado em um arquivo de recurso, PRD pode traduzir mensagens em outras línguas.
teor de-campo	campos de imagem exibe a partir de um conjunto de dados.
imagem	Apresenta uma imagem de ambos os recursos de um local ou uma URL.
elipse	Insere uma elipse.
retângulo	Insere um retângulo.

ELEMENTO	DESCRIÇÃO
na linha horizontal	Insere uma linha horizontal.
linha vertical	Insere uma linha vertical.
levantamento escala	Um mini-mapa que exibe os resultados da pesquisa em uma escala de 1-5. (Isso é configurável. O intervalo real é definido através de atributos).
gráfico	Insere um gráfico, que pode ser editado pelo editor gráfico.
simples códigos de barras	Traduz o conteúdo do campo em um código de barras que pode ser lido por leitores digitais.
bar sparkline	Um gráfico de mini-bar para ser utilizado em linha.
linha sparkline	Um gráfico de linha mini para ser usado em linha.
torta-sparkline	Um gráfico de pizza mini para ser usado em linha. Este tipo de campo pode também ser usado para criar indicadores semáforo em uma gestão sumário.
banda	Pode ser usado para agrupar e formato diferentes elementos.
externa-elemento-campo	Pode ser usado para carga externa sub-relatórios a partir de uma URL ou caminho.
sub-relatório	Insere um sub-relatório, que pode ser aberto em seu próprio PRD tela.

Criando Conjuntos de dados

A parte mais importante da criação de um relatório é determinar quais dados devem ser exibido e, em caso de PRD, como os dados são agrupados e agregados. Embora o agrupamento e agregação podem ser adicionados mais tarde, faz sentido pensar o design do relatório antes de começar a construir os conjuntos de dados. A PRD relatório pode conter apenas um conjunto de dados, mas os relatórios podem conter sub-relatórios com seus próprios dados conjuntos. Não é possível usar ou combinar dados de sub-relatórios no relatório principal.

PRD pode recuperar dados de muitas fontes de dados, você pode até usar o JavaScript como fonte de dados. A maneira mais comum para criar consultas, no entanto, é usar um Conexão JDBC ou os arquivos de metadados Pentaho. Quando um relatório é criado, Existem três maneiras de criar uma fonte de dados:

- Usando a opção Adicionar Fonte de Dados com a opção do menu Dados
- Botão direito do mouse no ícone DataSets na guia Dados sobre a direita da tela
- Clique no ícone do banco de dados na guia dados diretamente

Com qualquer um destes métodos você obterá uma lista de opções com JDBC e Metadados no topo. Estas duas opções serão usados com mais freqüência para que eles sejam explicado mais adiante no texto que segue.

Criando consultas SQL usando JDBC

A tela de definição de uma fonte de dados JDBC consiste de um painel com os disponíveis conexões, consultas disponíveis, o nome da consulta e da consulta propriamente dita. Você já criou o wcm_dwh conexão no Capítulo 3 para que possa ser selecionado aqui.
Após clicar no sinal de mais à direita do texto disponível consultas, uma
consulta vazia é criada uma nova com o nome Consulta 1. Se esta é a única consulta
você estará criando para o relatório deste nome é bom, mas aconselhamos que você sempre dar um nome significativo para as consultas que você está construindo. Você pode escrever o Consulta SQL diretamente no painel de consulta, mas não há uma alternativa muito melhor, que irá aparecer quando você clicar no pequeno lápis na direita. Isso abre uma
Query Designer gráfico, que é uma versão integrada do open source
SQLeonardo projeto. O Designer de Consulta oferece uma maneira fácil de criar SQL consultas, mas você precisará de algum conhecimento de SQL básico para fazê-lo. É por isso que
incluiu um primer SQL no Capítulo 7.

A guia de design, que é aberto por padrão, consiste em consulta visual
representação no canto superior esquerdo, as tabelas e exibições disponíveis no canto inferior esquerdo,
ea tela de um desenho à direita. Figura 13-4 mostra um exemplo da tela
com a consulta que será usada para o exemplo de imagens mais tarde neste capítulo.

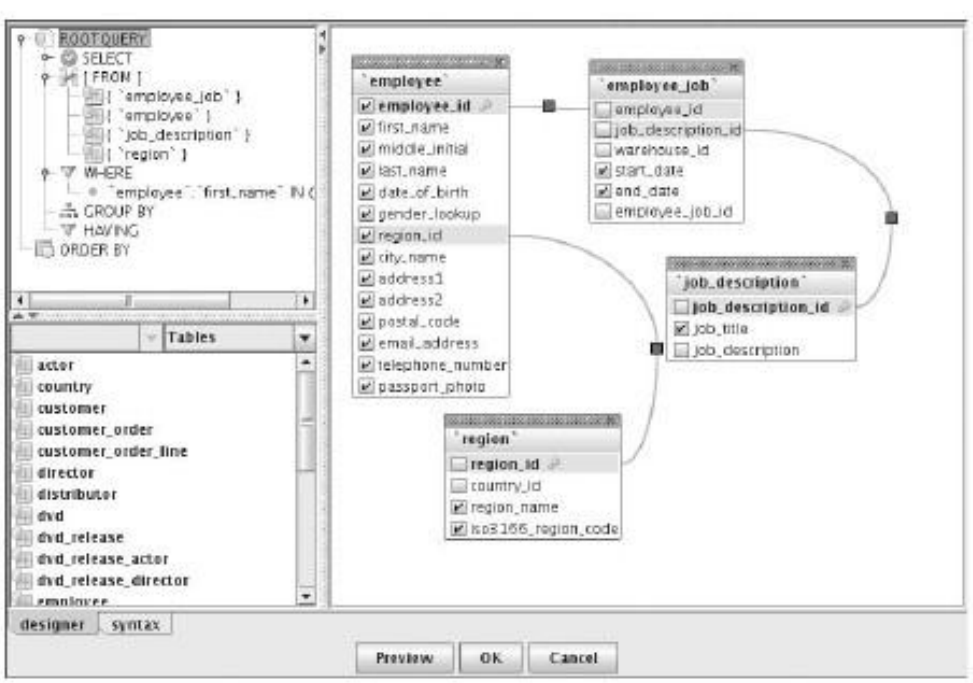


Figura 13-4: Designer de Consulta SQL tela

Primeiro você tem que decidir quais as tabelas a utilizar para a consulta. Você pode adicioná-los
para a tela à direita, arrastando-los lá ou simplesmente clicando duas vezes
o nome da tabela. Quando você adiciona uma segunda tabela você notará que a consulta Designer adiciona automaticamente a representação gráfica da chave estrangeira
as relações se tiverem sido definidos no banco de dados. Com certeza o jeito mais rápido para selecionar um esquema em estrela completa no criador de consultas é arrastar uma tabela de fatos
para a tela, clique no nome da tabela e selecione Abrir Todas as tabelas primárias
a partir do menu. Isto irá adicionar todas as tabelas de dimensão para a estrela
esquema porque eles são os quadros de chave primária para todos os _Casos campos na tabela de fatos. Por padrão, todos os campos da tabela são selecionados, para desmarcar-los, clique no
nome da tabela na tela e selecione Desmarcar tudo.

Note que isto só irá funcionar quando restrições de chaves estrangeiras foram definido. Com o MySQL utilizando MyISAM, isto não é possível a todos os que você tem que definir todas as relações manualmente.

NOTA Relações de chave estrangeira nem sempre existem no banco de dados, que significa que você tem que identificar as relações entre as tabelas mesmo. Basta clicar em uma coluna em uma tabela, mantenha o botão do mouse pressionado e mova o mouse para a coluna relacionada na outra tabela. Quando você soltar o botão do mouse, o relacionamento é adicionado. Botão direito do mouse sobre o pequeno quadrado vermelho no meio de um relacionamento permite-lhe seleccionar a função de edição. Isso traz a juntar-se editor onde você pode mudar de operador (=, >, < =>, <=, <>) e indicar a junção tipo, selecionando a partir do qual todos os valores da tabela devem ser recuperados.

Selecionar os campos que precisam ser parte do conjunto de resultados é uma questão de clicando nas caixas apropriadas. Adicionando cálculos leva um pouco de consideração, entretanto. Se tudo que você quer é uma soma de um campo, não marque a caixa de seleção, mas
botão direito do mouse na coluna e selecione Adicionar Expressão. As funções agregadas Contagem,Min,MaxE Soma estão disponíveis a partir da lista drop-down. Selecionando o Soma função para a coluna de receitas no fact_orders tabela cria o item sum ("fact_orders '.' receitas ') no conjunto de resultados. Embora PRD aceita
isso como um nome de campo, é melhor adicionar um alias significativa clicando com o função na lista de seleção, escolhendo Editar e adicionar o texto como Receita para a função. Do capítulo 7, você deve se lembrar que um agregado exige umpelo grupo declaração também. O Designer de Consulta não adiciona os campos para GRUPO seção automaticamente, mas estes podem ser facilmente arrastado lá.
Adicionando restrições à onde cláusula funciona de forma muito semelhante à adição expressões. Botão direito do mouse no nome da coluna na tela e selecione a opção adicionar condição em que. Isso abre o editor de conexão em que as restrições
pode ser definida. Não é nem uma opção de lista de valores nem qualquer outra forma de exibição de dados de uma única coluna que você terá que saber as entradas disponíveis antes de definir as condições.

DICA Quando você está criando um onde condição e não sei o valor correto para entrar, faça o seguinte:

1. Feche a tela do Query Designer, clicando em OK, e adicionar uma nova consulta à fonte de dados.

2. Abra a tela de Designer de Consulta e arraste a tabela que contém a coluna cujos valores que você precisa saber para a tela.

3. Desmarque todas as colunas e selecione a coluna da direita.

4. Direito do mouse no SELECT bem abaixo do cabeçalho ROOTQUERY e Escolha distintas.

5. Pressione Visualizar para visualizar os valores da coluna, você pode copiar os valores você quer, selecionando a linha (s) e pressionando Ctrl + C (não há botão direito do mouse opção).

6. Feche o designer, retire da lista de consulta de valores e reabrir o consulta original.

7. Edite o onde condição e colar os valores selecionados dentro (não for- começa a colocar aspas em torno do valor [s], se a seleção está em uma coluna de texto. Os valores numéricos não exigem aspas).

Você pode visualizar os resultados da consulta e do SQL que é gerado como resultado da consulta diretamente da tela do Query Designer. A guia de sintaxe em os interruptores canto inferior esquerdo da vista para o SQL gerado, ea opção de visualização executará a consulta e mostrar o conjunto de resultados em forma tabular. Quando terminar construção de sua consulta, pressione OK para fechar o Designer de Consulta e posteriormente
pressione OK para fechar o editor de fonte de dados. O novo conjunto de dados agora aparece na
Dados da guia PRD com todos os nomes de coluna expandido, como mostrado na Figura 13-5.

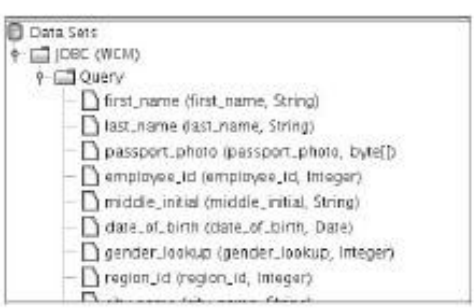


Figura 13-5: JDBC resultado de consulta

DICA Quando quiser adicionar uma tabela pela segunda vez para ser usado como um alias (para exemplo, para usar a dimensão de data para a data do pedido ea data da expedição), você

notar que PRD gera um aviso com o texto"A tabela já carregados e
aliasing desativado!"Basta fechar esta mensagem e mude para o separador de sintaxe do
Designer de Consulta após desmarcar todas as colunas. Adicione o nome de alias para o direito de
o nome da tabela. Não não OK, mas voltar para o guia designer. Você receberá um
mensagem dizendo que a sintaxe mudou: aceitar as mudanças ea tabela
nome é mudado para o novo apelido. Agora você pode adicionar a mesma tabela uma
pela segunda vez. Repita esse processo se aliases são necessários mais.

Criando consultas de metadados

Ao selecionar Metadados como fonte de dados, uma tela semelhante aos dados JDBC
janela de fonte aberta. Além da ligação correta, os dados de metadados
Fonte tela Editor exige a seleção de um arquivo XMI. Depois de selecionar o
XMI arquivo e abrir o editor de consultas, o MQL Query Builder é aberto para que
o modelo de negócios pode ser selecionado. As visões de negócios disponível dentro de um
modelo são exibidos como pastas e campos podem ser selecionados clicando no
nome do campo e, posteriormente, clicando na seta para adicionar o campo para a
Selecione, Condição, ou ORDER BY lista.

ATENÇÃO Em PRD, os dados têm que ser ordenadas do jeito que você quer que ele apareça em
o relatório usando as opções de classificação na consulta. Não há outra maneira de classificar os
dados depois de ter sido recuperado.

Criando um conjunto de dados usando o editor de MQL parece ser uma muito simples
processo e, à primeira vista ele é. O mesmo truque para obter valores a serem aplicados em
sua
condições pode trabalhar com o editor de consultas JDBC, mas você deve estar ciente de
as seguintes limitações:

- Nenhum expressões podem ser adicionados, por isso, quando um objeto não está
  disponível no
  modelo de metadados que você não pode adicioná-lo no editor de MQL. Em vez disso, o
  modelo
  precisa ser ampliado e exportada como um arquivo XMI novamente antes de
  usar o novo campo.
- Condições não podem ser agrupadas usando parênteses para forçar a avaliação
  precedência, mas são avaliados de acordo com a ordem expressão lógica
  onde E tem precedência sobre OU. Você pode experimentar por si mesmo: O
  expressão E um b ou C e d é avaliada como (A e b) ou (c e d),
  que poderia ser correto, mas se você queria ser A e (b ou c) e d
  você tem um problema porque o editor não pode MQL condições, este grupo
  caminho. A única maneira de condições corretamente grupo desta forma é reconstruir
  a consulta no editor JDBC vez e adicionar manualmente os colchetes.

O resultado da consulta MQL é o mesmo que com qualquer outra fonte de dados.
No primeiro nível é um conjunto de dados que mostra o tipo de fonte de dados (neste caso
Metadados) Com o nome da conexão entre parênteses. Aninhados abaixo do conjunto de
dados é

a consulta com o nome da consulta (na maioria dos casos, uma consulta ou apenas de consulta), e no
nível mais baixo dos campos disponíveis. A diferença entre um JDBC e Metadados conjunto de dados é a descrição das colunas de metadados, que mostra a
nome da coluna metadados de negócios, e não o nome da coluna do banco de dados real.

Exemplo de dados Set

Para as seguintes instruções que você vai usar um dado muito simples definido a partir do dim_date tabela. Os dados do exemplo dado pode ser criado usando ambos os
JDBC e editores de metadados e consiste dos campos ano4,quarter_name, month_numberE MONTH_NAME de qualquer tabela de dimensão de data. Ordem dados por ano4,quarter_nameE month_number e Verifique se a opção Distinct
é escolhido. Nenhuma condição pode ser adicionada ainda. Arraste as linhas do resultado para o
Detalhes da faixa e pressione o botão Preview (o olho''pouco''para a esquerda do lista de fontes drop-down na parte superior da tela de projeto). Figura 13-6 mostra o resultado parcial do presente exercício. Parabéns-você acabou de criar seu primeiro relatório!

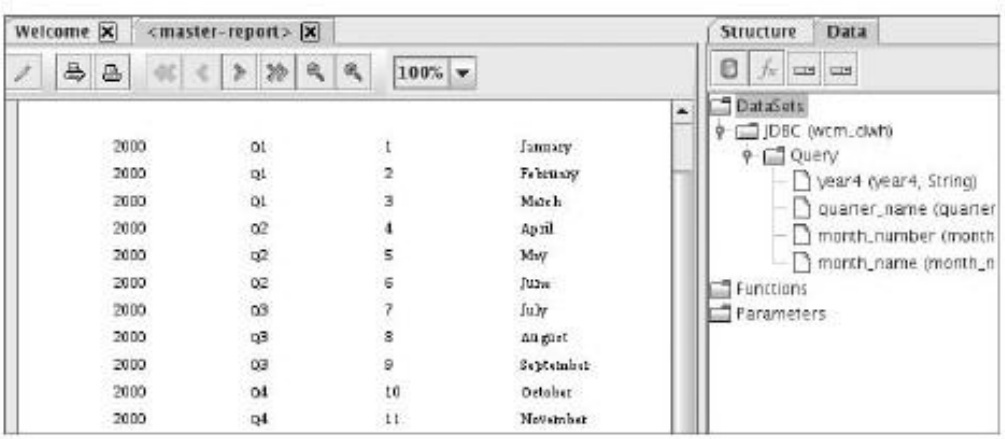


Figura 13-6: Primeiro Relatório Lista

Adicionando e Usando Parâmetros

Fixo consultas são muito bem para relatórios padrão, mas normalmente um pouco mais de interação
é necessária. Essa interação pode ser adicionado usando parâmetros que permitem que um usuário a escolher determinados valores cada vez que o relatório é executado. Adicionando parâmetros
É fácil usar o PRD Adicionar função de parâmetro. Você pode encontrar este sob a Dados do menu, clicando na árvore Parâmetros na guia Dados, ou simplesmente clicando no ícone de atalho na parte superior da guia Dados. Um parâmetro é normalmente com base num conjunto de valores que um usuário pode escolher, mas um parâmetro de texto livre é
como bem disponível. Todos os tipos de parâmetro e exigem uma lista predefinida de IDs e valores. A lista de valores deve vir de uma fonte de dados diferente
da consulta principal. Qualquer fonte de dados podem ser usados aqui e porque JDBC e

fontes de metadados já tenham sido abrangidos, vamos introduzir um novo tipo de fonte de dados aqui. PRD tem a capacidade de definir uma tabela personalizada, incluindo o conteúdo da tabela no relatório. Quando você adiciona um parâmetro e clique em no sinal de mais para criar uma nova fonte de dados, selecione Tabela. Isso inicia o Quadro Datasource editor, que permite definir IDs de costume e valores. O ID é o valor que será passado para a consulta, o valor será exibido como uma escolha para o usuário quando executar o relatório.

Além do parâmetro de texto básica, sete tipos de visualização e seleção são disponíveis. A lista drop-down, lista simples / múltipla, um botão de rádio e caixas são elementos básicos da interface do usuário que podem ser encontrados em muitas outras aplicações como também. Os dois últimos são bastante raros, o que é estranho, considerando a clareza ea facilidade de uso para usuários finais. Estes são os tipos simples e multi-botão, que apresentam uma banda de botões para escolher. Para o relatório de exemplo, você criará dois parâmetros, um para a selecção de um ano a partir de uma caixa suspensa, e um para selecionando um ou mais quartos com um parâmetro de multi-botão. Cada parâmetro vai buscar a sua própria consulta. Siga estes passos:

1. Adicionar um novo parâmetro para o Ano e adicionar um JDBC ou dados de metadados fonte. A fonte de dados contém uma consulta que seleciona os valores distintos da ano4 campo em qualquer dimensão de data.

2. Dê o nome qry_param_year à fonte de dados eo nome do parâmetro param_year, Insira o texto Selecione o ano como parâmetro Label Drop e selecione Down como tipo. Como a consulta contém uma única coluna, ano4, este é automaticamente selecionado para o ID e Valor.

3. Criar o Trimestre parâmetro. Adicionar um novo parâmetro param_quarter e adicionar um novo quadro DataSource com o nome tbl_param_quarter. Agora inserir os valores como mostrado na Figura 13-7.

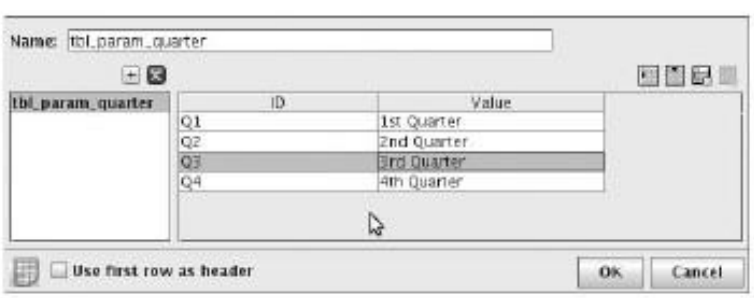


Figura 13-7: Parâmetro editor Tabela

4. Selecione Id como ID / Nome e valor como valor a partir das listas drop-down e certifique-se que java.lang.String é selecionado como o tipo de dados. O Adicionar tela de parâmetro agora parece que a tela na Figura 13-8.

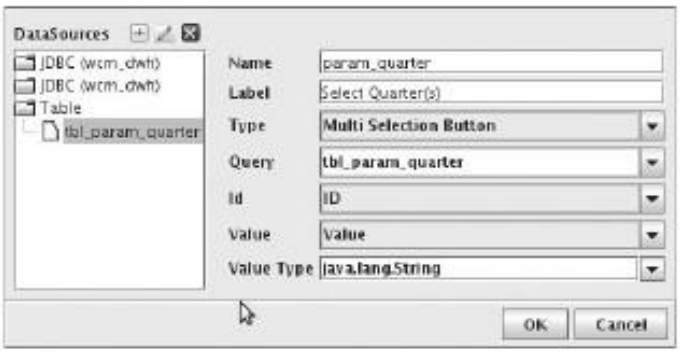


Figura 13-8: Adicione tela de parâmetro

Os parâmetros do recém-criado pode ser testado, executando uma previsão: a parâmetros será visível na parte superior da tela do relatório. Você pode selecionar e valores desmarcar, mas fora isso, nada acontece ainda. A etapa final implica associar os parâmetros para a consulta principal. Primeiro você precisa saber como referência o parâmetro em uma consulta. Usando uma fonte de dados JDBC este é simplesmente uma questão de colocar o parâmetro ID com chaves e antes los com um cifrão, como no $ {} Param_year. Você pode usar o gráfico designer para adicionar um onde condição para as colunas ano4 e quarter_name mas o texto pode ser digitado diretamente sem abrir o editor. O completo consulta deve ser semelhante ao seguinte:

```
SELECT DISTINCT
    "Dim_date_en_us '. Ano4,
    "Dim_date_en_us '. Quarter_name,
    "Dim_date_en_us '. Month_number,
    "Dim_date_en_us '. MONTH_NAME'
DA
    "Dim_date_en_us '
ONDE
    "Dim_date_en_us '. Ano4' = $ {} param_year
 E 'dim_date_en_us'. Quarter_name 'IN ($ {} param_quarter)
```

Como o parâmetro trimestre param_quarter pode ter vários valores que usar o EM operador e colocar o parâmetro entre parênteses. Quando a consulta é executado, isto se traduz em uma lista separada por vírgulas de valores. Agora você tem um relatório parametrizado e ao selecionar a opção de visualização, você verá uma tela vazia. Depois de selecionar um ano, e um ou mais quartos, e clicando em Atualização, os dados selecionado aparece na tela. Com a opção Autoupdate Também é possível ter o fogo consulta automaticamente quando um novo valor é selecionado. O resultado de seu trabalho até agora é exibido na Figura 13-9.

Para maior clareza, todas as consultas e os parâmetros foram ampliados, o que faz claro porque os nomes de parâmetro de consulta e devem ser significativas. Faz manter o relatório muito mais fácil quando você faz.

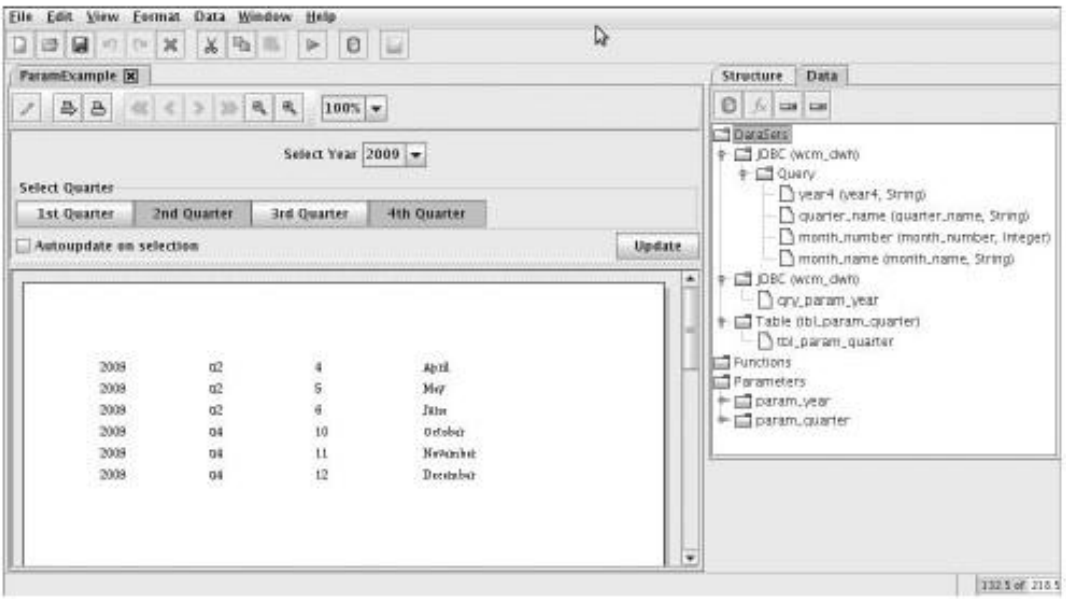


Figura 13-9: Exemplo de parâmetro de relatório

Layout e Formatação

A maioria das pessoas gastam horas incontáveis de trabalho no layout do relatório e mais ferramentas de comunicação não fazem isso muito fácil. PRD, no entanto, é baseada em algumas
smart princípios de design que torna muito fácil de alterar rapidamente o formato de o relatório. O primeiro e mais importante é herança de estilo (Capítulo 12 contém uma introdução básica aos princípios da herança.) Quando você olha a estrutura de um relatório no painel de estrutura, você vai notar que a estrutura do tem a forma de uma árvore aninhada. Esta não é uma coincidência: PRD segue esta hierarquia
quando mudar as propriedades do item. O guia de propriedade Style contém quatro colunas: Nome, Herdar valor e fórmula. A segunda coluna, herdar, está marcada para cada propriedade, por padrão. Isto significa que as configurações para a propriedade são herdadas de um nível mais elevado na hierarquia. Você pode tentar isso em um simples maneira: Basta selecionar a banda Detalhes no painel 'Estrutura. A família de fontes tem a caixa de seleção para herdar selecionados eo valor é Serif, que é o padrão fonte. Agora altere o valor para Verdana: Você perceberá que remove o PRD Herdar de seleção, e todos os itens que fazem parte da banda de detalhes têm agora a font-family Verdana definido automaticamente por herança. O nível mais alto na hierarquia é o nível do relatório mestre-e aqui você pode mudar o estilo de relatório de largura propriedades, como tipo de fonte, tamanho de fonte, cores de texto, e assim por diante.

Este princípio de herança não é restrito aos internos nos níveis de estrutura. A agrupamentos de adicionar-se funcionam da mesma maneira. Suponha que você queira criar um cabeçalho de uma linha de detalhes que consiste em quatro colunas. Você pode adicionar quatro
rótulos de colunas e formatá-los individualmente, mas é muito mais fácil para criar primeiro uma banda arrastando-os elementos de design e colocando as etiquetas dentro

a banda. Agora só a banda precisa ser formatado, todos os rótulos herdará os valores de propriedade da banda. É impossível cobrir todas as opções de formatação aqui, nos restringimos aos primeiros passos para ajudá-lo em seu caminho. Nós abrirá o relatório do ano-de-Semana pouco para os exemplos e começar por adicionando um cabeçalho para as colunas no relatório:

1. Navegue até o cabeçalho de detalhes na estrutura de árvore. Se a esconder sobre tela está marcada, desmarque-a para apresentar a banda na tela.
2. Arraste uma faixa da lista de objetos à esquerda para a banda de cabeçalho e detalhes posicioná-lo no topo.
3. Abra os atributos de estilo e definir a largura, que pode ser encontrado no tamanho e posição do grupo, a 100 por cento.
4. Definir o estilo seguintes atributos para a banda:
   - Font-family: Arial
   - Tamanho da fonte: 12
   - Bold: checked
   - Texto de cor: branco
   - Bg-cor: cinza escuro
5. Arraste quatro rótulos para a banda. Eles herdam automaticamente o estilo atributos que você acabou de criar para a banda. As guias de posicionamento pode ser usado para colocar o rótulo exatamente acima dos quatro colunas no relatório. Isso é é a seguinte: a formatação feita.
6. A única coisa que resta é definir os nomes certos para as colunas. Na Atributos guia você verá o valor propriedade digitar o cabeçalho da coluna, ou você pode usar as etiquetas para Wizard, que apresenta uma lista drop-down com as colunas disponíveis na consulta. Se o nome da coluna consulta é exatamente o que você gostaria que o cabeçalho para exibir, usar o assistente. Em qualquer outro caso, basta digitar os valores.

Além das propriedades de estilo, PRD contém um multi-tabbed para-tela esteira similar às telas de diálogo de formatação encontrados nos mais modernos processadores de texto. Basta selecionar um objeto em seu relatório e selecione Formatar Fonte do menu principal.

Cores de linha alternadas: Bandas da Linha

Para melhorar a legibilidade de um relatório tipo de lista, é uma boa idéia para uso alternativo colorir linha onde as linhas com um número de linha até obter uma cor diferente aqueles com um número de linha ímpar. Muitas ferramentas de relatórios exigem que você use

algo semelhante ao seguinte: a primeira utilização de um cálculo mod, em seguida, criar uma variável, e, finalmente, usar a formatação condicional com base na variável de destacar linhas ímpares e pares. Com PRD, este processo é ainda mais simples. Selecione a opção de formato de linha de bandas a partir do menu e escolha as cores necessárias para as linhas ímpares e pares. PRD, em seguida, cria uma linha de bandas de função, que é acessível a partir da guia Dados na pasta Funções. Basta arrastar a função de Detalhes da banda e tornar o campo invisível. Quando você clique em Visualizar, você ver as cores de linha alternadas aparecer, como mostrado na Figura 13-10.

Year	Quarter	Month number	Month name
2009	Q1	1	January
2009	Q1	2	February
2009	Q1	3	March
2009	Q2	4	April
2009	Q2	5	May
2009	Q2	6	June
2009	Q3	7	July
2009	Q3	8	August
2009	Q3	9	September
2009	Q4	10	October
2009	Q4	11	November
2009	Q4	12	December
2008	Q1	1	January

Figura 13-10: bandagem Row aplicada

Agrupando e resumindo dados

Um dos princípios básicos do PRD (ou qualquer outra ferramenta de comunicação em faixas) é o
agrupados apresentação de dados. Até agora, você criou as consultas necessárias para selecionar e filtrar os dados para o relatório. Agora é hora de trazer estrutura para o conjunto de resultados, adicionando grupos, cabeçalhos e sumários para o layout do relatório.

Adicionando e modificando grupos

Um grupo é usado para organizar o conteúdo em diferentes níveis, onde os sub-cabeçalhos e totais podem ser adicionados. Os grupos podem ser adicionados clicando em qualquer lugar a tela de design, clicando na seção Grupos na estrutura de relatório,
ou clicando no ícone Adicionar Grupo na guia Estrutura. Quando você adiciona um grupo, uma nova camada é adicionado à seção Grupos. Cada grupo pode ligar para um campo nos dados de relatórios de jogo, assim quando você quer criar um relatório com uma
agrupamento Ano-de-mês, é melhor quando esses campos já estão presentes no conjunto de dados em primeiro lugar, embora isso não seja necessário. Um grupo padrão é usado para
agrupar os dados de detalhe, como você pode ver na árvore de estrutura. Quando você abre o editor de grupo (o ícone mais à esquerda da esquerda na guia estrutura) este vazio

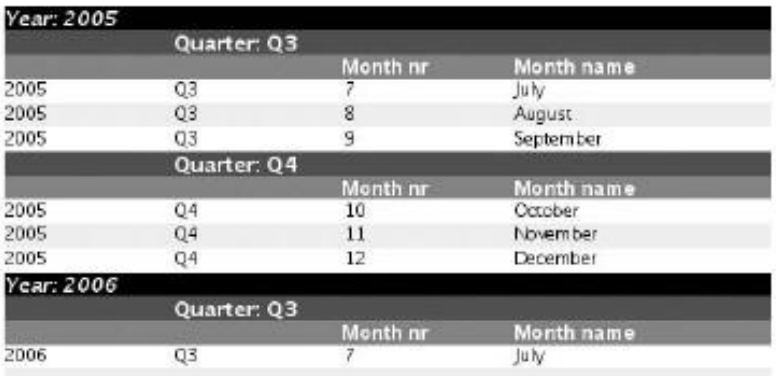
grupo padrão é exibido como uma linha em branco que faz referência a um campo vazio lista []. Este é o grupo padrão, que pode ser usada para o agrupamento, mas não pode removê-lo, embora você pode remover todos os grupos a partir do editor grupo.
Após fechar o editor e reabri-lo, você vai notar a linha em branco com o
grupo padrão novamente. Porque você não quer desperdiçar recursos, você vai usar este grupo padrão para o nível Trimestre grupo em seu relatório de exemplo. Usando nomes de grupo é necessária, caso contrário você não pode fazer referência os grupos da relatório. Então adicione QuarterGroup como o nome e selecione o campo quarter_name como o campo de grupo. Enquanto você ainda estiver nessa tela, digite um outro grupo e nome presente YearGroup. O campo de referência é a ano4. Clique na YearGroup linha e movê-lo, clicando na seta para cima. Figura 13-11 mostra o resultado até agora.

Figura 13-11: editor do Grupo

Quando você executar um preview de novo, você vai notar que os resultados têm agora um quebra em cada trimestre, e que os detalhes do cabeçalho é repetido para cada trimestre. Agora adicione uma nova banda em ambos os anos, o quarto eo cabeçalho e definir a banda largura de 100 por cento novamente. Verifique se o esquema de cor é diferente para cada um dos
os três níveis de agrupamento (Ano, Trimestre, detalhes).

NOTA Não é (ainda) possível arrastar e soltar objetos, tais como rótulos, campos ou gráficos de uma banda (cabeçalho) para outro, mas é possível copiar ou cortar a partir de uma faixa e colá-los em outro.

Agora você pode simplesmente mover o cabeçalho campos do cabeçalho de detalhes para a ano e quarto, respectivamente, cabeçalho, mas vamos usar um campo de mensagem em seu lugar. A
anos de campo de cabeçalho de mensagem obtém o valor Ano: $ (ano4); Cabeçalho no trimestre
campo da mensagem se o valor Bairro: $ (quarter_name). Agora você pode remover os campos de detalhe para ano4 e quarter_name também. Para ser capaz de selecionar mais de um ano e ver os efeitos do agrupamento que você adicionou, você pode alterar o parâmetro ano, em uma lista ou tipo multi-multi-botão e alterar a consulta condição para EM. Não se esqueça de adicionar o parâmetro entre colchetes. A relatório deve agora parecido com o exemplo na Figura 13-12.



Year: 2005			
	Quarter: Q3		
		Month nr	Month name
2005	Q3	7	July
2005	Q3	8	August
2005	Q3	9	September
	Quarter: Q4		
		Month nr	Month name
2005	Q4	10	October
2005	Q4	11	November
2005	Q4	12	December
Year: 2006			
	Quarter: Q3		
		Month nr	Month name
2006	Q3	7	July

Figura 13-12: Agrupamento aplicada

Usando funções

O passo final na criação de um relatório agrupado consiste na adição de textos de rodapé e campos de resumo. A fim de mostrar resumos como o número total de meses ou trimestres em um grupo, você precisa criar funções. No caso de você não tenha notado, você já criou uma função para aplicar a linha de bandas formatação, mas o que é feito de uma forma quase transparente. PRD contém muito mais funções que podem ser usados para melhorar o seu relatório. A lista dos funções disponíveis pode ser acessado na guia Dados, clicando em Funções ícone de atalho na parte superior da guia de dados, clicando na linha Funções dentro da guia Dados, ou selecionando Adicionar função no menu Dados. A lista contém expressões de função para as conversões, layout, conteúdo e agregação. As funções de agregação são os necessários para a adição de grupo e resumos total a um relatório e são divididos em Global e executando As funções de agregação. A distinção entre estes dois tipos de funções é o contexto no qual são aplicados: Funções globais sempre calcular o agregação no grupo para o qual eles são definidos. Executar funções calcular a totalização até o ponto ou sub-grupo em que estão colocado. Para ilustrar isso, vamos criar quatro funções para contar o número de linhas no relatório em diferentes níveis:

1. Adicionar um item de duração função Count (selecione Adicionar Função, clique duas vezes
   Correndo, selecione o conde [de duração] Função, e clique em Adicionar). Funções obter um nome padrão que consiste de seu nome da função e um apenso número de seqüência para mudar o nome para RunningCountYear. Selecione YearGroup como o Reset nome do grupo para garantir que a contagem começa a partir do início de cada ano encontrado.

2. Adicione um nome de ponto de função Count Total, este TotalCountYearE selecione YearGroup como o nome do grupo (selecione Adicionar Função, clique duas vezes em Soma
   Maria, Count selecione e clique em Adicionar).

3. Agora soma mais de duração e Total Item Count função eo nome
   eles RunningCountReport e TotalCountReport mas deixar o grupo
   e nomes de grupo redefinir vazio para ter a agregação calculados para todos os
   os valores no relatório.

4. Abra o Quarter Group Footer e arraste as quatro funções recém-criado
   na área de rodapé. Observe que isso cria um campo com um número de referência
   para a função, que agora faz parte da lista de campos.

5. Adicione quatro rótulos, bem como para indicar que o valor é representado por cada
   de campo. Figura 13-13 mostra que o grupo de funções e do relatório
   olhar de rodapé, como após a aplicação de todas as etapas anteriores.

Figura 13-13: Agregação exemplo funções

ATENÇÃO As funções são referenciados pelo nome para se certificar de que todas as funções tem um nome único dentro do relatório.

Como a prova do pudim está no comer, é hora de acertar o
Preview botão novamente e saber como essas funções são avaliadas pelo PRD.
Figura 13-14 mostra parte dos resultados, com uma selecção de dois anos (2008 e
2009) e todos os quatro trimestres.

Year: 2009			
	Quarter: Q2		
2009	Q2	4	April
2009	Q2	5	May
2009	Q2	6	June
Running Year	*6*		
TotalYear	*12*		
Running Report	*18*		
TotalReport	*24*		
	Quarter: Q3		
		Month nr	Month name
2009	Q3	7	July
2009	Q3	8	August
2009	Q3	9	September
Running Year	*9*		
TotalYear	*12*		
Running Report	*21*		
TotalReport	*24*		

Figura 13-14: resultados da agregação

Este valor é um bom exemplo de como as funções de agrupamento e agregação
trabalho e deve dar-lhe fundo o suficiente para construir seus próprios relatórios.

O que é mostrado é que para o Q2 de 2009, a contagem corrida é item 6, o que é correto porque cada quarto é composto por 3 linhas de mês, eo total no ano
grupo é de 12. A duração contagem Relatório mostra 18 para 2009 Q2 porque todos os 2008 linhas, bem como dos atuais 6 de 2009, fazer 18. Finalmente, o relatório mostra a contagem total
24 em cada grupo, que é o número de meses a 2 anos.

Você pode ser tentado, neste ponto, tentar adicionar cálculos baseados em esses campos nova função, por exemplo, para calcular um percentual do valor da linha de uma
grupo total. PRD não funciona dessa maneira, mas fornece funções out-of-the-box para estes cálculos também. Como último exemplo, em um quarto o ano-mês relatório, você pode adicionar uma função ItemPercentage para calcular o número do mês percentagem do número total do grupo mês. Como referência, você pode criar este grupo no mês Número total e exibi-lo no rodapé do grupo Bairro.
A função ItemPercentage é parte das funções de agregação de execução (Selecione Adicionar função, abra duração, e selecione Porcentagem do Total). Após o função foi adicionada, o campo para criar o percentual para o grupo e para determinar o valor total deve ser selecionado. Também é uma boa idéia para mudar o nome da função em algo significativo, como MonthPercentage. A especificação da função completa é apresentada na Figura 13-15.

Name	Value
Required	
Function Name	MonthPercentage
Field Name	month_number
Optional	
Reset on Group Name	QUARTER
Rounding Mode	4
Scale	2
Scale Result To 100	False
Dependency Level	0

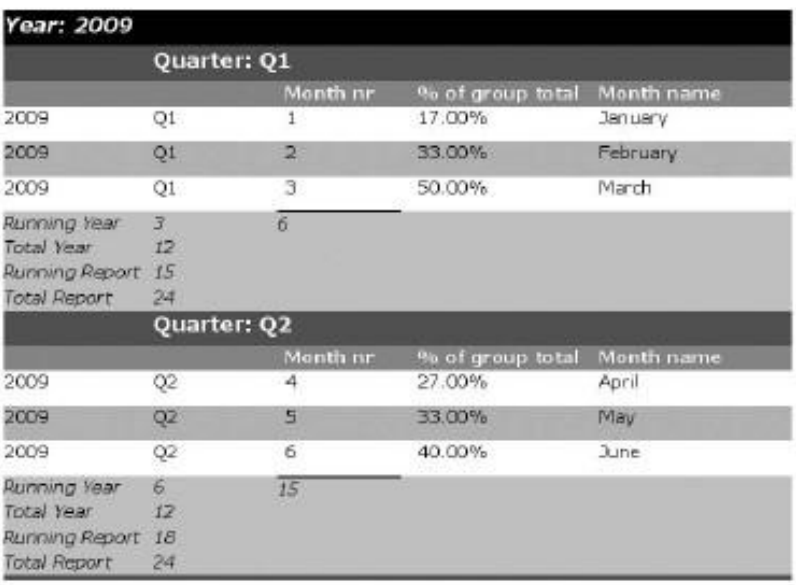
Figura 13-15: definição do percentual do item

Agora, a função pode ser adicionada aos detalhes no relatório. Para ter coluna exibida como uma porcentagem, altere o valor no formato do campo atributos em 00.0%. O resultado completo pode ser visto na Figura 13-16, embora este exemplo não é um cálculo muito útil no mundo real, ele faz um bom trabalho de explicar como o mecanismo de cálculo do PRD obras.

Usando fórmulas

Na seção anterior, você aprendeu como adicionar uma função ItemPercentage para calcular a porcentagem número do mês do total de número de grupo de mês. Se você já possui uma função que calcula o número total mês, você pode obter o mesmo resultado usando uma fórmula. Para fazer isso, você precisa criar um campo Número e especificar uma expressão que calcula a porcentagem item para o campo Fórmula propriedade. A sintaxe para fórmulas é o mesmo que

encontrados em programas de planilha como o Microsoft Excel e Open Office Calc: inserir um sinal de igual, seguido pela expressão. Para calcular a verdade ItemPercentage, é necessário dividir o número do mês pelo total de meses. Assim, na fórmula, a expressão que aparece atrás do sinal de igual que leia [Month_number] / [MonthTotal]. Note que essa [Month_number] e [MonthTotal] são eles próprios também expressões: [Month_number] referências um campo da consulta, e [MonthTotal] se refere a uma função Soma Total calcular a soma dos números do mês.



Year: 2009				
	Quarter: Q1			
		Month nr	% of group total	Month name
2009	Q1	1	17.00%	January
2009	Q1	2	33.00%	February
2009	Q1	3	50.00%	March
Running Year	3	6		
Total Year	12			
Running Report	15			
Total Report	24			
	Quarter: Q2			
		Month nr	% of group total	Month name
2009	Q2	4	27.00%	April
2009	Q2	5	33.00%	May
2009	Q2	6	40.00%	June
Running Year	6	15		
Total Year	12			
Running Report	18			
Total Report	24			

Figura 13-16: Número percentagens mês

Expressões em fórmulas Pentaho Reporting não estão limitados aos campos da conjunto de dados e referências a funções com nome, você também pode usar valores constantes,
comum operadores aritméticos, como + -, * E / e operadores lógicos como AND e OR. Além disso, há uma série de funções internas para tarefas como data do cálculo do tempo / e manipulação de cadeia. Cuidado, porém, que isso não é um programa de planilha: você só pode fazer referência a linha atual de dados, não o linha anterior ou seguinte. Você pode, no entanto, utilizar funções para determinar mínimos ou valores máximos dentro de um grupo, por exemplo, e fazer referência à função no nível de linha. As funções também podem fazer referência a resultados de outras funções, de modo
praticamente não há limites para o que você pode fazer com as funções, fórmulas e expressões.

NOTA Para uma visão completa da Pentaho Reporting Fórmulas, built-in funções e operadores, consulte a documentação do Pentaho em http://wiki.pentaho.com/display/Reporting/9.+Report+Designer + + Expressões Fórmula.

Adicionando gráficos e elementos gráficos

Uma imagem vale por mil palavras, e essa sabedoria comum é particularmente verdade no mundo de BI. Mostrando apenas os números muitas vezes não é suficiente quando você deseja obter uma visão imediata da evolução ou da distribuição dos dados. PRD, portanto, não somente fornecer uma interface para criar Pentaho Reporting, mas pode ser utilizadas para integrar JFreeChart também. Lembre-se que Pentaho Reporting e JFreeChart são dois projetos diferentes, que foram integrados em um única solução por Pentaho. Mais informações sobre JFreeChart pode ser encontrado em http://www.jfree.org/jfreechart.

Antes de um gráfico pode ser usado, os dados para ser visualizado no gráfico precisa estar presentes no relatório. PRD não é possível utilizar os campos de dados de uma consulta directa mas precisa de um provedor de dados especial, chamado função de cobrador que transforma o dados para uso em um gráfico. Existem seis funções de coletor e 14 tipos de gráfico disponíveis. Cada tipo de gráfico usa a sua função de coletor próprio, e alguns tipos de gráfico pode usar dois deles. Para um gráfico simples, basta ter uma coluna da série e um valor definido, enquanto que para um gráfico de barras empilhadas com um segundo eixo Y mais dados precisam ser passados para o gráfico. A vantagem do uso de coletores de dados que estão separados dos dados principal é definir a capacidade de criar gráficos diferentes a partir de um único conjunto de dados.

Captar o modo de gráficos obras e as opções de dados diferentes disponíveis Pode parecer difícil no começo, então vamos começar por um exemplo simples para ter uma noção de como as coisas funcionam. Este exemplo é baseado no banco de dados de exemplo Pentaho, e irá mostrar a receita ea quantidade por ano ordem e linha de produtos em várias maneiras.

NOTA Em uma situação da vida real, você deve iniciar com o design do relatório, que deve ser baseado em um requisito de negócio. Somente depois que você descreveu o que você quer alcançar com o novo relatório que você pode começar a pensar em que dados seria necessário e como o relatório deve ser construído. Os passos seguintes supor que os requisitos e fase de concepção já está concluído para que você possa agora começar a construir o relatório atual.

1. Depois de criar um novo relatório, use o PRD Designer de Consulta para adicionar um Fonte de dados com JDBC OrderYear,ProductLine,OrderRevenueE OrderQuantity. Para começar o ano a fim, é necessário o ano () função.

ATENÇÃO O PRD Query Designer permite que você adicione as expressões suma, min, max, e contagem mas podem ser alterados posteriormente. Tenha em atenção que quando você estiver usando funções não-agregado, como ano (), o campo deve ser adicionada ao grupo por cláusula. Quando o fizer, será acrescentado, incluindo o alias, o que não é correto. A ordenar por e pelo grupo cláusulas terão que ser ajustados manualmente para gerar os resultados corretos. Além disso, cuidado que cada vez que você entra na gráfica

Designer de Consulta, SQLeonardo irá gerar a sintaxe da consulta errado novamente traduzindo ano (OrderDate) em apenas ano.

2. A consulta correto para obter o conjunto de resultados que você precisa é exibido na seguinte bloco de código:

```
SELECT      ANO (o.orderdate) AS orderyear
,           p.productline
,           SUM (d.quantityordered)                    orderquantity AS
,           SUM (* d.priceeach d.quantityordered)      orderrevenue AS
DA          o ordersAS
INNER JOIN  OrderDetails AS d o.ordernumber ON         = D.ordernumber
INNER JOIN  productsAS p ON d.productcode              = P.productcode
GROUP BY    ano (o.orderdate)
,           p.productline
ORDER BY    ano (o.orderdate) ASC
,           ASC p.productline
```

3. A partir deste conjunto de resultados diferentes gráficos podem ser gerados em diferentes níveis de
detalhes. Os gráficos podem ser colocados no cabeçalho do relatório, mas você também pode usar
os grupos criados em um relatório. Ao fazer isso, você pode dar um alto nível Resumo na parte superior do relatório e prestação de informações detalhadas na repartições grupo. A idéia do projeto é o seguinte:

- Mostrar um gráfico de barras na parte superior do relatório com a quantidade total por linha de produtos discriminados por ano para destacar tendências de vendas para cada
linha de produtos ao longo do tempo.
- Para cada ano, apresentar a distribuição de renda ea quantidade por produto linha em percentagem e utilizando gráficos de pizza.
- Exibição do ano, a quantidade total ea receita total como cabeçalho de grupo.

4. Para fazer isso, criar primeiro um grupo baseado na OrderYear. Os gráficos podem ser adicionado por arrastar um objeto gráfico a partir da paleta na tela do projeto, neste caso para o cabeçalho do relatório. Para ter a propagação gráfico mais toda a largura da página, clique sobre o gráfico e definir o estilo a seguir atributos: x = 0,y = 0,width = 100%E height = 190.

NOTA Todos os objetos em um relatório pode ser dimensionada e posicionada em um absoluto ou um
forma relativa dentro de uma faixa relatório. O padrão no PRD está usando o posicionamento absoluto e dimensionar em pixels, por exemplo, quando uma carta é colocada em algum lugar em uma banda, o xe yvalores indicam a posição absoluta do canto superior esquerdo da gráfico e os largura e altura Os valores estão em pixels. Todos os valores podem ser alterados em valores relativos, adicionando o sinal de porcentagem (%).

5. Gráficos que são colocados sobre a tela pode ser editado clicando em o gráfico e escolha a opção editor gráfico. Isso abre a tela que é apresentado na Figura 13-17.

Figura 13-17: editor gráfico

A tela mostra os diferentes tipos de gráficos na parte superior, as propriedades do gráfico à esquerda e à definição da fonte de dados à direita. Esta parte da fonte de dados é vazio quando você cria um gráfico e você terá que selecionar uma das funções de coletor disponível para alimentar o gráfico. Cada tipo de gráfico tem uma ou duas disponíveis as funções de coletor, dependendo do tipo de gráfico. A fim de fazer as escolhas gráfico à direita, você vai precisar saber o que você pode fazer com os diversos funções de coletor:

- PieSet-Requer uma série (as fatias do bolo) e uma coluna de valor. Usado para gráficos de pizza e anel.
- CategorySet-Requer uma categoria (eixo X) e uma coluna de valor. Opcionalmente, uma série (avaria de uma categoria) podem ser adicionados. Usado para gráficos de barra, linha, torta multi-cachoeira e região.
- PivotCategorySet-Pivôs as colunas de valor para usá-los como categorias. Requer que você tem pelo menos duas colunas de valores (por exemplo, real e orçamento), que são traduzidas em categorias. Usado nos tipos de gráfico mesmo como CategorySets.
- TimeSeries-Requer uma coluna de data / hora para ser usado como uma categoria (Eixo X) e um valor. Uma série pode ser adicionada, opcionalmente. Usado em dispersão,
  XY bar, XY linha, área XY, e os gráficos de linhas prolongado XY.
- Permite-YYSeries duas medidas deve ser posicionado sobre o X e Y eixo a parcela relativa a dependência de duas variáveis. O valor da série é opcional. Pode ser usado nos gráficos mesmo que o coletor TimeSeries função.

- Somente XYZSeries para o gráfico de bolhas; precisa de três colunas de valores onde X e Y determinam a posição em relação ao eixo X e Y, e Z
  valor determina o tamanho das bolhas. Adicionando uma série permite um período de quatro
  visualização tridimensional.

Adicionando um gráfico de barras

O primeiro gráfico que deseja criar é baseado em uma CategorySet onde productline é usado como categoria, OrderYear como a série de colunas, e orderquantity como um valor da coluna. Observe o seguinte:

- Adicionando as colunas para a série de valores e requer um pouco mais clicando através do que abrir uma tela de seleção distintas. Parece que você poderia adicionar várias colunas para qualquer série ou um valor, mas quando o fizer, esta não tem qualquer efeito sobre o gráfico, ou ele simplesmente não vai exibir todos os valores
  em tudo.
- Certifique-se que o valor de uso de resumo somente é definido como False. O estilo atributos do lado esquerdo pode ser ajustado ao seu gosto.
- Quando um título tem de ser exibido, digite-pol Se você não quer um título, deixar o campo vazio. Não há título de exibição atributo como não existe para o lenda e os eixos X e Y.
- O resto das opções de formatação são bastante auto-explicativo, exceto por as fontes. As fontes podem ser ajustadas por meio de um texto em três partes divididas por traços onde a primeira parte consiste no nome da fonte, seguido do estilo e tamanho. A negrito, fonte Arial 12 pontos é codificado como Arial Bold-12.
- Ao usar a opção de visualização, você pode verificar se o gráfico partidas suas exigências.

Gráficos de pizza

Os gráficos de pizza que você deseja adicionar exigir um PieSetCollectorFunctionE porque você quer uma torta para mostrar a quantidade e as receitas do outro, ambos começam a sua função de coletor próprio. A coisa agradável sobre gráficos de pizza é que você pode mostrar rótulos para as fatias e indicar qual o conteúdo deve ser exibidos dentro do rótulo. Três valores podem ser exibidos: o texto real da etiqueta, o valor de corte, ea porcentagem da fatia. Eles podem ser acessados através da adição {0}, {1}, ou {2} em qualquer ordem ou combinação para o atributo formato de etiqueta onde pode adicionar o texto também. Se, por exemplo, você quer exibir o nome da fatia seguido pelo valor entre colchetes, basta digitar {0} ({1}).

Os gráficos de pizza no exemplo são colocados na Ano cabeçalho de grupo e exibir os valores para um determinado ano. O valor para o pelo grupo atributo

precisa ser definido para o YearGroup grupo, caso contrário, os gráficos não usará o agrupamento e todos irão mostrar os mesmos valores globais.

Com gráficos de pizza, você também pode explodir uma fatia (posicioná-lo fora do círculo por uma determinada porcentagem) ou exibir o gráfico em 3D. A combinação destas não é possível: não se pode explodir uma fatia em um gráfico de pizza 3D. Para selecionar a fatia de
explodir, você vai precisar conhecer o seu ID, que é determinada pela ordem de classificação na consulta. números de identificação começa com 0 então neste caso, a fatia Classic Cars ID 0 e será explodido. O montante explodir é indicado como uma percentagem
entre 0 e 1, para 0,1 dará um deslocamento de 10 por cento da fatia.

Você não pode usar o editor gráfico para o dimensionamento e gráficos de posicionamento, que é
parte do designer do relatório principal. Para colocar duas tabelas lado a lado e ter escala-los automaticamente, defina o valor X e Y para o gráfico da esquerda para 0 eo largura para 50 por cento. O gráfico à direita terá um valor de x e uma largura de 50 por cento. Combinados, esses dois quadros agora preencher a página da esquerda para a direita.

Para finalizar o projeto definido, há apenas uma coisa: o texto do cabeçalho do grupo ao ano, quantidade e receitas nele. Se você adicionar um campo de mensagem e deixar um pouco de espaço em branco no topo do cabeçalho de grupo, defina o campo a 100 por cento,
e mudar a cor de fundo para a mesma luz cinzenta como gráficos de pizza,
o cabeçalho do grupo será a aparência de um bloco único, que dá um muito profissional aparência do relatório. O campo mensagem é o valor:

```
Ano: $ quantidade (ORDERYEAR) Total $ (ORDERQUANTITY, ,#,###), número
  Total de receitas $ (ORDERREVENUE, ,$#,###.##) número
```

O relatório resultante é mostrado na Figura 13-18.

Trabalhando com imagens

PRD pode manipular dados de imagem de URLs, arquivos locais, ou colunas de banco de dados. Este
permite inserir logotipos de empresas, mas também para criar relatórios com o produto informações, incluindo imagens do produto, ou funcionário folhas, incluindo
uma foto de passaporte, se o que está disponível em um banco de dados corporativo. Para que o
seguindo o exemplo funcionar, você precisará de um logotipo em qualquer formato de arquivo de imagem (JPG,
PNG, GIF, BMP WMF ou SVG. TIFF provavelmente irá funcionar tão bem, mas isso não é garantido) para criar um cabeçalho do relatório, e um ou mais fotos em um banco de dados.
A tabela de funcionários no banco de dados contém uma amostra WCM passport_photo coluna do tipo LONGBLOB onde as imagens podem ser armazenadas. A tabela contém algumas
imagens de exemplo, mas você também pode substituí-los por conta própria ou usar um outro

Trabalhando com fotos

NOTA Na maioria dos bancos de dados modernos, um campo blob pode ser definida que pode armazenar
informação binária de um tipo arbitrário, incluindo fotos. No MySQL, você precisa

definir um campo do tipo BLOB (pequena, MÉDIO, ou LONGBLOB, dependendo do tamanho da os arquivos a serem armazenados). Obter uma imagem no banco de dados é fácil com o MySQL comando LOAD FILE que aponta para o arquivo a ser carregado. A instrução usada para fazer o upload de imagem Roland é:

```
UPDATE empregado
SETpassport_photo LOAD FILE = ('/ media / foto / photo_roland.jpg)
WHEREemployee_id = 22612
```

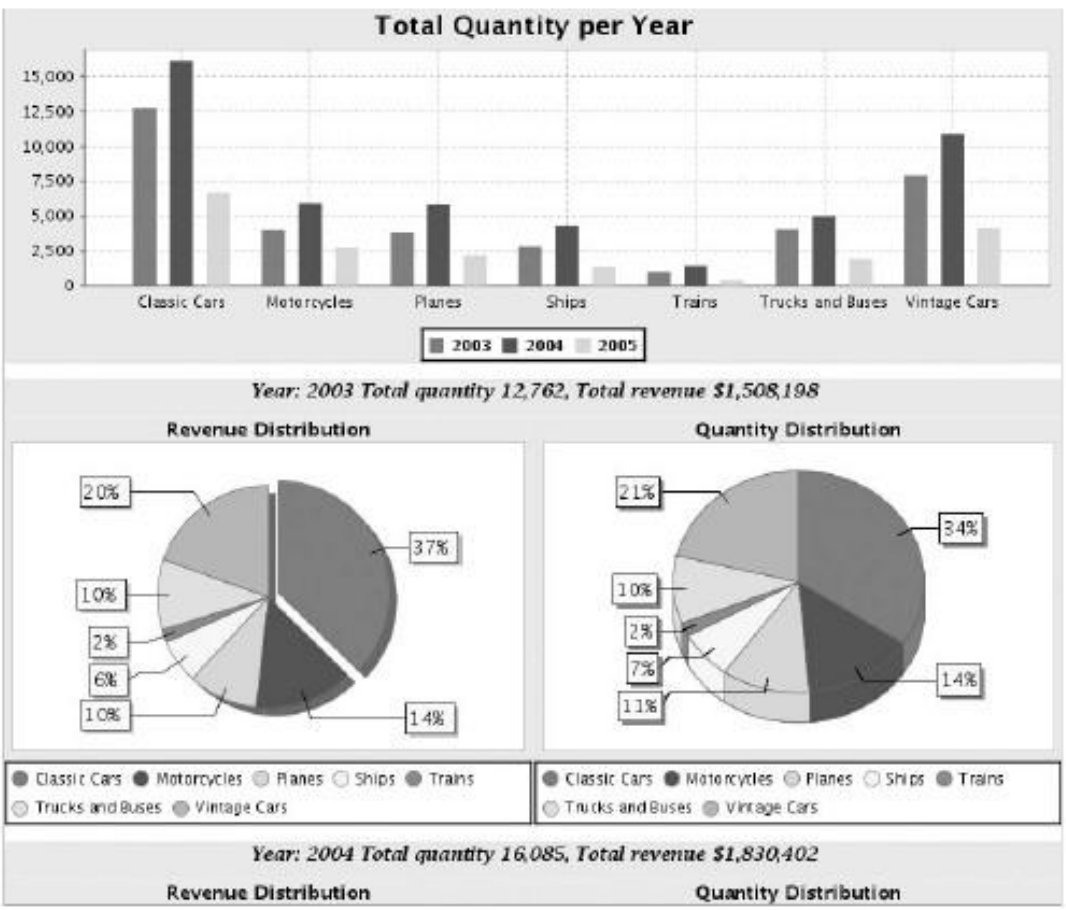


Figura 13-18: Exemplo de gráfico Relatório

Siga estes passos para criar o relatório mostrado na Figura 13-19:

1. Em PRD, iniciar um novo relatório e criar uma consulta para selecionar o trabalhador detalhes do banco de dados WCM, incluindo a coluna passport_photo.

2. Para exibir imagens de um arquivo ou URL, coloque um campo de imagem a partir do paleta no cabeçalho da página e adicionar um título assim.

3. Botão direito do mouse no campo da imagem e selecione Editar Conteúdo. A localização do arquivo
   O editor é aberto e você pode digitar uma URL ou procure um arquivo local. A
   de exemplo usa o logotipo WCM.
4. Adicionar um grupo Empregado e uso employee_id como campo de grupo. Isso permite que a organizar o relatório e para iniciar uma nova página para cada funcionário.

5. Agora você pode colocar todos os campos de detalhe de funcionário no cabeçalho do grupo, mas
   verifique se você usar um campo de conteúdo para a imagem.

ATENÇÃO PRD não reconhece o campo blob automaticamente assim que quando você acabou de arrastar o campo a partir do separador de dados para a tela torna-se um campo de texto, e não um domínio de conteúdo. Este é um bug conhecido (PRD-1394).

6. Para tanto o objeto de conteúdo e domínio de conteúdo, há dois atributos
   que provavelmente terá que ser definido: a escala e proporção de atributos. Ambos
   estão disponíveis quando você clica nos campos e também fazem parte da
   os atributos de estilo na vista de estrutura. Figura 13-19 mostra o resultado da
   Neste exercício imagem.

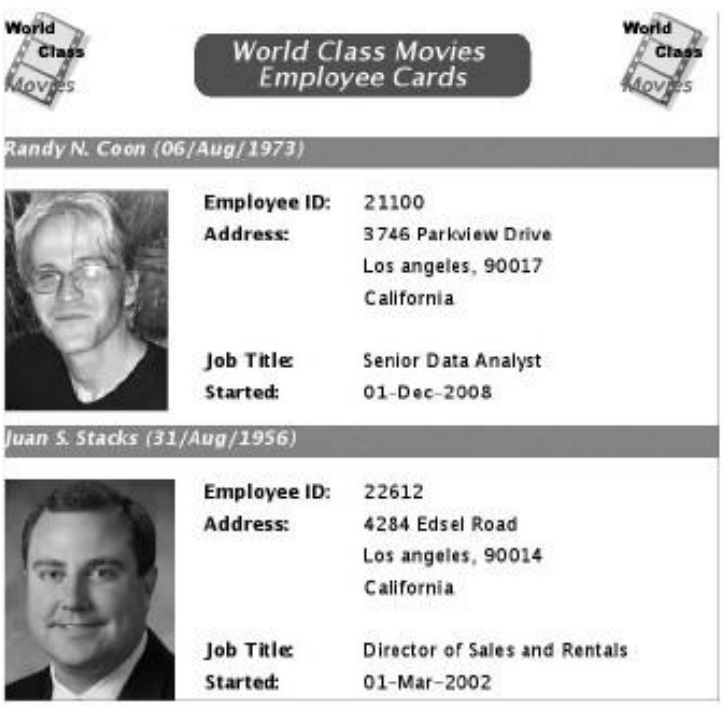


Figura 13-19: exemplos PRD Imagem

Aqui estão algumas dicas para criar estes relatórios:

- Use os campos de mensagem para os nomes, endereços ou qualquer outra informação que
  contém vários campos de um banco de dados. O valor para o campo de nome em Este exemplo é $ (First_name) $ (middle_initial) $ (last_name). Não há necessidade de espaços inserir explicitamente, eles serão captados automaticamente pelo PRD.
- Os campos de data devem ser formatados, caso contrário eles serão exibidos como plena valores de data e hora, incluindo o fuso horário. O formato da data está inscrita no atributo formato do campo data. Neste exemplo, o valor é dd-MMM-yyyy (Sem aspas).
- Os campos podem ser organizadas, alinhadas e distribuídas, utilizando o Organizar, Alinhar e Distribuir assistentes layout no menu Formatar. Você pode selecionar vários campos que precisam ser alinhados ou distribuídos uniformemente pela exploração
  pressionada a tecla Shift e clicar nos campos.

Trabalhando com sub-relatórios

Sub-relatórios pode adicionar uma dimensão extra a seu design do relatório e permitir exibir dados agrupados em um nível de agregação diferente ou a partir de dados diferentes fontes. Uma palavra de cautela, porém, antes sub-relatórios usando: um sub-relatório usa sua própria fonte de dados para que ele irá disparar uma segunda consulta ao banco de dados. Usando
vários sub-relatórios em um relatório mestre torna esta situação ainda pior. A melhor estratégia é criar seus relatórios usando como sub poucos, e, portanto, extra consultas, como possível. Com esta precaução em mente, vamos mostrar como usar como sub-relatórios que se destinam.

Um relatório do mestre PRD é utilizada para visualizar um único conjunto de dados. Em certo sentido, a
relatório inteiro age como uma única tabela. Isso significa que você não pode filtrar ou exibir apenas uma parte do conjunto de dados, que não por meio de tabelas ou grupos. Resultados de dois
quadros adjacentes, com uma estrutura semelhante, mas com conteúdo diferente, não é possível
sem o uso de sub-relatórios. Um rápido exemplo pode ilustrar isso. Suponha que você deseja exibir a parte superior e inferior a cinco clientes com base em suas receitas a lado. Isso exigiria colocar duas listas em um único relatório, que não é possível, PRD (além de usar SQL complexa para gerar as duas listas em o conjunto de dados). Com sub-relatórios, é muito fácil de fazer isso: o mestre relatório contém um Top 5 e um fundo de cinco sub-relatório, cada um com sua própria consulta, cabeçalhos e detalhes. Estes sub-relatórios podem tanto ser colocado no relatório cabeçalho, como mostrado na Figura 13-20.

Para construir este relatório, nós criamos um relatório novo mestre vazio, sem dados ou o layout e adicionou dois sub-relatórios no cabeçalho do relatório. Quando você arrasta um sub-relatório para a tela e clique duplo-lo, uma nova guia com um relatório

tela se abre. Um sub-relatório é semelhante a um relatório do mestre. Ele pode conter um conjunto de dados
fonte, diferentes grupos, bandas e gráficos. Há uma diferença notável,
no entanto: não é possível criar novos parâmetros para um sub-relatório.

Top 5		Bottom 5	
Name	Revenue	Name	Revenue
Euro+ Shopping ..	908,165	Boards & Toys Co.	9,080
Mini Gifts Distributors ..	651,878	Atelier graphique	24,033
Australian Collectors, ..	200,133	Auto-Moto Classics ..	26,309
Muscle Machine Inc	196,952	Frau da Collezione	28,780
La Rochelle Gifts	179,129	Microscale Inc.	32,946
Total:	2,136,257	Total:	121,148

Figura 13-20: Top 5 e inferior 5 clientes

PRD tem dois tipos de sub-relatórios: em faixas e inline. Você deve escolher que tipo você quer que você arrastar o objeto sub-relatório da paleta e colocá-lo na tela. Um faixas subrelatório toda a largura da banda onde o sub-relatório é colocado, enquanto um sub-relatório em linha pode ser colocada em qualquer lugar. Tenha cuidado ao usar sub-inline porque exigem mais memória e poder de processamento do que suas contrapartes em faixas. Outra coisa estar ciente de é a faixa em que o sub-relatório é colocado. Se o sub-relatório é colocado dentro do cabeçalho ou rodapé do relatório, ele será executado apenas uma vez. Quando
colocado dentro de um cabeçalho ou rodapé do grupo, ele será executado tantas vezes quanto
há valores do grupo, e se você posicionar um sub-relatório em detalhes, será executado para cada linha do relatório. Não é possível colocar um sub-relatório em o cabeçalho ou rodapé de página.

Passando valores de parâmetros para sub-relatórios

Quando um parâmetro é definido no relatório principal, que deve ser dada uma única ID pelo qual o parâmetro pode ser referenciado. O sub-relatório pode fazer referência o parâmetro de relatório mestre por meio de importação, mas sem novos parâmetros podem ser
definido em um sub-relatório. Você pode abrir a tela de importação de parâmetros por botão direito do mouse no ícone do parâmetro na guia Dados do sub-relatório ou clicar na Sub-relatório de parâmetro ícone na guia Dados. A tela permite que você Mapa parâmetros mestre relatório ao equivalentes sub-relatório. Você pode usar o mesmo nome para o comandante eo sub-relatório, mas nós preferimos usar um diferente nome para a clareza. Como exemplo, você pode expandir o superior / inferior e cinco relatório da seção anterior para incluir uma linha de produtos de seleção que você vai passar para o sub-relatório.

1. Primeiro, crie ambos os parâmetros no relatório mestre antes de abrir a sub-relatório.

2. Em seguida, no sub-relatório, abra o Sub-relatório editor de parâmetro e insira os nomes exteriores e interiores dos parâmetros. Os nomes exterior referem-se a nomes do relatório pai de parâmetro, os nomes interior será utilizado no sub-consulta. Figura 13-21 mostra a tela do editor de parâmetro com os valores inseridos.

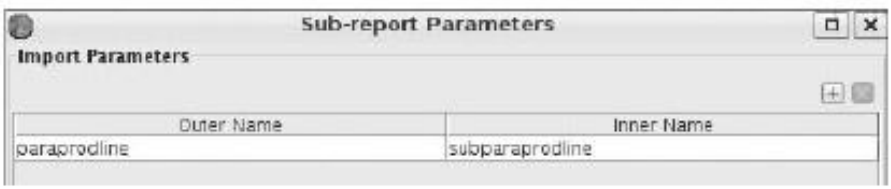


Figura 13-21: parâmetros de importação

3. A consulta para a lista Top 5 já pode ser estendido com as necessárias onde cláusulas e entrada de parâmetro, como mostra o seguinte código:

```
SELECT customers.customername
,       SUM (orderdetails.priceeach * orderdetails.quantityordered)
        como receitas
FROMorders
INNER JOIN OrderDetails
ONorders.ordernumber orderdetails.ordernumber =
INNER JOIN clientes
ONorders.customernumber customers.customernumber =
INNER JOIN produtos
ONorderdetails.productcode products.productcode =
WHEREproducts.productline = $ {} subparaprodline
GRUPO BYcustomers.customername
ORDEM BY2 DESC
LIMIT5
```

4. Agora você também pode incluir um campo de mensagem em seu relatório mestre exibir a linha de produtos selecionados com a expressão Selecionados Valor: $ (paraprodline). O resultado final deste exercício é apresentado no Figura 13-22.

Publicando e Exportando relatórios

Você pode publicar relatórios para o Pentaho BI Server a partir do menu principal pelo PRD Publicar escolhendo Arquivo, ou clicando no ícone Publicar situado à direita do no botão Salvar na barra de atalho. Isso abre a caixa de diálogo Publicar mostrado na Figura 13-23.

Para publicar, você deve configurar a configuração de publicação. A Publicar localização deve ser o nome de um diretório existente que residem sob o pentaho soluções diretório. A URL Publicar Web deve ser apontado para o seu

Servidor Pentaho BI. Para obter a senha de publicação, você deve usar a senha que foi definido no publisher_config.xml arquivo. Configurar essa senha é abordada no capítulo 3. Finalmente, você deve usar o ID de usuário e senha de um usuário que tem a função de Administrador ("Joe"e""senha de um padrão instalação).

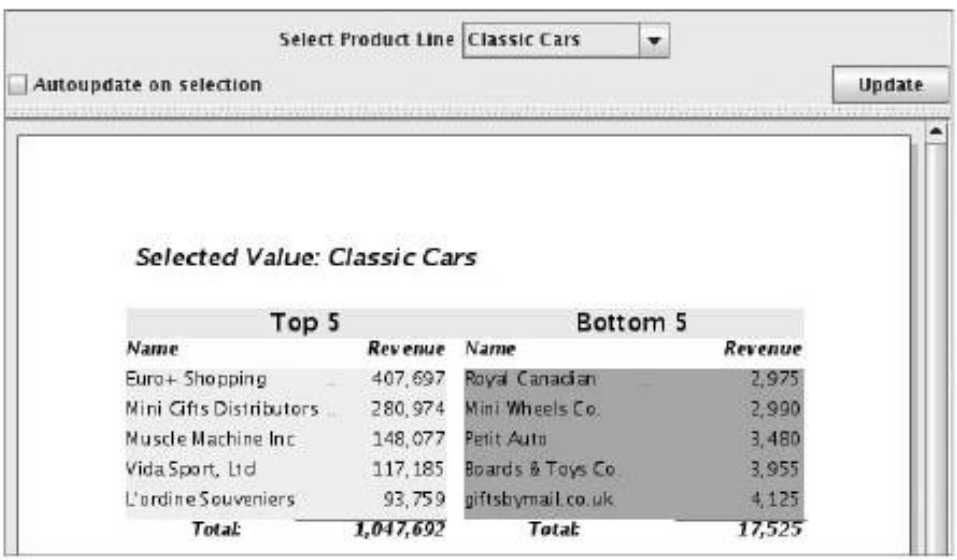


Figura 13-22: sub-relatórios parametrizados

Figura 13-23: A caixa de diálogo Publish

Atualizando os Metadados

Após a publicação do relatório para o servidor, você deve dizer ao servidor para recarregar os metadados para garantir que o usuário será presenteado com a versão mais recente do relatório, quando ele abre. Isso pode ser feito a partir do console do usuário através do menu, escolha Ferramentas Atualizar Cache Repositório de Metadados.

Alternativamente, você pode atualizar o cache usando o Servidor de Administração Console. Para atualizar o cache do Servidor de Administração console, vá para Na guia Serviços e pressione o botão Atualizar no painel Repositório Solution.

Exportando relatórios

relatórios Pentaho PRD pode ser exportada no mesmo formato da WAQR aplicação, com a opção extra de exportar o relatório no formato RTF. Antes exportar o relatório, pode ser visualizado em formato de exportação desejado selecionar o formato de saída (HTML, PDF, XLS, RTF, CSV) a partir do arquivo Preview opção do menu. O ícone de atalho visualização tem uma opção adicional chamado Preview como texto.

Resumo

Este capítulo apresenta a arquitetura básica de uma solução de relatórios e descreveram dois principais ferramentas para criação de relatórios em uma solução Pentaho, a
Web-based Ad Hoc ferramenta de relatório (WAQR) e os novos Pentaho Report Designer (PRD). Após oferecer um breve panorama da WAQR começamos relatórios prédio usando PRD e explicou os seguintes tópicos:

- Diferentes componentes do Report Designer Pentaho
- A estrutura de um relatório Pentaho e os diferentes elementos que podem ser usado dentro de um relatório
- Criando JDBC e Metadados consultas usando o Designer de Consulta
- Adicionando parâmetros para dar aos usuários a flexibilidade de escolher apenas um subconjunto de
os dados de um banco de dados
- As opções de layout e formatação disponíveis no PRD
- Como agrupar e resumir dados em um relatório
- Utilização das funções e expressões
- Como visualizar informações usando gráficos e diagramas
- Usando imagens de URLs, arquivos e tabelas do banco de dados
- Trabalhando com sub-relatórios para criar layouts de relatório complexo, constituído de vários conjuntos de dados
- Como criar parâmetros que podem ser passados para o conjunto de dados de um sub-relatório
- Como publicar os relatórios para o servidor Pentaho e certifique-se que o última versão é exibida para um usuário

Há muito mais a comunicação eo Pentaho Report Designer do que nós
poderia abranger neste capítulo. Pretende-se como uma base sólida para que você obtenha
começou, não como uma referência completa ou guia para o desenvolvimento.
Recomendamos que você
para visitar o Wiki e os fóruns on-line se você precisar de mais informações técnicas
sobre o produto.

CAPÍTULO

14

Agendamento, assinatura e de ruptura

Neste capítulo, vamos olhar para algumas das capacidades da plataforma Pentaho para distribuir conteúdo para o usuário final. Basicamente, existem três maneiras diferentes para usuários para executar sequências de ação:

- Imediata execução ou interativo-In Neste caso, a ação é executada imediatamente após a solicitações de usuários, eo usuário espera para a entrega dos de saída da ação, que marca o fim do pedido do usuário.

- Execução em segundo plano-Upon a solicitação do usuário, a ação é implicitamente, prevista para ser processada o mais rapidamente possível, e em vez de aguardando o resultado, a solicitação do usuário termina aqui. A execução efectiva da ação procede de forma assíncrona até entregar sua produção, que é então armazenado para que o usuário pode buscá-la em um momento posterior.

- Explícita-agendamento Este é semelhante à execução em segundo plano, mas em vez de programar a ação a ser realizada imediatamente, ele é executado de acordo com um cronograma pré-definido.

Você tem visto uma série de exemplos de execução imediata, em anteriores capítulos. Neste capítulo, você explora programação e como você pode usá-lo para organizar a entrega de conteúdo.

Agendamento

Pentaho fornece agendamento de serviços por intermédio da Empresa Quartz Job Scheduler, que é parte do projeto Open Symphony. O programador é um componente que permite que o servidor de BI para realizar tarefas em um tempo programado ou definido
intervalo de tempo.

NOTA Para mais informações sobre quartzo eo projeto Open Symphony, visite o
Abrir site no Symphony www.opensymphony.com/quartz/.

Há um número de casos de uso para o programador:

- Periódico de execução das tarefas de manutenção
- Execução de tarefas de ETL (como atualizar as tabelas de agregados)
- Executando tarefas demoradas, tais como relatórios de grande no fundo
- Distribuir a carga de trabalho do servidor ao longo do tempo
- Preparação de conteúdo para assinantes

Conceitos do Scheduler

Dois itens estão envolvidos na programação:

- Acronograma, que é uma regra ou conjunto de regras que especificam um momento ou série de momentos no tempo
- Um seqüência de ações (Ou grupo deles) que é executado no tempo ou vezes especificado no cronograma

Discutimos o conceito de seqüências de ação no capítulo 4, e você tem vi alguns exemplos deles em outros capítulos deste livro. No restante desta seção, descrevemos horários em detalhes.

Público e Agendas Privada

Existem dois tipos distintos de modelos: público e privado.
agendas públicas são visíveis para todos os usuários. Eles são principalmente usados para implementar assinatura.
horários Privada estão disponíveis para o administrador do servidor. São principalmente usados para implementar as tarefas de manutenção. Por exemplo, na página da guia Serviços da o Pentaho Server Administration Console (PAC), há um botão Agendar no painel de repositório de conteúdo. Ao pressionar este botão irá criar um grupo privado cronograma que será executado diariamente para limpar o repositório de conteúdo. Discutimos o repositório de conteúdo na próxima seção, mas o importante é que este é um exemplo de uma tarefa de manutenção programada que é governada por uma empresa privada cronograma.

Repositório de conteúdo

Normalmente, quando um usuário executa uma seqüência de ação, o resultado da ação seqüência é imediatamente entregue ao usuário final. Programado seqüências de ação funcionam de forma diferente. A produção de seqüências de ação programada é gerado no

fundo. A saída é guardado e preservado no que é chamado de conteúdo repositório, para que o usuário pode inspecionar o resultado mais tarde.

Fisicamente, o repositório de conteúdo é simplesmente um diretório no sistema de arquivos onde a saída seqüência de ação são armazenados em arquivos. Você pode encontrar a saída arquivos no conteúdo Diretório do sistema Pentaho solução.

A organização do repositório de conteúdo é semelhante à organização de a solução de repositório. Para cada solução Pentaho, há um conteúdo separado diretório do repositório. A estrutura da árvore embaixo de cada repositório de conteúdo espelha a estrutura de diretório de sua solução de correspondente no repositório, mas em vez de seqüências de ação, há um diretório partilha o nome do seqüência de ação. Nesses diretórios você pode encontrar um diretório assinaturas, que contém a saída a seqüência de ação para cada vez que ele correu de acordo com um cronograma.

Além dos diretórios para as soluções Pentaho, a raiz do conteúdo repositório também contém um diretório chamado fundo. Este diretório é usado para armazenar a produção de seqüências de ação que executam em segundo plano.

Na seção"Espaço do Usuário"mais adiante neste capítulo, você aprenderá como o usuário final pode acessar o conteúdo do repositório de conteúdo.

Criação e manutenção de agendas com o Pentaho Console de Administração

Você pode criar e manter programas de uso da Administração Pentaho Console para trabalhar com horários, aponte seu navegador para a casa do PAC página. Assumindo que você está executando na máquina local, você pode encontrar este em http://localhost:8099/. Ative a página de administração e clique no Agendador de guia. Você pode ver uma lista de todas as agendas públicas e privadas mantidas pelo servidor de BI Pentaho. Isso é mostrado na Figura 14-1.

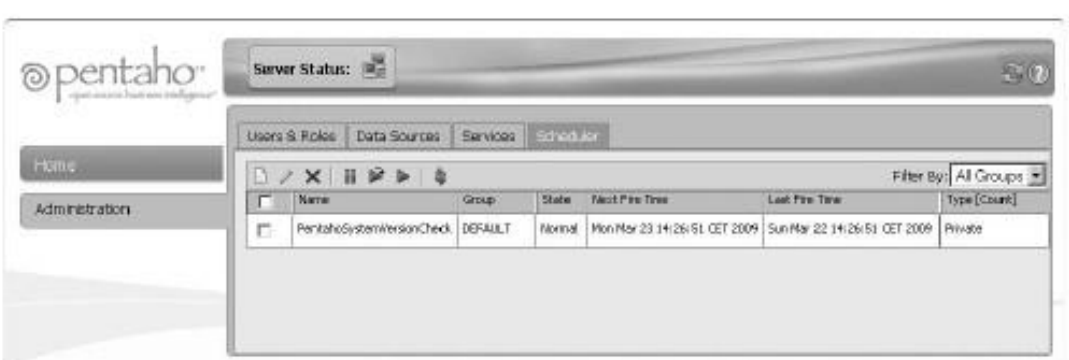


Figura 14-1: A ficha de Programador na Administração Pentaho Console

Por padrão, há uma programação chamada PentahoSystemVersionCheck. Este é um programa privado que é usado para verificar periodicamente se há uma nova versão do Pentaho está disponível.

Criar uma Nova Agenda

Clique no primeiro botão da barra de ferramentas para criar um novo agendamento. O Criador Agenda
diálogo aparece como mostrado na Figura 14-2.

Figura 14-2: A Agenda de diálogo Criador

Se você quiser criar uma agenda pública, selecione a opção Agenda Pública. Isso garante que todos os usuários podem ver a programação e que lhes permite usar esse cronograma de inscrições.

Use o campo Nome para introduzir um nome único para a agenda. Tenha em atenção que para agendas públicas, este nome será apresentado para os usuários finais, sempre que tem que escolher um horário. Use nomes que são claras e concisas.

Use o campo Grupo para especificar um nome que descreve o tipo de sched-ule é isso. Por exemplo, você pode criar grupos de acordo com o departamento (Site do armazém), localização (Edmonton, Los Angeles), área temática (Sales, Marketing), ou no tempo (diário, semanal).

A barra de ferramentas na página Scheduler (veja a Figura 14-1) inclui uma caixa de listagem que permite que o administrador do filtro para todas as programações pertencentes à mesma grupo. Esta característica de conveniência permite o administrador a trabalhar rapidamente com todos os horários pertencentes ao mesmo grupo.

NOTA Mesmo se você achar que não precisa organizar programações em grupos, que são
Continua a ser necessário digitar algo no campo Grupo. O mesmo vale para o
Descrição campo, embora seja inteiramente descritiva, ainda é necessária.

O campo Descrição permite inserir um texto descritivo. Você deve usar este campo para breve documento a finalidade da programação. Ao invés de somar até as propriedades que têm a ver diretamente com a programação em si (como tempo), este campo deve descrever os tipos de relatórios que são executados de acordo a esta programação e da audiência para que a lista se destina.

A caixa de listagem Recorrência contém uma série de opções para especificar quantas vezes
a programação é acionado. Você pode escolher entre uma ampla gama de intervalos, todo o caminho de segundos até anual. Há também uma opção Executar uma vez para agendar ações one-shot. Se a flexibilidade oferecida por essas opções ainda não é suficiente, você pode escolher a opção Cron para especificar recorrência na forma de umcron string.

Usando expressões CRON

O cron é um agendador de tarefas bem conhecidas para sistemas UNIX-like. Para cron, recorrer-ocorrência é especificado na forma de uma seqüência que denota seis ou sete campos de um valor data / hora, separados por espaços em branco. O planejador de quartzo usados por Pen-Taho oferece suporte para cron expressões.

Em cron expressões, da esquerda para a direita os campos de data e hora são: segundos, minutos, horas, dias do mês, mês, dia da semana, ano. Os campos podem con-manter valores inteiros simples, mas também é possível especificar listas de valores, intervalos e curingas, como todos, por último, e muito mais. Por exemplo, o seguinte cron expressão denota um retorno às 01h00, cada segunda terça-feira de cada mês:

0 0 1? * 3 # 2 *

(Em zero segundos, zero minutos, uma hora, independentemente do dia do mês, cada mês, a segunda ocorrência do dia da semana em terceiro lugar, para cada ano).

Você pode encontrar mais informações sobre cron expressões em quartzo na http:// quartzo www.opensymphony.com / wikidocs / CronTriggers Tutorial.html

Para todas as opções disponíveis da opção de Recorrência (exceto Cron) você pode especificar um tempo Comece escolhendo o valor apropriado na hora, minuto, e segundo caixas de listagem.

Para todas as opções de recorrência, exceto executado uma vez e Cron, você pode especificar como
muitas vezes o cronograma deve ser acionado. O Widget de especificar as valor aparece na tela assim que você selecione uma opção na lista de recorrência caixa. Por exemplo, para um retorno de segundos, minutos, horas e dias em que você pode inserir um número para indicar quantos segundos, minutos, e assim por diante entre as execuções subseqüentes do cronograma.

O semanário e opções de recorrência mensal apoio mais avançado possibilidades para especificar o intervalo. Por exemplo, a mensal e diário opções de recorrência aparecer um widget que permite que você defina o agendamento para

cada segunda-feira de cada mês, ea opção de recorrência semanal permite que você para especificar a quais dias da semana o horário se aplica.

Finalmente, você pode especificar uma data de início ou um intervalo de datas de início e fim. Estas datas
determinar o período de tempo em que o cronograma se repete. Depois de especificar a programação, clique em OK para salvá-lo e fechar o diálogo.

Correndo Horários

Apesar de horários são, normalmente provocada pela passagem do tempo, você pode também executá-los manualmente a partir do Console de Administração Pentaho. Esta é úteis para fins de teste, e também para se recuperar de erros.

Para executar manualmente as seqüências de ação ligada a uma agenda, selecione o caixa que aparece logo antes de cada programação. Em seguida, clique no botão Play (O triângulo verde) na barra de ferramentas.

Suspensão e retomada de Horários

Às vezes é útil para evitar temporariamente que um cronograma de execução. Para exemplo, você pode, temporariamente, a necessidade de recursos do sistema para fazer algumas outras
tarefa, ou você pode precisar para implantar várias seqüências de ação novo ou modificado e você deseja evitar que o cronograma de ser acionado enquanto estiver em
no meio de uma implantação.

Para esses efeitos, você pode suspender temporariamente uma agenda. Suspendendo o cronograma evita as acções conexas de ser executado. Uma suspensão
agenda permanecerá suspensa até que você configurá-lo para continuar. Ambas as ações pode ser feito de dentro do PAC.

Para suspender um ou mais modelos, selecione-os respectivos
na página da guia Horários no servidor de Administração Console. Depois de fazer a seleção desejada, pressione o botão Pausa na barra de ferramentas. Isto irá suspender as programações por tempo indeterminado, eo cronograma de Estado vai mostrar como Suspensa. Isso é mostrado na Figura 14-3.

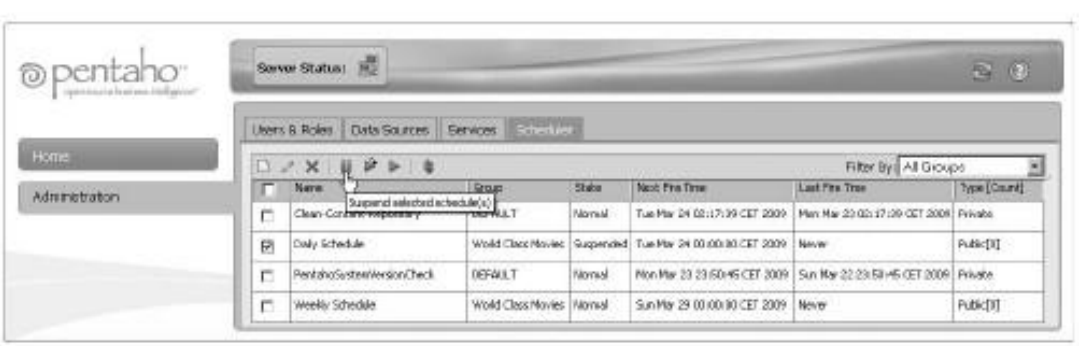


Figura 14-3: Suspendendo um cronograma

Note-se que o Estado coluna na tabela de horários tem o valor Suspensa, Considerando que todos os outros modelos têm o valor Normal.

Para retomar a programação suspensa, selecione o calendário que deseja retomar e pressione o botão Continuar (o triângulo verde com a seta).

Excluindo agendas

Você pode apagar horários marcando a caixa associados e, em seguida, pressionando o botão com o X vermelho""na barra de ferramentas. Antes de excluir um público programação, é uma boa idéia para verificar primeiro se o cronograma está sendo usado pelos assinantes. Você pode ver isso na visão calendário apresentado pelo PAC, porque a coluna [Contagem] Tipo mostra se o calendário é público, bem como o número de assinantes.

Você é sempre solicitado para confirmar a remoção de um cronograma. Figura 14-4 mostra como uma tentativa de remover o programa"minutos"usando o Administração Console dispara um pedido de confirmação, e também mostra que há ainda é um assinante.

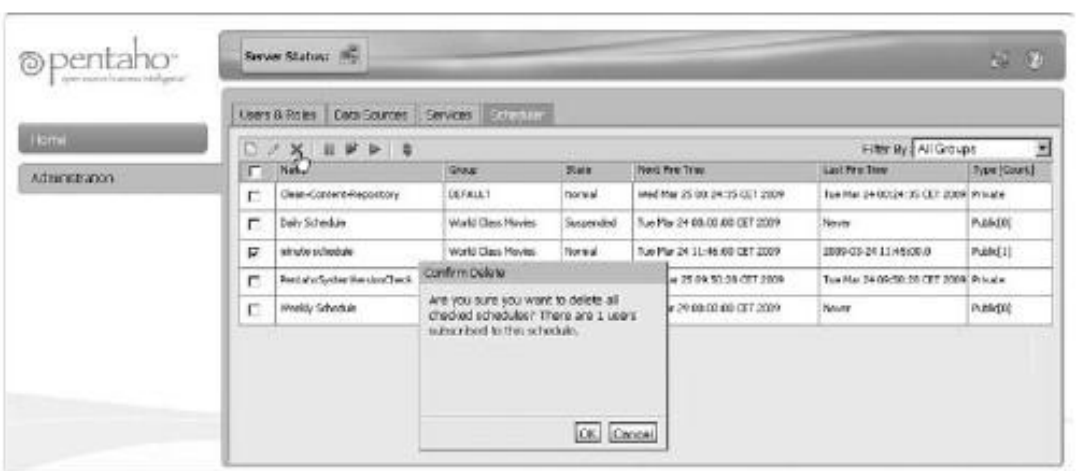


Figura 14-4: Excluindo uma programação dispara um pedido de confirmação

Programação do Programador com seqüências de ação

Além da interface gráfica oferecida pela Administração Console, você também pode trabalhar com o Agendador de uma maneira mais programática através de sequências de ação.

sequências de ação oferecem uma série de ações processo que permite que você trabalhe com o Scheduler. Estes podem ser encontrados no submenu Scheduler do Processo menu Acção. O menu Scheduler é mostrado na Figura 14-5.

As ações processo Scheduler pode ser usado para criar uma alternativa para o Programador de interface oferecida pelo console de administração. Um uso válido caso para reconstruir esse tipo de funcionalidade é que ela permitiria que os usuários finais de exercer mais controle sobre o conteúdo programado dentro de autorização do servidor quadro.

Figura 14-5: O Agendador menu ação do processo

As seguintes seções descrevem resumidamente as ações processo Scheduler.

Adicionar tarefa

A Adicionar tarefa acção processo cria um novo agendamento privada e associados com uma seqüência de ação. A Adicionar tarefa ação do processo tem o seguinte parâmetros:

- Nome do trabalho-O nome do trabalho de Quartz. Isto é aproximadamente equivalente ao
  o nome do programa em Pentaho. (A plataforma Pentaho adiciona alguns abstração para o programador de quartzo. Uma programação Pentaho é realmente um combinação de um trabalho de quartzo e um gatilho de quartzo).
- XAction Arquivo-A nome do arquivo (incluindo o . Xaction extensão) da seqüência de ação que deve ser associado a esse cronograma.
- Solução Nome-O nome da solução Pentaho que contém o seqüência de ação que está associado com o cronograma.
- Solução Subdireção-A subdiretório (se houver), onde a ação arquivo de seqüência reside.
- Trigger Nome-O nome do gatilho de quartzo. Em Quartz, o gatilho contém todas as informações de tempo que pode ser usado para determinar quando um trabalho deve ser executado. Um disparo de quartzo refere-se a um trabalho de quartzo (mas um quartzo
  trabalho pode ser desencadeada por vários gatilhos). Em Pentaho, a distinção entre gatilho e trabalho é oculto. Em vez disso, os associados Pentaho um emprego com um cronograma e se refere a isso como um agendamento. Também veja a descrição
  do parâmetro Nome do trabalho.
- Trigger Type-Este é um botão para escolher entre um trigger simples ou um cron gatilho. Trigger para disparar cron gatilhos é especificado através de um cron string. Para simples gatilhos de disparo é determinado por um valor de intervalo, que determina em que momento no futuro, o gatilho é acionado e uma repetição

contagem. Simples gatilhos não suportam reincidência, e são assim úteis apenas para"one-shot"eventos, tais como a execução do fundo.

- Cron String-Aplica para cron gatilhos. Especifica o cron seqüência de caracteres que rege a execução do trabalho.
- Trigger Interval-Aplica a simples gatilhos. Isso especifica a quantidade de tempo antes que o gatilho será disparado, especificado em microsegundos. Para exemplo, um minuto a partir de agora seria especificado como 60.000 (60 segundos vezes 1000), e uma hora a partir de agora seria especificado como 3,6 milhões (60 minutos por 60 segundos vezes 1.000 microsegundos).
- Trigger Count-Aplica a simples gatilhos. Em tempo de disparar tiros, o trabalho serão executados repetidamente de acordo com esse valor. Na maioria dos casos, este valor deve ser 1 (um) para executar o trabalho apenas quando o intervalo já passado.

Figura 14-6 mostra o design de uma seqüência de ações simples que solicita parâmetros de trabalho usando uma Prompt / Secure Filtro ação do processo e, em seguida, alimenta
estes parâmetros em um Adicionar tarefa processo de ação.

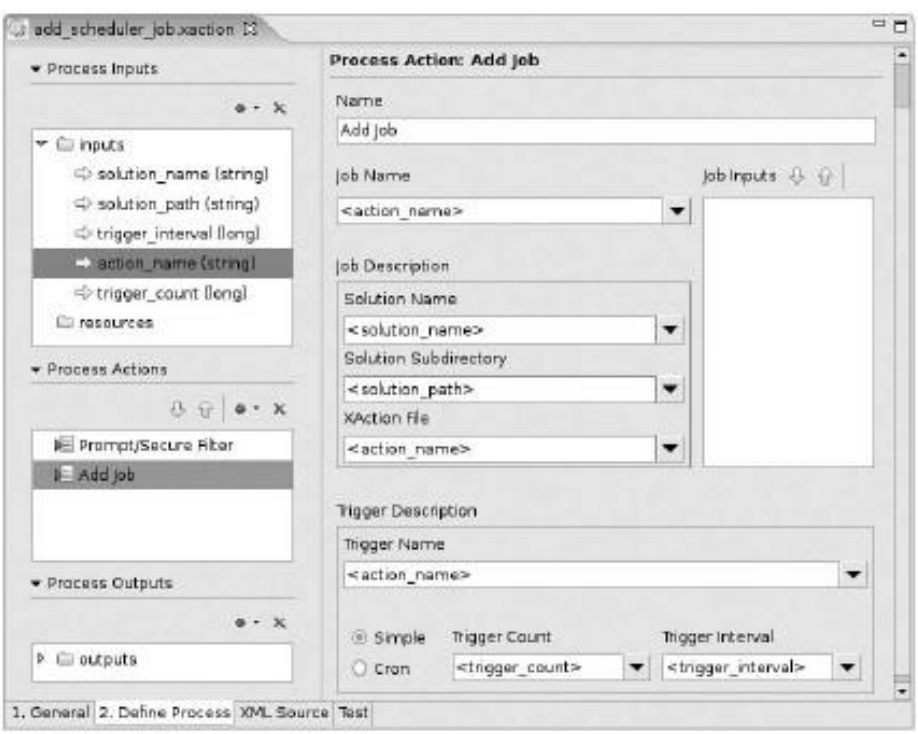


Figura 14-6: Um agendador de tarefas muito simples que é acessível através de uma seqüência de ação

Esta seqüência de ação simples faz uma prova de conceito para a construção de uma schedul-
baraço que podem ser acessíveis aos usuários finais dentro da autorização quadro do Servidor Pentaho.

Trabalho suspender, reiniciar Trabalho, Emprego e Excluir

A Suspender Trabalho,Currículo de EmpregoE Excluir Trabalho ações processo de aceitar apenas um Nome do trabalho parâmetro que deve ser usado para especificar o nome de um Pentaho cronograma. (Para uma discussão a respeito das programações Pentaho e quartzo Jobs, consulte a seção anterior sobre a Adicionar tarefa processo de ação). A ações, em seguida, executar a ação implícita do seu nome para o Schedule. A Suspender Trabalho ação irá interromper o processo especificado Anexo, o Currículo de Emprego ação do processo vai continuar, e os Excluir Trabalho ação será permanentemente remover o Schedule (e todas as suas ações associadas).

Ações Outros Processo Scheduler

Além das ações processo que acabamos de discutir, há um pouco mais Agendador de ações relacionadas com processo:

- Suspender Scheduler -Pausa a execução de todo o Agendador de indeterminar o tempo
- Resume Scheduler -Termina o Agendador de pausa
- Status Scheduler Indica-se o Agendador de suspensão ou em execução.
- Empregos lista agendada -Lista todos os horários privada

Essas etapas não são muito úteis quando comparado com as funções equivalentes-dade oferecida pelo console de administração. Os passos que o relatório da Programador status são bastante limitadas, e se você acha que precisa deles para construir um interface do usuário personalizada de programação, você é provavelmente melhor fora diretamente consultando o banco de dados de quartzo.

Programador Alternativas

O Agendador incorporado na plataforma Pentaho ternos muitos fins de agendamento. No entanto, ainda pode haver vantagens para agendamento de tarefas usando as ferramentas que são independentes do servidor de BI Pentaho. Aqui estão algumas razões que podem se por:

- Estas tarefas ETL- pode ser de longa duração, e pode colocar carga sobre o grave servidor. A plataforma, o programador, e todas as suas tarefas agendadas são todos executados na mesma instância do Java Virtual Machine, e para o desempenho e as razões de disponibilidade você pode querer execução separada de tarefas ETL do restante do servidor de BI.
- Política-Para razões de manutenção e sustentabilidade, a empresa pode padronizaram em uma ferramenta de programação particular, e isso pode impedir de usar Pentaho's built-in Agendador.

Se você só precisa agendar recursos para estes tipos de efeitos e você
não requerem nenhuma das funcionalidades Pentaho que constroem em cima da vegetação nativa
Scheduler, você deve pensar em ir com uma solução mais leve.

Em seguida, discutir brevemente as soluções de programação mais comuns de trabalho para sistemas baseados em UNIX e Windows.

Sistemas baseados em Unix: Cron

sistemas baseados em UNIX geralmente oferecem algum tipo de cron implementação. Normalmente, este é criado como parte da instalação do sistema operacional, e você
não deve ter que instalar nada para começar a trabalhar.

Você tem que especificar as tarefas que você deseja programar. Isso é feito simplesmente adicionar entradas para o emprego em um arquivo especial chamado de crontab
(Para"tabela"cron).
Esse arquivo está normalmente localizado no / Etc / crontab, Mas pode haver diferenças dependendo do sabor de UNIX. Em muitos casos, há uma crontab utilitário
disponíveis, que facilita a manutenção cron entradas.

O real cron as entradas são compostas de cron string (que foi discutido
anteriormente para o Quartz), seguido pelo comando. Há uma diferença notável,
rença entre UNIX cron cordas e quartzo cron strings: UNIX cron
cadeias não suportam um campo de segundos. Em vez disso, o campo é menor tempo minutos.

A linha a seguir no crontab arquivo deverá agendar o daily_load
Pentaho trabalho de integração de dados para executar todos os dias à meia-noite:

```
# M h dow dom seg                comando
  0 0 ***                        / Opt / pentaho / PDI / arquivo kitchen.sh / home wcm / PDI
/ Daily_load.kjb
```

NOTA Para obter mais opções no cron e crontab usando, consulte o seu funcionamento sistema de documentação. man crontab geralmente é um bom começo. Há também muitos recursos on-line que oferecem bons exemplos e explicações sobre o cron.

Windows: o de utilidade pública e do Agendador de Tarefas

Os usuários do Windows podem usar o utilitário ou menos o Agendador de Tarefas.

O utilitário está disponível a partir da linha de comando. Aqui está um exemplo simples que ilustra como programar a execução de um trabalho em lotes para executar cada dia meia-noite:

```
às 00:00 / every: M, T, W, Th, F, S, Su "D: \ pentaho \ pdi \ daily_job.bat"
```

Em vez de fornecer um comando muito diretamente para a em linha de comando, é geralmente melhor que escrever um arquivo em lotes (. Morcego arquivo) e ter ao executar isso. (Este
técnica pode, naturalmente, ser aplicado também em sistemas UNIX-like, onde um iria escrever um bater ou sh script.)

O Windows também oferece uma interface gráfica para agendamento. Você pode encontrar Agendador de Tarefas do Windows no Painel de Controle ou no menu Iniciar navegar até Iniciar Programas Acessórios Ferramentas do sistema Agendador de Tarefas.

NOTA Para obter mais informações sobre o em comando eo agendador de tarefas, por favor ir para a http://support.microsoft.com/ e procure por"no"comando e ""Agendador de Tarefas.

Contexto de execução e assinatura

Contexto de execução e assinatura são duas aplicações especiais de agendamento. Nesta seção, descrevemos estas características em detalhes.

Como funciona a execução em segundo plano

execução de fundo é uma característica conveniente que permite aos usuários executar seqüências de ação sem aguardar o resultado. Ao invés de iniciar a ação seqüência e esperando por ele para encerrar, uma agenda privada é criada para programar a ação ocorra o mais rapidamente possível. Após a conclusão, o saída da seqüência de ação é armazenado para que o usuário pode consultá-lo em um momento posterior.

Para executar uma seqüência de ação em segundo plano, você pode simplesmente clicar com o botão direito ela e escolha a opção Executar em segundo plano. Este aparece como uma caixa de mensagem o mostrado na Figura 14-7.

Figura 14-7: Um aviso antes de executar em segundo plano

Conforme sugerido pela caixa de mensagens, executando em segundo plano, desta forma é Não há suporte para seqüências de ação que exigem parâmetros de entrada, mas há uma alternativa.

Se você criar uma seqüência de ação que pede ao utilizador para parâmetros de entrada usando um Prompt / Secure Filtro Processo de ação, um botão de rádio é automaticamente apresentada, que permite ao usuário escolher imediatos ou de fundo execução. Um alerta simples / Secure filtro de entrada é mostrado na Figura 14-8. A Executado em segundo plano botão é visível na parte inferior.

Figura 14-8: Prompt para executar em segundo plano

Assinatura Como funciona

Assinatura é um recurso que permite aos usuários finais para receber conteúdo de acordo com BI
com uma programação predefinida. Em Pentaho, a subscrição é construído imediatamente após a
topo da Scheduler, permitindo que os usuários finais a atribuir seqüências de ação para agendas públicas. Como tal, a assinatura é uma extensão da programação característica.

Alguns requisitos devem ser cumpridos para permitir inscrição:

- O administrador do servidor deve definir uma ou mais agendas públicas. Nós discutida a criação de cronogramas utilizando o Server Console de Administração em na seção anterior. Lembre-se que você tem que selecionar o público Agenda checkbox.
- Para cada seqüência de ação, o administrador deve especificar quais os horários o usuário tem permissão para escolher quando se inscrever.
- Um usuário que deseja se inscrever com um relatório deve ter o privilégio de executar e agendar o relatório.

Permitir que usuários se inscrevam

O administrador do servidor deve especificar a que agendas públicas, o utilizador pode atribuir uma seqüência de ação em particular do usuário dentro do console. O servidor administrador pode fazer isso a partir do console de usuário (não o Servidor de Administração Console) clicando com a seqüência de ação e escolha Propriedades.

Escolhendo o item de menu Propriedades abre a caixa de diálogo Propriedades. Na de diálogo Propriedades, ative a página da guia Avançado. Na página da guia Avançado você tem que primeiro verificar o Uso Público checkbox Anexos. Depois, você pode selecionar uma agenda a partir da lista de horários disponíveis no lado esquerdo da guia página e use o botão para movê-lo para a lista Horários atual sobre o direito lado da janela. Isso é mostrado na Figura 14-9.

Isso permite que todos os usuários que têm o privilégio de fazê-lo para agendar esta ação seqüência.

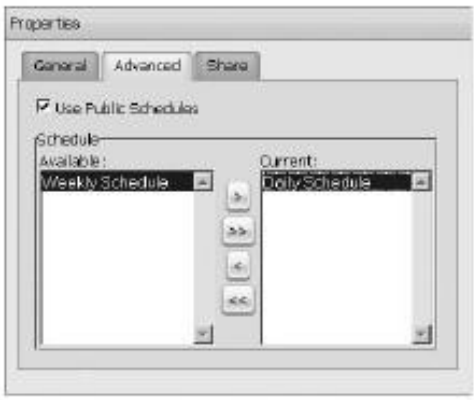


Figura 14-9: Fazendo uma seqüência de ação registráveis

Concessão de execução e cronograma Privilégios

Os usuários precisam ter, pelo menos, executar e agendar os privilégios de uma ação seqüência antes que eles possam assiná-lo. Não importa se esses privilégios eram concedidos directamente ou através de um papel.

Para conceder privilégios sobre uma seqüência de ação particular, botão direito do mouse e escolha a opção Compartilhamento no menu de contexto. Na caixa de diálogo que aparece, o de ações ficha já está ativado. (Outra maneira de chegar aqui é escolher o Propriedades opção no menu de contexto e, em seguida, ativar manualmente o compartilhamento página da guia.)

Na metade superior da página da guia de ações na caixa de diálogo, você pode ver todos os disponíveis usuários e funções. Na metade inferior, os privilégios que são actualmente configurado para o papel ou o utilizador que está atualmente selecionado são exibidas. Figura 14-10 ilustra isso.

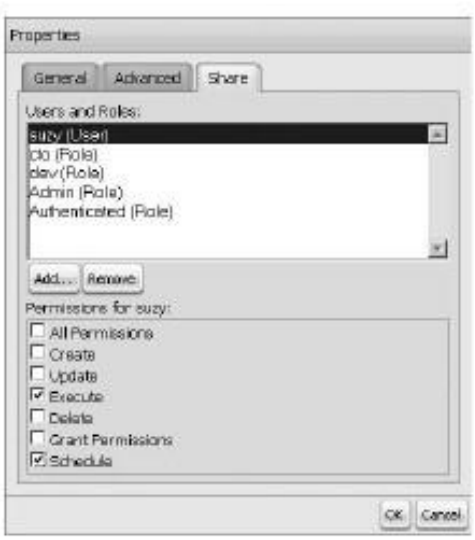


Figura 14-10: Concessão de Execução e Programação privilégios

A subscrição efectiva

No console do usuário, os usuários finais podem criar uma assinatura com o botão direito do mouse sobre um
seqüência de ação e escolhendo a opção Schedule no menu de contexto. Se a seqüência de ação é registrável eo usuário tem privilégios suficientes, o página web vai carregar uma caixa de diálogo como o mostrado na Figura 14-11.

Figura 14-11: Um diálogo simples assinatura

O diálogo mostrado na figura 14-11 é o mais simples possível assinatura diálogo. No diálogo, o usuário precisa digitar um nome no nome do relatório caixa de texto para identificar o relatório agendado. (Você pode digitar no máximo 25 caracteres,
que não é muita coisa.)

Com a programação caixa de listagem, o usuário pode selecionar uma das programações oferecidos. Estas são todas as programações que foram associados com a seqüência de ação pelo administrador do servidor.

Se a seqüência de ação utilizado na inscrição requer entrada do usuário, o diálogo de inscrição também apresenta os parâmetros do relatório, que pode então ser preenchido pelo usuário. Por exemplo, a figura 14-12 ilustra como um prompt usado para configurar o formato de saída do relatório são combinados com as instruções para o parâmetros de assinatura. O alerta para o formato de saída é adicionado explicitamente na seqüência de ação, ao passo que o alerta para a inscrição é apresentado automaticamente.

Figura 14-12: O prompt regular para os parâmetros é automaticamente combinadas com avisos para a subscrição

A entrada do usuário é armazenado junto com a escolha da programação e da ação seqüência na assinatura. Quando o cronograma de execução dos gatilhos seqüência de ação, os valores armazenados são usados como valores de parâmetro para a ação
seqüência.

Como foi explicado na subseção anterior, o usuário deve ter o Executar e da Tabela de privilégio para a seqüência de ação, a fim de subscrevê-lo.

Se o usuário não tem o privilégio de Programação, este é indicado por uma mensagem caixa como o mostrado na Figura 14-13.

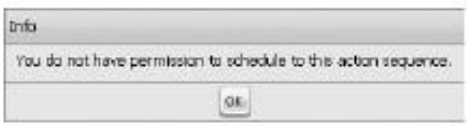


Figura 14-13: Uma caixa de mensagem indicando que o usuário não tem permissão para agendar o relatório

NOTA No momento da redação deste texto, ele aparece como se é possível inscrever-se seqüências de ação que não tenham sido feitas registrável pelo servidor administrador. Se você clique direito do mouse sobre uma tal seqüência de ação, você pode escolher
Inscrever-se que aparece um Agendador de diálogo Criador. No entanto, o preenchimento do diálogo e confirmando que sempre leva a uma mensagem de erro.

Espaço de Trabalho do Usuário

Com os recursos para execução de programação e de fundo, surge a necessidade para que o usuário seja capaz de controlar e monitorar o conteúdo programado. Além disso, tem que haver algum meio para que o usuário acessar os resultados gerados por qualquer agendada ações ou de fundo.

Pentaho oferece uma vista por usuário chamado Workspace que permite aos usuários gerenciar sua parte do repositório de conteúdo e seus horários.

Visualizando o Conteúdo da Área de Trabalho

No console do usuário, os usuários podem revisar o conteúdo de sua área de trabalho pessoal escolhendo Exibir espaço de trabalho a partir do menu. Na página Espaço de Trabalho, os usuários podem monitorar e controlar todas as seqüências de ação que estão agendadas ou
executados em segundo plano em seu nome. Figura 14-14 mostra um exemplo de usuário espaço de trabalho.

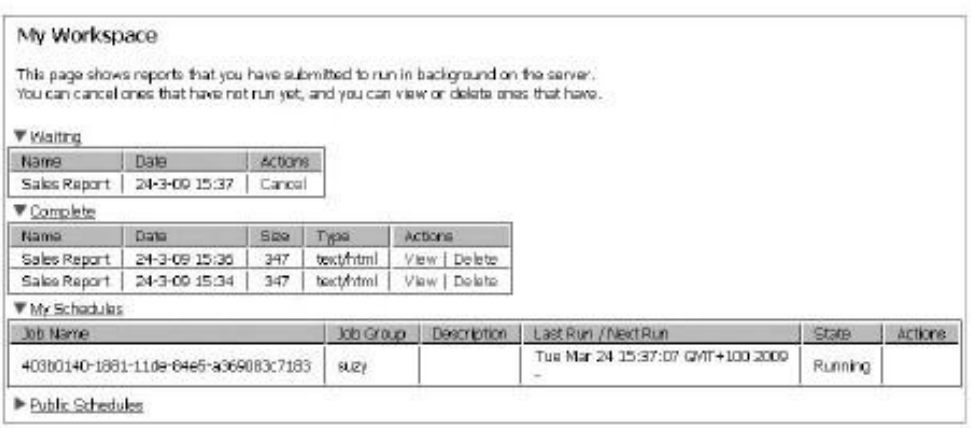


Figura 14-14: Um espaço de trabalho do usuário

A página Workspace mostra uma série de painéis contendo todas as medidas seqüências regulares em nome do usuário. Cada painel de ações pode ser col- caducado ou expandida clicando no triângulo imediatamente antes do painel títulos.

A espera, completa e painéis Meu Horários

A espera, completa e painéis Meu Horários podem ser utilizados para a gestão da acção seqüências que são executados em segundo plano.

- O painel de espera contém uma linha para cada seqüência de ação que é atualmente em execução em segundo plano em nome do usuário atual. A nome e data de início são mostrados para cada trabalho. Um link Cancelar está disponível para
interromper o trabalho. Por exemplo, na Figura 14-14 você pode ver que um trabalho chamado
""Relatório de Vendas iniciou a execução em segundo plano em 24-3-09 15:37. Em o tempo a página Workspace estava sendo visto, o trabalho ainda estava correndo.
- O painel de listas completas de todas as seqüências de ação que foram encomendados para
execução em segundo plano e que já se acabaram. Para cada ação seqüência, o nome e data de execução são mostrados. Além disso, o resultado da execução da ação está disponível a partir do painel completo: O coluna de tamanho indica o número de bytes na saída, e do tipo indica o tipo de saída foi armazenado. Há também um link para Ver inspecionar o resultado, e um link Delete para removê-lo permanentemente da espaço de trabalho. Na Figura 14-14, dois resultados de execução em segundo plano antes
do relatório de vendas estão disponíveis.
- O meu painel de listas também mostra as seqüências de ação que foram iniciado. Além das informações mostradas no painel de espera, o Meu painel Horários apresenta informações mais detalhadas (como o estado) sobre a seqüência de ação de execução.

O Painel de Agendas Públicas

O público painel Horários mostra todas as assinaturas do usuário atual. Para cada assinatura, uma série de links estão disponíveis.

- O Run Now link executa a seqüência de ação imediatamente. Esta é útil no caso de haver uma súbita necessidade de obter o resultado antes do programação regular. Como o Run Now link, o link Executar e Arquivo executa a seqüência de ação antes do previsto. No entanto, neste caso, a seqüência de ação é executada em segundo plano eo resultado armazenado no repositório de conteúdo.

- Com o link Editar, você pode alterar os detalhes da inscrição, tais como o cronograma. Se o relatório aceita parâmetros, estes podem ser modificados

a partir deste link, também. A figura 14-15 ilustra o que acontece quando você clica no link Editar.

- O link Delete remove permanentemente a assinatura. Quando você clica ela, uma caixa de mensagem aparece e pede confirmação para excluir a assinatura. Confirmando isto irá remover apenas o usuário atual assinatura, mas o cronograma real próprio público.

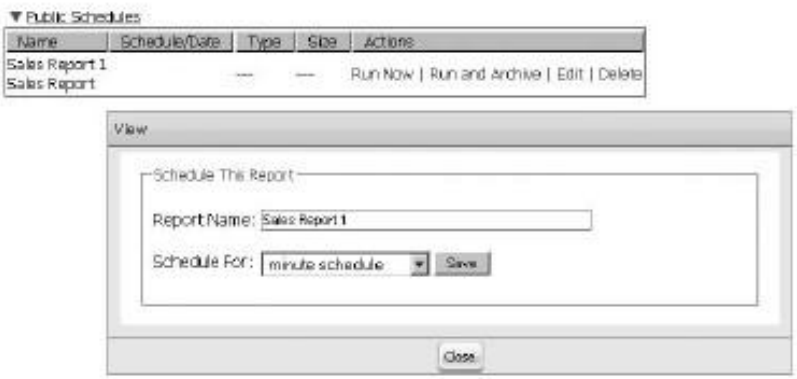


Figura 14-15: Editando uma subscrição existente

A saída de inscrição normalmente aparece logo abaixo da assinatura entrada no painel Horários Pública. Figura 14-16 mostra o Relatório de Vendas 1 assinatura, e imediatamente abaixo, são três linhas que correspondem a uma execução bem-sucedida da seqüência de ação. Usando o link View, os usuários pode baixar o resultado da ação, e com o link Excluir, os usuários podem remover permanentemente o conteúdo do seu espaço de trabalho.

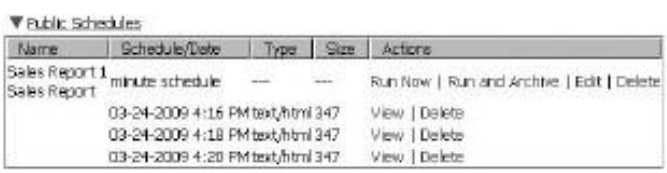


Figura 14-16: Uma assinatura e sua saída

Área de trabalho do administrador do servidor

espaço de trabalho do administrador do servidor, mostrado na Figura 14-17, tem um adicional de
Todos os Horários painel. Este painel mostra todos os horários e lhes permite ser executado imediatamente, suspensos, e retomada. A mesma funcionalidade está disponível a partir do Servidor de Administração do console, mas é repetido aqui como um conveniência recurso.

Figura 14-17: espaço de trabalho do administrador do servidor

Limpando a área de trabalho

Quando deixado sem vigilância, o repositório de conteúdo será em algum momento, preencher o
disco. Os usuários podem excluir seus próprios itens do repositório de conteúdo, mas o servidor
Os administradores não devem se basear nisso. Para qualquer configuração de produção, os administradores
deve ter o cuidado de monitorar o repositório de conteúdo.

Na página da guia Serviços do Console de Administração, alguns itens são disponíveis para ajudar a gerenciar o repositório de conteúdo. O repositório de conteúdo painel de serviço é mostrado na Figura 14-18.

Figura 14-18: serviços Console de Administração para gerenciar o repositório de conteúdo

Pressionando o botão Executar irá executar o clean_repository ação seqüência localizado no Admin Pentaho solução. Isto irá remover todos os arquivos no repositório de conteúdo que são mais de 180 dias. Se você gosta, você pode editar o clean_repository seqüência de ações e modificar o número de dias.

Apertando o botão irá executar o Schedule schedule_clean ação seqüência, que também reside no Admin Pentaho solução. Esta acção usa a seqüência Adicionar tarefa ação do processo para instalar um programa para o diário clean-up do repositório de conteúdo. Como você poderia esperar, esta recai sobre o clean_repository para fazer a limpeza real.

Ruptura

Em um contexto de BI, ruptura é por lotes de produção, baseados em dados e distribuição de BI de conteúdo, como relatórios e gráficos.

Um exemplo muito simples de rompimento seria a criação de relatórios da expedição para cada armazém e enviar a saída do relatório para o correspondente Warehouse Manager. Outro exemplo típico é o envio de cada cliente um e-mail com um anexo mostrando uma visão geral de todas as rendas e as compras efectuadas durante o ano passado.

NOTA A expressão""estouro origina-se da época em que grandes relatórios foram impresso todo em papel, a impressão e precisaria ser dividido em separado peças para seus respectivos destinatários.

Nestes dois exemplos, o relatório é executado, mas o conteúdo eo contexto do produção é feita sob medida para cada destinatário.

Implementação de ruptura em Pentaho

A edição da comunidade do Servidor Pentaho não oferece nativa rebentar recursos. No entanto, com alguma criatividade e esforço, estourando pode ser imple- complementados por looping através de um conjunto de dados e usando os dados da linha atual executar um relatório e envia os resultados para o receptor apropriado (s). Os seguintes seção o orienta através de um exemplo que estouram no Pentaho.

Exemplo de ruptura: Aluguel lembrete E-mails

Considere o caso de uso de lembrete de aluguer e-mails: a cada semana, de Classe Mundial Filmes envia lembretes de e-mail aos seus clientes para notificá-los que DVDs deverão ser devolvidos (ou adquiridos), durante a semana seguinte. Porque este é essencialmente um processo operacional, escolhemos como base para este exemplo diretamente no banco de dados WCM, e não o data warehouse. (Em uma configuração de prática, você seria mais provável ainda usar uma instância de banco de dados separado, como uma replicação escravo, a fim de descarregar outras operações, como a entrada do pedido).

Os passos seguintes descrevem como isso funciona:

1. Obter um conjunto de resultados de todos os clientes que ordenou DVDs que deverão ser retornados durante a semana seguinte.
2. Loop sobre as linhas no conjunto de resultados. Para cada linha:
   - Obter a lista atual de títulos de DVD, que deverão ser devolvidos durante na semana seguinte para o cliente atual.
   - Executar um relatório que lista os títulos de DVD e as respectivas datas de vencimento.

- Conecte a saída do relatório para uma mensagem de correio electrónico e enviá-lo à cliente.

No restante desta seção, descrevemos como implementar isso em um Pentaho seqüência de ação. Para seguir com o exemplo, você vai precisar Pentaho Design Studio (para criar a seqüência de ação) eo Relatório de Pentaho Designer (para criar o relatório).

Passo 1: encontrar clientes com DVDs que são Entrega esta semana

O primeiro passo é encontrar todos os clientes que precisam receber um lembrete. Um maneira de abordar este problema é olhar para todos os DVDs que são devidos durante o Na semana seguinte, e depois de olhar para cima os dados do cliente através da correspondente
da ordem.

Use o SQL na listagem 14-1 para fazer isso.

Listing 14-1: Encontrar clientes com DVDs que são devidos

```
SELECTcustomer_id
,       MAX (c.first_name) first_name
,       MAX (c.last_name) last_name
,       MAX (c.balance) equilíbrio
,       MAX (c.email_address) email_address
,       CAST (
          GROUP_CONCAT (
            col.customer_order_line_id
          ) AS CHAR (512)
        ) Customer_order_lines
FROMwcm.customerc
INNER JOIN wcm.customer_orderco
ONc.customer_id = co.customer_id
INNER JOIN wcm.customer_order_linecol
ONco.customer_order_id = col.customer_order_id
WHEREcol.return_due_date
Entre {} report_date
E report_date {} + INTERVAL {} report_interval SEMANA
GRUPO BYc.customer_id
```

Algumas coisas são dignas de nota sobre essa consulta. Na consulta, você pode veja {} Report_date e {} Report_interval. As chaves não são válidas SQL. Ao contrário, eles delimitar parâmetros ao nível de uma seqüência de ação. Em vez de incorporar uma chamada para CURRENT_DATE () ou algo parecido em a consulta, nós preferimos a exteriorizar isso em um parâmetro. Isso fará com que muito mais fácil depois para iniciar a seqüência de ação de alguns dias à frente ou atraso. Nós gostamos de ter a {} Report_interval por um motivo semelhante. Ao torná-la disponível como um parâmetro, podemos facilmente decidir mais tarde para

enviar e-mails uma vez a cada duas ou três semanas, em vez de a cada semana. Embora seja impossível para parametrizar tudo (ou pelo menos não seria muito prático), esses itens são os prováveis candidatos para a mudança por causa da mudança de negócios decisões.

A consulta usa uma GROUP BY customer_id cláusula. Porque customer_id é o chave primária da cliente tabela, essa consulta irá produzir, por definição, apenas um linha por cliente. Na SELECT lista, que você deseja recuperar todos os tipos de dados de o cliente, e para evitar confusão, você deve adicionar explicitamente MAX agregada funções para todas as outras colunas que você precisa do cliente. (Tecnicamente, você não precisa fazer isso, por causa da GROUP BY em customer_id você pode ser se todas as outras colunas da cliente tabela terá apenas um valor por distintos customer_id. No entanto, você deve aplicar MAX () de qualquer forma para a clareza e portabilidade).

Um aspecto final desta consulta é a aplicação do GROUP_CONCAT função. Esta é uma função específica do MySQL agregado que as linhas de grupos usando con-corda encadeamento, por padrão, separando os valores com uma vírgula. Neste caso, a única argumento é o customer_order_line_id, O que significa GROUP_CONCAT função irá gerar uma lista separada por vírgulas dos customer_order_line_idS que correspondem aos DVDs que deverão ser devolvidos. Você vai usar isso mais tarde na na seqüência de ação para executar uma consulta para o relatório.

Para colocar isto em seqüências de ação, você pode usar o Relacional processo acção. Você pode encontrá-lo sob a obter dados a partir do menu. Na Consulta propriedade, digite a instrução SQL. Para o nome do conjunto de resultados, os clientes que você digita e em
Além disso, você definir explicitamente cada coluna do conjunto de resultados no Resultado Definir grade Colunas. Isto tornará mais fácil para se referir a itens específicos no conjunto de resultados. Você também tem que especificar os parâmetros de ação para a seqüência de entrada
{} Report_date e {} Report_interval parâmetros.

A seqüência de ação é mostrado na Figura 14-19.

Passo 2: looping através dos clientes

A segunda etapa da seqüência de ação é um Loop processo de ação. Isso itera através de todos os clientes que encontramos na primeira etapa da seqüência de ação. Figura 14-20 mostra como usar o Loop ação processual na ação de ruptura seqüência. Na figura, o Loop ação do processo é marcado Loop em Ação Clientes.

Em si, a Loop ação do processo não é muito interessante. No entanto, torna-se interessante em virtude das ações que são colocados dentro dela. Essas ações são repetido para cada iteração do loop. Na Figura 14-20, os restantes três ações aparecem recuadas em relação ao ciclo. Assim, estas acções serão repetido para cada linha da clientes conjunto de resultados.

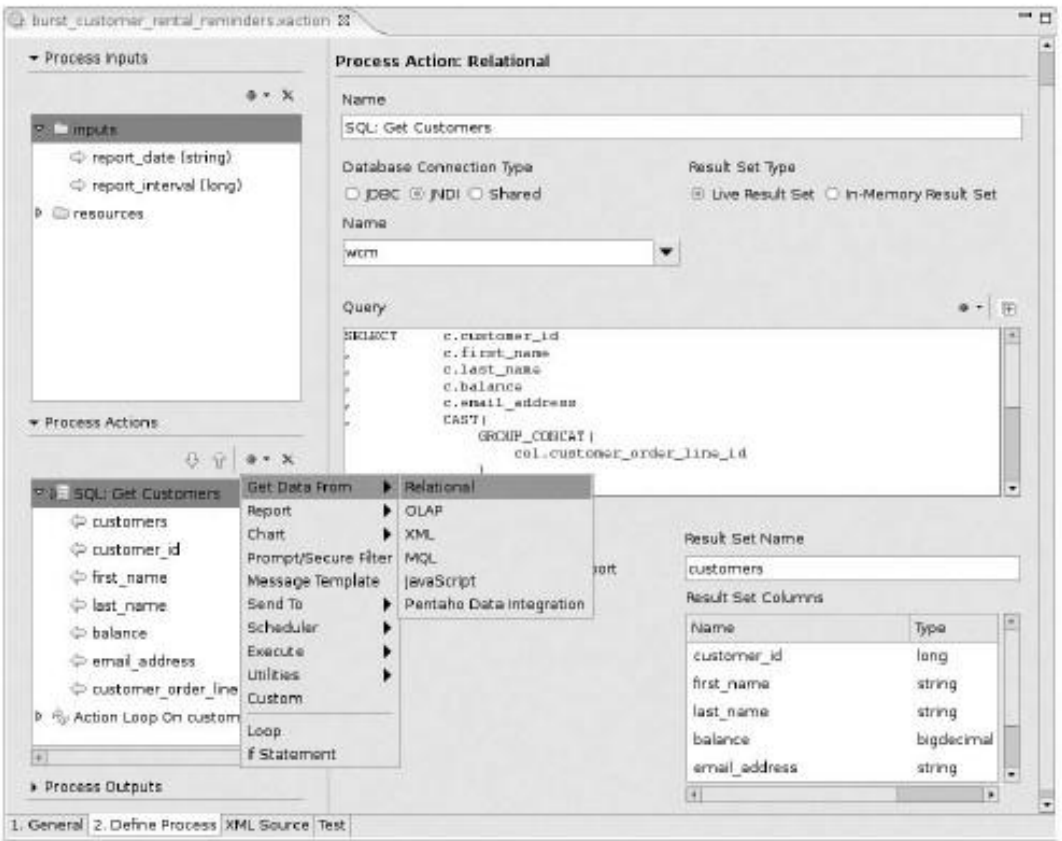


Figura 14-19: Encontrar clientes usando uma etapa da ação processo relacional

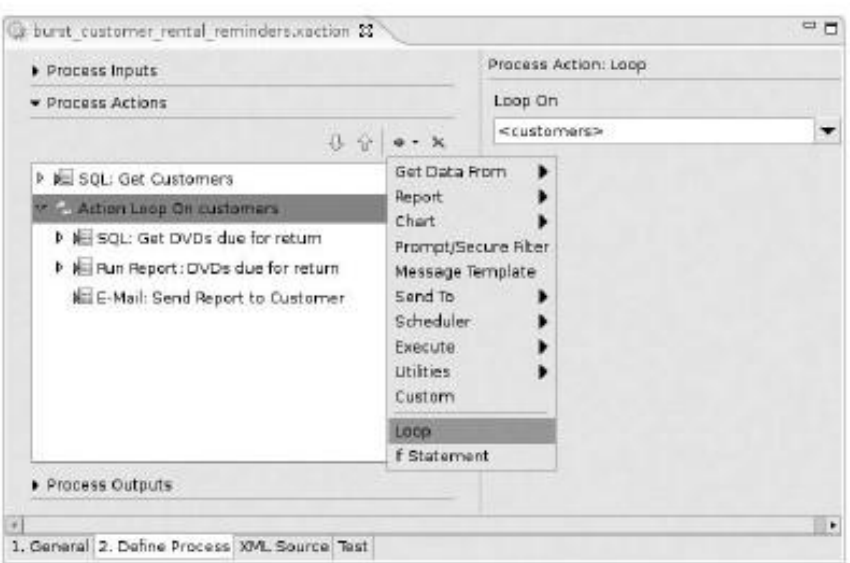


Figura 14-20: Loop através de um conjunto de resultados

Passo 3: Primeiros DVDs que deverão ser devolvidos

Dentro do loop, você precisa descobrir quais DVDs deverão ser devolvidos para a iteração atual cliente. Felizmente, isso é muito fácil! Porque você incluiu o GROUP_CONCAT expressão sobre o customer_order_line_id coluna em sua clientes consulta, você pode usar um muito simples e eficiente EM consulta para procurar as linhas relevantes do customer_order_line tabela. De o tabela customer_order_line, Você pode facilmente procurar o correspondente DVDs na dvd_release tabela utilizando uma INNER JOIN.

Mais uma vez, você usa um Relacional ação do processo, como você fez para os clientes. Para os DVDs, foi utilizada a consulta mostrada na listagem 14-2.

Listagem 14-2: Encontrar DVDs que são devidos

```
SELECT      {} Customer_id customer_id
,           first_name 'first_name {}'
,           last_name '{} last_name "
,           {} Equilíbrio equilíbrio
,           {} Report_date report_date
,           d.title
,           col.return_due_date
,           col.rental_price
,           col.purchase_price
DA          col wcm.customer_order_line
INNER JOIN  wcm.dvd_released
ON          col.dvd_release_id = d.dvd_release_id
ONDE        col.customer_order_line_id IN ({} customer_order_lines)
```

Na ONDE cláusula, você vê o {} Customer_order_lines parâmetro, que é usado para localizar o customer_order_line linhas. Lembre-se que o GROUP_CONCAT função usada na consulta do cliente retorna uma lista separada por vírgulas customer_order_id valores.

Além da {} Customer_order_lines parâmetro, você copiou o cus-Tomer dados da iteração atual cliente para o SELECT lista. Você precisa dados de clientes de qualquer maneira para personalizar o relatório, e embora você possa obter as mesmas informações de programação, alargando o SQL estado mento com INNER JOINS para customer_order e cliente, Você pode muito bem beneficiar do fato de que a obra já está executada.

Passo 4: Executando o relatório lembrete

Agora que você tem dados sobre o cliente e os DVDs, você pode processar um relatório. Para este efeito, você precisa de um relatório muito simples, que nós criamos com o Designer de Relatórios Pentaho. Não vamos descrever este processo em pormenor, mas o ponto principal deste relatório é que todos os dados relativos à alugado

DVDs, como o DVD de título, aluguel de dados de vencimento, preço de aluguel, eo preço de compra são
prestados no item Band. Os dados repetidos no item Band é marcado
utilizando rótulos estáticos que aparecem no cabeçalho do relatório. Os dados referentes ao o relatório em si (como o report_date) Ou um cliente seu recebimento (como first_name,last_name,equilíbrioE report_data) E são colocados na página Cabeçalho. O design do relatório é mostrado na Figura 14-21.

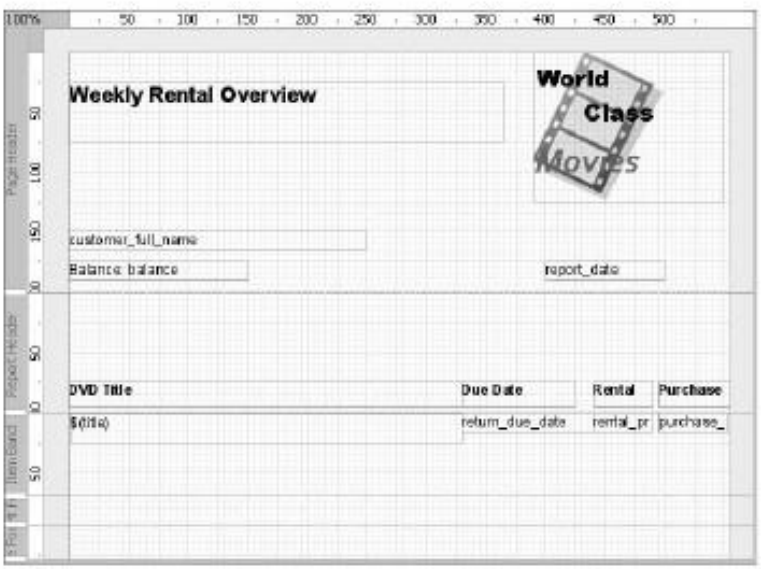


Figura 14-21: O projeto para o arrendamento lembrete Relatório

Da perspectiva do design do relatório, as coisas são bastante simples se
enviar o relatório como um anexo em PDF com uma mensagem de correio electrónico. No entanto,
Se você quiser enviar o relatório em si como o corpo de um email HTML, você
pode querer configurar a saída de HTML especialmente para esse fim. Você pode
acesso a configuração do relatório através do painel de propriedades mostradas no relatório Figura 14-22.

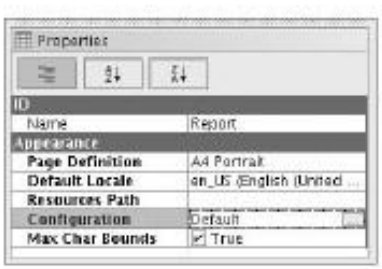


Figura 14-22: Acessando as propriedades do relatório no painel Propriedades

Na janela de propriedades, você pode ativar o saída de tabela html
configuração e explicitamente suprimir folhas de estilo externas, ou mesmo a força estilos
a ser escrita em linha. Para relatórios baseados na Web normal, folhas de estilo externas
geralmente são preferíveis, e um estilo inline deve ser evitado. No entanto,
folhas de estilo externas e estilos fora-de-linha pode impedir alguns clientes de e-mail
da prestação corretamente o relatório, é por isso que explicitamente se desviam
entre as opções padrão neste caso. Substituindo esses padrões é mostrada na
Figura 14-23.

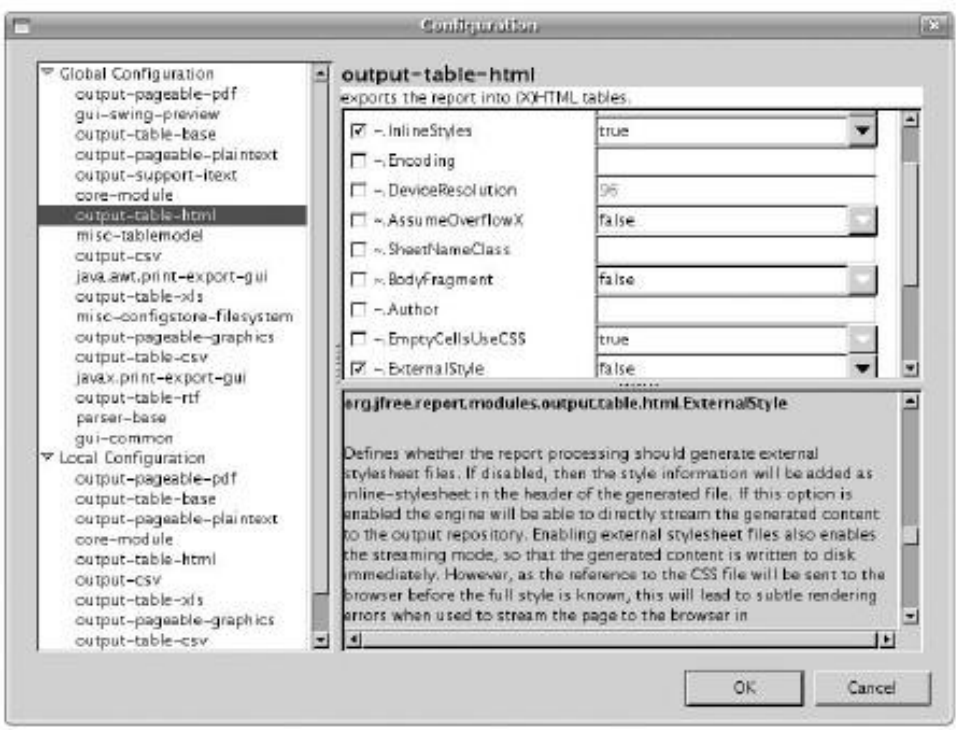


Figura 14-23: Forçar estilo interno e estilos inline para HTML e-mail

O relatório é chamada com o Pentaho Report ação do processo, que reside
sob o submenu relatório. Figura 14-24 mostra que a ação do processo
parece.

Note-se que o <dvds> parâmetro é utilizado para o relatório de dados. Este é o nome
do resultado do conjunto de parâmetros a partir da ação do processo anterior. Note também a
Nome da saída do relatório, que é definida como rental_reminder_report. Você precisa
referem-se a isso no próximo passo, onde você vai enviar a saída do relatório em um e-mail.

Passo 5: o envio do relatório via e-mail

O passo final na seqüência de ação toma conta da entrega ao cliente.
Para fazer isso, use um EMAIL processo de ação. Isto pode ser alcançado a partir do Enviar
Para menu, como mostrado na Figura 14-25.

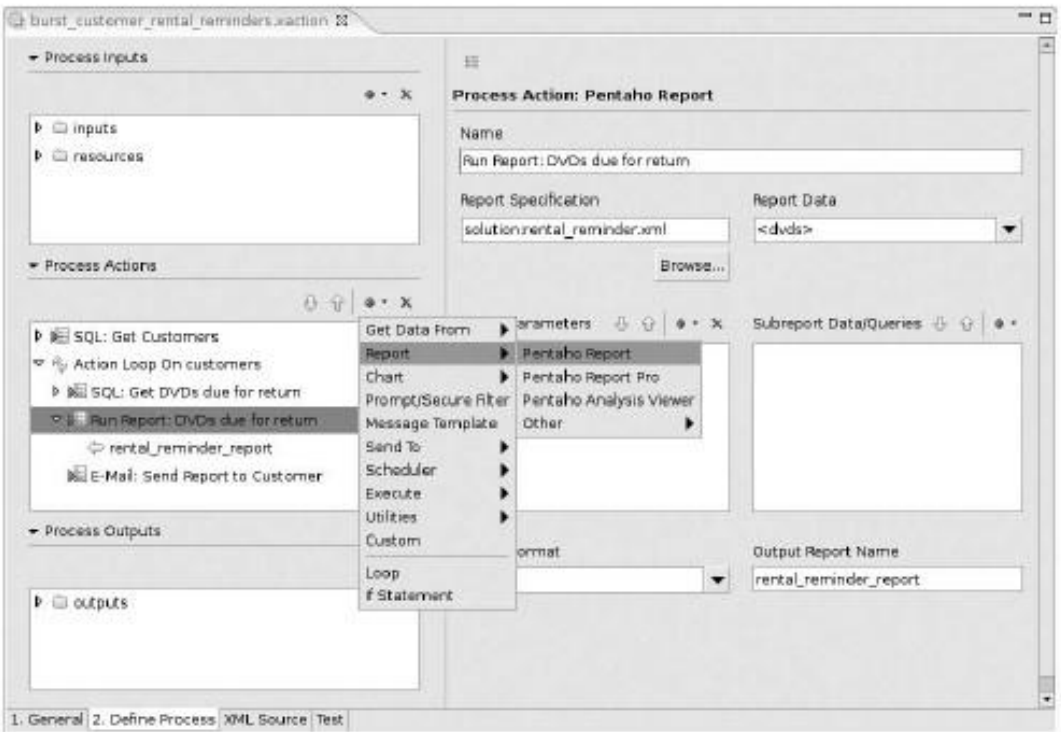


Figura 14-24: Chamando o Relatório lembrete

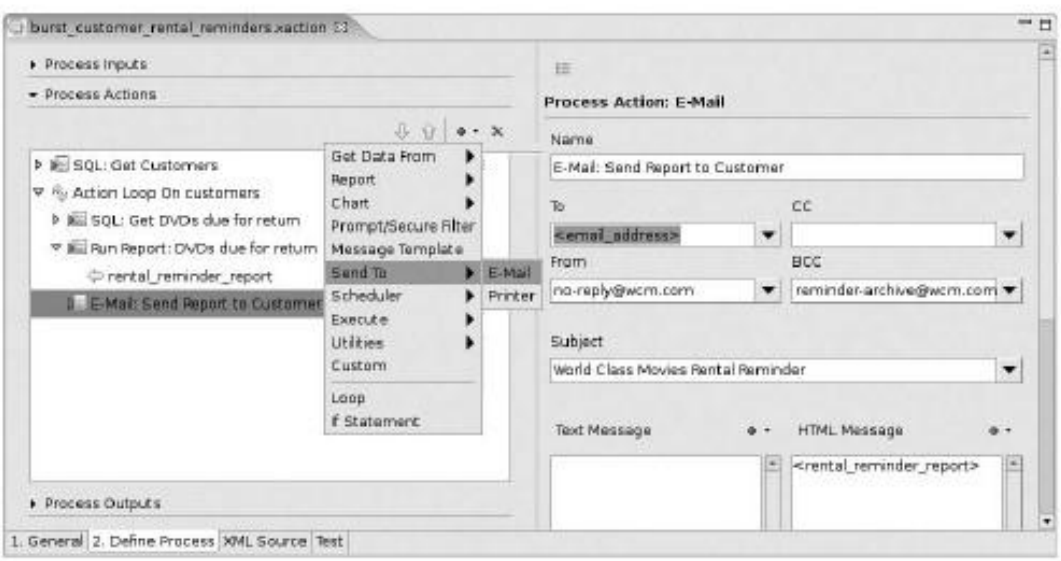


Figura 14-25: O envio do relatório, como HTML e-mail

Observe que você use o <email-address> a partir dos dados de clientes do exterior conjunto de resultados como um parâmetro para o resolver. Para a mensagem HTML, utilize o <rental-reminder-report> parâmetro, que é a saída do relatório real criada no processo de ação anterior.

Outras implementações de ruptura

O exemplo já visto é simplesmente uma forma de execução de ruptura. Não pode haver casos em que esta abordagem funciona muito bem. Pode haver outros casos onde esta abordagem não é tão útil (tal como quando o resultset exterior é grande, e consultar o para o resultset interno é complexo e caro). Aqui são apenas algumas outras possibilidades para a aplicação do relatório de ruptura com base na Pentaho BI Suite:

- Em vez de enviar e-mail HTML, você pode configurar o relatório (ou a ação processo de chamada do relatório) para gerar um PDF. Neste caso, a saída do relatório será anexado ao e-mail ao invés de enviado como o corpo. Com o E-mail ação do processo dessa mudança é quase trivial. Você pode basta mover o <rental-reminder-report> parâmetro a partir do HTML campo de mensagem para a grade Anexos e escreva uma mensagem de texto estático em Mensagem de Texto de campo.

- Se você quiser enviar apenas uma mensagem de correio electrónico simples, você pode usar o Mensagem Modelo ação do processo de mesclar um modelo de texto estático com o conjunto de resultados

- Você pode até escolher entre e-mail de texto simples, HTML e-mail ou um arquivo PDF penhora, dependendo das preferências do usuário. Para fazer isso, você iria expandir a consulta com um campo que detém a preferência do usuário para PDF, texto simples, ou HTML. Dependendo do valor do campo, você pode usar um Se Declaração ação do processo para enviar o tipo correto de e-mail.

- Em vez de construir seqüências de ação complexo, você poderia escolher para implemento muito da lógica dentro de uma transformação Pentaho Data Integration. Esta abordagem é especialmente útil quando você precisa simplesmente de enviar um mensagem de e-mail, e podem ignorar gerar um relatório completo. Um extra vantagem de usar PDI é que você pode usar as suas capacidades de clustering para distribuir a carga entre múltiplas máquinas.

- Depois de trazer PDI na mistura, você pode ser capaz de eliminar a necessidade para consultas de banco de dados totalmente aninhados. Em vez disso, você pode escrever apenas uma consulta que inclui dados de clientes e títulos de DVD, e usar a PDI transformação que diretamente em uma estrutura aninhada, como XML ou JSON. Você seria então necessário para analisar a utilização de uma obter dados de / XML ou Javascript etapa, mas esta ainda pode ser mais rápido do que espera para uma consulta de banco de dados.

Resumo

Neste capítulo, investigamos alguns dos métodos mais avançados de fornecimento de conteúdos de inteligência comercial para usuários finais. Neste capítulo, você Aprendi a:

- Criar e manter horários usando o console de administração
- Use sequências de ação para programaticamente criar agendas
- Uso externo soluções de programação, tais como cron e em
- Permitir que as seqüências de ação a ser registráveis, associando-as com um cronograma
- Conceder os privilégios apropriados para os utilizadores (e funções), que lhes permitam subscrever para os itens de ação
- Monitorar e gerenciar o conteúdo inscrito no espaço de trabalho do usuário
- Limpe o repositório de conteúdo
- Implementar explodindo em uma seqüência de ação

CAPÍTULO

15

Soluções OLAP Utilizando Pentaho Analysis Services

Pentaho Analysis Services fornece os recursos OLAP do Pentaho
Plataforma. Capítulo 8 breve introdução aos conceitos de OLAP, ROLAP, MOLAP,
e HOLAP. Este capítulo é dedicado à obtenção de soluções (R) para cima e OLAP
execução na plataforma Pentaho BI Pentaho usando Analysis Services (PAS).
PAS lhe permite analisar os dados de forma interativa a partir da data warehouse
fornecendo uma interface de referência cruzada de estilo em que as diferentes dimensões, tais como
tempo, produto e cliente pode ser colocado. Ao contrário de uma ferramenta de comunicação, não há
preciso primeiro definir uma consulta, obter os resultados e formatar esses, embora esta
É possível, se desejar. Um front end OLAP fornece uma interface intuitiva ponto-e-clique
ou interface drag-and-drop que irá automaticamente recuperar e formatar dados
com base na seleção feita por um usuário. Ela permite a rápida zoom em
certas partes do cubo de dados, também chamado drill down, ou agregar detalhes
um nível de síntese, também chamado drill-up. Você pode aplicar condições para apenas olhar
partes de um cubo, que também é chamado corte. Trocando informações de linhas
a colunas ou vice-versa, finalmente é como o giro do cubo e olhá-lo de
ângulos diferentes, que também é chamado corte. Executando drill up, drill down,
Fatie e pique, e fazendo tudo isso de uma alta velocidade, é o que forma interativa
OLAP distingue de outros tipos de análises e relatórios e permite uma
usuário para analisar rapidamente os dados e encontrar exceções ou ter uma visão de negócios
desempenho.

Neste capítulo, começamos por descrever a arquitetura da Pentaho Analysis
Serviços, seguido por uma breve introdução ao MDX, que é de facto
linguagem padrão de consulta OLAP. A seguir, vamos explicar em detalhe como criar
e implementar cubos OLAP para o motor ROLAP Mondrian, que é o coração
do Analysis Services Pentaho. Também explicamos como navegar estes cubos

usando o front-end JPivot. Finalmente, discutiremos como você pode usar o Pentaho designer de agregação para melhorar o desempenho do OLAP.

Visão Geral dos Serviços de Análise Pentaho

PAS é composto dos seguintes componentes:

- JPivot front-end-análise JPivot é uma ferramenta de análise baseada em Java que serve como a interface real do usuário para trabalhar com cubos OLAP.
- Mondrian motor ROLAP-A motor recebe a partir de consultas MDX front-end ferramentas como JPivot, e responde enviando uma mul- resultado provisório-set.
- Schema Workbench-Este é a ferramenta visual para projetar e testar Mondrian esquemas cubo. Mondrian usa esses esquemas cubo para interpretar MDX e traduzi-lo em consultas SQL para recuperar os dados de uma RDBMS.
- Agregado Designer-A ferramenta visual para geração de tabelas de agregados para acelerar o desempenho do motor analítico.

NOTA Em 2009, um projeto da comunidade foi iniciada a construção de uma nova geração Pentaho Analysis Tool (PAT), que visa substituir JPivot no futuro. No momento da isto foi escrito, o PAT está ainda na sua fase inicial de desenvolvimento assim para o resto do Neste capítulo, vamos furar a JPivot. Se você quiser dar uma olhada no PAT, visite o página do projeto na casa http://code.google.com/p/pentahoanalysistool.

Arquitetura

Figura 15-1 mostra uma visão esquemática dos componentes do PAS e os seus relacionamentos.

Primeiro, vamos resumir os elementos e as interações mostrado na Figura 15-1. A seguinte seqüência de eventos descreve o que acontece quando o usuário final usa um típico Pentaho aplicação OLAP:

1. O navegador do usuário final de internet faz uma solicitação HTTP para ler, ver ou drill down em uma tabela dinâmica OLAP. (Em Pentaho, isso normalmente resulta em a execução de uma seqüência de ação, que neste caso foi construído para chamar JPivot).
2. O servlet JPivot recebe o pedido e converte-lo em um MDX consulta. A consulta MDX é enviado para o motor de Mondrian.
3. Mondrian interpreta a consulta MDX e traduz isso em um ou mais Consultas SQL. Esta técnica especial é referido como ROLAP, que representa Relational OLAP. (Neste capítulo, nos referimos ao termo OLAP para maior clareza, embora no contexto de Mondrian, a técnica atual é mais apropriadamente chamada ROLAP).

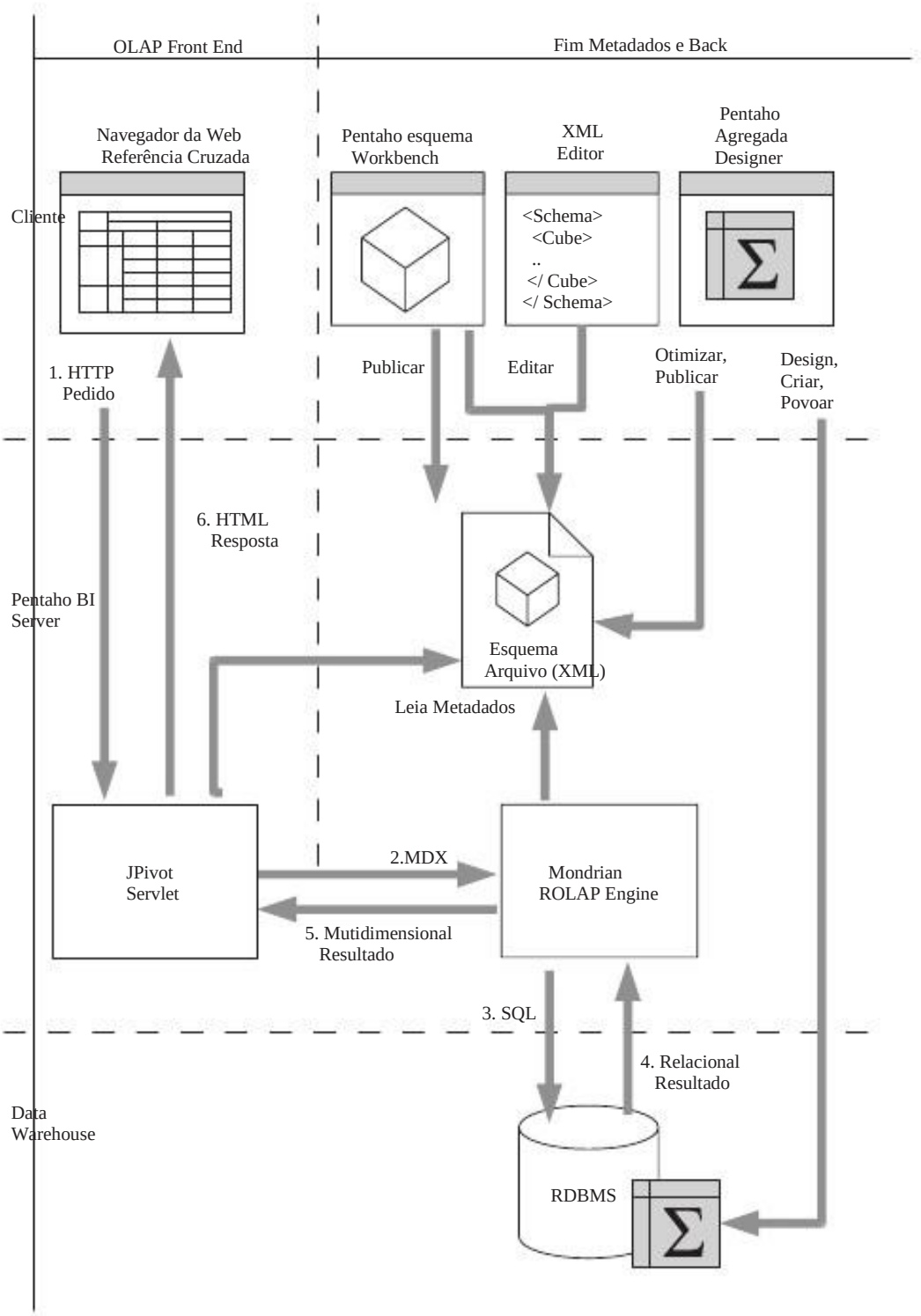


Figura 15-1: Visão geral dos componentes Pentaho OLAP

4. O Relational Database Management System (RDBMS) executa o
consultas emitidas por Mondrian. Mondrian recebe tabular (relacional)
os resultados.

5. Mondrian processa os resultados recebidos do RDBMS e traduz
los a um conjunto de resultados multidimensional. Esta é realmente a consulta MDX
resultado da consulta MDX enviado para Mondrian na Etapa 2.

6. JPivot utiliza o resultado multidimensional para renderizar uma página HTML para
exibir os dados. Este é então enviado para o navegador onde é mostrado
o usuário.

Esquema

A estrutura central mostrado na Figura 15-1 é o esquema. O esquema é essencialmente
um documento XML que descreve um ou mais cubos multidimensionais. A
cubos também descrever o mapeamento das dimensões do cubo e as medidas de
tabelas e colunas em um banco de dados relacional. Para Mondrian, o esquema é a chave
em traduzir a consulta MDX para consultas SQL.

Ferramentas de projeto de esquema

A parte superior direita da figura 15-1 mostra uma série de componentes que fazem
não participa diretamente na interação anteriormente resumidas. Estes todos
representa o design e desenvolvimento de ferramentas usadas para construir ou melhorar
Mon-
drian esquemas. Um esquema de Mondrian é o mapeamento entre o relacional
eo modelo multi-dimensional. Esse mapeamento é usado para ajudar a traduzir
Consultas MDX para consultas SQL, e para transformar os resultados relacional recebidos em
resposta às consultas SQL para resultados multi-dimensional. O multidimensional
modelo, composto de dimensões, hierarquias e medidas, é criada primeiro e
o modelo relacional é mapeado para o esquema. Porque você vai trabalhar
com um esquema em estrela como a origem do esquema de Mondrian, esta será uma grande
simples processo.

Pentaho Schema Workbench oferece uma interface gráfica para criar Mon-
drian esquemas. Além disso, Pentaho Schema Workbench pode publicar esquemas
ao Servidor Pentaho, que em seguida, armazena-los na solução de repositório. Uma vez
armazenados na solução de repositório, os esquemas podem ser usados pelo servidor
motor de Mondrian como um back-end para serviços OLAP.

Pentaho Schema Workbench é apenas uma ferramenta que você pode usar para criar
esquemas.
Você também pode usar um editor de texto ou editor XML para gravar o esquema homem-
vamente, é por isso que o incluiu na Figura 15-1, ao lado de Pentaho
Schema Workbench. Você pode publicar esquemas escritos manualmente usando Pen-
Bancada Taho, ou movendo o arquivo XML que contém o esquema para o
Diretório da solução Pentaho desejada no sistema de arquivos.

Tabelas agregadas

O Pentaho Designer Agregado (PAD) é uma ferramenta que pode ajudá-lo a geração e preenchimento de tabelas de agregação. Mondrian pode aproveitar de tabelas de agregados para gerar consultas SQL eficientes que podem ser consideravelmente
melhorar o desempenho. Assim, o PAD análises do banco de dados de back-end, gerando as declarações adequadas SQL para criar e preencher as tabelas de agregação. Em mesmo tempo, PAD modifica o arquivo de esquema, que é necessário para Mondrian usar as tabelas de agregação.

Este capítulo aborda todas as ferramentas necessárias para desenvolver e utilizar soluções OLAP,
mas antes de podermos fazer isso precisamos cobrir alguns conceitos básicos como MDX ea estrutura de um cubo OLAP.

MDX Primer

MDX é a abreviação de expressões Multi Dimensional, que é um linguagem que é especialmente concebido para consultar bancos de dados OLAP. MDX é um padrão de facto originalmente desenvolvido pela Microsoft.

NOTA MDX é um padrão criado e implementado pela primeira vez pela Microsoft no servidor da Microsoft Analysis, que é fornecido como parte do SQL Server RDBMS. Após sua introdução, MDX foi amplamente adotado por outros fabricantes como a OLAP linguagem de consulta.

Atualmente não há consórcio ou comitê estabelece normas fora da Microsoft, que mantém uma especificação MDX. documentação de referência só está disponível como parte da documentação do produto SQL Server. A referência atual documentação pode ser encontrada em http://msdn.microsoft.com/en-us/library/ms145506.aspx.

Você pode encontrar mais informações e documentação no site de um dos inventores da linguagem MDX, Mosha Pasumansky: http://www.mosha/.com MSOLAP. Outra excelente fonte é a MDX série Essentials no www.databasejournal.com.

Além desses recursos on-line, você pode ler mais sobre MDX MDX Soluções: Com o Microsoft SQL Server Analysis Services 2005 e Hyperion Essbase, por G. Spofford, S. Harinath, C. Webb, Huang Hai D. e F. Civardi (Wiley, 2006, ISBN: 978-0-471-74808-3).

Em certo sentido, MDX é o modelo multidimensional que o SQL é o modelo relacional. Consultas SQL definir operações em tabelas do banco de dados para recuperar
um conjunto de linhas, enquanto as consultas MDX operar em um cubo e entregar um multidimensional-
multidimensional coleção de células.

Embora você não precisa ser um especialista em MDX, que ajuda a saber o
noções básicas antes de começar a construir os cubos de Mondrian. Além disso, se você quiser
criar soluções analíticas que excedem o padrão de perfuração e de filtros
do front-end JPivot, você precisará modificar o MDX gerado a si mesmo.
Mas antes de explorar o poder do MDX, é uma boa idéia para cobrir algumas de base
conceitos OLAP para ajudar você a entender melhor a sintaxe.

Cubos, dimensões e medidas

Quando você primeiro encontro MDX, um dos aspectos mais confusa é a
terminologia utilizada. Alguns termos podem parecer familiares e referem-se os conceitos utilizados na
modelagem dimensional. No entanto, alguns conceitos são totalmente diferentes e parecem
inábil no início. Nesse sentido, e as subseções seguintes, vamos dar uma simplificada
versão do modelo de WCM e usar isso para ilustrar todos estes conceitos.

O Conceito de Cubo

A maioria dos livros sobre OLAP MDX e começar por apresentar uma representação visual
de um cubo tridimensional. Por que devemos romper com esta tradição? Então, vamos
apresentar alguns dos principais conceitos com a Figura 15-2, que mostra um simples
cubo construído a partir do armazém de dados WCM.

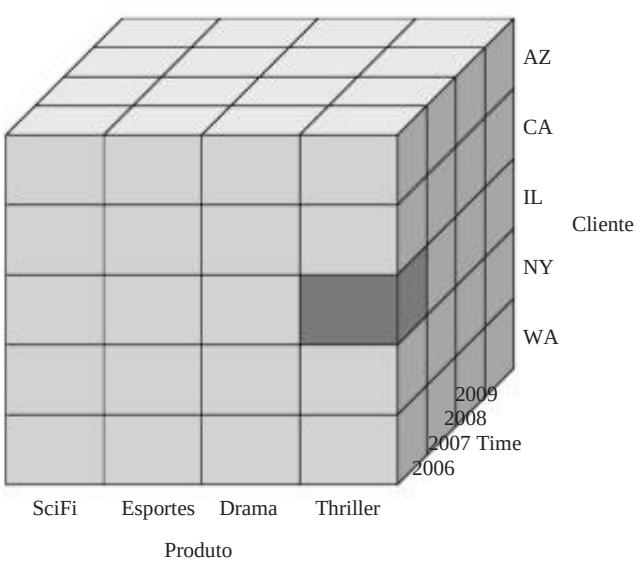


Figura 15-2: cubo tridimensional

O cubo mostrado na Figura 15-2 é composto por um tempo, produto e cliente
dimensão, cada uma delas colocada num dos lados do cubo. As dimensões são
a pedra angular de cubos OLAP: um cubo é simplesmente uma coleção de vários
dimensões.

Além dimensões com os dados de identificação, você precisa de algo a relatar, tais como vendas ou receitas de aluguel, custo, ou o número de aluguéis. Estes valores são chamada medidas. Um aspecto importante das medidas é que elas não repre- enviou um único valor. Ao contrário, eles representam um agregada valor. As medidas acabamos de mencionar são tipicamente agregados por soma. Em MDX, a recolha, ção de medidas que fazem um tipo especial de dimensão é chamada medida dimensão.

Agora, dê uma olhada no cubo da Figura 15-2, é claro, há mais de um produto, um cliente e um ponto no tempo. Na verdade, cada cliente, produto, eo dia em que o cubo tem uma intersecção chamada tupla. Cada tupla, que pode ser uma célula individual ou de uma secção no cubo, pode conter uma ou medidas mais.

NOTA Porque nem todo cliente compra todos os produtos disponíveis em cada dia, um lote de tuplas não existem fisicamente em um cubo OLAP. Isso é chamado esparsidade mas porque os dados são recuperados de um banco de dados relacional, onde apenas os dados existentes são armazenados, esparsidade não representa um problema.

Analogia Star Schema

Há uma analogia entre estes conceitos eo modelo MDX-dimensional conceitos de esquema ing e estrela discutido nos capítulos 6 e 7, respectivamente. A termo""dimensão refere-se ao mesmo conceito, em todos estes domínios. Um MDX cubo é análogo a um esquema de estrela, e as medidas são análogos aos fatos. Para clareza, Figura 15-3 mostra como um esquema em estrela simples podem ser mapeados para um cubo como o mostrado na Figura 15-2.

A correspondência entre o cubo mostrado na Figura 15-2 ea estrela esquema mostrado na Figura 15-3 é apenas uma consequência prática da estrela esquema de técnica de modelagem: embora seja possível organizar as coisas para que MDX cubos, dimensões e medidas correspondem diretamente a um esquema em estrela, tabelas de dimensão, e uma tabela de fatos, respectivamente, isso não é necessário ou implícita.

Mencionamos a analogia porque, na prática, alguns mapeamento que descreve como um cubo pode ser construído a partir dos dados no data warehouse sempre existe. No caso dos motores ROLAP, como Mondrian, esse mapeamento é realmente muito apertado, como os dados do cubo é construído on-the-fly, consultando a banco de dados.

Cubo Visualization

Não há limite prático para o número de dimensões que podem ser usados para construir um cubo. No entanto, a maioria das ferramentas que são projetadas para apresentar visual representações de cubos OLAP para os usuários finais podem apresentar apenas duas dimensões.

Normalmente, este assume a forma de um tabela cruzada, também conhecido como tabela de referência cruzada ou um tabela dinâmica.

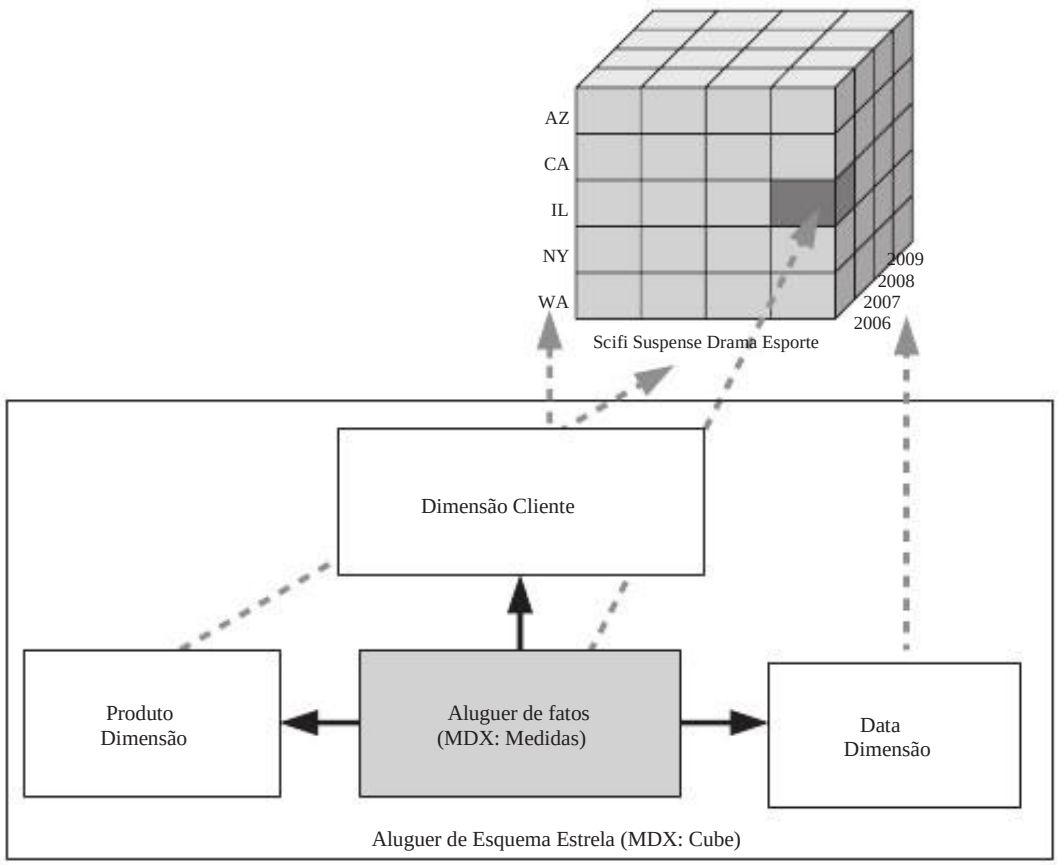


Figura 15-3: Um cubo como um esquema em estrela

No caso da tabela de referência cruzada, as duas dimensões da tela são horizontais e eixos verticais da tabela. Dimensões de um cubo multidimensional pode ser combinadas, produzindo dois conjuntos construídos a partir de uma ou mais dimensões. Estes podem ser mapeados para os dois eixos da tabela de referência cruzada. Figura 15-4 ilustra uma possível mapeamento do cubo mostrado na Figura 15-2 a uma tabela de referência cruzada.

Na Figura 15-4, o DVD e as dimensões são combinados Data e aparecem como colunas da tabela de referência cruzada (o eixo horizontal). A dimensão do cliente aparece como linhas na tabela de referência cruzada (o eixo vertical).

Hierarquias, níveis, e membros

Para o seu cubo para ser útil, você precisa mais do que apenas valores individuais no interseções dimensão individual. Você precisa encontrar uma maneira de agregar os dados entre as várias dimensões. Para este efeito, as dimensões são organizadas em uma ou mais hierarquias.

Hierarquias

Uma hierarquia é uma estrutura de árvore que pode ser usado para recuperar dados de o cubo em diferentes níveis de agregação. A maneira mais fácil e mais usado

exemplo é a dimensão de data com a hierarquia do Ano-de-mês-dia.
Para ilustrar o conceito da hierarquia e do MDX terminologia correspondente, Figura 15-5 mostra uma parte ampliada hierarquia Ano-de-mês-dia de um data de dimensão.

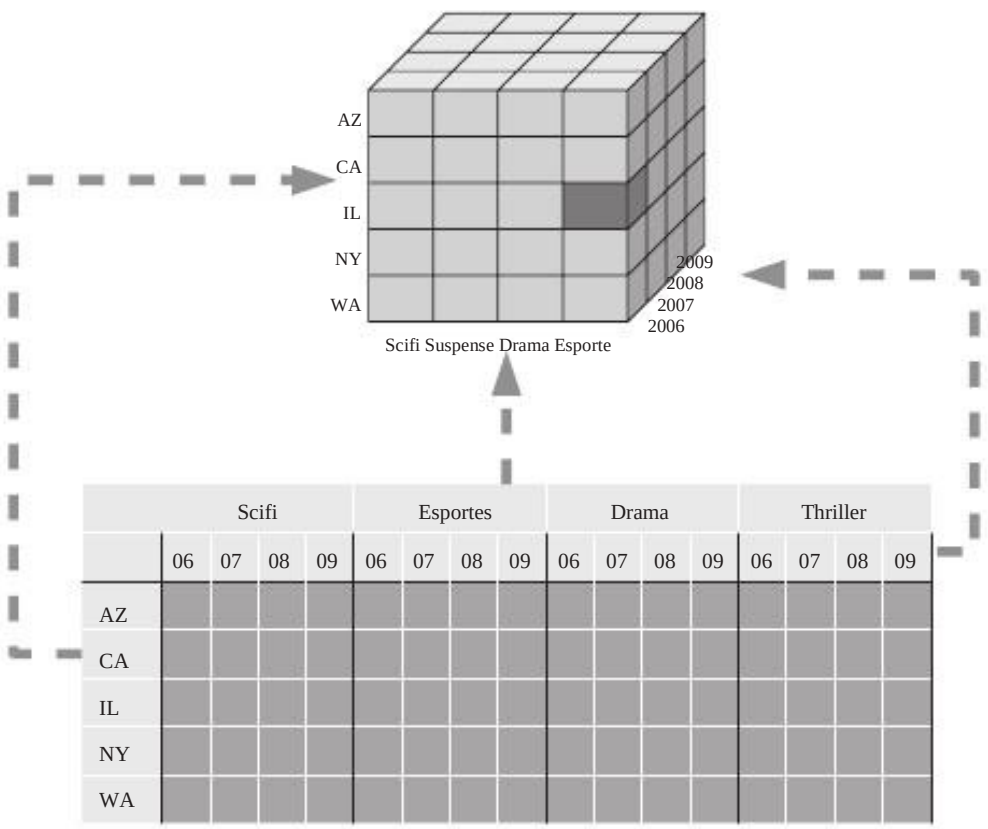


Figura 15-4: Referência cruzada a visualização de um cubo

Níveis e deputados

A hierarquia consiste níveis, e cada nível tem uma ou mais membros. Assim, em Figura 15-5, ano, mês, trimestre, e de dia são níveis e os itens nomeados dentro de cada nível são os membros. Por exemplo, os anos 2008, 2009 e 2010 são os membros ao nível do ano, e os quartos Q1 através Q4 são membros ao nível Quarter. A ordem dos níveis hierárquicos reflete a organização: Ano representa um nível superior de agregação de Bairro, e Mês representa um menor nível de agregação de Bairro.

Uma coisa importante a se notar é que os membros dentro de um nível não sobrepõem-se: dentro do nível do ano, os membros são todos diferentes um do outro. Mas o mesmo é verdadeiro para os membros abaixo do nível do ano: ambos os anos 2009 e 2010 pode ter um membro chamado Q1, a nível Bairro, mas, apesar com o mesmo nome, estes membros são distintos um do outro. Este garante que a agregação de dados ao longo de um nível produz um resultado consistente.

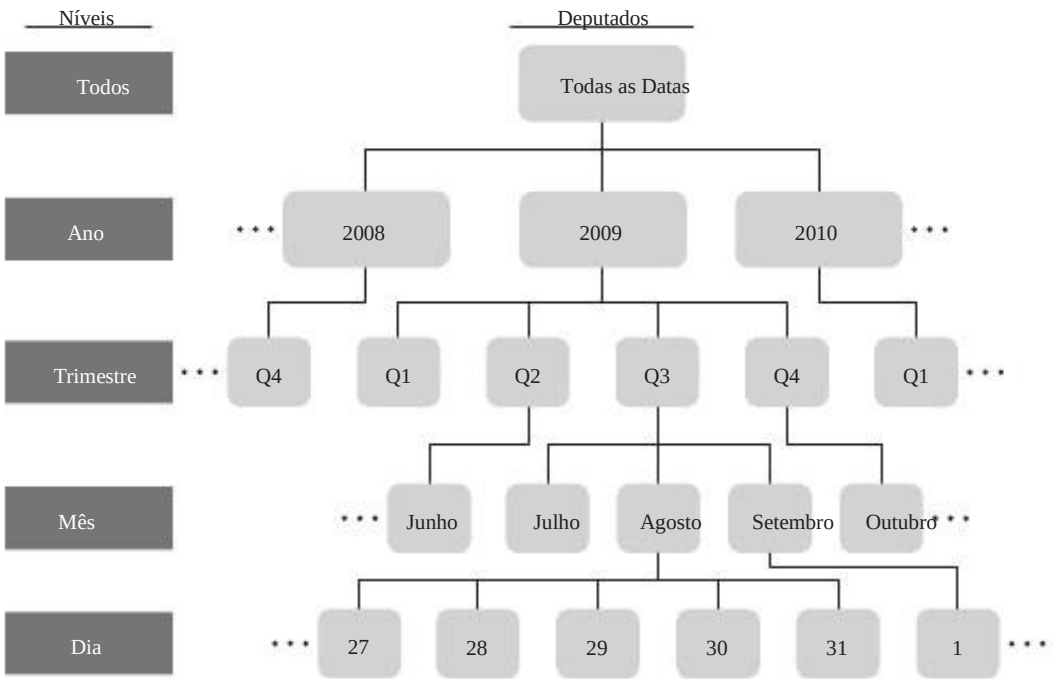


Figura 15-5: Um ano-de-mês-dia hierarquia em uma dimensão de data

Usando a sintaxe MDX, um membro pode ser sempre referenciados usando o caminho de membro de pleno direito constituído da dimensão, nível e nome do membro, como em [Data]. [Ano]. [2008]. Às vezes, os membros podem ser referenciados dentro de um hierarquia sem nomear explicitamente o nível que o membro pertence, isto é, enquanto o nome do membro é único. Assim, no exemplo anterior, [2008] funcionaria tão bem se este é um nome único.

DICA Uma das primeiras coisas que você vai perceber quando olhar para instruções MDX é o uso extensivo de colchetes. Isso não é necessário, mas se seu nomes dos membros conter espaços ou números, ou se o nome é uma palavra reservada MDX, usando o botão [] parênteses é obrigatório. A melhor prática é sempre usar os colchetes.

O nível de todos, todos os Estados e os Estados-padrão

Hierarquias podem ter uma especial todos os níveis para além do expressamente definido níveis. O nível de todas é a raiz conceitual da hierarquia. É especial porque não é explicitamente definido. Em vez do motor OLAP vez deriva-lo. É contém exatamente um membro, o todos os membros, que é construída através da unificação todos os membros pertencentes ao mais alto nível de agregação definido. Na Figura 15-5, o nível de todos é rotulado como Todos e Todos os membros é marcado todas as datas.

Hierarquias também pode especificar o membro padrão, que é usado quando os membros não são explicitamente especificado. Isto é importante observar que, como você geralmente não explicitamente todas as dimensões em uma consulta MDX. Normalmente, a todos os membros

denominação é utilizada como membro padrão. O efeito disto é que as medidas são automaticamente agregadas ao mais alto nível em todas as dimensões que não são explicitamente usada na consulta MDX.

Define-Membros

Para recuperar todos os membros de um nível, basta adicionar o MEMBROS palavra-chave ao nível, como no [Data]. [Meses]. [Ano]. MEMBROS. Esta expressão avalia a todos os membros do nível do ano como um membro do conjunto. Você também pode especificar explicitamente um conjunto
enumerando os nomes dos membros de uma lista separada por vírgulas em crespos suspensórios. Os colchetes indicam o conjunto, como no {[Ano]. [2007], [ano]. [2008], [Ano]. [2009]}.

Várias hierarquias

As dimensões podem conter múltiplas hierarquias, oferecendo múltiplos analítica ou caminhos de navegação dentro de uma dimensão. Normalmente, várias hierarquias servir negócios diferentes necessidades. Por exemplo, em uma dimensão de data, uma Trimestre-hierarquia mês-dia é usado para análise de vendas, e um adicional de Ano-de-semana, Dia da hierarquia pode ser usado para executar a análise de ordem de compra semanal. hierarquias múltiplas para a mesma dimensão podem ter alguns níveis comum: neste caso específico, o nível do ano tem o mesmo significado em ambas as hierarquias. No entanto, as definições nível das hierarquias diferem em algum lugar no line-isto é, de fato, a ponto de ter várias hierarquias. Por exemplo, o duas hierarquias data diferem uma da outra para todos os níveis abaixo do nível do ano.

Em MDX você pode atribuir uma hierarquia específica a ser parte do di-Sion referência. Por exemplo, suponha que você tenha um Data dimensão com duas hierarquias separadas: uma para o ano-de-mês-dia e outra para Ano-Dia-de-semana, chamado Meses e Semanas , respectivamente. Neste caso, a ex-Sion [Date.Months]. [2000]. [1] [1]. refere-se a 1 mês em um trimestre no ano 2000, enquanto que [Date.Weeks]. [2000]. [1] [1]. refere-se ao 1 º dia da 1 ª semana do ano de 2000. Observe como o nome da hierarquia é anexado ao nome da dimensão diretamente, usando a notação de ponto dentro dos colchetes.

Se você omitir uma especificação de hierarquia, a hierarquia padrão, que é o primeiro hierarquia especificada para a dimensão, é usado.

Relacionamentos Familiares Cube

Dentro de cada hierarquia em uma dimensão de cubo a ordem dos membros que constituem-bros é predefinido automaticamente com base na sua natural ordem alfabética, numérica ou ordenação temporal. Os níveis dentro de uma hierarquia não são automaticamente ordenada. Eles precisam ser colocados em uma certa ordem, indo de um resumo a um nível detalhado. Cada membro de um nível inferior na hierarquia podem pertencer a

apenas um membro de nível superior. Como exemplo, pense de produtos: um produto pode apenas fazer parte de um grupo de produtos, mas um grupo de produtos pode conter vários produtos. Para começar com o fim dos níveis, as relações entre diferentes níveis é chamado de pai-filho relacionamento. Em uma dimensão de data, o nível do ano é o pai do nível do Bairro, que por sua vez, é o nível criança de ano.

Ele fica mais interessante quando você olha para os membros do nível. Agora, o ano de 2009 tem quatro filhos, Q1, Q2, Q3 e Q4, e de cada trimestre, por sua vez tem três filhos, Jan, Fev, Mar e para o Q1, e assim por diante. Cada membro também pode ter irmãos e primos. Os irmãos são membros relacionados no mesmo nível, mas com o outro progenitor, eles só podem ser abordadas, apontando para o primeiro ou o último irmão ou por recuperar todos os irmãos. A posição de um primo no nível correspondente para o membro de referência. Suponha que você quer olhar para o primo Q1 2009 em 2008, o membro resultante será 1 º trimestre de 2008. A semelhança entre filhos e irmãos, por um lado, e primos, por outro, é que as crianças e os irmãos só podem ser tratadas como um grupo ou especificando explicitamente o membros do primeiro ou último, mas primos precisam ser abordadas, especificando o pai.

Muitas mais maneiras de referenciar células relacionadas estão disponíveis, você pode, por exemplo, usar a palavra-chave descendentes para recuperar um conjunto de membros que desce do membro de partida. Isto também vai exigir que você adicione o distância para o membro pai, então a função do MDX Descendentes ([2008], 2) recupera todos os meses de 2008, caso a hierarquia é o Ano-de-mês. Neste caso, o nível do mês é a dois passos a partir do nível do ano.

Relativo relações de tempo

Quando você está trabalhando com uma dimensão de tempo, funções especiais relativos são disponíveis, que são baseadas em relações familiares entre os membros anteriores dentro de uma hierarquia. Na maioria das soluções de BI, os utilizadores estão interessados em comparar um período com o mesmo período no ano anterior. Embora você possa usar o primo função para chegar a este período de comparação, o ParallelPeriod função é mais adequada para isso. A função recebe três argumentos: a hierarquia número do nível dos períodos de volta, e membro da dimensão do tempo. Por exemplo, Se você está interessado no período de outubro de 2009 em paralelo nos últimos ano, você usaria ParallelPeriod ([Ano], 1 [Data]. [Mês]. [200910]), que retornaria [200810].

Year-to-date é também uma característica essencial das soluções de BI, e os Acumulado do ano função fornece essa informação. De cada período de um ano, se você está lidando com os trimestres ou meses, você pode recuperar o Acumulado do ano membros muito facilmente. Quando usar Acumulado do ano ([Time]. [Mês]. [200903]), Os membros [200901],[200902], e [200903] são retornados como um conjunto.

A sintaxe MDX Query

À primeira vista, pode parecer familiar MDX se você já conhece o SQL
linguagem. As aparências enganam, no entanto, existem algumas semelhanças, mas o
diferenças são ainda maiores.

Basic Consulta MDX

Assim como você conseguir dados de seu cubo em uma resultado MDX? Vamos começar com
o mais simples de consultas simples MDX:

```
SELECT
DA wcm_sales
```

O resultado da consulta dessa consulta é o total do cubo da medida padrão
da wcm_orders cubo. Claro, isso não é muito útil e só é bom
para testar se funciona sua conexão, mas este exemplo ilustra-se
trar alguns conceitos-chave do MDX, então vamos dar uma olhada neste simples
declaração.

O cubo a partir do qual os dados a serem recuperados é especificado no MDX
DA cláusula. Como tal, é algo análogo ao SQL DA cláusula, que
especifica uma tabela ou intermediário juntar o resultado que fornece os dados para o resto da
a consulta.

A SELECT palavra-chave é especificada, mas porque ele não é seguido por um
lista de expressões, a medida padrão é usado. Neste caso, a medida padrão é
Receita. Porque as medidas são agregadas, essa consulta retorna a soma de todos
receitas do cubo chamado wcm_orders.

Eixos: em linhas e ON COLUNAS

MDX pode representar informações sobre vários eixos. Geralmente, esses são os ROWS
e COLUNAS eixos. Outros eixos padrão PÁGINAS,SEÇÕESE CAPÍTULOS,
mas a maioria das ferramentas de visualização OLAP (incluindo JPivot) não pode trabalhar com eixos
além das duas primeiras, muito menos a 128 eixos possível o padrão MDX
permite.

Na maioria dos casos, você vai trabalhar com apenas dois conjuntos de membros de cada vez, uma
exibido no eixo de colunas e um no eixo linhas. A expressão MDX
portanto, usa a seguinte sintaxe genérica:

```
SELECT <member collection> em colunas,
    <member collection> ON ROWS
DA <cubename>
```

Na verdade, as palavras-chave ROWS e COLUNAS são sinônimos para o padrão numeradas notação, COLUNAS é a abreviatura para AXIS (0)E ROWS representa EIXO (1). Você pode até omitir a palavra-chave AXIS e simplesmente usar 0 e 1. Assim, no exemplo anterior também pode ser escrita como:

```
SELECT <member collection> ON AXIS (0),
    <member collection> EM 1
DA <cubename>
```

Embora a ordem real das especificações eixo é irrelevante, deve especificar uma série sem intervalos consecutivos de eixos, a começar AXIS (0) (O COLUNAS eixo). Por isso, é possível especificar apenas o primeiro EM COLUNAS linha, mas você não pode ignorá-lo e só usar ON ROWS. Embora você poderia inverter a ordem dos e ON ROWS linhas, seria confuso, é por isso que não recomendamos o uso dessa ordenação.

Para obter uma tela vertical de linhas somente com uma única medida, use o seguinte consulta:

```
SELECT [Measures]. [Receita] em colunas,
    <member collection> ON ROWS
DA <cubename>
```

Olhando para uma parte dos dados

Você pode se perguntar se uma consulta MDX também pode conter um ONDE cláusula, e na verdade ele pode. É o chamado fatiador porque restringe os dados que pode ser acessado através do cubo para um subconjunto específico (a fatia). Embora o ONDE palavra-chave é o mesmo que no SQL regular, o comportamento é bastante diferente, como o nome""slicer implica. A consulta MDX a seguir os limites da análise a o ano de 2008, usando um cortador de:

```
SELECT [Measures]. [Receita] em colunas,
    <member collection> ON ROWS
DA <cubename>
WHERE [Data]. [Ano]. [2008]
```

Assim, em vez de especificar uma condição, como se estivesse em SQL, a MDX ONDE cláusula exige que você especifique a parcela dos dados que você gostaria para incluir em termos de uma especificação membro da dimensão. Você pode especificar uma lista separada por vírgulas de nomes de membros desde que você coloque a lista parênteses. Assim, a consulta MDX seguintes limites o resultado para o ano de 2008 e os clientes do Canadá:

```
SELECT [Measures]. [Receita] ON COLUNAS
DA <cubename>
WHERE ([Data]. [Ano]. [2007], [cliente]. [No país]. [CA])
```

Dimensão em apenas um eixo

E agora a má notícia: você não pode usar a mesma dimensão em mais de um eixo. Por exemplo, se você usar a dimensão de data já ON ROWS ou ON COLUNAS, Não pode ser usado no fatiador. A consulta a seguir, portanto, não:

```
SELECT [Measures]. [Receita] em colunas,
    [Data]. [Bairro]. Deputados ON ROWS
DA <cubename>
ONDE [Data]. [Ano]. [2008]
```

A maioria das pessoas usadas para SQL achar isto difícil de entender, mas é simplesmente uma dos meandros MDX que você vai ter que se acostumar. Se você só quer olhar trimestres em 2008, você pode fazer isso usando o CRIANÇAS palavra-chave o membro 2008 do nível do ano:

```
SELECT [Measures]. [Receita] em colunas,
    [Data]. [2008]. Crianças ON ROWS
DA <cubename>
```

MDX permite a utilização de múltiplas dimensões em um eixo, mas lembre-se que uma dimensão particular pode aparecer em apenas um dos eixos (ou o slicer).

Mais exemplos MDX: um simples cubo

Para o restante da explicação MDX e exemplos, é melhor ter um modelo específico. Nos parágrafos seguintes, vamos construir o wcm_orders cubo, que tem a estrutura conforme listado na Tabela 15-1.

Você vai aprender como criar um cubo Mondrian para este modelo mais tarde, mas por agora usá-lo para continuar explorando MDX.

A função FILTER

A FILTRO função permite-lhe limitar a escolha dos membros do cubo. Você pode usar o ONDE cláusula apenas a uma parte da fatia do cubo, mas o que se quiser para limitar os resultados de títulos de DVD que têm receita acima de R $ 10.000 em 2008, ou consulte apenas os títulos dos filmes que mostraram um aumento de 10 por cento na receita em relação ao mês anterior? Estes são apenas dois exemplos onde a FILTRO função pode ser útil. FILTRO é uma função MDX que usa dois argumentos: um conjunto e um pesquisa condição. O primeiro exemplo traduz a declaração MDX a seguir:

```
SELECT [no país]. Membros em colunas,
    FILTRO (
      [Título]. Associados,
      [Receita]> 10000
    ) ON ROWS
[Wcm_orders] FROM
WHERE [ano]. [2008]
```

A instrução anterior inclui um filtro e um fatiador. Este último assume precedência sobre os outros elementos, o que significa que o filtro eo conjunto de resultados são limitadas a 2008 a fatia do cubo.

Tabela 15-1: WCM_orders_cube

DIMENSÃO	HIERARQUIA	NÍVEL
Data da Ordem local	Meses	Ano Trimestre Mês Data
Data da Ordem local	Semanas	Ano Semana Data
Cliente		País Região Cidade CEP Nome
DVD		Gênero Título
Medidas		Receita Quantidade RentalDuration

A função ORDEM

Se você executar a consulta do exemplo anterior, os resultados mantêm a sua natural ordem, que está em ordem alfabética dentro da hierarquia. Isso significa que o filme títulos são ordenados alfabeticamente, mas o gênero que sempre precede a ordenação. (Mais tarde neste capítulo, quando explicamos como construir o cubo, você aprenderá como para controlar a ordem de membro de dimensão na definição do cubo).

Em muitos casos, você quer resultados ordem baseada em uma medida, por exemplo, talvez você queira ordenar os resultados em faturamento, de alto a baixo. Você usa o ORDEM função para realizar exatamente isso. A ORDEM função recebe três parâmetros: o conjunto de ser ordenada, a expressão a fim de, eo tipo ordem, o que pode levar de quatro valores possíveis: ASC,DESC,BASCE BDESC. A Bnos últimos dois valores forças para romper as hierarquias para permitir uma total ordem de classificação, por isso, se você quer os resultados do exemplo anterior classificada em ordem decrescente, use a seguinte consulta:

```
SELECT [no país]. Membros em colunas,
    DESPACHO (
      FILTRO ([título]. Associados, [Receita]> 10000),
      [Receita]
```

```
        BDESC
    )ON ROWS
    [Wcm_orders]
DA  [Ano]. [2008]
ONDE
```

Agora, os resultados serão classificados com os filmes mais vendidos no topo da da lista.

Usando TopCount e BOTTOMCOUNT

Não há outro caminho para chegar aos melhores filmes (ou pior) que vendem, e que é por usando o TopCount e BOTTOMCOUNT funções. Essas funções são realmente um abreviação para o uso do CABEÇA e CAUDA funções que podem ser usados para limitar um resultado conjunto ordenado como o do exemplo anterior, mas fazer a consulta de um muito mais simples de escrever.

A TopCount e BOTTOMCOUNT funções usam três argumentos: um conjunto, o número de membros a serem exibidos, e à medida de ordem no. Aqui é a consulta que recupera os 10 melhores filmes de 2008:

```
SELECT [no país]. Membros em colunas,
      TopCount (
        [Título]. Deputados
      , 10
      [Receita]
      ) ON ROWS
[Wcm_orders] FROM
WHERE [ano]. [2008]
```

Combinando Dimensões: A função Crossjoin

MDX não se limita à exibição de uma dimensão para cada eixo, mas para colocar multi-plo dimensões em um único eixo, você deve usar o função. Este função é semelhante à CROSS JOIN tipo de associação no SQL simples e cria uma combi-nação de todos os valores a partir de duas dimensões. Por exemplo, usando Crossjoin, Você pode colocar os dois [Cliente]. [País] e [DVD]. [Gênero] no mesmo eixo, e uso [Data Local Ordem]. [Ano] e [Measures]. [Quantidade] sobre os outros:

```
SELECT Crossjoin (
        {[Data Local Ordem]. [Ano].} Deputados
      {[Measures]. [Quantidade]}
      ) Em colunas,
       Crossjoin (
        {[Cliente]. [No país].} Deputados
      {[DVD]. [Gênero].} Deputados
      ) ON ROWS
[Wcm_orders] FROM
```

Usando não vazia

Um comumente usado na construção de queries MDX é Não vazio. Ele força a consulta para retornar apenas os resultados que contêm um valor. Porque é altamente improvável que todos os

cliente compra todos os produtos em cada dia do ano, o MDX média
consulta retorna uma grande quantidade de resultados vazio que você começar a perfurar abaixo em detalhes.
A Não vazio frase pode ser colocada na frente do conjunto em qualquer consulta dimensões. Quando um Crossjoin é usado, o Não vazio palavra-chave é colocada em frente, como no exemplo a seguir:

```
SELECT Não vazio Crossjoin ({[Data Local Ordem]. [Ano].} Membros,
{[Measures]. [Quantidade]}) EM COLUNAS
```

Na verdade, este é o comportamento padrão do JPivot na plataforma Web Pentaho. A barra de ferramentas JPivot contém um botão rotulado Suprimir linhas vazias / Colunas que faz com que as consultas MDX gerado para incluir Não vazio. Por padrão, esse recurso é ativado.

Trabalhando com conjuntos e a cláusula WITH

No início deste capítulo, fazemos uma breve usou o termo conjuntos e as chavetas necessário especificá-los. Um conjunto é uma coleção de membros de uma dimensão e é geralmente definida implicitamente. Por exemplo, quando você usa [Ano]. MEMBROS em um dos eixos, você está usando um conjunto, porque a MEMBROS operador retorna um conjunto. De fato, muitas das funções MDX que você viu até agora retornar conjuntos: CRIANÇAS, IRMÃOS,DESCENDENTESE Acumulado do ano todos se comportam assim.

O problema com essas funções é que elas sempre voltam a completa coleção de membros dentro de um nível. Às vezes, isso não é conveniente para fazer uma análise específica por parte apenas de uma dimensão. Usando o COM cláusula, você pode criar seus próprios jogos para uso em consultas MDX. Suponha que você queira combinar alguns gêneros de filme em uma CoolMovies conjunto que contém apenas alguns dos
os gêneros no banco de dados. CoolMovies não é uma entidade reconhecida no seu modelo, mas você pode criá-la de qualquer jeito usando a seguinte sintaxe:

```
COM SET [o conjunto de nome] como '{definição}'
```

Traduzido em um exemplo CoolMovies, este se tornaria

```
COM
SET [CoolMovies]
AS '{[DVD]. [All DVD]. [Ação / Aventura], [vídeo]. [All DVD]. [Fantasia]
   [DVD]. [Todos DVD]. [SciFi], [vídeo]. [Todos DVD]. [Suspense / Thriller]
   [DVD]. [Todos DVD]. [Suspense]} '
SELECT não vazio {[CoolMovies]} em colunas,
     NÃO Crossjoin EMPTY ([Data Local Ordem]. [Todos Ordem Local Datas]
                [Cliente]. [Todos os Clientes]) ON ROWS
[Wcm_orders] FROM
WHERE [Measures]. [Quantidade]
```

Como você pode ver no exemplo anterior, você precisa usar chaves em torno de seu conjunto nomeado quando você usá-lo no SELECT declaração, sob pena de

não é reconhecido como um conjunto e um erro é gerado. Observe também que foram incluídos
a definição de membro de pleno direito nos conjuntos. Nesse caso, esse não teria sido necessário porque o membro [Suspense] e os outros gêneros são únicos
em nosso cubo, de modo {[Ação / Aventura] [Fantasia], [ficção científica], Suspense [/ Suspense], [] Thriller} funciona tão bem.

Usando membros calculados

A última parte deste primer MDX abrange a criação de membros calculados. A cal-
membro calculada é um membro de dimensão cujo valor é calculado em tempo de execução usando uma expressão especificada e pode ser definida como membros de dimensão regular-sões ou como membros da dimensão de medidas. Mondrian permite que você
adicionar membros calculados diretamente no esquema, mas esse recurso usa stan-
Dard sintaxe MDX calculado também. Somente definições são armazenadas porque o valor da expressão é determinada em tempo de execução. Como conseqüência, calculado membros não ocupam espaço em um cubo, mas que exigem poder de computação extra.

Um exemplo muito simples de um membro calculado é uma pequena variação no CoolMovies conjunto nomeado que você já viu antes:

```
COM
MEMBROS [Measures]. [CoolMoviesRevenue]
AS '[Ação / Aventura] + [Fantasia] + [SciFi] + [Suspense / Thriller] + [Suspense] "
SELECT vazio {[Measures] NÃO. [CoolMoviesRevenue]
            [Measures]. [Receita]} em colunas,
     NON EMPTY {[no país].} Deputados ON ROWS
[Wcm_orders] FROM
WHERE [ano]. [2008]
```

Você também pode usar os membros calculados para os totais dos membros do grupo, por exemplo
para criar uma coleção recente e histórica de anos:

```
COM
MEMBROS [Ordem Local Data]. [Todas as datas]. [Histórico]
AS      "SUM ([Ano] [2000]:. [Ano] [2005]). '
MEMBROS [Ordem Local Data]. [Todas as datas]. [Recentes]
AS      "SUM ([Ano] [2006]:. [Ano] [2010]). '
SELECT  NON EMPTY {[Ordem Local Data]. [Todas as datas]. [Histórico]
                   [Ordem Local Data]. [Todas as datas]. [Recentes]} em colunas,
        NON EMPTY {[no país].} Deputados ON ROWS
        [Wcm_orders]

DA
```

Apesar de existirem inúmeros outros possíveis usos e exemplos de cálculo membros, isso lhe dá uma idéia de como você pode usá-los para ampliar o
o conteúdo disponível de um cubo. Para mais informações sobre Mondrian específicas MDX, consulte a documentação on-line em http://mondrian.pentaho.org/documentação / mdx.php. E porque MDX foi originalmente desenvolvido pela

Microsoft, um número enorme de referências on-line e os livros estão disponíveis neste assunto, embora muito poucos livros são dedicados a MDX sozinho.

ATENÇÃO Desde a sua criação e adoção por muitos fornecedores OLAP, Microsoft acrescentou várias extensões à especificação MDX. Algumas funções e exemplos pode não funcionar com Mondrian. Quando em dúvida, verifique a linha Mondrian MDX referência em http://mondrian.pentaho.org/documentation/schema.php.

Criando esquemas Mondrian

Nós já discutimos o esquema e como se relaciona com os diferentes componentes da Pentaho Analysis Services quando discutimos Figura 15-1. Nesta seção, vamos mostrar como criar esquemas de Mondrian.

Embora esta seção fornece informações detalhadas sobre a criação de Mon- esquemas drian, não podemos cobrir todos os ângulos possíveis. Para uma completa ref- cia, você deve consultar a documentação oficial. Você pode encontrá-lo em http://mondrian.pentaho.org/documentation/schema.php.

Começando com Pentaho esquema Workbench

Nós mencionamos que os esquemas de Mondrian são definidos como documentos XML e que Pentaho esquema Workbench (PSW) oferece uma interface gráfica do usuário para editar esses esquemas. Também mencionou que é possível editar os esquemas manualmente utilizando um editor XML, IDE, ou até mesmo um editor de texto simples.

Ambas as abordagens são igualmente válidas e úteis no seu próprio caminho. Para essa razão, nós pagamos a atenção para PSW, bem como para o formato XML usado para denotam esquemas Mondrian.

Baixando Mondrian

Mondrian é mantido em seu próprio projeto de www.sourceforge.net. Você pode encontrá-lo no site SourceForge simplesmente procurando por Mondrian. A site do projeto Mondrian fornece os binários e código-fonte do Mondrian motor em si, bem como Pentaho esquema Workbench eo agregado Pentaho Designer. Por enquanto, você precisa baixar somente Pentaho Schema Workbench. Mais tarde, você também vai ter um olhar para o designer de agregação, então você pode bem baixe-o agora, também.

NOTA Você não precisa fazer o download do software do motor Mondrian em si: esta é já incluídos no Pentaho BI Server. Você precisa baixar o motor em si só se você quiser atualizar o mecanismo de Mondrian ou implantar Mondrian separadamente sem Servidor Pentaho. No entanto, estes casos de uso estão fora do escopo deste livro, e não são discutidos aqui.

Esquema de Instalação do Pentaho Workbench

Pentaho Schema Workbench é distribuído como Zip. e . Tar.gz arquivos.
Os usuários do Windows devem baixar a Zip. arquivo, e usuários de baseados em UNIX
sistema deve começar a . Tar.gz arquivo. Após o download, você precisará descompactá
o arquivo. Isso produz um único diretório chamado esquema da bancada contendo todos os
o software. Nós nos referimos a este diretório como o diretório home PSW ou simplesmente
PSW casa. Você deve mover o diretório para algum lugar que faça sentido para
seu sistema. Por exemplo, usuários do Windows pode movê-la para C: \ Arquivos de Programas
e usuários do Linux pode movê-la para / Opt / ou o diretório home do usuário atual.

Depois de desempacotar PSW, você precisa colocar qualquer driver JDBC Jar. arquivos que
você pode ter de ligar para o data warehouse no motoristas diretório. Este
diretório podem ser encontrados imediatamente abaixo do diretório home PSW.

A partir do esquema Pentaho Workbench

Pentaho Schema Workbench é iniciado com um shell script. Shell scripts para di-
diferentes plataformas estão localizadas diretamente no diretório home PSW. No Windows
sistemas, você precisa dar um duplo clique no workbench.bat arquivo para iniciar PSW. Em um
sistema baseado em UNIX, você precisa iniciar o workbench.sh script. Você pode precisar
para fazer o workbench.sh script executável em primeiro lugar. Se você estiver executando uma
gráfica
desktop como o GNOME, você pode fazer isso simplesmente clicando com o script
e escolha Propriedades. De lá você pode tornar o arquivo executável. Alterna-
mente, você pode torná-lo executável a partir do terminal, usando o chmod comando:

```
chmod ug + x workbench.sh
```

Depois de iniciar o script, você verá a janela do aplicativo PSW (ver
Figura 15-6).

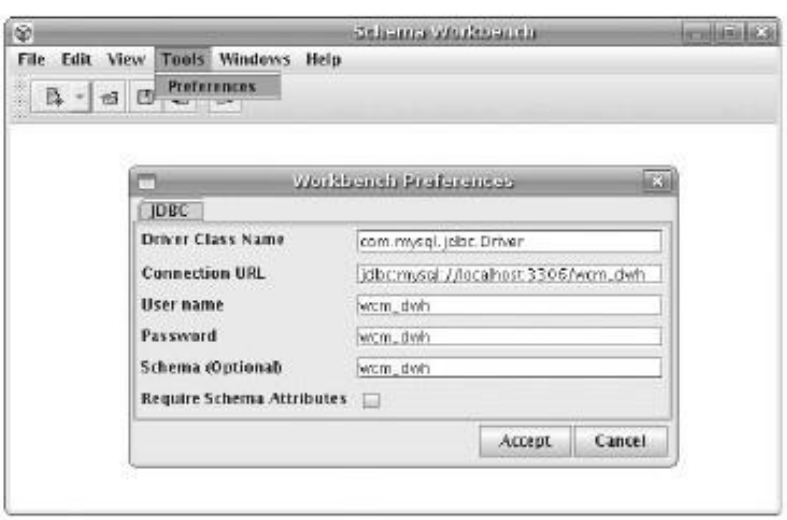


Figura 15-6: O Pentaho esquema janela do aplicativo Workbench

Estabelecendo uma conexão

Depois de iniciar PSW, você deve estabelecer uma conexão com o banco de dados. Você pode abrir a janela de conexão com o banco através do menu principal Preferências. Alternativamente, você pode pressionar as ferramentas Preferenceschoosing botão da barra, que está localizado na extremidade direita da barra de ferramentas do aplicativo.
Uma janela aparecerá onde você pode preencher os parâmetros de conexão JDBC. O menu de contexto e de diálogo são mostrados na Figura 15-6.
Você precisa preencher as seguintes propriedades na caixa de diálogo:

- nome do driver de classe presente é o nome da classe Java do driver JDBC que você será utilizado. Para o MySQL, este é com.mysql.jdbc.Driver.
- URL de Conexão-Este é o seqdeconexao usado para manter contato com o banco de dados servidor. Supondo que você deseja se conectar ao wcm_dwh banco de dados em um local MySQL instância em execução na porta padrão 3306, A URL é jdbc: mysql: / / Localhost: 3306/wcm_dwh.
- Nome de usuário e senha A credenciais do usuário do banco de dados que conecta ao banco de dados.

Após o preenchimento do diálogo, você pode clicar no botão Aceitar para estabelecer a conexão. Se uma caixa de mensagem aparece informando que o banco de dados conexão não pôde ser estabelecida, você deve verificar os parâmetros que você sup-dobraram. No entanto, mesmo se você especificou os parâmetros corretos, você ainda pode receber uma mensagem informando que a classe driver não pode ser encontrada (ver Figura 15-7).

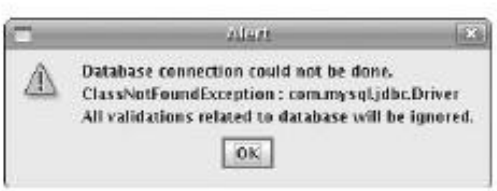


Figura 15-7: A mensagem de erro indicando o driver JDBC não foi carregado

Neste caso, você deve ter certeza que você colocou o Jar. arquivo contendo o driver JDBC que você precisa no motoristas sob o diretório home PSW diretório (como mencionado anteriormente, na subseção de instalação). Você precisa PSW reiniciar para pegar qualquer novo Jar. arquivos que você colocou no motoristas diretório.

NOTA Se você não conseguir estabelecer uma conexão com o banco, você ainda pode usar o PSW para definir um esquema. No entanto, o processo será um pouco mais difícil porque os recursos tais como caixas de lista suspensa para escolher as tabelas e colunas não vai funcionar.

JDBC Explorer

Depois de estabelecer a conexão, você pode abrir uma janela do Explorer para JDBC ver o conteúdo do banco de dados. Você pode abrir o Explorer, escolhendo JDBC File Explorer New JDBC. Isso é mostrado na Figura 15-8.

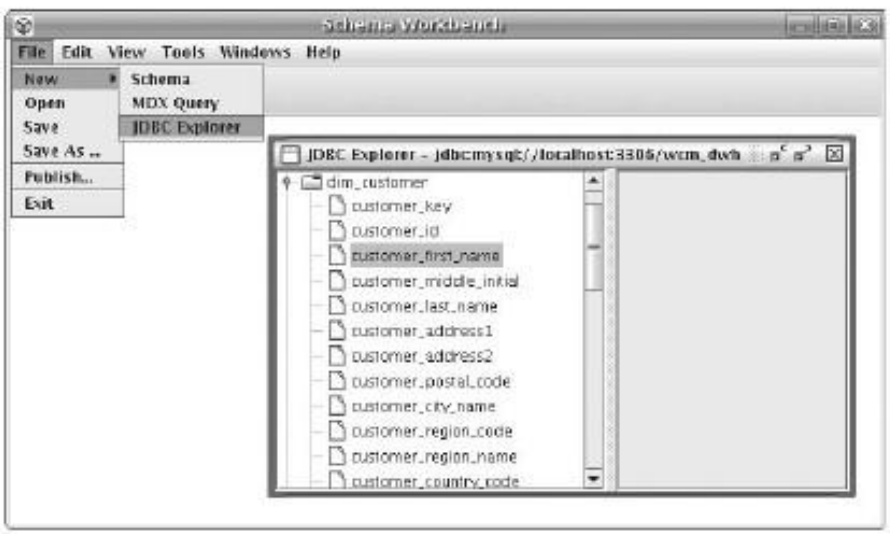


Figura 15-8: Abrindo o Explorer JDBC

O Explorer JDBC consiste de uma árvore que exibe as tabelas que você pode acesso a partir da conexão atual como pastas. Você pode expandir as pastas para veja o que as colunas da tabela contém. O Explorer não oferece nenhuma funcionalidade JDBC além disso. No entanto, como você verá mais adiante, este é apenas o suficiente para torná-lo um pouco mais fácil para construir seus cubos.

Usando o editor de esquema

Os esquemas são criados e editados usando o editor do esquema. Nesta subseção, vamos destacar brevemente algumas das características do editor do esquema.

Criando um novo esquema

Use o menu Arquivo e escolha Novo esquema para abrir o editor do esquema. O editor do esquema tem uma árvore no lado esquerdo, mostrando o conteúdo da esquema. Inicialmente, isso será praticamente vazia, exceto para o nó do esquema, que é a raiz de todo o esquema. No lado direito, o editor de esquema tem um espaço de trabalho onde você pode editar elementos no esquema. O editor do esquema é mostrado na Figura 15-9.

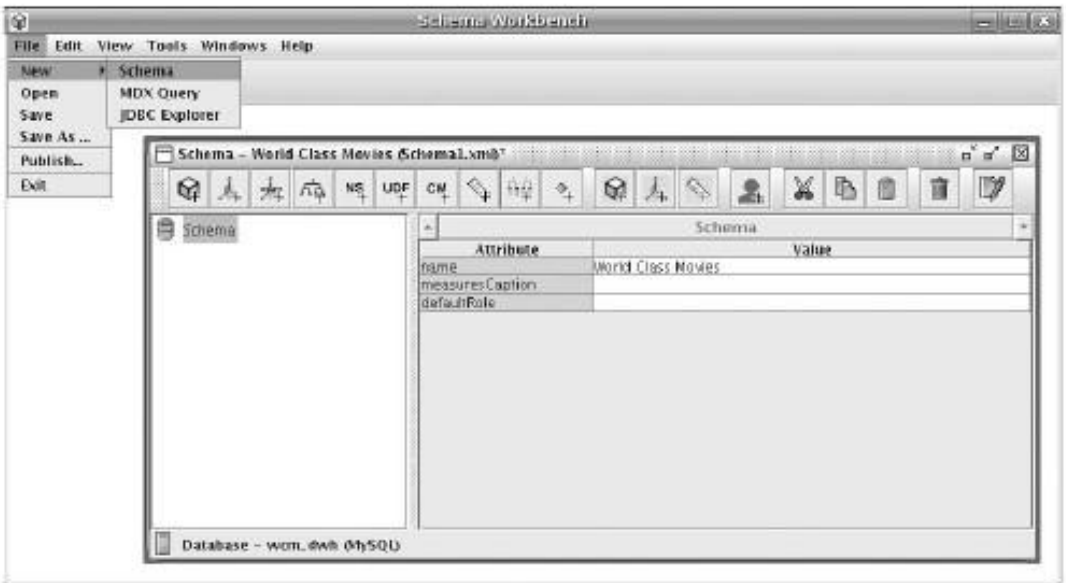


Figura 15-9: O editor do esquema

No topo do editor do esquema uma barra de ferramentas contém diversos botões. A maioria dos
destes botões estão lá para que você possa adicionar novos itens para o esquema. Todas essas
botões residem no lado esquerdo da barra de ferramentas, e você vai reconhecê-los por
no sinal de mais (+) sinal no ícone de botão. No lado esquerdo da barra de ferramentas são algumas
botões de Recortar, Copiar, Colar e Apagar. Você estará usando um destes botões
muito no restante desta seção, mas se você quiser saber mais agora
sobre esses botões, basta passar o ponteiro do mouse sobre os botões para ler as
dica.

Além dos botões da barra de ferramentas, você pode usar o menu de contexto do
exibição de árvore para realizar as mesmas tarefas. Simplesmente o botão direito do mouse em qualquer nó da árvore
vista. O menu de contexto irá aparecer e lhe oferecer todas as ações disponíveis para
esse nó particular.

Salvando o esquema em disco

Depois de criar um esquema, é uma boa idéia para salvá-lo. Basta clicar no ícone do disquete
na barra de ferramentas da janela do aplicativo e você será solicitado a especificar
um local. Por agora, basta escolher o nome eo local que você encontrar
conveniente, por exemplo, Schema1.xml. Seu desktop é uma escolha razoável
enquanto você está trabalhando com este capítulo.

Durante o restante deste capítulo, não vamos mencionar que você deve
salvar o seu trabalho o tempo todo. Em vez disso, você deve decidir por si mesmo quando e
se você deve salvar seu trabalho. Ao desenvolver um esquema, é uma boa

idéia de usar um sistema de controle de versão como o SVN para controlar as mudanças ou reverter para uma versão anterior caso você precise.

Edição de objeto Atributos

Para editar um esquema especial de itens, tais como o esquema em si, um único clique no item na exibição de árvore para selecioná-lo. Os itens selecionados estão destacadas em azul. Aviso
que este é o caso do esquema mostrado na Figura 15-9.

A área de trabalho automaticamente mostra os atributos disponíveis para os selecionados item em uma grade de atributos. A grade de atributos tem duas colunas: a primeira coluna Atributo é rotulado e lista os nomes de atributos, a segunda coluna é rotulado Valor e contém o valor do atributo. A grade de atributos é também visível na Figura 15-9.

Para saber mais sobre um determinado atributo, passe o ponteiro do mouse sobre o nome do atributo. Uma dica aparecerá, exibindo um texto que descreve o finalidade do atributo, bem como o valor padrão para o atributo.

Para editar um valor de atributo, coloque o cursor na coluna valor da atributos da rede e digite o valor desejado. Por exemplo, na Figura 15-9 nós entrou World Class Filmes como o nome desse esquema.

Alterar Edit Mode

O esquema é representado como um documento XML. Se você gosta, você pode alternar para inspecionar a representação XML do esquema, a qualquer momento, alternando o modo de edição. Você pode fazer isso clicando no botão no lado direito da barra de ferramentas (ver Figura 15-10).

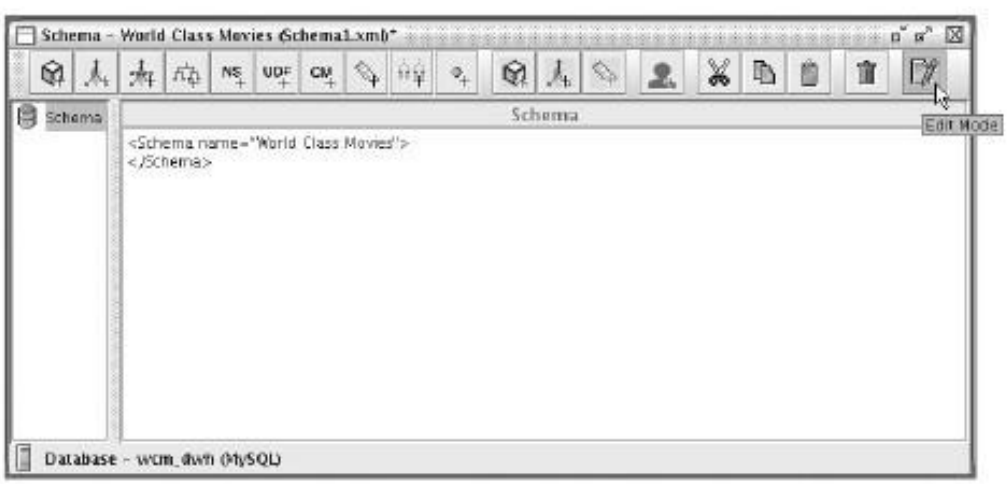


Figura 15-10: Usando o modo de edição para ler a representação XML do esquema

O XML exibida na área de trabalho corresponde exatamente com o selecionado item na exibição de árvore, que neste caso é o próprio esquema. Como você pode ver, em seu estado atual a representação XML do esquema consiste somente nos abertura e fechamento de tags <Schema> elemento.

NOTA Cuidado que a visão XML-que é apenas uma visão. Você não pode editar o XML diretamente de dentro do PSW.

Criação e edição de um esquema básico

Nesta subseção, vamos criar um esquema que contém o cubo como costumávamos nos exemplos da sintaxe de consulta MDX. Antes de começar, você pode querer Números de revisão 15-2 e 15-3, bem como a Tabela 15-1. Como dissemos em nosso discussão da Figura 15-3, há uma analogia entre cubos e esquemas em estrela, e para esse efeito vamos basear nosso cubo no esquema Ordens da estrela, que foi desenvolvido no capítulo 8 e mostrado na Figura 8-6.

Esquema Básico tarefas de edição

No restante desta seção, descrevemos todas as tarefas comuns de edição de esquema em detalhe. Antes de fazer isso, é bom ter uma visão geral das diferentes tarefas e subtarefas envolvidas, como é fácil perder de vista o quadro maior.

As tarefas podem ser resumidas como segue:

- Criando um esquema
- Criando cubos
  - Escolhendo uma tabela de fatos
  - Adicionando medidas
- Criando (compartilhado) dimensões
  - Editando a hierarquia padrão e escolha uma tabela de dimensão
  - Definição de níveis de hierarquia
  - Opcionalmente, a adição de mais dimensões
- Associando dimensões com cubos

Criando um Cubo

Você pode adicionar um cubo, clicando no botão Adicionar Cubo, que é o primeiro botão na barra de ferramentas. O novo cubo se torna visível na exibição em árvore sob a esquema como um novo nó com um ícone de cubo. Após adicionar o cubo, você pode editar suas propriedades (ver Figura 15-11).

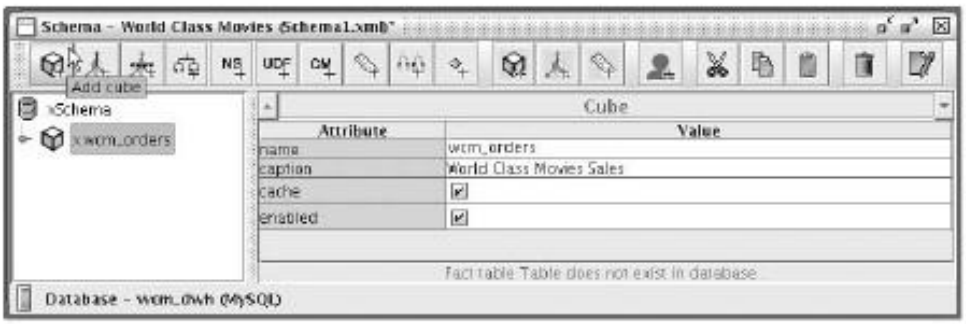


Figura 15-11: Criando um cubo

cubos Mondrian pode ter os seguintes atributos:

- nome -Especifica o nome que será usado em consultas MDX para se referir a este cubo. Esse nome deve ser exclusivo dentro do esquema. Para o nosso exemplo cubo, que deve especificar wcm_orders aqui.
- legenda Especifica um nome de exibição, que será utilizado pelo usuário interface para apresentar o cubo para o usuário final. Neste caso foi utilizado Mundial Filmes Vendas classe.
- cache Controla se os dados da tabela de fatos devem ser armazenados em cache.
- habilitado Controles de Mondrian se deve carregar ou ignorar o cubo.

Há um outro atributo que não é actualmente suportado por PSW:

- DefaultMeasure -Opcionalmente, você pode especificar o nome de um dos do cubo medidas aqui fazer explicitamente que a medida padrão.

Você pode encontrar mais informações sobre as medidas nas próximas duas subseções.

Note que, quando adicionado o wcm_orders cubo, um pouco vermelho X ícone apareceu à esquerda do esquema e os ícones do cubo. O ícone de X vermelho indica que há é algum erro ou errada em algum lugar ou abaixo desse particular nó. Quando um problema é detectado por PSW, que borbulha",''causar qualquer acima de nós que o nó para exibir o ícone de X vermelho também. Isso permite que você ver se algo está errado imediatamente. Se um ícone de X vermelho é exibido, algo está errado.

Sempre que você vê esses indicadores X, você deve procurar uma mensagem a vermelho na parte inferior do espaço de trabalho. Esta mensagem fornece uma descrição textual do problema e, geralmente, lhe dá uma boa idéia sobre que parte do esquema que você deve olhar para corrigir o problema. Uma mensagem diferente aparece dependendo de qual nó é selecionado na exibição em árvore, então se você quiser ver o razão pela qual um determinado nó tem um ícone de X vermelho, selecioná-lo primeiro.

Escolher uma Mesa de Fato

No caso da Figura 15-11, você pode dizer que deve haver algo de errado com o cubo, ou algum nó abaixo dele. A mensagem de erro indica um desconfiguração da tabela de fatos é a fonte do problema. O nó do cubo inicialmente recolhido, e se expandi-lo, você vai notar que ele contém uma nó tabela. Este nó da tabela representa a tabela de fatos em que o cubo é construído. Neste caso, seu cubo deve ser baseada na fact_orders mesa, é por isso que você definir o nome da tabela usando a caixa de lista suspensa (ver Figura 15-12).

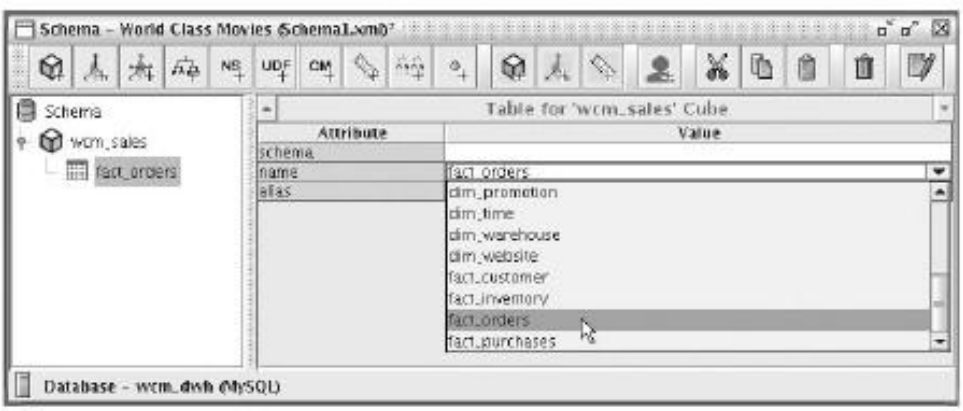


Figura 15-12: Escolhendo a tabela de fatos

Note-se que os pequenos ícones X desaparecem após você digitar o nome de uma já existente
tabela. Neste caso, o elemento da tabela foi errada, e optando por um
tabela de fatos existentes, você resolve esse problema. Isto imediatamente faz com que o cubo e
esquema de nós, acima da tabela nó a ser fixo também.

O nome da tabela é tudo que você precisa para configurar para o quadro do cubo elemento. No entanto, incluir uma descrição dos seus atributos para ser completo:

- esquema -O identificador do esquema do banco de dados que contém o fato tabela. Quando não especificado explicitamente, o esquema padrão do banco de dados conexão é usada.

- nome -O nome da tabela de fatos. Quando conectado a um banco de dados, o editor de propriedade fornece uma caixa de listagem drop-down como o mostrado na Figura 15-11, que permite a você escolher qualquer um dos quadros do padrão esquema. Observe que esse nome é o identificador do SQL-ao contrário do nome do o cubo, esta propriedade de nome não tem qualquer influência sobre qualquer MDX consultas.

- alias -Este é o apelido da tabela que será usada para essa tabela quando gerar instruções SQL. Pode ser útil para especificar isso no caso de você deseja depurar as instruções SQL geradas por Mondrian.

Adicionando Medidas

Agora que você configurou um quadro de verdade, você pode adicionar algumas medidas. Para adicionar
medidas, primeiro selecione o cubo (ou de sua tabela de verdade) na exibição em árvore. Em seguida, clique no
Medida botão Adicionar da barra de ferramentas. A partir da esquerda, este é o primeiro botão com
o ícone de régua e no sinal de mais. Alternativamente, você pode botão direito do mouse o cubo e
escolha a opção Adicionar Medida a partir do menu de contexto (ver Figura 15-13).

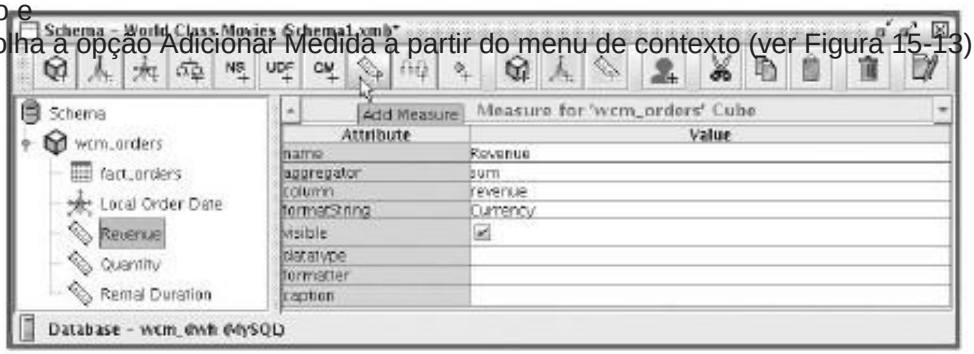


Figura 15-13: Adicionando as medidas da tabela de fatos do cubo

Por enquanto vamos ficar perto do desenho da tabela de fatos e adicione o
Receitas, Quantidade e Duração medidas de arrendamento. Estes correspondem diretamente ao quantidade,receitasE rental_duration colunas na fact_orders
tabela.

A ordem em que você especificar as medidas é significativa: implicitamente, o primeira medida no cubo é considerado a medida padrão. Se você gosta, você pode substituir esse especificando explicitamente o nome da medida padrão na DefaultMeasure atributo do cubo. Atualmente, esse atributo não é suportado por PSW, e você terá que editar manualmente o esquema XML e modificar o tag cubo de abertura para incluir uma DefaultMeasure atributo da seguinte forma:

```
<Cube DefaultMeasure="Quantity" ...>
...
</ Cube>
```

As medidas se podem ter os seguintes atributos:

- nome -O identificador que será usado para se referir a esta medida em MDX consultas. Este deve ser exclusivo dentro do cubo.
- agregador -O nome da função que é usada para agregar os medida. A grade atributo oferece uma caixa de listagem drop-down de onde você pode escolher um dos soma,contagem,min,max,avgE distinta da contagem. Para o Receita medidas e quantidade, você deve escolher o soma agregador. Para medir a duração do arrendamento, o avg agregador é a mais útil escolha.

- coluna -O nome de uma coluna da tabela do cubo fato. Ao con-
  conectado ao banco de dados, o editor de atributos oferece uma caixa de listagem drop-down
  a partir do qual você pode escolher a coluna.
- formatString -Aqui você pode especificar o nome do formato predefinido
  Moeda ou um padrão de seqüência de caracteres que especifica como o valor da medida será
  formatada para exibição. seqüências de formato são discutidos em mais detalhes posteriormente
  nesta seção.
- visível -Um sinalizador que especifica se a medida é exibida para o
  usuário final na interface do usuário.
- datatype -Aqui você pode usar uma caixa de listagem drop-down para escolher String, Numéricos,Inteiro,Boolean,Data,TimeOu Timestamp. Ao retornar
  dados, o tipo de dados especificado será usado para retornar dados do resultado da MDX.
- formatador -Você pode usar esse atributo para especificar um personalizado para-
  matéria. formatadores de célula personalizado deve implementar a interface Java mondrian.olap.CellFormatter.
- legenda -Especifica o nome de exibição que é usado para apresentar esta medida
  na interface do usuário. Se você deixar este campo em branco, o nome da medida é apresentado em seu lugar.

Adicionando dimensões

Os esquemas de Mondrian pode conter dimensões em dois locais:

- Dentro do cubo que possui""a dimensão-Estes dimensões são
  chamada dimensão privada, porque eles são conhecidos apenas para o cubo que ela contém e não pode ser usado fora do cubo envolvente.
- Dentro do próprio esquema-Estes são dimensões compartilhadas e pode ser
  vezes associado com vários cubos, e / ou múltiplos com a mesma
  cubo. dimensões compartilhados são excelentes para a execução do role-playing dimensões.

Geralmente, recomendamos que você use sempre dimensões compartilhadas ao invés de dimensões particulares. Embora o processo de criação privada e compartilhada dimensões é semelhante, a capacidade de reutilização compartilhada dimensões fornece uma benefício considerável, mesmo no curto prazo.

Para criar uma dimensão compartilhada, primeiro selecione o esquema. (Para criar uma empresa privada
dimensão, selecione o cubo que irá conter a dimensão vez.) Em seguida, clique o botão Adicionar dimensão na barra de ferramentas do editor do esquema. Este botão é segundo botão na barra de ferramentas, ao lado do botão Adicionar cubo.

Você pode definir os seguintes atributos de dimensões:

- nome -Para as dimensões particulares, o nome se refere a esta dimensão em MDX
  consultas. Para obter as dimensões compartilhadas, o nome se refere à dimensão quando

você está associando-a com um cubo. Para dimensões particulares, o nome do deve ser único entre todas as outras dimensões utilizadas pelo cubo. Para compartilhada dimensões, o nome deve ser exclusivo dentro do esquema. O compartilhada dimensão mostrado na Figura 15-14 usa data como o nome.

- foreignKey -Se esta é uma dimensão particular, este é o nome de um coluna da mesa do cubo que se refere à tabela de dimensão que corresponde a esta dimensão.
- tipo Se a sua dimensão é hora ou data relacionada, você deve usar TimeDimension. Isso permite que você use o tempo padrão e MDX funções de data. Caso contrário, use StandardDimension.
- usagePrefix -Isso se aplica apenas às dimensões privadas.
- legenda -Este é um nome de exibição usado para apresentar essa dimensão ao usuário final através da interface do usuário.

Figura 15-14 mostra como adicionar uma dimensão compartilhada chamada Data neste maneira.

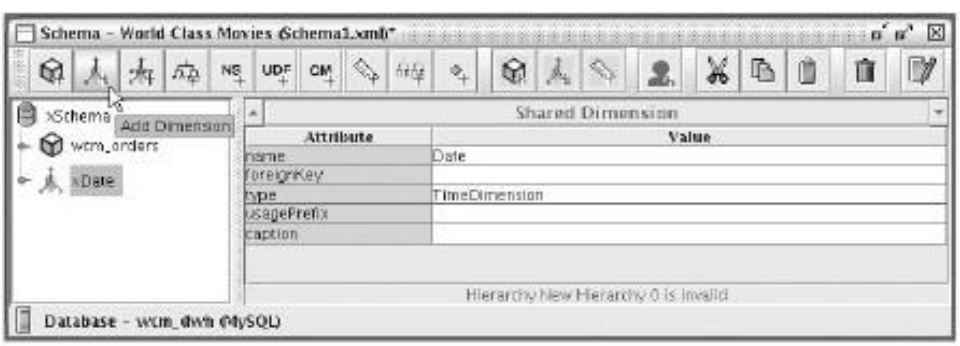


Figura 15-14: Adicionando uma dimensão compartilhada Data

Adicionando a dimensão faz com que X vermelho ícones para aparecer novamente. O erro mensagem na parte inferior do espaço de trabalho indica que há algo errado com uma hierarquia, assim que adicionar e hierarquias edição é o tema do próximo subseção.

Adicionando e hierarquias Editar e escolha de tabelas de dimensão

Quando você cria uma dimensão, uma nova hierarquia também é criado. Você pode vê-lo quando você expandir o nó dimensão. Além disso, o nó de uma tabela automaticamente é criado sob o nó de hierarquia. Antes de editar o nó de hierarquia, é melhor para configurar o nó da tabela subjacente.

O nó da tabela representa a tabela de dimensão que irá entregar os valores para os níveis da hierarquia. O procedimento para configurar o quadro é exatamente o o mesmo que o processo de escolha de uma tabela de fatos de um cubo, que foi

descrito anteriormente nesta seção. Para a dimensão de data criada no ano anterior subseção, você tem que escolher o dim_date_en_us Dimensão da tabela.

Depois de escolher a tabela de dimensão, você pode configurar a própria hierarquia. Hierarquias de apoio os seguintes atributos:

- nome -O nome usado em consultas MDX para se referir à hierarquia. Deve ser único dentro da dimensão. Omitir o nome faz com que a hierarquia para obter o mesmo nome de sua dimensão. Além disso, esta hierarquia é o padrão de hierarquia.
- legenda -O nome que é utilizado para apresentar essa hierarquia para o usuário final na interface do usuário.
- hasAll -Um indicador que indica se a hierarquia deve ter um todo nível com um membro de tudo, por exemplo, um único membro na parte superior da hierarquia que representa todos os outros membros. Normalmente você deve deixar isto.
- AllMemberName -Se hasAll estiver ativado, este parâmetro especifica o identificador MDX que está a ser utilizado para todos os membros. Quando este for omitido, a todos os membros nome é derivado automaticamente como Todos <nome da hierarquia>.
- allMemberCaption -Se hasAll estiver ativada, você pode usar isto para especificar o nome que será utilizado para apresentar a todos os membros para o usuário final na interface do usuário.
- allLevelName -O nome usado para se referir ao nível de todas as consultas MDX.
- DefaultMember -O nome do membro padrão. Se isso não for especificado, em seguida, a todos os membros serão utilizados como membro padrão se a hierarquia tem um Todos os membros. Isso geralmente é exatamente o comportamento desejado. Quando especificado, esse membro será utilizado quando um membro é esperado, mas não explicitamente especificados na consulta MDX. Se o membro padrão não é especificado eo hasAll Bandeira é desativado, o primeiro membro do primeiro nível na hierarquia será utilizado como membro padrão.
- memberReaderClass -Nome de uma classe personalizada membro leitor. Este especicados classe deve implementar mondrian.rolap.MemberReader. Normalmente você não especificar um leitor de membro do cliente, mas vamos ler o Mondrian membros do RDBMS acordo com os mapeamentos do esquema do banco de dados. Configurar esta opção é um recurso avançado que está além do escopo este livro.
- primaryKeyTable -Pode ser usado para especificar o nome da tabela da que esta hierarquia consultas seus membros. Se não for especificado, os membros são consultado a partir da tabela da hierarquia. Você normalmente pode deixar em branco se você está criando um cubo em cima de um esquema em estrela como neste exemplo. A flexibilidade para especificar um nome de tabela aqui é necessária quando se lida com esquemas floco floco de neve ou dimensões.

- primaryKey -Normalmente, você deve usar isso para especificar o nome do coluna de chave primária da tabela esta hierarquia de dimensão. Para ser exato: esta é o nome da coluna da tabela de dimensão que está referenciado pelas linhas na tabela de fatos. Esta deve ser uma coluna na dimensão desta hierarquia de tabela.

Para configurar a primeira hierarquia da dimensão Data, você precisa especificar as seguintes propriedades:

- nome -Isso deve ser meses, de acordo com os homens de design de cubo citadas na Tabela 15-1.
- hasAll -Esse deve ser ativado.
- primaryKey -Isso deve ser definido para date_key, Que é a chave primária da dim_date_en_us tabela de dimensão, que configurado para este hierarquia.

O projeto para essa hierarquia é mostrado na Figura 15-15.

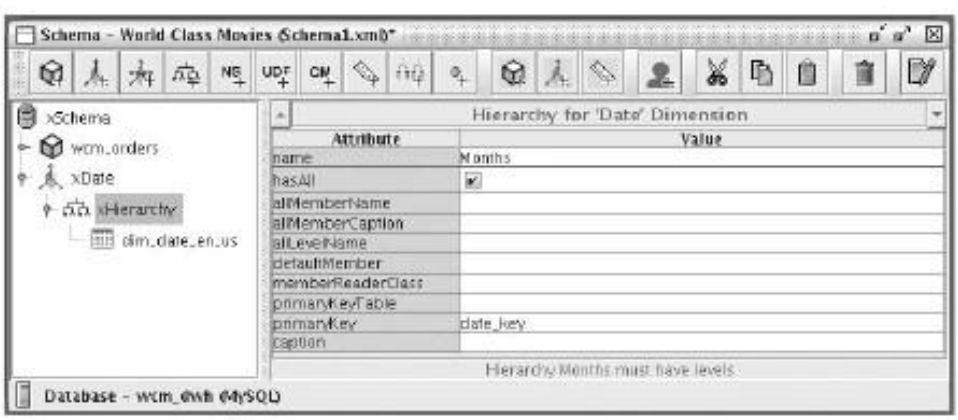


Figura 15-15: O desenho da hierarquia Meses da dimensão Data

De acordo com o design do cubo mostrado na Tabela 15-1, você também deve adicionar uma hierarquia semanas. Para adicionar a segunda hierarquia, você pode selecionar o nó de dimensão de data e clique no botão Adicionar na hierarquia barra de ferramentas ou botão direito do mouse o nó de dimensão de data e escolher a hierarquia Adicionar opção no menu de contexto. Como você viu na edição da hierarquia Meses, a hierarquia Semanas novo já contém um nó de tabela. Você pode configurar a nó tabela exatamente como você configurou a hierarquia meses, e apontá-lo para o dim_date_en_us Dimensão da tabela. Por último, os atributos de configuração para a hierarquia semanas deve ser idêntico ao da hierarquia Meses, exceto que o valor da sua nome atributo deve ser semanas, e não Meses.

Adição de níveis de hierarquia

Agora que você criou as hierarquias, você deve definir seus níveis. Você pode adicionar um nível a uma hierarquia existente ou selecionando-o e clicando em Nível botão Add na barra de ferramentas ou clicando com a hierarquia e escolhendo a opção Adicionar menu Nível. Depois de adicionar um nível, você pode editar o seguintes atributos:

- nome -O nome que é usado para se referir a este nível em consultas MDX.
- tabela -O nome da tabela que contém as colunas onde o dados de dimensão é armazenada para este nível. Quando não especificado, da hierarquia tabela de dimensão será utilizado. Esta é a situação normal para esquemas em estrela como o usado neste exemplo. Você precisa especificar uma tabela específica só quando se trata de esquemas floco de neve.
- coluna -O nome da coluna que representa o identificador de membro para este nível. Esta deve corresponder à mesa este nível (veja a tabela atributo).
- NameColumn -O nome da coluna que contém o nome deste nível. Quando não especificado, o valor do nome propriedade é usada. Normalmente você deve deixar este campo em branco.
- parentColumn -Isso se aplica somente a tipos especiais de hierarquias chamado pai-filho hierarquias. Normalmente você deixar este campo em branco, mas se você está lidando com uma hierarquia pai-filho, você usa esse atributo para especificar o coluna que faz referência a membros do pai.
- nullParentValue -Quando se tratar de uma relação pai-filho, você pode usar esse atributo para especificar que valor indica o membro pai não existe. Deixe em branco quando não lidar com pais e filhos hierarquias.
- ordinalColumn -Esse atributo pode ser usado para especificar qual coluna define como os valores dos membros devem ser separados por padrão. Você deve especificar esta sempre a ordem de classificação natural dos próprios membros não cabe a ordem de classificação desejada. Se a ordem de classificação natural da membros é também a ordem de classificação desejada, você pode deixar este campo em branco.
  Às vezes, você ainda pode especificar uma coluna aqui se o ordinalColumn tem um tipo de dados mais adequada para classificar a coluna que fornece o valores de membro.
- tipo -O tipo de dados dos valores membro. Isto é usado para controlar se e como os valores devem ser citados na geração de SQL a partir de consultas MDX.
- uniqueMembers -A bandeira que indica se todos os membros a este nível ter valores únicos. Isto é sempre verdadeiro para o primeiro nível (não contando o nível de todos) de qualquer hierarquia. Se você sabe que é verdade para qualquer um dos níveis subseqüentes, você pode especificá-lo lá também, e isso pode permitir que

Mondrian para gerar consultas SQL mais eficiente. Não permitir que isso se você não está 100 por cento se os valores são únicos, pois pode causar resultados incorretos para ser devolvido.

- levelType -Se você deixar este campo em branco, será assumido este é um regular nível, que é o valor correto para a maioria das dimensões. Dimensões que foram configurados para ser do tipo TimeDimension deve especificar um dos tipos pré-definidos para os níveis TimeDimension: TimeYears, TimeQuarters, TimeMonths, TimeWeeks e TimeDays. Para TimeDimensions, especificar os levelType é um pré-requisito para o uso correto do Mondrian Data / Hora funções como acumulado do ano.

- HideMemberIf -Isso determina os casos em que um membro deve ser oculto. Normalmente, você pode deixar este campo em branco, que é equivalente à configuração o valor para Nunca. Neste caso, o membro é sempre mostrado.
- approxRowCount -O número estimado de membros a este nível. Specifying uma boa estimativa para este atributo pode permitir Mondrian fazer melhores decisões sobre a forma de consulta e / ou o cache de dados, que podem melhorar o desempenho.

- legenda -O nome que será utilizado para apresentar a este nível para o usuário a interface do usuário. Quando não especificado, o nome do nível será usado.
- captionColumn -Aqui você pode especificar quais colunas de dimensão ao nível dotabela Comissão deverá ser usado para apresentar os membros para o usuário final. Quando não especificado, o identificador de membro será utilizado. (Veja o atributo de coluna Para obter mais informações sobre este assunto.)
- formatador -Isso pode ser usado para especificar um formatador de costume, bem como já discutimos as medidas.

Agora que nós discutimos os atributos possíveis dos níveis, que realmente pode adicioná-los. Tabelas 15-2 e 15-3 mostra os níveis que você precisa para criar para o Meses e hierarquias semanas, respectivamente, e como você deve configurar seus atributos.

Tabela 15-2: Os níveis da hierarquia Meses

NOME	LEVELTYPE	COLUNA	CAPTIONCOLUMN	UNIQUEMEMBERS
Ano	TimeYears	ano4		habilitado
Trimestre	TimeQuarters	trimestre_number	trimestre_name	deficientes
Mês	TimeMonths	Mês_number	Mês_abbreviation	deficientes
Dia	TimeDays	day_in_month		deficientes

Tabela 15-3 Os níveis da hierarquia Semanas

COLUNA LEVELTYPE NOME			CAPTIONCOLUMN	UNIQUEMEMBERS
Ano	TimeDays	ano4		habilitado
Semana	TimeWeeks	week_in_year		deficientes
Dia	TimeDays	day_in_week	day_abbreviation	deficientes

O resultado deve ser algo como a Figura 15-16.

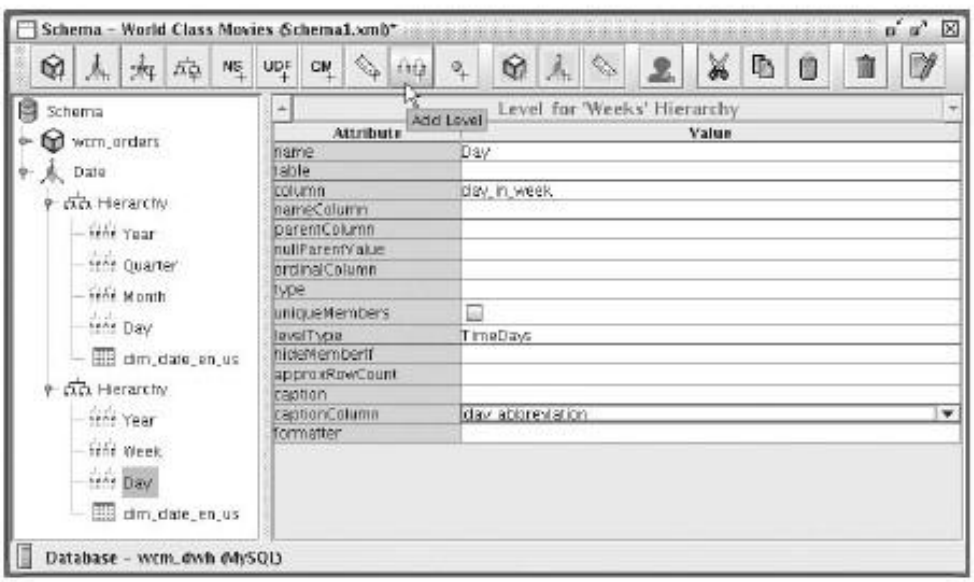


Figura 15-16: Níveis para os meses e hierarquias Semanas

Associando Cubos com Dimensões compartilhadas

Nas seções anteriores, você construiu uma compartilhado dimensão Data. Antes de usar que, no wcm_orders cubo, você precisa se associar a dimensão eo cubo.
Em esquemas de Mondrian, a associação entre um cubo e uma dimensão compartilhada é chamado de dimensão de uso. Para adicionar uma dimensão de uso, selecione o cubo e clique no botão Adicionar utilização Dimension na barra de ferramentas, ou clique com o botão
cubo e escolher a opção de adicionar a dimensão de uso do menu de contexto.
Você pode editar os seguintes atributos para um uso dimensão:

- nome -O nome que será usado para se referir à dimensão em MDX
  consultas. Esse nome não precisa ser idêntico ao nome do compartilhadas
  dimensão própria. Em vez disso, você deve fornecer um nome (ou um apelido, se preferir)
  que reflete o propósito específico da dimensão em relação ao cubo.
  A capacidade de fornecer um nome específico, único cada vez que você usar um compartilhado
  dimensão é efetivamente um caso de execução dimensões role-playing.

- foreignKey -O nome da coluna na tabela do cubo que referências a chave primária da tabela de dimensão. Lembre-se que você tinha que especificar a coluna de chave primária da tabela de dimensão como a primaryKey atributo das hierarquias; bem, esta é a contrapartida, definindo o fato final da tabela do relacionamento.

- fonte -Este é o nome da dimensão compartilhada.

- nível -Aqui você pode especificar um nome de nível da dimensão compartilhada que serão unidos contra a tabela de fatos do cubo. Para esquemas estrela, que deve normalmente ser deixado em branco.

- legenda -O nome usado para apresentar a dimensão para o usuário final pela interface do usuário. Se deixado em branco, o valor da nome atributo ser usado.

No caso da dimensão Data, você precisa configurar o uso da dimensão como se segue:

- nome -Aqui você deve especificar a data da ordem local. A date_dimension mesa sobre a qual basear a dimensão de data no esquema podem desempenhar um número de papéis em relação ao fact_orders mesa sobre a qual você base do cubo. Por agora, limitar a data da ordem local, que se reflecte no nome.

- foreignKey -Aqui você deve especificar o local_order_date_key coluna da fact_orders tabela de acordo com o papel desempenhado pela Data dimensão.

- fonte -Aqui você deve especificar a data, que é o nome que identifica sua dimensão no esquema.

O resultado é mostrado na Figura 15-17.

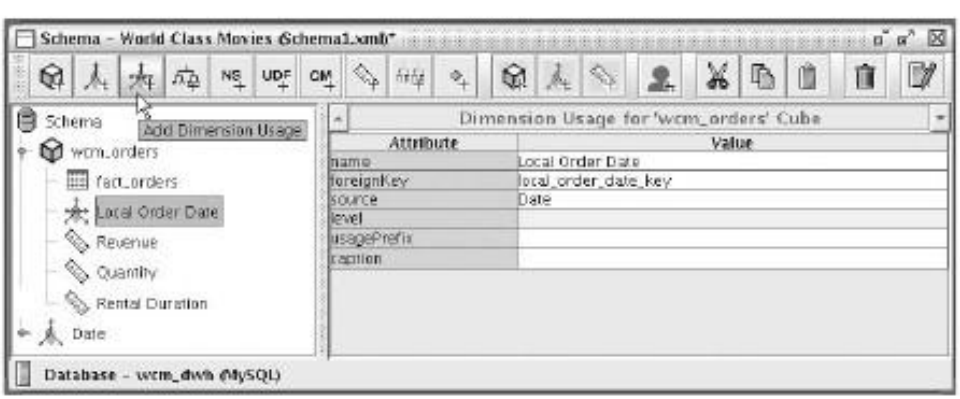


Figura 15-17: O uso da dimensão associando o cubo com a dimensão de data como Data da Ordem local

Adicionando as Dimensões e DVD ao Cliente

Para completar o esquema, você precisa adicionar uma dimensão compartilhada para o DVD e
dimensões do cliente e associá-las com a wcm_orders cubo. Até agora,
você deve ser capaz de completar estes passos em seu próprio país. Para resumir, você deve repetir os passos seguintes para ambas as dimensões:

- Adicionar e editar uma dimensão.
- Escolha uma tabela de dimensão.
- Edite a hierarquia.
- Adicionar e editar os níveis de hierarquia.
- o cubo associado com a dimensão.

Em comparação com a adição da dimensão Data, acrescentando que o DVD eo Cliente dimensões envolve algumas pequenas diferenças que realmente simplificar as coisas:

- O DVD e as dimensões do cliente ter apenas uma hierarquia, em vez de dois.
- Como essas dimensões só têm uma hierarquia, você deve sair em branco o nome atributo de suas respectivas hierarquias.
- Sua tipo atributo da dimensão deve especificar StandardDimension (O padrão) ao invés de TimeDimension.
- Da mesma forma, porque os níveis também são normais, você não deve especificar o seu levelType atributo.

Para completar, a lista a seguir fornece uma visão geral da configuração da dimensão do DVD e sua associação com o wcm_orders cubo:

- nome: DVD
- tipo: StandardDimension
- Dimensão da tabela:
  - nome:dim_dvd_release
- Hierarquia:
  - hasAll:verdade
  - primaryKey:dvd_release_key
  - Nível de Gênero:
    - nome: Gênero
    - coluna:dvd_release_genre
    - uniqueMembers:verdade

- Nível Título:
  - nome: Gênero
  - coluna:dvd_release_genre
  - uniqueMembers:falsa
- Uso da Dimensão:
  - nome: DVD
  - foreignKey:dvd_release_key
  - fonte: DVD

A lista a seguir fornece uma visão geral da configuração do Cliente dimensão e sua associação com o wcm_orders cubo:

- nome: Cliente
- tipo: StandardDimension
- Dimensão da tabela:
  - nome:dim_customer
- Hierarquia:
  - hasAll:verdade
  - primaryKey:customer_key
  - Nível nacional:
    - nome: País
    - coluna:customer_country_code
    - uniqueMembers:verdade
    - captionColumn:customer_country_name
  - Nível Região:
    - nomeRegião:
    - coluna:customer_region_code
    - uniqueMembers:verdade
    - captionColumn:customer_region_name
  - Nível da cidade:
    - nome: Cidade
    - coluna:customer_city_name
  - Código Postal:
    - nome: Código Postal
    - coluna:customer_postal_code

- Nome:
  - nomeNome:
  - coluna:customer_id
  - captionColumn:customer_last_name
- Uso da Dimensão:
  - nome: Cliente
  - foreignKey:customer_key
  - fonte: Cliente

XML Listagem

Para completar, incluir a fonte XML do esquema Mondrian aqui
como listagem 15-1. Fizemos alguns (insignificante) mudanças para tornar o código mais compacta. Você pode usar isso para comparação com seu próprio esquema:

Listagem 15-1: XML de origem do esquema de Mondrian

```
Esquema_nome name="World Classe Movies">
 <Dimension Type="TimeDimension" name="Date">
  <Hierarchy Name="Months" hasAll="true" primaryKey="date_key">
   <table Name="dim_date_en_us"/>
   <Nome Nível = "Ano" coluna = "ano4" uniqueMembers = "true"
        levelType = "TimeYears" />
   <Nome Nível = "Bairro" coluna = "quarter_number" uniqueMembers = "false"
        levelType = "TimeQuarters" captionColumn = "quarter_name" />
   <Nome Nível = "Mês" coluna = "month_number" uniqueMembers = "false"
        levelType = "TimeMonths" captionColumn = "month_abbreviation" />
   <Nome Nível = "Dia" coluna = "day_in_month" uniqueMembers = "false"
        levelType = "TimeDays" />
  </ Hierarquia>
  <Hierarchy Name="Weeks" hasAll="true" primaryKey="date_key">
   <table Name="dim_date_en_us" schema="" alias=""/>
   <Nome Nível = "Ano" coluna = "ano4" uniqueMembers = "true"
        levelType = "TimeYears" />
   <Nome Nível = "Semana" coluna = "week_in_year" uniqueMembers = "false"
        levelType = "TimeWeeks" />
   <Nome Nível = "Dia" coluna = "day_in_week" uniqueMembers = "false"
        levelType = "TimeDays" captionColumn = "day_abbreviation" />
  </ Hierarquia>
 </ Dimension>
 <Dimension Type="StandardDimension" name="DVD">
  <Hierarchy HasAll="true" primaryKey="dvd_release_key">
   <table Name="dim_dvd_release" schema="" alias=""/>
   <Level Name="Genre" column="dvd_release_genre" uniqueMembers="true"/>
   <Level Name="title" column="dvd_release_title" uniqueMembers="false"/>
  </ Hierarquia>
 </ Dimension>
```

```
 <Dimension Type="StandardDimension" name="Customer">
  <Hierarchy HasAll="true" primaryKey="customer_key">
   <table Name="dim_customer" schema="" alias=""/>
   <Nome Nível = "País" coluna = "customer_country_code" uniqueMembers = "true"
         captionColumn = "customer_country_name" />
   <Nome Nível = "Região" coluna = "customer_region_code" uniqueMembers = "true"
         captionColumn = "customer_region_name" />
   <Level Name="City" column="customer_city_name" uniqueMembers="false"/>
   <Nome Nível = "Código Postal" coluna = "customer_postal_code"
         uniqueMembers = "false" />
   <Nome = Nível coluna "Nome" = "customer_id" uniqueMembers = "false"
         captionColumn = "customer_last_name" />
  </ Hierarquia>
 </ Dimension>
 <Nome do Cubo = "wcm_orders" caption = "World Class Vendas Filmes" cache = "true"
enabled => "true"
  <table Name="fact_orders"/>
  <Fonte DimensionUsage = "Data" name = "Local Ordem Data"
foreignKey = "local_order_date_key" />
  <DimensionUsage Source="DVD" name="DVD" foreignKey="dvd_release_key"/>
  <DimensionUsage Source="Customer" name="Customer" foreignKey="customer_key"/>
  <Nome da Medida = "Receita" coluna = "receita" formatString = "Moeda"
          agregador = "soma" visible = "true" />
  <Nome da Medida = "Quantidade", coluna = "quantidade"
          agregador = "soma" visible = "true" />
  <Nome da Medida = "Aluguer de Duração" coluna = "rental_duration" formatString = "# .00"
          agregador = "avg" visible = "true" />
 </ Cube>
</ Schema>
```

Testes e Implantação

Agora que você criou o esquema, você está quase pronto para usá-lo. Você deve publicar o cubo para o Pentaho BI Server antes que você possa usá-lo para construir OLAP aplicações. Mas antes de implantá-lo, você pode querer fazer alguma preliminar primeiro teste.

Usando a ferramenta de consulta MDX

PSW inclui uma ferramenta de consulta básica MDX. Esta não é adequado como um usuário final
ferramenta de relatórios, mas é bastante útil para testar se o seu cubo é funcional. Você também pode usá-lo como uma ajuda no desenvolvimento de consultas MDX para uso direto na sua
soluções OLAP.

Você pode chamar a ferramenta de consulta MDX a partir do menu principal: File Nova MDX. O editor MDX aparece em sua própria janela. Se acontecer de você tem um editor do esquema aberto, a ferramenta de consulta MDX tenta conectar-se a banco de dados subjacente, bem como carregar a definição do esquema.

Isso é mostrado na Figura 15-18.

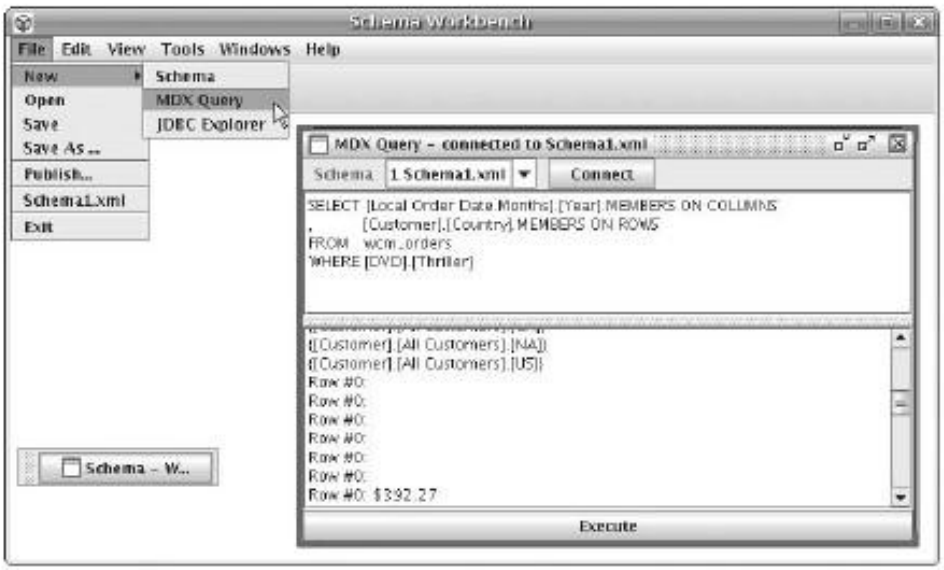


Figura 15-18: A ferramenta de consulta MDX

A ferramenta de consulta MDX tem um painel superior e um painel inferior. Você pode entrar Consultas MDX no painel superior. A consulta MDX é executada quando você clica Execute o botão que aparece na parte inferior da ferramenta de consulta MDX. A resultados são mostrados na parte inferior do painel.

Agora você pode experimentar algumas das consultas MDX discutido anteriormente neste capítulo.

Publicando o Cubo de

Quando estiver satisfeito com o design do cubo, você pode publicar o cubo para a Pentaho BI Server. Para chamar a publicar diálogo, verifique se você tem um esquema aberto no editor do esquema. Ative a janela do editor do esquema. Em seguida, escolha Publicar no menu principal, ea caixa de diálogo aparece, como mostrado infile Figura 15-19.

Para a URL, especifique o endereço web do Pentaho BI Server para o qual você deseja publicar o esquema. Você deve usar a senha do editor que especificado no servidor da publisher_config.xml arquivo. Para o usuário e senha, especifique as credenciais de qualquer usuário que tem o privilégio de publicar.

Quando você confirmar a caixa de diálogo, uma conexão é feita com a solução repositório, o que pode levar algum tempo. Se a conexão for bem sucedida, uma caixa de diálogo
Parece que lhe permite ver a solução de servidor de repositório. Escolha o caminho adequado, por exemplo, / WCM / Análise.

Na parte inferior da janela, você pode especificar quais dados do servidor JNDI fontes para usar no lado do servidor para executar as consultas SQL. Se você ativar o caixa de seleção Registrar XMLA Datasource, o esquema é registrado como um OLAP

fonte de dados no servidor. Isso permite que a interface de usuário para exibir uma lista de esquemas.

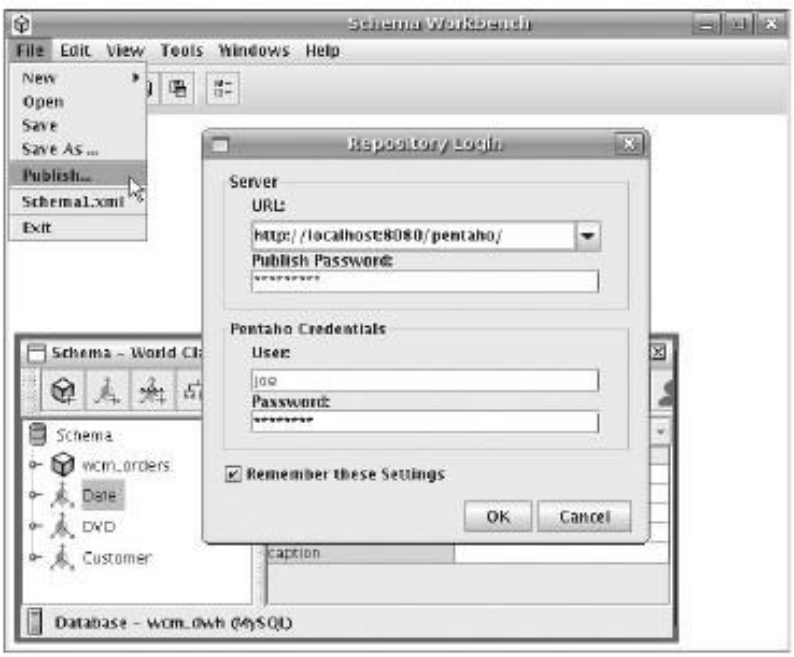


Figura 15-19: Publicar um esquema para a solução de repositório

Tópicos Design esquema nós não cobrimos

Nas subseções anteriores, você aprendeu a construir uma base Mondrian esquema e cubo. Embora nós cobrimos um lote de terreno, a este respeito, há muitos temas que nós não cobrimos. Aqui está uma lista parcial:

- Calculado-membros Mondrian cubos podem conter definições de calcu-membros lada.
- Funções e controle de acesso de esquema e elementos do cubo pode ser associado com papéis de conceder apenas determinados grupos de usuários acesso a particular elementos do cubo.
- Trabalhando com dimensões snowflaked-nos baseado inteiramente nosso exemplo em um esquema em estrela simples, usando apenas um-para-um entre cada tabela de dimensão ea sua respectiva dimensão do cubo. Mas é Mondrian capaz de muito mais mapeamentos avançados que permitem que você crie cubos em esquemas floco de neve.
- A formatação condicional-nos brevemente discutida a formatação de strings, mas por-esteiras pode realmente ser especificado como fórmulas que retornam uma marcação especial
que pode ser processado pela interface do usuário. Isso permite que você faça as coisas tais como medidas de exibir em cores diferentes, dependendo de seu valor.

- Internacionalização-Muitos dos atributos de elementos do esquema podem ser dado um valor variável, dependente de localidade do formulário %% Propertyname. Os valores reais para essas propriedades podem ser definidas em separado propriedades. arquivo. Isso pode ser usado para implementar internacionalizado OLAP aplicativos usando o adequado propriedades. arquivo para uma determinada localidade. Você pode, por exemplo, usar isso para escolher entre o Inglês eo Francês tabela de dimensão de data, como discutido nos capítulos anteriores ..

- Funções definidas pelo usuário-Você pode criar sua própria função como Java classes e importar estes no seu esquema.

Se você está interessado nestas funcionalidades, consulte o Mondrian original documentação em mondrian.pentaho.org.

Visualizando Cubos Mondrian com JPivot

Após a publicação do cubo para a solução Pentaho repositório pode ser usado para construir aplicações de análise. Nesta seção, você vai aprender a criar dinâmicas tabelas dinâmicas para procurar e analisar cubos Mondrian. Através destes pivô tabelas, os utilizadores empresariais podem perfurar para cima e para baixo, e cortar e cortar os dados.

Introdução à vista da análise

O usuário do console do Pentaho BI Server oferece a possibilidade de criar uma análise de vista, que é essencialmente uma mesa JPivot cruz no topo de um Mondrian cubo, envolvido em um processo de ação Pentaho. Para criar uma visão nova análise, clique no ícone da visão de análise na barra de ferramentas ou na página inicial do espaço de trabalho.
Você será solicitado a escolher um esquema, e dentro do esquema, um cubo. Este é mostrado na Figura 15-20.

Depois de escolher o esquema e cubo e confirmar a caixa de diálogo, uma tabela dinâmica aparece. Inicialmente, os membros padrão de todas as dimensões são exibidas no eixo vertical, ea medida padrão é exibido no eixo horizontal.
Lembre-se que, normalmente, o membro padrão é o membro mais, assim que o resultado é uma tabela com um único valor no mais alto nível de agregação possível. Esta é mostrados na Figura 15-21.

Se você gosta, você pode salvar a visão de análise para uso posterior clicando em um dos o disquete ícones na barra de ferramentas do Usuário Pentaho Console. Você será solicitado a fornecer um local dentro do repositório, bem como um nome.

No restante desta seção, discutiremos os métodos que permitem obter radicalmente diferentes visões sobre os dados na vista de análise. Ao salvar o tabela dinâmica, o estado da tabela também será salvo, permitindo-lhe obter um visão específica sobre os dados que você está interessado pol Se você vê algo que você gosta,
salvá-lo usando um novo nome.

Figura 15-20: Criação de uma visão de análise

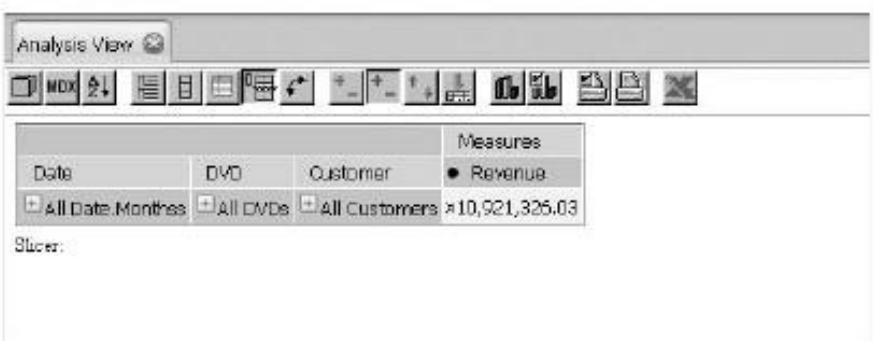


Figura 15-21: A tabela dinâmica padrão

Usando a Barra de Ferramentas JPivot

Além da tabela dinâmica, JPivot fornece uma barra de ferramentas na parte superior da página.
Nesta barra você encontrará uma série de ações interessantes. A barra de ferramentas mostrados na Figura 15-22.

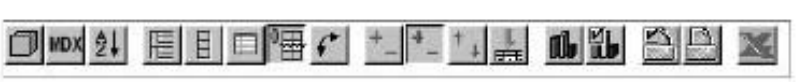


Figura 15-22: barra de ferramentas do JPivot

Discutimos alguns dos botões da barra nas seções seguintes.

Perfuração

Uma das grandes características de uma tabela dinâmica é que ela é interativa e dinâmica. Uma característica típica, a este respeito é a sua perfuração funcionalidade. A perfuração é essencialmente
uma forma de navegação de dados onde o usuário navega de um nível de agregação para outro. Apoio à navegação de dados é provavelmente a razão mais importante para exigir uma organização hierarquizada em níveis de dimensões.

Na Figura 15-21, você pode ter notado o pouco +ícones do todos os membros das dimensões. Clicando em um ícone de adição expande o membro, revelando os membros do próximo nível (o que representa um menor nível de agregação). Ao mesmo tempo, as medidas são re-agregada para se ajustar ao nível do conjunto de novas dos membros revelado. Este tipo de ação é conhecida como perfuração para baixo, como você está navegando a partir do atual nível de agregação de um membro um menor nível de agregação aplicáveis aos membros do seu filho. Por exemplo, se você toma a dimensão de data na tabela dinâmica mostrado na Figura 15-21 e broca descer um nível a partir do membro Datas Todos ao nível do ano, revelando a Ano membro individual, você teria algo parecido com a Figura 15-23.

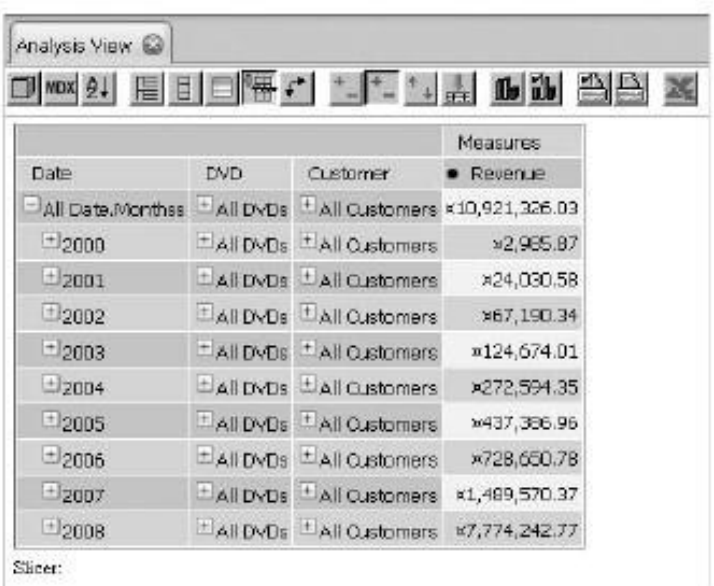


Figura 15-23: Perfuração para baixo de todo o ano

Perfuração Sabores

Há uma série de maneiras diferentes ação de um usuário de perfuração pode resultar em um resultado de perfuração em particular. Os métodos são:

- Broca-Membros
- Broca posição

- Substituir Drill
- Perfurar

O método de perfuração é controlado através dos botões da barra de ferramentas mostrada na
Figura 15-24.

Figura 15-24: A barra de botões método de perfuração

Os três primeiros métodos são aplicáveis às dimensões. Você pode usar apenas uma destes métodos ao mesmo tempo. O quarto método é aplicável às medidas. Pode ser ativado de forma independente dos outros três métodos.

Broca-Membros e posição da broca

Por padrão, usa o JPivot broca membro acção. Com esta ação, a perfuração de um exemplo de um membro também é aplicado a todas as outras instâncias desse membro. Esse comportamento é ativado, alternando o primeiro da barra de ferramentas do método de perfuração
botões.

Para ver broca membro em ação, verifique se este botão da barra de ferramentas é alternado.

Agora, suponha que você começar a perfurar na tabela mostrada na Figura 15-23 no All Clientes membro do ano 2000. A ação da broca será aplicada a
todas as outras ocorrências de todos os clientes, eo resultado seria algo
como a Figura 15-25.

			Measures
Date	DVD	Customer	Revenue
All Date.Monthss	All DVDs	All Customers	10,921,326.03
		Canada	752,988.55
		United States	10,168,337.48
2000	All DVDs	All Customers	2,985.87
		United States	2,985.87
2001	All DVDs	All Cust...	
		All Customers	1,489,570.37
		Canada	80,279.18
		United States	1,409,291.19
2008	All DVDs	All Customers	7,774,242.77
		Canada	658,570.07
		United States	7,115,672.70

Figura 15-25: Broca membros sobre todos os clientes de 2000 também irá expandir os clientes todos membro para todos os outros anos

Como você pode ver na Figura 15-25, a perfuração para baixo para todos os clientes do ano 2000 revela os membros Cliente no próximo nível para todos os anos, não apenas os ano de 2000.

O sabor da broca segunda broca posição. Isso pode ser habilitado, alternando o segundo a perfuração botões da barra de ferramentas do método. Neste caso, a perfuração ocorre diretamente na instância membro e não é aplicada a quaisquer outras instâncias desse membro. Se você usar esse método para detalhar a Clientes Todos membro durante o ano de 2000, a tabela mostrada na Figura 15-23, apenas a membros inferiores dos clientes do ano 2000 são revelados, e eles permanecer oculto para todos os outros anos.

Substituir Drill

Com o broca substituir método, o membro passa a ter perfurado com a broca resultado. Você pode habilitar este método, alternando a barra de ferramentas terceiro método de perfuração, mas, ton. Por exemplo, suponha que você tenha uma tabela como a mostrada na Figura 15-21. Se você usou broca em substituir o membro clientes, o resultado seria como a Figura 15-26.

			Measures
Date	DVD	Customer	Revenue
All Date.Months	All DVDs	Canada	$752,988.55
		United States	$10,168,337.48

Figura 15-26: Broca substituir remove o membro perfurados com o resultado da broca

Como você pode ver na Figura 15-26, o membro Todos os clientes que já se foi. Em vez você vê os membros no novo nível.

Perfurar

Considerando que todos os métodos de perfuração discutido anteriormente aplicáveis às dimensões, perfurar se aplica às medidas. Uma broca através de ação recupera o detalhe linhas (linhas da tabela verdade) correspondente à soma medida enrolado valor, apresentando os resultados em uma tabela separada. Perfure através podem ser ativados mudando o quarto botão da barra de perfuração método.

O Navigator OLAP

O Navigator OLAP é uma interface gráfica que permite que você controle como os mapas JPivot o cubo para a tabela dinâmica. Você pode usá-lo para controlar quais dimensão é mapeada para qual o eixo, como as múltiplas dimensões em um dos eixos são ordenada, e que fatia do cubo é usado em análise.

Você pode abrir o Navegador de OLAP, clicando no botão da barra de ferramentas com o ícone pequeno cubo. Isso é mostrado na Figura 15-27.

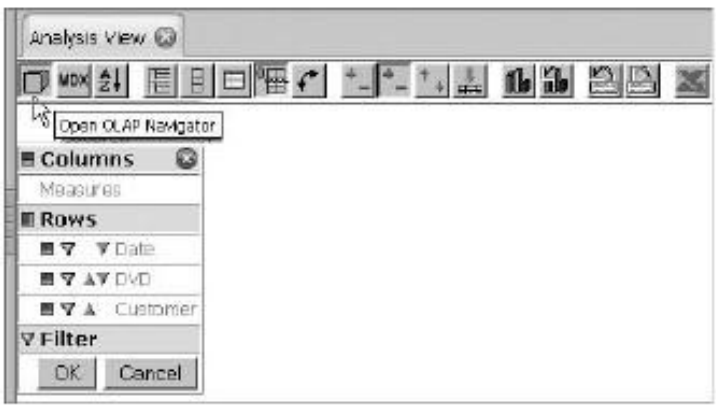


Figura 15-27: O Navigator OLAP

Figura 15-27 tem os seguintes elementos:

- A seção Colunas, atualmente apenas com medidas
- Uma seção de Linhas, agora com todas as dimensões
- Uma seção de filtro, que está atualmente vazia

No restante desta seção, vamos explorar estes diferentes elementos da OLAP Navigator, e veja como você pode usá-los para obter cruz diferente seções de dados do cubo.

Controlando a veiculação de dimensões em eixos

Em OLAP Navigator, você pode mover uma dimensão para outra, clicando no eixo ícone quadrado pouco antes da dimensão. Por exemplo, ao clicar no ícone quadrado que é exibido logo antes da dimensão Cliente move o Cliente dimensão a partir do eixo de linhas para o eixo Colunas. (Veja a Figura 15-28).

Figura 15-28: Movendo a dimensão Cliente para as colunas do eixo

Você também pode alterar a ordem das dimensões dentro de um eixo. Para fazer isso, clique em os pequenos ícones triangulares para mover a posição de uma dimensão. Por exemplo, se

você clica no ícone pequeno triângulo apontando para cima antes que a dimensão do cliente, ele é movido para uma posição acima. Pode obter o mesmo efeito clicando no pequeno direito no ícone triangular para baixo antes da dimensão Medidas. O resultado situação é mostrada na Figura 15-29.

Figura 15-29: Mudar a ordem das medidas e dimensões do cliente

Note que apenas temos estado a edição no Navigator OLAP-se o pivô tabela em si não mudou. Se você clicar em OK na OLAP Navigator, o OLAP Navigator está oculto e que a página é recarregada automaticamente para refletir a mudança nos eixos. O resultado é mostrado na Figura 15-30.

		Customer
		All Customers
		Measures
Date	DVD	Revenue
All Date.Months	All DVDs	$10,921,326.03
2000	All DVDs	$2,985.87
2001	All DVDs	$24,090.58
2002	All DVDs	$67,190.34
2003	All DVDs	$124,674.01
2004	All DVDs	$272,594.35
2005	All DVDs	$437,386.96
2006	All DVDs	$728,650.78
2007	All DVDs	$1,489,570.37
2008	All DVDs	$7,774,242.77

Figura 15-30: Cliente dimensão no eixo horizontal

Fatias com o Navigator OLAP

O Navigator OLAP não se limita a manipulação de coluna e linha de eixo. Você também pode usá-lo para especificar o fatiador. Lembre-se que o slicer corresponde à MDX ONDE cláusula e pode ser usado para mostrar apenas um subconjunto particular ou uma fatia""
dos dados. Vamos usar o slice para olhar apenas para um grupo específico de DVD. Para fazê-lo, primeiro você deve reabrir o Navigator OLAP. Deve parecer exatamente como Figura 15-29 neste momento. Agora clique no ícone da direita pouco antes do funil de DVD

dimensão. Isso faz com que a dimensão do DVD para passar para o slicer, como mostrado na Figura 15-31.

Figura 15-31: Movendo a dimensão de DVD para o slicer

Agora, embora você moveu a dimensão do fatiador, fatia"seu"ainda contém todos os dados do DVD. Isso é compreensível que o membro todos é o membro padrão eo membro contém todos os DVDs. Para ver uma parte da todos os DVDs, você precisa especificar quais membros da dimensão de DVD que você está dentro interessado em fazê-lo, clique na palavra DVD. Você recebe um menu como o mostrados na Figura 15-32.

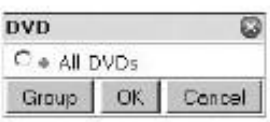


Figura 15-32: O cortador de DVD

Quando você clica no ícone de adição, a dimensão DVD expande dentro do OLAP Navigator, mostrando seus membros. Você pode então escolher um membro para definir da fatia. Selecione a fatia Ação / Aventura. Isso é mostrado na Figura 15-33.

Figura 15-33: Escolhendo uma fatia gênero

Lembre-se que você deve clicar em OK para fechar o Navegador OLAP antes de qualquer mudanças serão visíveis na tabela dinâmica.

Especificando Estados jogos com o Navigator OLAP

Você acabou de aprender como você pode especificar um determinado membro do slicer. Você
também pode especificar membros em particular sobre as colunas e linhas de eixos.

1. Abra o Navegador OLAP, e clique em cliente.
2. Expandir o Cliente, e selecionar apenas Canadá e Estados Unidos.
3. Clique em OK para confirmar.
4. Em seguida, clique em Data. Expanda Data, e escolher apenas 2007 e 2008.
5. Clique em OK para confirmar.

As seleções são mostrados na Figura 15-34.

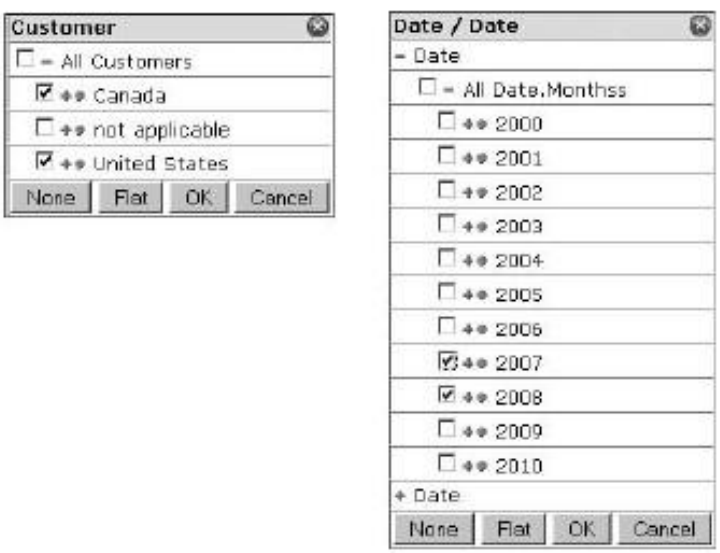


Figura 15-34: Escolhendo membros em particular para os eixos

Se você fechar o navegador OLAP, a tabela dinâmica é atualizada a aparência Figura 15-35.

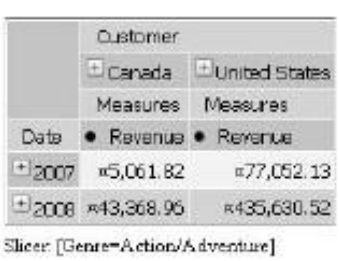


Figura 15-35: Anos de 2007 e 2008, os clientes do Canadá e dos Estados Unidos, em fatias por Ação / Aventura

Resultados de várias medidas

Podemos usar o OLAP Navigator também para mostrar mais medidas. Abra o OLAP Navigator, clique e Medidas. Um menu aparece mostrando todas as medidas definidas no cubo (ver Figura 15-36).

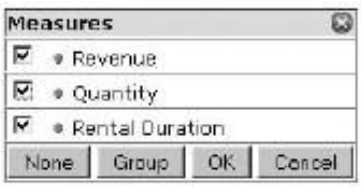


Figura 15-36: Seleção de múltiplas medidas

Agora selecione as restantes medidas e clique em OK para fechar a OLAP Navigator. O resultado é mostrado na Figura 15-37.

	Customer					
	Canada			United States		
	Measures			Measures		
Date	Revenue	Quantity	Rental Duration	Revenue	Quantity	Rental Duration
2007	$5,061.82	614	17.13	$77,052.13	10,839	17.39
2008	$43,368.96	4,627	15.35	$435,630.52	54,218	16.29

Slicer: [Genre=Action/Adventure]

Figura 15-37: Uma tabela dinâmica com múltiplas medidas

Diversos recursos

Nesta seção, discutimos brevemente uma série de recursos úteis que JPivot não abrangidos em outros lugares.

Painel de Consulta MDX

Você pode alternar a visibilidade do painel de consulta MDX usando o segundo botão na barra de ferramentas JPivot (que diz MDX). Este painel de consulta contém o MDX consulta que representa o estado atual da opinião de análise.

O painel de consulta MDX é muito útil para estudar e aprender a sintaxe MDX. Navegando em uma tabela dinâmica e interpretar o resultado é bastante prestados straightfor-ala para a maioria das pessoas, mais do que a compreensão dos resultados baseados em texto
devolvido pelo editor MDX PSW.

Você também pode usar o painel de consulta MDX para adicionar temporariamente uma definição para
um membro calculado e imediatamente inspecionar os resultados. Você pode voltar e para trás algumas vezes entre o painel e MDX a tabela dinâmica até que você esteja satisfeito. Você pode então colocar o código para este membro calculado diretamente dentro do cubo.

PDF e Excel Exportar

O ponto de vista de análise permite exportar o resultado processado para PDF e formatos do Microsoft Excel com um único clique do mouse. Para exportar para o Excel, basta clicar no botão da barra de ferramentas JPivot com o ícone do Excel. Ele está localizado na extremidade direita da barra de ferramentas JPivot.

Gráfico

JPivot também oferece um recurso gráfico. Você exibir o gráfico, alternando o Mostrar o botão da barra de ferramentas Gráfico (o ícone barchart na metade direita do JPivot barra de ferramentas). Você pode controlar a maneira como o seu aspecto gráfico usando a configuração de gráfico botão, localizado no lado direito da barra de ferramentas Show Chart. A alternância do Gráfico botão Config exibe um gráfico de Propriedades do formulário onde você pode especificar todos os tipos de propriedades do gráfico, tais como o tipo de gráfico e tamanho do gráfico. Nós não podemos possivelmente cobrir JPivot gráficos na íntegra. Em vez disso, recomendamos que você experiência com esse recurso. Para começar, nós fornecemos instruções para criar um gráfico como o mostrado na Figura 15-38.

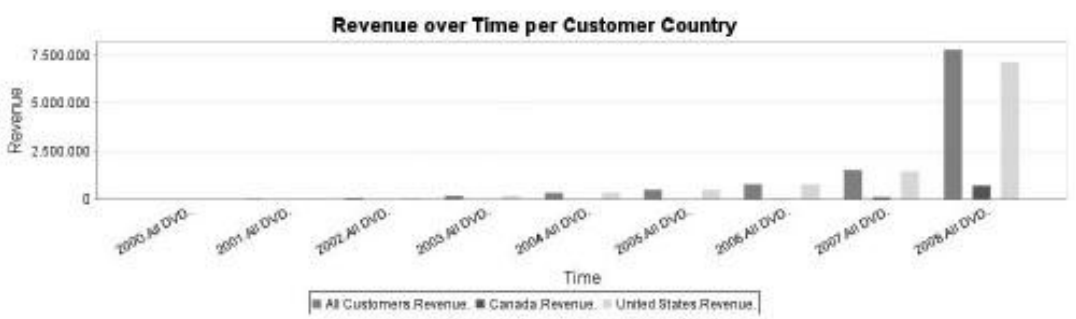


Figura 15-38: Um gráfico mostrando JPivot receitas sobre dividir o tempo dos clientes por país

O gráfico mostrado na Figura 15-38 fornece uma introspecção no desenvolvimento de receitas volvimentos ao longo do tempo, divididos por país de residência dos clientes. O gráfico dados que é visualizado no gráfico é mostrado na Figura 15-39.

Você pode obter esse ponto de vista sobre os dados, colocando a data e DVD dimensões no eixo vertical ea dimensão do cliente na horizontal eixo. Certifique-se de expandir o membro Todos os clientes para dividir os dados de acordo para o país de residência do adquirente. Além disso, você deve remover a Todos membro da dimensão de data. (Se você não remover o membro Tudo aqui, As receitas do primeiro par de anos será pouco visível em comparação com os grandes receitas acumuladas para o membro Datas Todos).

		Customer		
		All Customers	Canada	United States
		Measures	Measures	Measures
Date	DVD	Revenue	Revenue	Revenue
2000	All DVD	2,985.87		2,985.87
2001	All DVD	24,030.58		24,030.58
2002	All DVD	67,190.34		67,190.34
2003	All DVD	124,674.01		124,674.01
2004	All DVD	272,594.35		272,594.35
2005	All DVD	437,386.96		437,386.96
2006	All DVD	728,650.78	14,139.30	714,511.48
2007	All DVD	1,489,570.37	80,279.18	1,409,291.19
2008	All DVD	7,774,242.77	658,570.07	7,115,672.70

Figura 15-39: Dados para o gráfico mostrado na Figura 15-38

Figura 15-40 mostra a configuração do gráfico. Entramos em algumas propriedades para configurar o gráfico:

- Tipo de gráfico: barra vertical. Existe uma grande variedade de tipos de gráficos disponíveis.
- Habilitar o detalhamento: marcada. Isso permite que o usuário clique sobre as barras No quadro para examinar as linhas de detalhes subjacentes.

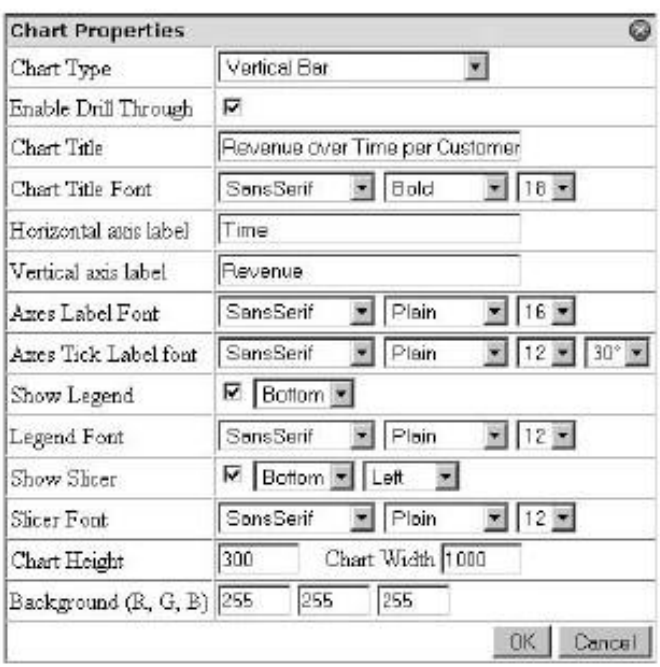


Figura 15-40: Gráfico propriedades para o gráfico mostrado na Figura 15-38

- Eixo horizontal no rótulo: Tempo; rótulo do eixo vertical: Receitas. Introduzir um rótulo torna muito mais fácil para os usuários a interpretar o gráfico.
- Gráfico de Altura, largura do gráfico: a altura e largura em pixels da tabela. Você vai descobrir que muitas vezes você precisa de experiência com esses valores para obter
um quadro razoavelmente legível.

Melhorando o desempenho usando o Pentaho Designer Aggregate

Mondrian faz um ótimo trabalho de resultados em cache, dando-lhe acesso rápido a obtidas anteriormente membros do cubo. Com grandes bases de dados, no entanto, é preciso tempo considerável navegar um cubo, porque você está trabalhando com um ROLAP ferramenta que precisa para recuperar dados de grandes tabelas de fatos detalhados. Para evitar
ter de digitalizar milhões de linhas de cada vez que um nível de dimensão nova for seleccionada,
você pode pré-agregados os dados para Mondrian. No capítulo 6, explicamos os conceitos de tabelas de agregados e visões materializadas como o desempenho impulsionadores e mencionou que, até à data, nenhum dos bancos de dados open source apoia o último. Bem, isso pode ser verdade para os bancos de dados em si, mas o comportamento de Mondrian usando tabelas agregadas chega muito perto de ter materializada com a opinião. Há uma diferença notável, porém, e que é a manutenção automática dos quadros agregado de cada vez que novos dados é carregados para o data warehouse. Uma maneira de resolver este problema é estender sua Chaleira empregos para atualizar as tabelas Mondrian globalmente no final de carga os dados detalhados para o data warehouse. Usando a data de limpeza e colunas de hora ou uma identificação de lote torna mais fácil identificar alterado ou inserido registros e usar isso como fonte para a atualização ou inserção de registros na agregados.

Agregação de Benefícios

O único benefício do uso de agregados é melhorar o desempenho da consulta, ou em no caso de Mondrian, ad-hoc de velocidade. Tabelas agregadas isso limitando o número de registros a serem verificados em uma consulta. Dê uma olhada Figura 15-41 para ver os efeitos de agregação. O lado esquerdo do diagrama mostra uma versão simplificada da tabela de fatos a fim WCM com 1.359.267 linhas porque o nível de transacção mais baixos do que os dados são armazenados na tabela. Agora olhar para o canto superior direito do diagrama onde criamos tabelas de agregados para o fato de as linhas detalhadas a nível do mês, o DVD Gênero e país. Este resulta em uma diminuição drástica do tamanho da mesa, porque você só precisa de 9.120 linhas
para armazenar todas as combinações possíveis de dados a estes níveis.

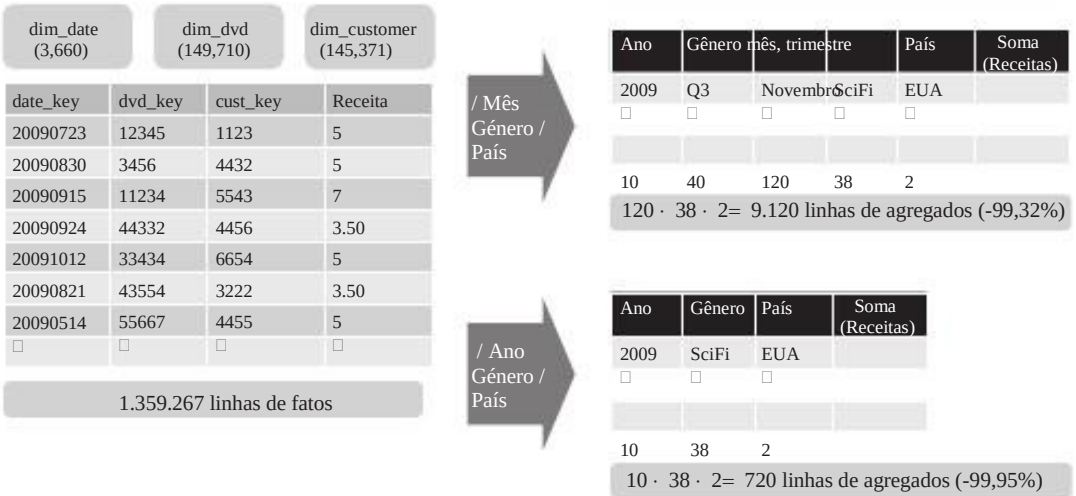


Figura 15-41: exemplos de Agregação

Uma redução adicional pode ser obtida quando você omitir o trimestre e nível do mês todo. Você ainda pode criar um relatório guia valioso cruz ou visão de análise com base nesses 720 linhas de dados, que é uma redução na tamanho de cerca de 99,95 por cento. Consultas contra essas tabelas agregadas produzir sub-segundo os tempos de resposta e usuários muito feliz. A questão é, como você incorporar esse conhecimento na sua configuração de Mondrian?

Estendendo Mondrian com tabelas agregadas

Antes havia uma ferramenta automatizada para projetar as tabelas dos agregados, Foi um processo manual. A próxima seção irá abranger agregados Pentaho Designer, embora você provavelmente usar esta ferramenta para desenhar e agregados de construção, que você precisa saber o que está acontecendo sob o capô no caso de algo correr mal. O exemplo que mostramos aqui é baseado na superior direito da tabela agregada na Figura 15-41. O primeiro passo é criar um novo tabela no esquema do data warehouse:

```
CREATE TABLE agg_wcm_orders_1 (
   Dim_date_Year CHAR (4),
   Dim_date_Quarter CHAR (2),
   Dim_date_Month CHAR (3),

   Dim_dvd_release_Genre VARCHAR (35),
   DUPLO fact_orders_Revenue,
   fact_orders_fact_count INTEGER)
```

Como você continuar, manter os seguintes pontos em mente:

- A tabela de exemplo agregado não contém apenas as colunas que você precisa para análise, mas tem um extra fact_count coluna também. O conteúdo deste

coluna indica como linhas facto muitos foram resumidos na
linha de agregação. Mondrian também tem esta coluna extra para verificar se o
tabela é um agregado e pode ser usado como tal.

- Além da dim_date_Month, O e dim_date
_Year colunas estão incluídos na tabela de agregados. Esta inclusão permite
Mondrian para obter todas as informações necessárias a partir da dimensão
tabela de agregados e não requer que você inclua o dim_date tabela
a consulta. Esta é também referida como dimensão do colapso ou cumulativo
conformado.
- A fact_orders_Revenue coluna é de um tipo diferente do que o receitas
coluna no fact_orders tabela. Porque vocês são a síntese dos dados, você
precisa ter certeza de que os valores somados ainda se encaixam as colunas no
nível agregado.
- O diagrama na Figura 15-41 mostra todas as dimensões em ambos os agregados
tabelas. Isso não precisa ser o caso: a coluna Gênero poderia ter
foi deixado de fora de uma tabela de agregados e ainda seria válido
agregada. Ao fazer isso, criaríamos um agregado com um perdido
dimensão (Gênero).

A consulta SQL mostrado na Lista 15-2 pode ser usado para carregar o agregado
tabela.

Listagem 15-2: Consulta SQL para carregar a tabela de agregado

```
INSERT INTO agg_wcm_orders_1 (
        dim_date_year
, Dim_date_quarter
, Dim_date_month
, Dim_customer_country
, Dim_dvd_release_Genre
, Fact_orders_revenue
, Fact_orders_fact_count)
dim_date_year SELECTdim_date.year4AS
, Dim_date_quarter dim_date.quarter_nameAS
, Dim_date.month_abbreviationAS dim_date_month
, Dim_customer.customer_country_code AS dim_customer_country
, Dim_dvd_release.dvd_release_genre dim_dvd_release_genre AS
, Fact_orders_revenue SUM (fact_orders.revenue) AS
, COUNT (*) AS fact_orders_fact_count
FROMfact_orders AS f
INNER JOIN dim_dateAS d ON f.local_order_date_key d.date_key =
INNER JOIN dim_customerAS c ON f.customer_key c.customer_key =
INNER JOIN dim_dvd_release AS r ON f.dvd_release_key r.dvd_release_key =
GRUPO BYd.year4
, D.quarter_name
, D.month_abbreviation
, C.customer_country_code
, R.dvd_release_genre
```

DICA Um agregado que é baseado em um outro agregado, por exemplo, uma tabela com Só no ano, país, e de receitas deve ser carregado a partir da agregação, não de a tabela de fatos base. Isso pode resultar em um ganho de desempenho considerável.

Depois disto, foi configurado, você pode usar o PSW novamente para especificar o tabela de agregados em seu cubo existente. Quando você expande o wcm_orders cubo e clique com o botão direito fact_orders, Abre um menu com três opções. O primeiro permite que você declare uma nova tabela agregada eo terceiro permite que você explicitamente excluir uma tabela. A segunda opção é realmente o mais interessante uma vez que permite definir um padrão que Mondrian usa para determinar quais tabelas no banco de dados são agregados que podem ser usados. Explicando o necessário expressões regulares e como funciona este recurso está além do escopo deste livro, mas este tópico é abordado em profundidade no site da Pentaho em Mondrian

notes.php http://mondrian.pentaho.org/documentation/developer # Agg _ _regras padrão.

Neste exemplo, ficar com a primeira opção, que permite que você adicione a agg_wcm _orders_1 nome agregado. Isso muda a fact_orders e adiciona um ícone

nível chamado de agregação Nome, que por sua vez, oferece um menu do botão direito do seu
próprio. Aqui você precisa especificar a coluna contagem agregada fato (neste caso: fact_orders_fact_count), Os níveis de agregação, e medidas agregadas. A níveis e as medidas têm dois atributos: a coluna eo nome do item.
A coluna se refere ao nome da coluna na tabela de agregados, mas o nome é a especificação MDX da instância de nível para [Cliente]. [País] para o customer_country coluna e [Data Local Ordem]. [Mês] para o month_number coluna. Um exemplo é mostrado na Figura 15-42.

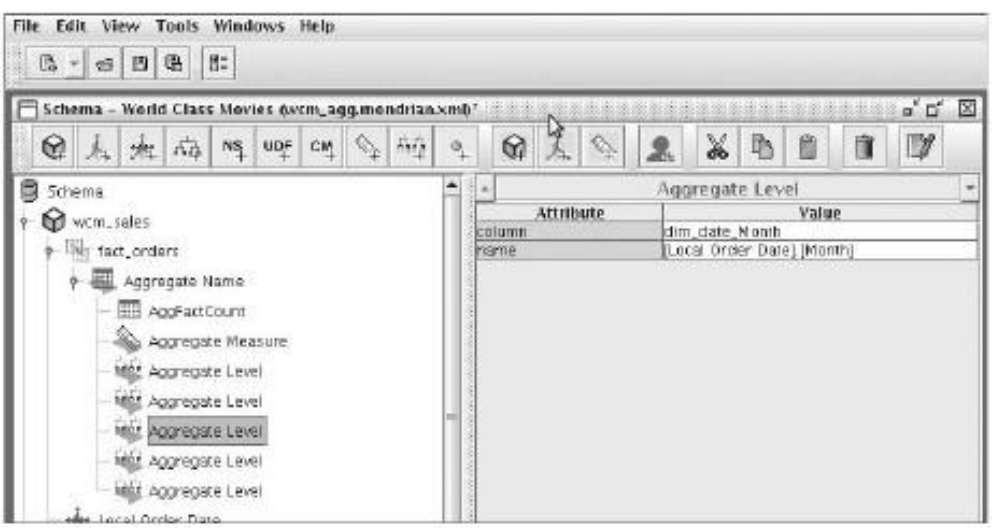


Figura 15-42: Nível de agregação de especificação

Até agora usamos apenas as colunas de agregados em nossas mesas agregado, mas você pode criar uma tabela resumida, onde uma ou mais colunas ainda contêm a chave de dimensão detalhada. Nesse caso, não há necessidade de recolher o dimensão e incluem os diferentes níveis da tabela agregada, como no Ano-Quartas-de-Semana exemplo. Se você quiser ser capaz de perfurar até o nível de dia, combinado com um agregado do país e do Gênero, você poderia substituir o Trimestre, ano, mês e colunas com um único date_key e especificar que o uma peça de agregação dos Negócios Estrangeiros. A coluna agg seria, neste caso, o date_key, ea coluna de fato seria o original local_order_date_key coluna.

HABILITAÇÃO tabelas agregadas

A utilização de tabelas agregadas em Mondrian é desabilitada por padrão. Você pode habilitar deste recurso, definindo os valores dos mondrian.rolap.aggregates.Use e mondrian.rolap.aggregates.Read para True em Mondrian propriedades do arquivo. Este arquivo pode ser encontrado no diretório <Pentaho instalar > Dir / biserver-ce/pentaho-solutions/system/Mondrian. Se o propriedades não estão presentes, basta adicionar as seguintes linhas para este ficheiro:

```
mondrian.rolap.aggregates.Use = 1
mondrian.rolap.aggregates.Read = 1
```

Pentaho Designer Aggregate

O parágrafo anterior cobriu a configuração manual de tabelas de agregados dentro Pentaho. Agora que você tem uma compreensão básica do que está acontecendo quando você está criando tabelas agregadas, é hora de tornar a vida um pouco mais fácil, usando Designer Pentaho Agregado, ou DAP. PAD oferece uma série de vantagens sobre o processo manual de criação de tabelas de agregação. O mais importante eles é que PAD é capaz de selecionar os melhores automagicamente tabelas agregadas para o esquema que você está tentando otimizar. Quando você começar a PAD, ele pede um banco de dados de conexão. Depois que é configurado, você precisa selecionar um Mondrian arquivo de esquema, e depois de aplicar a escolha, o cubo para trabalhar pode ser aberto. PAD, em seguida, valida o modelo de dados subjacente, verificando que todas as tabelas possuem chaves primárias e que não estão presentes os valores nulos nas chaves estrangeiras colunas. PAD avisa quando a tabela de fato não contém uma chave primária.
Você pode usar o DAP para fazer qualquer um dos seguintes:

- Crie a sua definição de tabelas agregadas manualmente clicando no botão Adicionar botão do lado direito e, posteriormente, selecionar os níveis de agregação à esquerda. Cada nova tabela agregada é adicionado à lista na parte inferior direita da a tela e pode ser editado clicando-lo.

- Pré-visualizar, executar ou exportar o script de criação para os agregados da Exportação e tela de publicação, que pode ser aberto clicando no botão Exportar botão.
- Execute ou exportar os scripts de carga para os quadros agregados a partir do mesmo Exportação tela.
- Publicar o esquema Mondrian ajustado a um servidor Pentaho ou exportar o arquivo para uma pasta no disco.

Estas são características muito úteis, mas a verdadeira diversão começa quando você clica no
Advisor botão. Aqui você pode especificar o número de agregados PAD pode tentar gerar e quanto tempo o conselheiro de consultas podem ser executados. Neste simples exemplo, você pode dizer que agrega 15 e 60 segundos e clique em Recomendar. PAD, em seguida, tenta determinar todos os agregados com base no design do cubo e do contagens coluna do modelo de dados subjacente. Figura 15-43 mostra os resultados deste exercício para o wcm_orders cubo.

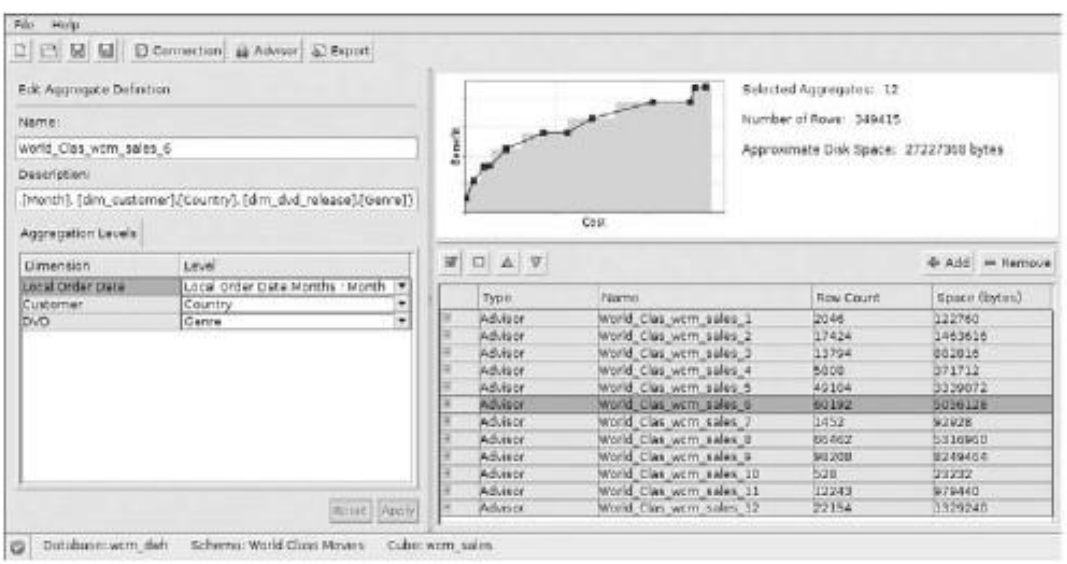


Figura 15-43: tabelas geradas Aggregate

Esta ainda é uma recomendação criado por PAD, que você pode aprovar ou rejeitar. Outra opção é para ajustar manualmente as tabelas de Conselheiro ou descartar a que você sente não será necessário. Você pode observar os efeitos das suas escolhas imediatamente no diagrama. Até agora, nunca encontrei a necessidade de contradizer os resultados conselheiro e você será pressionada duramente para encontrar uma melhor optimizado conjunto de
tabelas agregadas que não PAD.

Quando você olha para a contagem das tabelas conselheiro, você provavelmente encontrará que o número para o conjunto criado manualmente no parágrafo anterior

é muito maior que na Figura 15-41 DAP, porque também tem os níveis (Todos) em consideração, assim que os números apresentados são o máximo absoluto. A manual agg_wcm_sales_1 tabela tem um número de linhas PAD calculado de 60.192 Considerando que, na realidade, a tabela contém apenas 4.672 linhas, o que é ainda menos manualmente o número previsto.

Quando estiver satisfeito com o modelo criado, você pode usar novamente a exportação e publicar opções para criar e carregar as tabelas e publicar a actualização Mondrian esquema para o Servidor Pentaho.

Soluções Alternativas

Usando tabelas agregadas tem uma desvantagem considerável: as tabelas têm de ser atualizada cada vez que os dados no armazém de dados é atualizado. Em menor ambientes, esse é fácil de conseguir, usando os scripts de carga exportada para criar postos de trabalho simples Pentaho Data Integration. Mas não seria bom se tudo isso era desnecessária, pois seu banco de dados foi rápido o suficiente para entregar o desejado tempos de resposta, sem o uso de tabelas de agregação? No capítulo 5, mencionamos algumas dessas alternativas, mas é bom fazer isso de novo aqui. LucidDB, Infobright MonetDB e são todos muito rápidos, bases de dados da coluna-orientado. Estes bases de dados foram especialmente concebidos para o tipo de trabalho encontradas em um OLAP ambiente e valem a pena olhar e avaliar. No momento da escrever ainda havia problemas com o driver JDBC MonetDB e Mondrian, mas LucidDB em obras particulares muito bem em conjunto com Mondrian. A LucidDB na Wiki http://pub.eigenbase.org/wiki/LucidDbOlap descreve como configurar isso.

Resumo

Este capítulo mostrou como criar soluções OLAP para o BI Pentaho Plataforma. Os tópicos abordados incluíram o seguinte:

- Introdução aos conceitos e sintaxe MDX
- Criar soluções multidimensionais usando o Pentaho esquema Trabalho banco
- Criando visões análise utilizando o front-end JPivot
- Melhorar o desempenho OLAP usando tabelas agregadas
- Usando o Pentaho Designer agregadas para criar tabelas agregadas

CAPÍTULO

16

Data Mining com Weka

livros populares, como Moneyball, Freakonomics, e Competir no Analytics
têm aumentado o interesse em utilizar o Analytics para obter uma vantagem competitiva. Este capítulo explica como algumas das técnicas mais populares do trabalho analítico
e como eles podem ser aplicados em cenários da vida real. analistas de negócios e profissionais de BI estão acostumados a apresentar relatórios sobre o desempenho organizacional.
Até agora, a maioria das pessoas está familiarizada com o uso de ferramentas de BI e OLAP relatório, para identificar exceções, e responder a perguntas básicas. O desafio para muitas pessoas é que novas perguntas exigem novas formas de olhar para os dados. técnicas de Reporting e OLAP são boas quando os tipos de perguntas são
bem estabelecida, e para explicar a atividade passada ou atual. Estas técnicas
não pode ser usada para compreender as complexas relações, explorar grandes volumes de dados detalhados, ou prever a atividade futura. Mineração de dados (incluindo visualização e texto analytics) fornece os meios para realizar tarefas que não são possíveis
Estas análises avançadas geralmente não são utilizados porque
à sua complexidade e custo assumido. A verdade é que muitas técnicas
pode ser aplicada de forma simples, e muitas vezes com custo relativamente baixo, às vezes livre-tools. Uma das ferramentas mais populares é Pentaho Data Mining (PDM),
mais conhecido como Weka (rima com""a Meca), que é o tema da
capítulo atual.

Embora mineração de dados é um termo familiar utilizado para designar o assunto em mãos,
Algumas pessoas preferem chamá-lo aprendizado de máquina ou (Automatizado) descoberta de conhecimento.
aprendizagem de máquina é realmente um assunto mais amplo do que a mineração de dados, mas o
termos são em grande parte intercambiáveis. Para este livro, nos ater ao termo
mineração de dados.

Data Mining Primer

A mineração de dados tem sido um assunto de muita confusão e às vezes até mistério: o uso de computadores para a tomada de decisão automatizada ou simulem o processo de pensamento humano, usando redes neurais tem um efeito assustador sobre algumas pessoas. Para dizer a verdade, não há nada misterioso ou assustador mineração de dados, mas é verdade que a natureza percebida complicado assusta muitos as pessoas longe dela. Isso é uma pena porque a mineração de dados acrescenta um lote de potência
a sua caixa de ferramentas de análise, embora exija um certo investimento de aprendizagem.

O que exatamente é a mineração de dados? A mineração de dados é muitas vezes considerada uma
mistura de estatística, inteligência artificial, banco de dados e pesquisa, mas que é não é realmente uma definição. Uma definição cético pode até ser que a mineração de dados é igual a estatísticas mais marketing. O ceticismo de lado, a seguinte definição é comumente citado:

> Data Mining é o processo não trivial de identificação de romance, válidos, potencialmente padrões úteis e, finalmente, compreensíveis em dados.
>
> UM Fayyad, G. Shapiro-Piatetsky e Smyth P.,"De
> Data Mining para Descoberta de Conhecimento: Uma Visão Geral", em
> Avanços na Descoberta de Conhecimento e Mineração de Dados, edição
> U. Fayyad M, G. Shapiro Piatetsky, P. Smyth, e
> R. Uthurusamy, Imprensa AAAI / MIT Press, pp 1-34, 1996.

Esta definição nos diz várias coisas importantes. Em primeiro lugar, trata-se de descobrir padrões nos dados que precisam ser compreendidos. A mineração de dados envolve o trabalho
com grandes conjuntos de dados com milhões de linhas, por vezes, e centenas de colunas, que são de alguma forma ou de outra relacionados uns aos outros. Os padrões que queremos a descobrir necessidade de ser compreensível, portanto, os resultados têm de fazer sentido. A resultados também deve ser nova, o que significa que eles devem dizer-nos algo que nós não sabia antes, e os resultados devem ser válidos dentro de um contexto específico. Além disso, as conversações sobre a definição de um processo não-trivial, o que significa que não estamos fazendo isso por diversão, mas para resolver um negócio, social ou científica problema. aplicações de negócios comuns de mineração de dados são a detecção de fraudes, marketing direto e retenção de clientes. Mas você sabia que a mineração de dados
É também uma ferramenta importante na agricultura, onde é usado para diagnosticar de soja doenças ou marcar a qualidade dos cogumelos? Outro exemplo é o da
área das ciências da vida, onde a mineração de dados é utilizado para prever a taxa de sobrevivência
embriões. Na verdade, existem muitas áreas onde a mineração de dados poderia ser útil, mas no nosso caso, vamos nos concentrar em aplicações de marketing e vendas com base em
os dados no banco de dados mundial de filmes de classe.

Processo de Data Mining

Uma maneira fácil de começar é olhar para a mineração de dados como um processo. De fato, em
1996, um grupo de empresas industriais decidiu desenvolver uma mineração de dados

metodologia conhecida atualmente como CRISP-DM, abreviação de Cruz Indústria Stan-
Processo dard para Mineração de Dados (um guia de referência completa está disponível online em
http://www.crisp-dm.org). O fluxo de trabalho de base que tinham acordado é exibido
na Figura 16-1, que mostra o fluxo natural das etapas que compõem um conjunto de dados
mineração processo.

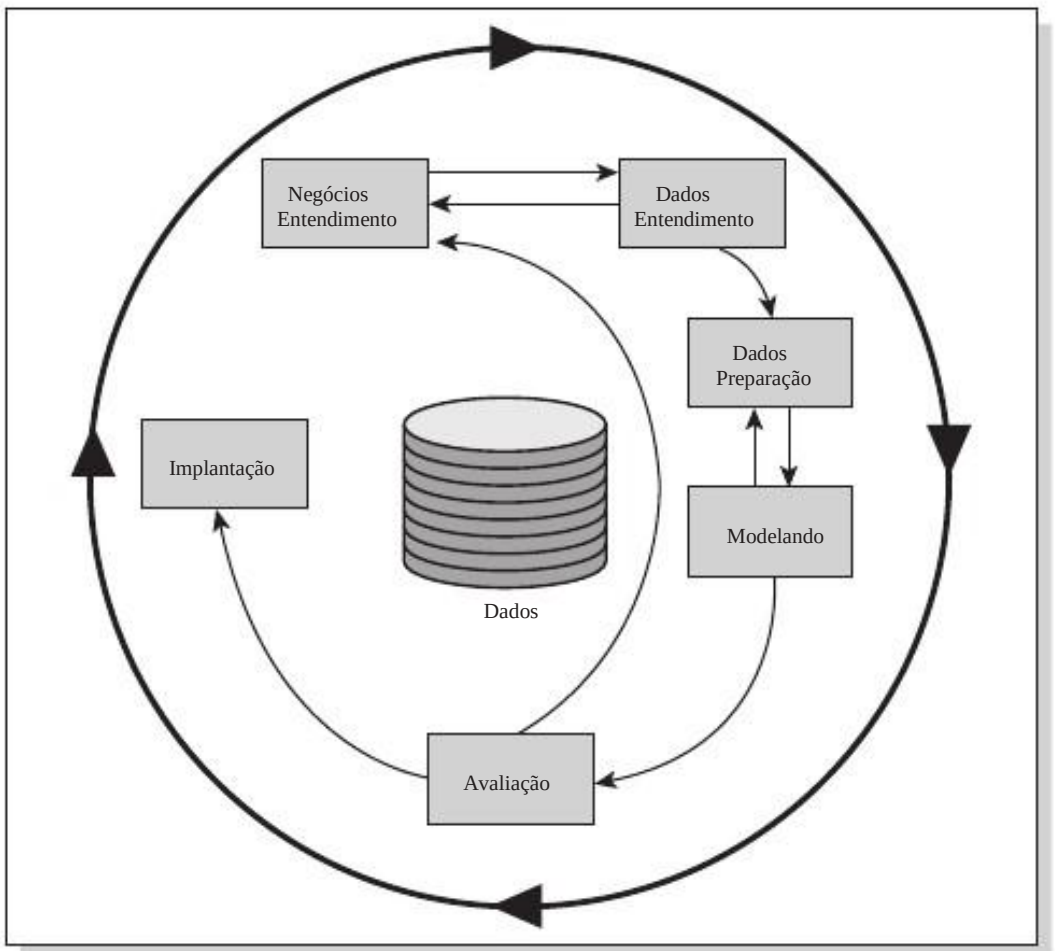


Figura 16-1: método CRISP-DM

Este modelo de processo mostra que há muita sobreposição entre BI e dados
atividades de mineração:

- A mineração de dados, como acontece com o BI, é um ciclo contínuo de atividades com o máximo
  objetivo de tomar melhores decisões
- Os dados são o centro de todas as atividades, sem dados, não há BI e não há dados
  mineração
- As três primeiras etapas do modelo de processo é muito semelhante à construção de uma
  data warehouse. Conhecimento do negócio, compreensão dos dados, e dados
  preparação foram objecto dos capítulos 5-11, e tudo o descrito
  não pode ser utilizada para mineração de dados também.

- Tal como em projetos de BI, cerca de 70 a 80 por cento do esforço é gasto no primeiro três atividades. Obtendo os dados certos e obter os dados corretos é ainda a parte mais demorada de qualquer projeto de mineração de dados como é para um BI projeto.

A diferença começa logo que você pode começar com a fase de modelagem. Modelagem significa que você considera os diferentes modelos e escolher o melhor com base no seu desempenho preditivo, ou seja, você está procurando o melhor ajuste.""Em um cenário do mundo real isso envolve testar diferentes soluções com diferentes conjuntos de dados de exemplo, tentando encontrar um modelo que produz os resultados mais estáveis
entre as diferentes amostras. E por resultados, significa que o modelo é capaz prever o resultado com base em dados de entrada. O objetivo da mineração de dados não é para explicar todas as relações possíveis entre os dados, mas é mais voltado para encontrar uma solução prática para prever certos resultados.

Data Mining Toolset

A natureza preditivo de mineração de dados é exatamente o que o torna tão atraente para um grande número de usuários de negócios e que a distingue de BI processos regulares e ferramentas. BI, como já cobriu nos capítulos anteriores deste livro, foi sobre os relatórios e análise sobre o desempenho passado, e comparando as metas com as medidas reais. A mineração de dados adiciona a capacidade de fazer previsões sobre o desempenho futuro, dentro de certos limites de confiança. As ferramentas para fazer isso são modelos e algoritmos.

As tarefas que queremos realizar com a mineração de dados são diferentes do que temos visto até agora, mas a maioria das metas pode ser alcançada usando as seguintes quatro
categorias principais: classificação, associação, agregação, de regressão e. Cada um dos Estes serão explicados nas seções seguintes. O que todas essas categorias têm em comum é o fato de que eles tentam prever ou explicar certos resultados (O alvo ou desconhecido valor) com base nos dados disponíveis.

Classificação

Classificação é o processo de dividir um conjunto de dados em mutuamente exclusivos grupos de tal forma que os membros de cada grupo são os"fechar"possível grupos entre si e diferentes como"muito"possível um do outro,
onde a distância é medida com relação a variável específica (s) que você está tentando de prever. Por exemplo, um típico problema de classificação é dividir um banco de dados de clientes em grupos que são tão homogéneos quanto possível no que diz respeito à uma variável de retenção com valores corredor de fundo e leaver. Você também pode tentar a classificação
clientes com base em sua classe de receita, como você vai descobrir mais tarde, a fim de prever a rentabilidade possível de um novo cliente.

A classificação começa por definir a variável de desfecho, neste caso, por exemplo, retention_group ou revenue_class. Para treinar o modelo, você

preciso de um conjunto de dados com diferentes variáveis e um resultado conhecido. Na retenção
caso você precisaria de um conjunto de dados composto de clientes leais que têm
foram produtos de encomenda de pelo menos dois anos, combinado com os clientes que
rescindido seu contrato, em algum momento no tempo. Para ambos os grupos que você
necessitam de informações extras, como dados demográficos, dados geográficos, ea ordem
da história. Com base nestes dados você pode então determinar quais variáveis contribuem
mais para esse resultado, para passar meses em última instância do cliente, os itens
adquiridos durante um determinado período, ou talvez o grupo de renda. Determinação da
impacto de cada uma dessas variáveis, ou qualquer combinação deles é um típico
extracção de dados da tarefa. Depois de concluído o processo (que trem o modelo), você
é capaz de validar o modelo através da execução do algoritmo de classificação contra
um conjunto maior de dados, que é chamado testar o modelo. Isso permite que você verifique
se o modelo satisfaz suas necessidades (o que poderia ser"cliente
grupo de lucratividade é previram corretamente em 95% ou mais dos casos") e se
isso acontecer, ele pode ser usado para os futuros clientes da pontuação também.

Porque a classificação é uma das aplicações mais utilizadas da mineração de dados
(Eo mais fácil de obter seus pés molhados), vamos trabalhar com um par de
exemplos mais adiante neste capítulo.

Clustering

Cluster ou segmentação do banco de dados é muito semelhante à classificação na
sentido de tentar identificar quais os elementos de um conjunto de dados ação ordinária
características e agrupá-los com base nessas características. A
diferença mais notável, entretanto, é que com o agrupamento deixar o algoritmo
determinar os grupos com base em uma seleção ou até mesmo todos os dados disponíveis, enquanto
com a classificação que você já tem os grupos definidos. Esta diferença também é
identificados como supervisionada (Classificação) versus não supervisionado (Clustering) de aprendizagem.
Clustering tem uma natureza mais exploratória, porque quando o data-mining
algoritmo determinou que os grupos de itens juntos, você pode
olhar o raciocínio por trás dele. A ferramenta de mineração de dados vai mostrar como ele fez a sua
decisão e quais as variáveis que mais contribuíram para os diferentes grupos. Agora
cabe ao analista investigar os diferentes grupos, porque eles são diferentes,
e como isso pode ser usado para tomar melhores decisões. Se os clusters encontrados comprovam
para ser útil, eles podem então ser utilizados como classes em uma análise de classificação.

Associação

Com a associação que você tente descobrir qual é a relação entre duas ou mais
elementos de dados em um conjunto de dados. Uma das melhores aplicações conhecidas da associação
é a análise de cesta de mercado, onde um grande número de transações são verificados para
determinar quais os artigos são comprados em que combinação. Um exemplo famoso
disso é o caso das fraldas e cerveja que você possa ter ouvido ou lido. A
grande cadeia de varejo analisados milhões de transações de vendas a partir do ponto de vendas

sistema e encontrou uma forte correlação entre as vendas de cerveja e fraldas.
A explicação que chegou foi que os jovens pais a caminho de casa
do trabalho pegou os mantimentos. E desde que eles não poderiam sair com seus
amigos por causa do bebê, que pegou um pacote de cerveja para beber em casa.
Como resultado, a cadeia de varejo começaram a posição dos dois produtos junto a cada
outros nas lojas para aumentar as vendas e serviço melhor seus clientes. Ou pelo menos,
essa é a história. Descobriu-se finalmente que houve de fato uma correlação,
mas não uma relação causal, sob a forma"compra fralda compra cerveja."A lição
aqui, e um dos erros mais freqüentes feitas em mineração de dados, é confuso
correlação de causalidade.

Numéricos de previsão (Regressão)

Os três categorias de aplicações de mineração de dados específicos prever
classes (valores nominais), que são valores não-numéricos, como boa,ruim,
Ou leaver. Muitas vezes, você quer prever um resultado numérico com base no passado
eventos, que é um estilo diferente de aplicação, porque agora você não tem
valores discretos como resultado, mas uma gama infinita de valores numéricos possíveis.
Pense em prever o desempenho de um computador baseado em componentes
e as configurações utilizadas, ou a estimativa de receita anual de um cliente baseado
sobre as características do cliente e comportamento de compra. Nós não cobriremos
regressão
neste capítulo, mas o"Leitura", secção no fim do capítulo
contém algumas excelentes referências se você quiser aprofundar mais o assunto.

Algoritmos de mineração de dados

As quatro categorias de aplicações de mineração de dados introduzidos no anterior
seções são a classificação de alto nível das ferramentas disponíveis para a mineração de
dados.
Você pode achar categorizações diferentes em outras fontes, mas as categorias que
uso aqui são baseados na divisão que o criador do Weka escolheu. O Weka
ferramentas, por conseguinte, também contêm estas mesmas quatro categorias básicas de
mineração de dados.

Categorias não são suficientes embora, para cada categoria são vários algoritmos
fornecidas para fazer o trabalho. Assim, a classificação é um tipo específico de dados
mineração, mas a classificação dos dados, precisamos de instrumentos tais como árvores de
decisão ou
algoritmos regra. Algoritmos são o que tornam uma ferramenta de mineração de dados
poderoso. Em
geral, os mineiros, os dados não se preocupam com interfaces de fantasia ou visualização
agradável
ferramentas, enquanto os algoritmos são suficientemente poderoso de uma previsão
ponto de vista. A maioria dos algoritmos populares estão disponíveis em domínio público
como eles são o resultado de pesquisas científicas divulgadas. Weka contém mais de 45
algoritmos.

ferramentas proprietárias de mineração de dados freqüentemente contêm melhorar ou
ampliar as versões
de algoritmos disponíveis publicamente. Um bom exemplo é o algoritmo de árvore de decisão,
conhecido como C4.5, que é o laborioso básico de qualquer minerador de dados. Um ajuste
versão deste algoritmo está disponível no Weka como o classificador de árvore de J48, mas
o sucessor do C4.5, que é conhecido como C5.0, só está disponível a partir de um

vendedor comercial. Porque esta nova versão apresenta melhorias tanto na velocidade e consumo de memória (e pequenas melhorias na precisão), muitas organizações decidem adquirir este software para obter melhores resultados de seus esforços de mineração de dados. Embora C5 tem reforço embutido, você pode começar um aumento de desempenho semelhante com Weka por J48 combinando e AdaBoost meta classificador.

Treinamento e teste

Dois termos que são freqüentemente confundidos com os outros será usada extensivamente quando você começa com um processo de mineração de dados: treinamento e teste. Formação é o processo de criação do modelo de mineração de dados; testes está usando um modelo obtidos a partir da fase de treinamento para validar a qualidade do modelo. A dados utilizados para essas duas atividades não devem ser os mesmos. Para tomar esta ainda mais: não deve haver qualquer sobreposição do conjunto de dados usados para treinar- ção e do conjunto de dados utilizados para testes. (Há exceções possível esta regra, mas eles não serão abordados aqui.) E para tornar ainda mais complicada: os dados utilizados para a formação deve ser uma seleção aleatória de con- tamanho considerável a partir do conjunto total de dados. Uma boa regra é que dois terços dos dados deve ser utilizado para o treinamento, eo restante (um terço) para o teste. Isto é normalmente referido como o método de validação""'do modelo de avaliação.

Você pode enfrentar novos desafios ao selecionar os dados corretos para treinar o seu modelo. Por exemplo, embora você pode querer, não é viável a utilização de 100 mil linhas de dados para construir um modelo, por isso, se o conjunto de dados é muito grande você deve considerar uma amostra do que antes de você começar a construir o seu modelo. Neste capítulo vamos mostrar como isso pode ser conseguido através de Pentaho Data Integration (PDI).

Estratificada de validação cruzada

Os conjuntos de dados tem um casal de possíveis problemas que tornam difícil de usá-los para construir um modelo de. A primeira questão é a aleatoriedade dos dados selecionados e possibilidade de que os valores de classe são desigualmente distribuídos entre a formação eo conjunto de teste. A solução para isso é estratificar os dados, o que garante que ambos os conjuntos de treinamento e de teste têm a mesma distribuição de valores de classe como os dados como um todo. O segundo problema ocorre quando os dados são limitados. A validação método (com estratificação) só vai uma parte do caminho para a protecção contra representação desigual de entrada e saída de valores em conjuntos de treinamento e teste. A forma mais geral para compensar qualquer viés causado pela amostra particular escolhido para a validação é usar uma técnica estatística chamada validação cruzada. Em validação cruzada, os dados são divididos em um número fixo de mutuamente excludentes partições ou dobras. Suponha que você use quatro dobras que significa que os dados são dividida em quatro pedaços aproximadamente iguais. Uma vezes é usado para testes

e os outros três para a formação. Este processo é repetido com uma dobra diferente de ensaio e outras três para a formação até todas as dobras foram utilizados em cada função. Usando estratificados em dez vezes a validação cruzada se tornou a forma padrão de predizer o desempenho de um algoritmo de aprendizado em aprendizado de máquina. Testes extensivos em muitos conjuntos de dados de benchmark mostram que dez é sobre o número certo de dobras para obter a melhor estimativa de precisão.

O Weka Workbench

Weka é uma ferramenta de mineração de dados, originalmente desenvolvido na Universidade de Waikato
na Nova Zelândia. O nome é um acrônimo para Waikato Ambiente para
Conhecimento de análise, mas também é o nome de um pássaro, que é agora o projeto mascote. Weka começou como um projeto financiado pelo governo em 1993 e seu objetivo "foi desenvolver um estado da arte da bancada de ferramentas de mineração de dados."Embora
Pentaho adotou a ferramenta Weka como seu mecanismo de mineração de dados, é a única parte
da plataforma de BI que ainda não está disponível no download do Pentaho regular sites. Além disso, Weka ainda não faz parte do nightly builds em Hudson, e o código fonte não pode ser descarregado a partir dos repositórios SVN Pentaho.
A integração na plataforma Pentaho é limitada a um plugin especial para o Chaleira para chamar um algoritmo Weka pontuação. No entanto, porque a ferramenta está em uma classe de sua
próprios, e provavelmente irá ser utilizado por especialistas, isto não vai representar um grande problema.
Em 1996, a primeira versão pública (2,1) foi lançado e em 1999 a versão 3 (100 por cento escrito em Java) foi lançado. A atual versão 3.6 é um incremental liberação ainda com base no código 3.0, tornando Weka, provavelmente, o mais maduro parte da plataforma de BI Pentaho.

Weka consiste de três ferramentas diferentes, cada um dos quais pode ser usado indepen- forma independente, mas, quando combinados, fazem uma plataforma de mineração de dados muito poderosa.
(Na verdade existem quatro ferramentas, mas duvidamos que você jamais vai usar o Cliente simples, que é uma interface de linha de comando para Weka) As principais ferramentas são as seguintes:

- O Explorer ponto de partida para se familiarizar com Weka e dados mineração. Explorer permite uma maneira fácil e exploratório (daí o nome) do trabalhar com conjuntos de dados. Também oferece uma ampla gama de funcionalidades.
- Experimentador intencionados para a criação e execução de experimentos maior onde vários conjuntos de dados e vários algoritmos podem ser adicionados simultanea- simultaneamente. Os resultados do experimento pode ser comparado com cada , para determinar quais foram os resultados (estatística) melhor do que outros.
- KnowledgeFlow-A mais recente adição ao conjunto de ferramentas Weka pode ser usado
para construir fluxos de trabalho de mineração de dados completo semelhante ao dos fluxos que são
familiarizado com a Pentaho Data Integration ou Design Studio.

Formatos de entrada Weka

Antes de analisar os dados, ele deve estar preparado para o uso em Weka. Weka pode ler dados de múltiplas fontes, incluindo diretamente de um banco de dados JDBC e CSV. Weka também tem seus próprios formatos de arquivos nativos. O primeiro é chamado ARFF (Attribute relação File Format), que é um formato de arquivo baseado em texto, mas com metadados adicionados para que Weka sabe que tipo de dados está no arquivo. Listagem 16-1 mostra um exemplo de um arquivo ARFF. Como você pode ver, ele contém o relação (o sujeito da análise), todos os atributos utilizados, incluindo tanto os possíveis valores ou o tipo de dados e os dados em si.

Listagem 16-1: ARFF formato com os dados meteorológicos

```
@ Tempo relativamente

@ Atributo      perspectivas {ensolarado, nublado e chuvoso}
@ Atributo      temperatura real
@ Atributo      umidade real
@ Atributo      ventoso {verdadeiro, falso}
@ Atributo      jogar {sim, não}


@ Dados
ensolarado, 85,85, FALSE, não
ensolarado, 80,90, TRUE, não
nublado, 83,86, false, yes
chuvosa, 70,96, false, yes
chuvosa, 68,80, false, yes
chuvosa, 65,70, TRUE, não
nublado, 64,65, TRUE, sim
ensolarado, 72,95, FALSE, não
ensolarado, 69,70, false, yes
chuvosa, 75,80, false, yes
ensolarado, 75,70, TRUE, sim
nublado, 72,90, TRUE, sim
nublado, 81,75, false, yes
chuvosa, 71,91, TRUE, não
```

O segundo formato é chamado XRFF (eXtensible atributo-relação de formato de arquivo) e é uma extensão baseada em XML do formato ARFF. Ambos ARFF e XRFF arquivos podem ser abertos de forma arquivado também. XRFF tem uma vantagem mais ARFF padrão que permite que o atributo da classe a ser especificada na arquivo. arquivos padrão ARFF não especificar um atributo de classe e deixá-lo para o usuário para selecionar uma via interface gráfica ou através de uma opção se estiver usando o Weka interface de linha de comando. Na Listagem 16-1, o atributo jogar é considerada classe (variável de desfecho), o usuário teria que estar ciente disso e escolher este atributo explicitamente. XRFF, por outro lado, permite que uma classe padrão

atributo a ser definido no arquivo. Este atributo, em seguida, é escolhido automaticamente Weka na GUI ou interface de linha de comando. Claro, isso não impede que o usuário selecionar manualmente outro atributo a ser tratada como a classe se o desejarem. Finalmente, XRFF permite adicionar o atributo e instância pesos (ARFF só suporta pesos exemplo), que permite equilibrar a importância de cada atributo em um resultado.

Há também um par de outros formatos de arquivo que pode lidar com Weka, como como C4.5, que é um formato de arquivo padrão semelhante ao ARFF mas o atributo descrições e os dados são divididos em dois arquivos separados. Com o LibSVM (Biblioteca de Support Vector Machine), Luz SVM, e SBI (binário serializado Instâncias), Weka é capaz de manipular dados de saída de mineração de dados de outros ou pacotes de estatística também.

Configurando conexões de banco de dados Weka

Como mencionado, Weka oferece a opção de ler dados de bancos de dados usando a interface JDBC. Antes que você possa usar este recurso algum esforço adicional de instalação é necessária porque não há out-of-the-box de suporte de banco de dados. Para ver o comportamento padrão do Weka, ao tentar utilizar um banco de dados, inicie o Weka a partir da primeira linha de comando.

DICA Um programa em Java normalmente pode ser iniciado através da abertura de uma tela do terminal ou pressionando as teclas Alt + F2 e digitando o comando

jar name> <jarfile - java

No caso do Weka, este se tornaria

java-jar weka.jar

Se você selecionar Explorer a partir da GUI Chooser, você verá a seguinte erros na consola:

```
Registrando --- --- Editores Weka
Erro, não CLASSPATH - RmiJdbc.RJDriver: Tentando adicionar JDBC driver?
Erro, não CLASSPATH - jdbc.idbDriver: Tentando adicionar JDBC driver?
Erro, não CLASSPATH - org.gjt.mm.mysql.Driver: Tentando adicionar JDBC driver?
Erro, não CLASSPATH - com.mckoi.JDBCDriver: Tentando adicionar JDBC driver?
Erro, não CLASSPATH - org.hsqldb.jdbcDriver: Tentando adicionar JDBC driver?
```

Isso significa que Weka não consegue encontrar esses drivers, o que é bom, desde que você não precisa se conectar a um banco de dados. Você gostaria de ler dados a partir do WCM armazém de dados, então algumas modificações precisam ser feitas. Primeiro, você necessidade de alargar o caminho de classe. O caminho de classe é uma variável de ambiente chamada CLASSPATH onde o programa pode localizar as classes Java necessárias para um determinado tarefa. Neste caso você precisa especificar a localização de um driver JDBC, então adicionar o

MySql driver para o CLASSPATH variável. Use o driver que já está disponível no Pentaho BI Suite e digite o seguinte comando no comando linha (observe que todo o comando deve ser em uma única linha):

```
Export CLASSPATH = $ CLASSPATH: / opt / pentaho / biserver-ce /
tomcat/common/lib/mysql-connector-java-5.0.7.jar
```

Se você quiser usar outro banco de dados você pode substituir o mysql conector para o condutor para o seu banco de dados. Adicionar mais de um banco de dados requer alargamento do caminho de classe com os drivers extra. Basta adicioná-los ao comando como demonstrado anteriormente, separados por dois pontos (:).

NOTA Um entradas de caminho de Linux de classe são separadas por dois pontos; no Windows ponto e vírgula (;) é usado.

O próximo passo é ter certeza de que Weka pode traduzir todos os MySQL tipos de dados utilizados em nosso armazém de dados. Porque nós definimos algum inteiro colunas como unsigned (que Weka não pode ler por padrão), você precisa modificar as configurações de banco de dados em um arquivo chamado DatabaseUtils.prop. E, enquanto você está nisso, você também pode se livrar do incômodo classe mensagens de erro de caminho.

Weka olha para o . Prop arquivos nos seguintes três localidades, e na na seguinte ordem: o diretório atual, o diretório home do usuário e, finalmente, no subdiretório do experimento weka.jar arquivo. Embora você possa alterar esta última diretamente, recomendamos fortemente contra ela, logo que reinstalar ou atualizar a instalação Weka atual todas as alterações serão perdidas. A primeira opção também não é o melhor porque isso iria forçá-lo para sempre Weka iniciar a partir do diretório Weka. Portanto, você precisa criar o arquivo no seu diretório home.

Embora você não deseja modificar o arquivo jar, contém amostra DatabaseUtils.props arquivos para os bancos de dados mais comuns, incluindo o MySQL. A maneira mais fácil de conseguir este arquivo em seu diretório home é abrir o weka.jar arquivo e navegue até a DatabaseUtils arquivos como mostrado na Figura 16-2.

O arquivo que você precisa é aquele com o . Mysql extensão, e porque não pode ser copiados a partir do jar diretamente, abri-lo primeiro com um duplo clique no arquivo. O arquivo abre em um editor de texto padrão para que você possa fazer as modificações necessárias. A linha com o driver JDBC deve ser alterada para:

```
com.mysql.jdbc.Driver jdbcDriver =
```

ea URL JDBC para

```
jdbcURL = jdbc: mysql: / / localhost: 3306/wcm_dwh
```

Isto pressupõe que você estiver executando o servidor de banco de dados na máquina local. Abaixo, a URL que você vai ver os tipos de dados específicos que têm sido comentadas

para fora. Estes são os padrões Weka e você não quer mudá-los. Agora você deve adicionar os mapeamentos inteiro sem sinal algum lugar abaixo desta seção com as seguintes linhas:

```
TINYINT_UNSIGNED = 3
SMALLINT_UNSIGNED 4 =
INTEGER_UNSIGNED = 5
```

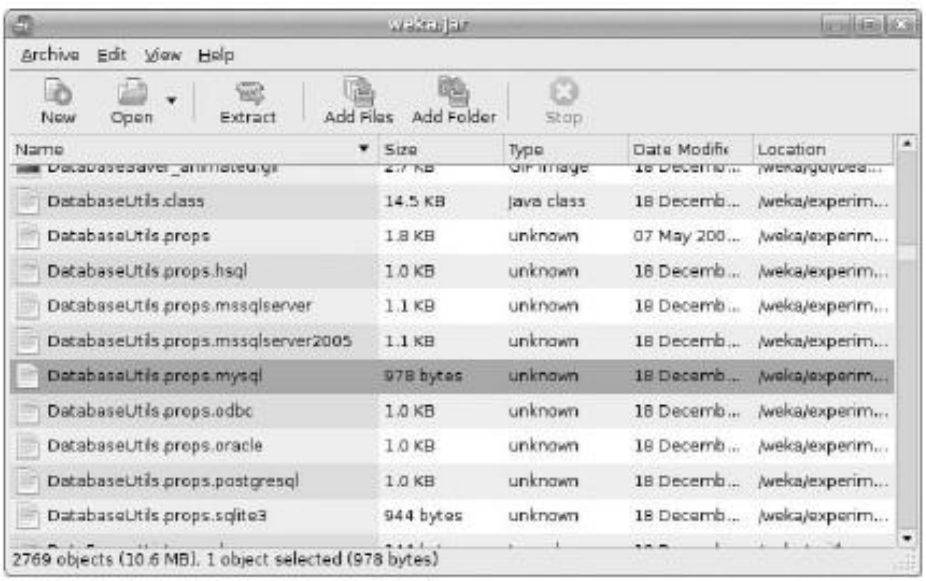


Figura 16-2: Weka.jar DatabaseUtils arquivos

Em seguida, você deve salvar o arquivo como DatabaseUtils.props em sua casa diretório. Uma descrição completa das propriedades do DatabaseUtils adereços. arquivo podem ser encontradas on-line em http://weka.wiki.sourceforge.net/experimento weka DatabaseUtils.props. Na próxima seção, vamos explicar como Weka deve ser iniciado através de um caminho de classe personalizado.

Começando Weka

Se você seguiu as instruções do parágrafo anterior e você começa Weka novamente, executando o java - jar weka.jar comando e selecione Explorer a partir do seletor de GUI, você vai notar que ainda há uma mensagem de erro à esquerda:

```
Registrando --- --- Editores Weka
Erro, não CLASSPATH - com.mysql.jdbc.Driver: Tentando adicionar JDBC driver?
```

Novamente, não há nada para se preocupar, é apenas um aviso de que o drivers de banco de dados não estão no caminho de classe ainda. CLASSPATH é uma ambiente variável, o que significa que pode ser facilmente estendida. Como você quer ler dados de um banco de dados MySQL, você deve adicionar o driver do MySQL para a classe

caminho. Use o driver que já está disponível no Pentaho BI Suite e introduzir o seguinte comando na linha de comando (o comando deve estar em uma única linha):

```
Export CLASSPATH = $ CLASSPATH: / opt / pentaho / biserver-ce /
tomcat/common/lib/mysql-connector-java-5.0.7.jar
```

A fim de utilizar o caminho de classe ajustada, você tem que explicitamente se referem a ele quando
a partir Weka. Se você usar o -Jar opção novamente, o CLASSPATH variável será substituídos. Você também precisa especificar a classe principal deve começar com Weka, e você pode ajustar a quantidade de memória de reserva para Weka, ao mesmo tempo. O seleccionador de GUI agora pode ser iniciado com o seguinte comando, mas você necessidade de estar no weka diretório para que ele funcione:

```
java-Xmx128m-classpath $ CLASSPATH: weka.gui.GUIChooser weka.jar
```

A melhor maneira é usar o seguinte comando com o Weka completo caminho incluído (novamente, este comando deve ser em uma única linha):

```
java-Xmx128m-classpath $ CLASSPATH: / opt / pentaho /
weka.gui.GUIChooser weka-3-6-0/weka.jar
```

Agora quando você iniciar o Explorer a partir do seletor de GUI, o erro do MySQL se foi e você pode usar a opção Open DB para se conectar aos dados WCM Armazém banco de dados. O último comando pode também ser copiado para um novo lançador
para torná-lo parte do menu Pentaho. A -Xmx parâmetro utilizado no comando especifica a quantidade de memória que Weka vai usar. Ao trabalhar com grandes conjuntos de dados, você deve definir esse valor a um nível elevado para evitar a mensagem de erro
na Figura 16-3.

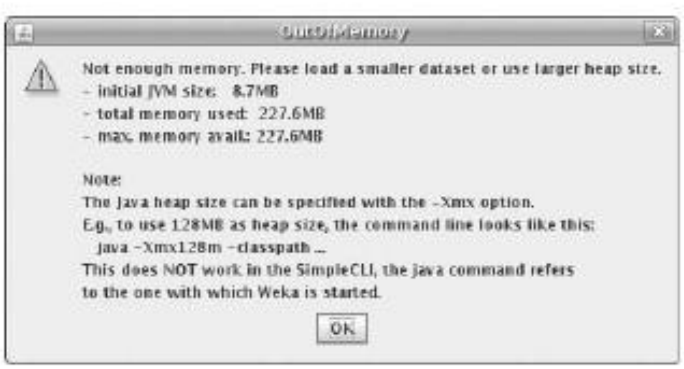


Figura 16-3: Weka erro de memória

Após esse erro, Weka fecha, e nenhum dos dados são salvos, então é melhor usar tanta memória como você pode poupar (por exemplo, usar -Xmx1024m para o início com um gigabyte de RAM).

O Weka Explorer

O Weka Explorer pode ser iniciado a partir do seletor GUI ou directamente a partir do linha de comando usando o seguinte comando:

```
java-Xmx128m-classpath $ CLASSPATH: / opt / pentaho /
weka.gui.explorer.Explorer weka-3-6-0/weka.jar
```

Agora, você provavelmente vai querer saber se a conexão com o banco tem sido configurado corretamente, então clique em Abrir DB para iniciar o Visualizador de SQL. Se tudo funciona da maneira que deveria, a URL do banco de dados é agora visível na conexão painel. Em seguida, você precisa especificar o usuário e senha, clicando no usuário opção, após o qual você pode se conectar ao banco de dados. A caixa de Informação na parte inferior do Visualizador de SQL irá mostrar uma mensagem que a conexão bem-sucedida, e você pode inserir uma consulta no painel Consulta. Porque este painel não apoiá-lo, por escrito, a sua consulta de forma alguma em tudo, talvez seja uma boa idéia para desenvolver a sua primeira consulta em outro ambiente, como o MySQL Query Browser, Esquilo, ou SQLeonardo. Como exemplo, digite a consulta `selecione * A partir de dim_date`. Se tudo foi configurado corretamente, os resultados devem ser exibida agora no painel de fundo. Se não, revise a configuração anterior etapas.

O Weka Explorer é uma aplicação de mineração de dados completo em si mesmo e lhe permite obter dados a partir de arquivos em diversos formatos, a partir de um banco de dados consulta, ou de uma URL. Ela inclui ainda uma opção para gerar um conjunto artificial de dados para trabalhar (jogar) com, que pode ser útil se você quiser comparar diferentes abordagens e algoritmos sem ficar distraído pelo conteúdo ou o significado dos dados. Pré-processamento inclui também a capacidade de aplicar filtros para os dados, tanto em atributos (colunas) e em instâncias (linhas). Para classificação, agrupamento, associação e predição numérica, a bancada oferece guias especiais, onde as quatro categorias de mineração de dados pode ser usada, cada com sua própria coleção de algoritmos e opções. A classificação e agrupamento modelos também podem ser salvos para uso fora da bancada Weka, como nós mostrarei adiante. Finalmente, há uma opção de visualização, que é exibido na Figura 16-4.

NOTA Apesar de um bom add-on, as chances são que você não vai pegar Weka para suas capacidades de visualização espectaculares, porque há muitas outras ferramentas que são mais adequados para esta finalidade. Uma alternativa melhor é usar RWeka, que contém tanto Weka ea biblioteca R estatísticos, incluindo RGraph.

No último exemplo deste capítulo, você verá o Explorer usado para criar um modelo para o PDI Weka Scoring plugin.

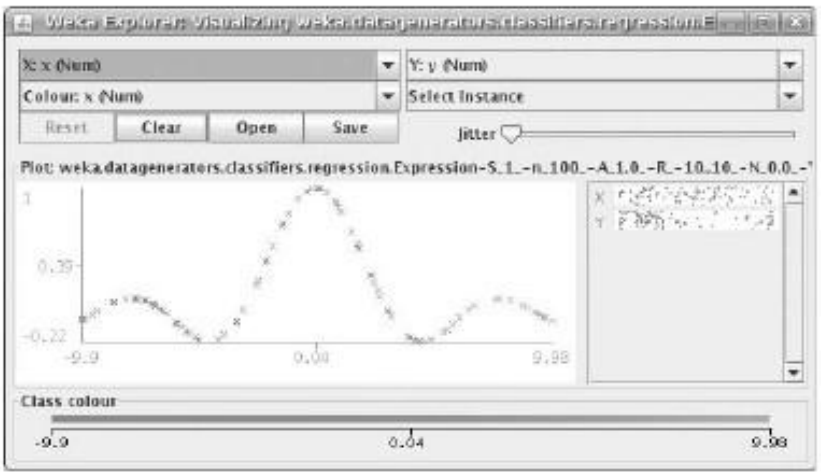


Figura 16-4: visualização de Regressão

O experimentador Weka

O experimentador tem dois modos de operação simples e avançados, e permite que você execute uma série de algoritmos de mineração de dados qualquer número de vezes contra uma série de conjuntos de dados. Isto também significa que é uma ferramenta para a mais avançada usuários, porque não dar-lhe feedback imediato sobre os resultados de uma experimento. Há uma série de vantagens em usar esta ferramenta:

- Várias operações podem ser executadas em um único lote.
- Um número de iterações pode ser definido, forçando o algoritmo para executar várias tempos. Correr dez iterações com um resultado dez vezes validação cruzada na execução do classificador mesmo cem vezes, tornando mais resultados estatisticamente válidos.
- Todos os resultados são gravados em um arquivo CSV (ou ARFF) ou tabela de banco de dados para fácil análise e comparação dos diferentes algoritmos.
- Os resultados podem ser analisados e testados utilizando qualquer campo comparação dos arquivo de resultados. Figura 16-5 mostra a tela Analisar depois de executar um Zeror e J48 algoritmo no mesmo conjunto de dados. É fácil ver que a decisão algoritmo de árvore fornece resultados muito melhores.
- Experimente configurações podem ser salvas e modificados.
- As notas podem ser adicionados a cada experimento.

Embora o pesquisador não será a primeira parte do conjunto de ferramentas que Weka você vai estar usando, ele é um complemento poderoso e, certamente, algo que vale mais investigação.

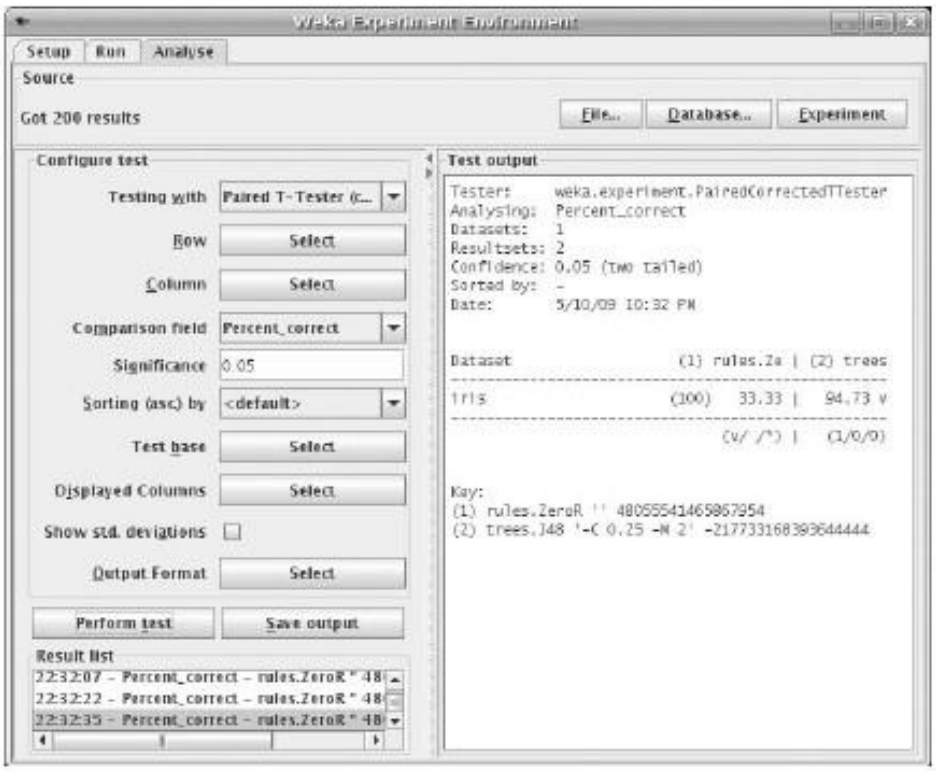


Figura 16-5: análise Experimentador

Weka KnowledgeFlow

O KnowledgeFlow Weka é como um explorador de esteróides, na verdade, assemelha-se a forma como algumas das principais ferramentas comerciais de apoio à criação de mineração de dados
fluxos e processos de mineração de dados. Quando você sabe o caminho de volta Explorer, KnowledgeFlow irá realizar alguns segredos como o arranjo dos ferramentas disponíveis é criada da mesma forma e todos os algoritmos do Explorer são Também disponível na ferramenta KnowledgeFlow. A principal diferença é a gráfica layout do fluxo de trabalho, que também permite a ramificação mais complexa que apoiado pelo Explorer. Figura 16-6 mostra um exemplo simples de um classificação de fluxo.

O desenvolvimento de KnowledgeFlow ainda está em curso, mas para além da A interface visualmente atraentes existem algumas outras coisas que o diferenciam a partir do Explorer:

- KnowledgeFlow pode processar múltiplas transmissões em paralelo.
- Os dados podem ser processados de forma incremental, bem como em lote.
- Os filtros podem ser encadeados (Explorer pode manipular apenas um de cada vez).
- A arquitetura de plugin está disponível, assim como o PDI, o que torna uma ambiente extensível.

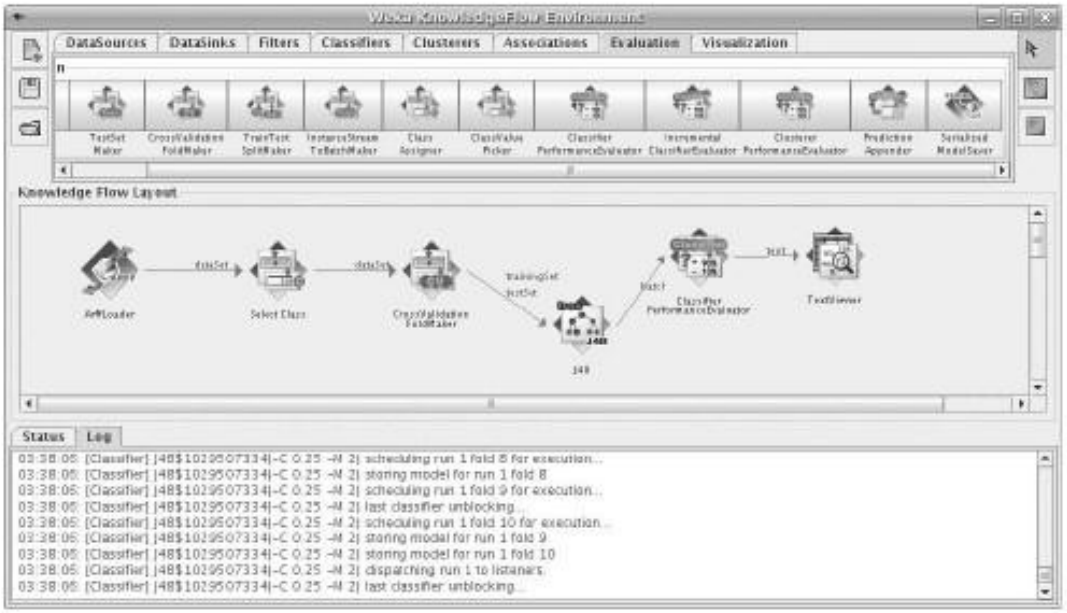


Figura 16-6: exemplo KnowledgeFlow

Mais informações sobre KnowledgeFlow pode ser encontrado no Weka Wiki.

Usando Weka com Pentaho

Como explicamos na introdução deste capítulo, Weka é um baixo acoplamento parte da plataforma de BI Pentaho. Há duas partes da plataforma onde algum tipo de integração existe, com PDI ter maior apoio ao Weka funcionalidade fornecida por plugins extra. Embora haja apenas um plugin que realmente usa os modelos construídos com o Explorador Weka, existem dois outros que ajudam você a preparar e fornecer dados de mineração de dados:

- Weka Scoring-Ativa a utilização de um modelo Weka de dentro de um PDI transformação. É uma ferramenta que permite que os modelos de classificação e agrupamento
  criado com Weka para ser usado em uma transformação PDI. Scoring simplesmente significa que as linhas de entrada de dados são marcados (variável de desfecho é determinado) durante a transformação. A pontuação plugin é capaz de anexar um rótulo previsto (classificação / agrupamento), número (regressão), ou distribuição de probabilidade (classificação / agrupamento) para uma linha de dados.
- ARFF-Output Cria um arquivo no formato ARFF (ver listagem 16-1) a ser utilizado em Weka. Os formatos de campo de entrada de PDI são mapeados para ARFF formato usando as seguintes regras:
  - String: nominal

- Number, Integer: número
- Data: data

- Reservatório de amostragem-A ferramenta que lhe permite obter uma seleção aleatória de linhas de um fluxo de dados de entrada Chaleira. Todas as linhas têm o mesmo chance de ser selecionado (por amostragem uniforme). Esta etapa é particularmente útil em conjunto com a etapa de saída ARFF, a fim de gerar um dados adequados de tamanho definido para ser usado pelo Weka. A partir do PDI versão 3.2, este é não mais um plugin, mas faz parte das etapas regulares.

Adicionando Plugins Weka PDI

Antes que você possa usar qualquer um dos plugins descrito até agora, eles precisam ser adicionados para uma instalação existente PDI. Além disso, o weka.jar arquivo precisa ser adicionado também. Os passos seguintes descrevem como isso funciona.

1. Faça o download do plugin do Wiki em Pentaho http://wiki.pentaho mostrar .com / EAI / Lista + de + Disponível + + Pentaho Data Integration + + Plug-Ins.

2. Descompacte o zip, cada arquivo zip irá criar um novo subdiretório no localização actual.

3. Criar um novo subdiretório para cada plugin no âmbito do / Opt/Pentaho/data-integração / plugins / etapas diretório e nome de cada um de acordo com o plugin.

4. Copie os arquivos do subdiretório implantação do descompactou download de arquivos em subdiretórios recém-criado a partir da Etapa 3.

5. Copie o arquivo weka.jar Weka a partir do diretório principal (no nosso caso, / Opt/Pentaho/weka-3-6-0) Para os subdiretórios do plugin passos para Pontuar Weka e as etapas de saída ARFF.

6. Reinicie Spoon (PDI).

Spoon deve ter agora os novos passos disponíveis: contém a pasta de saída a etapa de saída ARFF, a pasta contém o Transform Weka etapa de pontuação, ea pasta contém Estatísticas do Reservatório etapa de amostragem.

NOTA O plugin zipado arquivos contêm mais do que apenas a extensão de implantação. A documentação do usuário para a etapa está disponível como um arquivo PDF do subdiretório / Doc.

Começando com Weka e PDI

A parte final do capítulo consiste de um fluxo de trabalho completo que envolve dados preparação com PDI, a criação de modelos usando Weka, e processamento de dados utilizando o placar Weka plugin. A WekaScoring.pdf documento incluído plugin no arquivo zip contém um exemplo simples de como usar a pontuação

funcionalidade no PDI e trabalha com uma amostra de dados set (pendigits), que é
também faz parte do arquivo zip. Você pode usar esse documento como uma referência, porque nós
irá utilizar o mesmo fluxo de trabalho e funcionalidade aqui, mas vamos usar um
conjunto de dados diferente que mais se assemelha a dados que temos disponíveis no WCM data warehouse.

Aquisição de Dados e Preparação

Tal como no BI, a maioria do esforço de mineração de dados é gasto na obtenção de dados de boa qualidade
para trabalhar. Felizmente, você pode pular algumas etapas e construir sobre o trabalho
já feito por outros. O conjunto de dados que você irá usar é extraído os EUA
Os dados do censo e já contém uma seleção significativa. Ela pode ser recuperada
a partir de http://archive.ics.uci.edu/ml/datasets.html (Ou no site
para este livro) e é chamada de dados Adulto set.1 É bastante grande, o que permite
você dividi-la primeiro usando o Reservatório etapa de amostragem, e está em um sistema inutilizável
formato, que obriga a utilização PDI para preparar os dados e convertê-lo para
um formato ARFF primeiro. O conjunto de dados contém 14 atributos demográficos, além de a variável de classe, e pode ser usado para prever se uma pessoa com certas características faz mais ou menos de 50.000 dólares por ano. Você precisa de três arquivos para este exercício:

- adult.names contém a descrição dos dados e os resultados da
  algoritmos utilizados anteriormente. Note que o algoritmo do Weka J48 não está listado aqui, você pode realmente comparar este com o C4.5 original.
- adult.data contém os dados que você irá utilizar para o treinamento do modelo.
- adult.test contém o conjunto de teste para validar o modelo.

Baixe os três arquivos em um diretório de escolha (ou criar um trabalho e utilizar o
etapa HTTP para lê-los directamente a partir da URL). Nós adicionamos um diretório de dados para a raiz Pentaho para este exercício (/ Opt / pentaho / dados). Você quer fazer o seguinte:

- Mesclar os dados e arquivo de teste.
- Use o Sampler Reservatório para extrair 10 mil linhas.
- Exportar os dados da amostra como um arquivo
  ARFF.

A completa transformação é mostrado na Figura 16-7.
Vamos explicar brevemente como criar este, a transformação completa pode ser encontradas no site para este livro em www.wiley.com / go / pentahosolutions.

1. Em primeiro lugar, começar a colher e criar uma nova transformação. Comece por criar um
   CSV entrada com o adult.data arquivo (você verá porquê mais adiante). Porque este arquivo não tem uma linha de cabeçalho, você terá que especificar os nomes de campo

[1]Assunção, A. & Newman, D. J. (2007). UCI Machine Learning Repository [http://www.ics.uci
. Edu /~mlearn] / MLRepository.html. Irvine, CA: Universidade da Califórnia, Escola de Informação e Ciência da Computação.

manualmente. Eles podem ser encontrados no nomes. arquivo. Agora ajuste o CSV etapa de entrada da seguinte forma:

- Excluir o Recinto especificação.
- Desmarque a Cabeçalho da linha atual? checkbox
- Selecione o valor ambos para todos os campos sob tipo Trim.

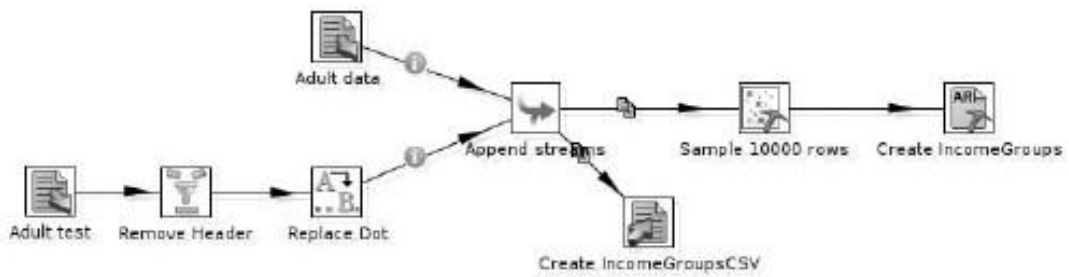


Figura 16-7: Adult2IncomeGroup

2. Duplicar o passo e mudar o arquivo de entrada da segunda via adult.test. A primeira linha deste arquivo é um valor único texto, o que torna impossível PDI para determinar a especificação de arquivo. Você já precisou que, em o passo de entrada original, e porque os arquivos têm uma estrutura idêntica, você não precisa fazer mais nada aqui.

3. Agora, o adult.test arquivo tem dois problemas. O primeiro é o quase vazio linha na parte superior, eo segundo é o campo de classe que tem um ponto no final, o que significa que se você não se livrar do que você vai acabar com quatro em vez de dois valores de classe. O cabeçalho pode ser removido usando um filtro de passo de linhas com a condição grupo de renda IS NOT NULL (Na verdade, qualquer campo, exceto o primeiro vai fazer aqui).

4. Para livrar-se do ponto, use um Replace na etapa de cordas e simplesmente o jogo no campo de fluxo para grupo de renda e valorizar a pesquisa para"."E deixar o Substitua com valor vazio. Agora, as duas correntes podem ser acrescentados e a saída pode ser enviada tanto para a etapa de saída um arquivo CSV (este contém todos os linhas em formato CSV) e um reservatório etapa de exemplo. Esta última é definida como 10.000.

5. Finalmente, a saída ARFF pode ser inserido como etapa final do transformação. Tudo que você precisa fazer neste último passo é entrar no arquivo e o nome da relação. Depois de executar a transformação, você deve ter um arquivo CSV chamado IncomeGroups.csv com 48.842 linhas e um arquivo ARFF chamada IncomeGroups.arff com 10.000 linhas de dados.

NOTA Há uma explicação simples para o fato de que o cliente WCM dados não puderam ser utilizados para este exemplo: é gerado aleatoriamente. Como um resultado, qualquer tentativa de agrupar ou classificar os clientes com base na receita, o filme informações, ou outras características falhará.

Como criar e salvar o modelo

Você pode agora iniciar o Weka Explorer, escolhendo Explorer a partir da GUIChooser ou iniciando-lo diretamente na linha de comando. Clique em Abrir Arquivo e selecione a IncomeGroups.arff arquivo que você criou anteriormente. O arquivo será ser carregado e estatísticas descritivas para os atributos mostrados na Preprocess painel, como você pode ver na Figura 16-8.

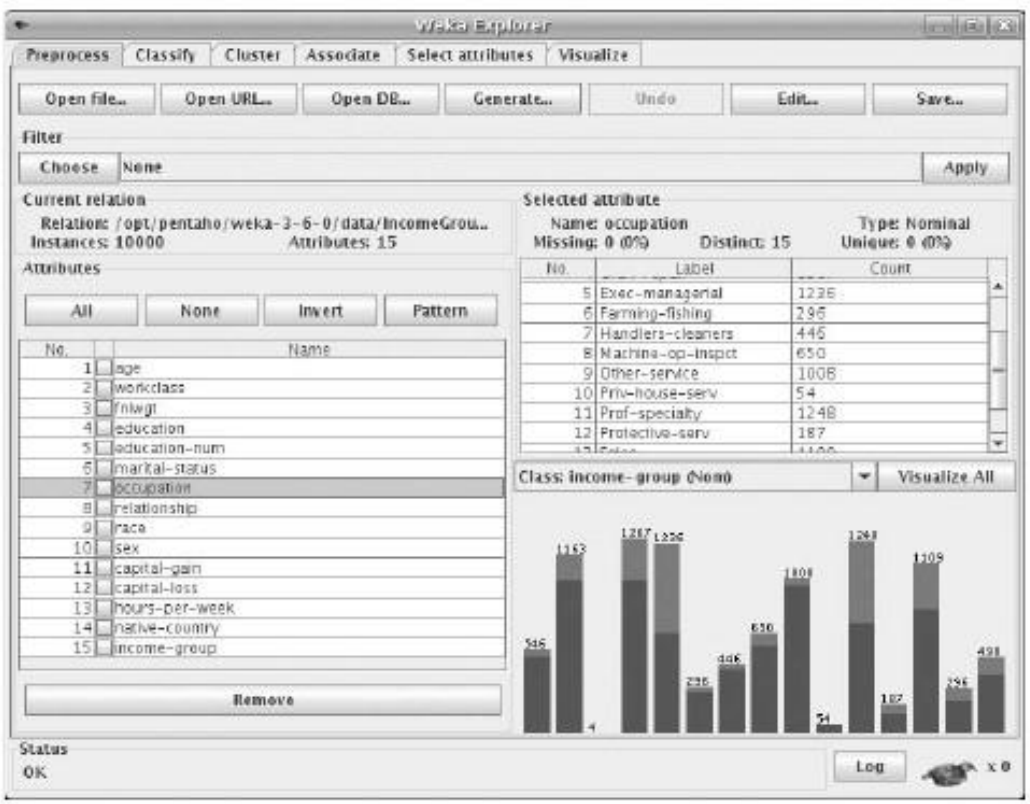


Figura 16-8: Loaded ARFF arquivo no Explorer

Você também pode ver que o nome Explorer para esta ferramenta é uma boa, pois permite que você navegue através dos dados e fazer algumas suposições sobre o assunto. Por exemplo, se você quiser fazer mais de US $ 50 mil por ano (lembre-se, neste são dados a partir de 1995!), é melhor ser um executivo ou um profissional, porque eles têm as gamas mais elevadas (> R $ 50 mil). Outra área de interesse é o sexo divisão na fêmea cerca de um terço e dois terços do sexo masculino, o que significa que mais homens do que mulheres ganham um salário. O que você vai notar também é que mais homens do que as mulheres ganham mais de US $ 50 mil por ano. Ser um homem casado também
ajuda. Assim, mesmo sem correr um classificador, você já pode dizer que estes atributos têm um grande impacto sobre a classe resultado. Isso nem sempre é o caso, porém, em muitos casos, a mineração de dados fornece a visão que não pode ser obtidas de outra forma.

O próximo passo é escolher um classificador. Lembre-se que esta é uma classificação conjunto de dados com um resultado conhecido (a classe). Se não ficou claro o que estávamos
procurando, nós poderíamos usar a opção de cluster para descobrir quais as instâncias da dados são""mais perto uns dos outros do que outros. O classificador, vamos escolher é J48, que está no Weka-classificadores-árvores pasta quando você clicar em Escolher. Depois selecionar o classificador, os parâmetros para o algoritmo pode ser definido clicando sobre o nome do classificador, que abre um editor de objetos. Deixar todos os valores em sua configurações padrão aqui. No Teste de opções, selecione validação cruzada com 10 dobras (O valor padrão), verifique se o grupo de renda é selecionado como a classe, e
clique em Iniciar para executar o classificador. A barra de status na parte inferior da janela irá mostrar a dobra que está executado e depois de correr dez vezes (o
número de dobras), os resultados serão mostrados na saída do classificador, como exibido na Figura 16-9.

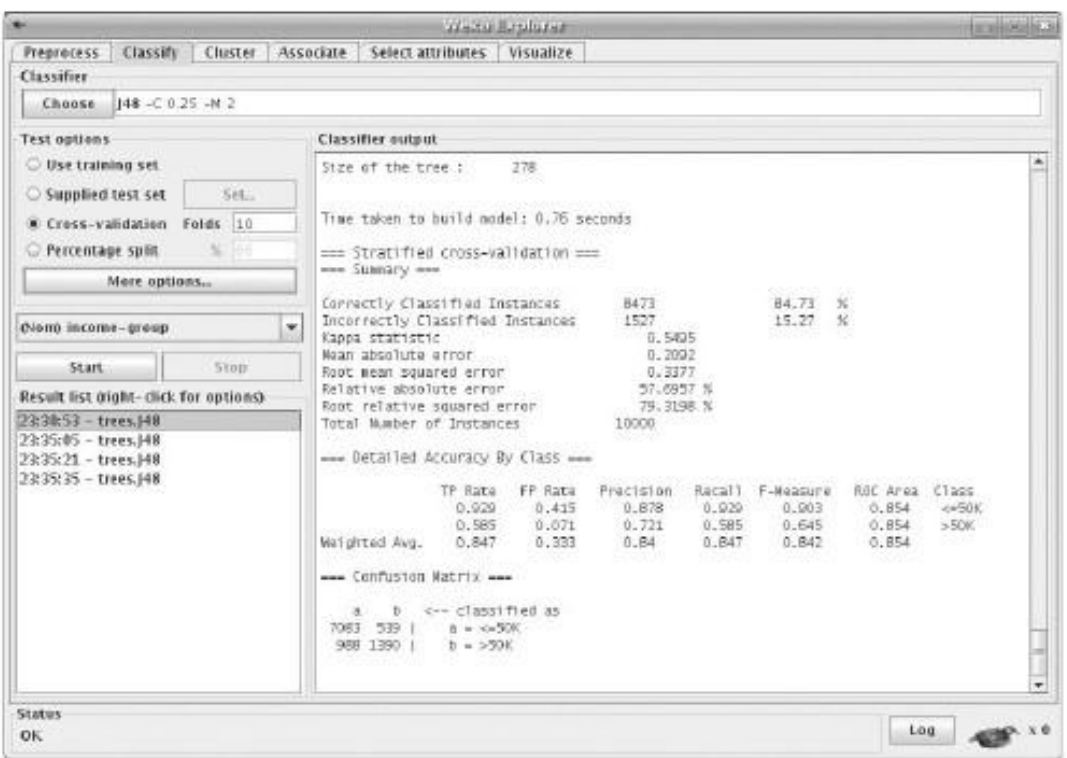


Figura 16-9: J48 saída do classificador

Além dos dez vezes validação cruzada, você pode experimentar com outras Teste opções para verificar se os resultados de melhor qualidade pode ser obtida através de um dividir porcentagem ou apenas usando o conjunto de treinamento. Na verdade, o conjunto de treinamento
produz resultados melhores neste caso: que classifica uma percentagem mais elevada corretamente, mas
os resultados não são muito confiáveis, porque usamos os mesmos dados para a formação

e testes. Como explicamos anteriormente, você deve tentar evitar isso, sempre que possível.

Exportando o classificador treinado agora é só alguns cliques de mouse de distância. Quando
você direito-clique no item da lista de resultados do modelo que você gostaria de exportação, um submenu é mostrado com a opção de salvar o modelo em algum lugar no média. Salve o modelo no diretório de dados Weka e nomeá-la J48. A . Modelo extensão é adicionada automaticamente.

Utilizando o Weka Scoring Plugin

Usando o modelo de formação em PDI dados nova pontuação é simplesmente uma questão de
configurando o placar Weka plugin para carregar e aplicar o arquivo de modelo que você criou na seção anterior.

1. Você usará o conjunto completo de dados de adultos que você salva como um arquivo CSV anterior.
   Para fazer isso, crie uma nova transformação em Spoon e adicione uma entrada CSV passo que lê o IncomeGroups.csv arquivo do seu diretório de dados.
2. O Weka passo seguinte pontuação necessária podem ser arrastados a partir da transforma
   para a tela. Em seguida, editar o Weka etapa de pontuação e selecione o J48.model arquivo como o de carga / modelo de importação. Depois de selecionar o modelo, os campos
   mapeamento guia mostra como os campos a partir da etapa de entrada CSV foram mapeadas para os atributos do modelo J48. Neste caso, não há erros, mas se havia, os tipos de campo deve ser ajustado na etapa de entrada CSV, não aqui. A guia modelo irá mostrar a árvore que foi o resultado do modelo formação.

NOTA O atributo de classe grupo de renda faz parte do arquivo CSV, mas não é necessários para o placar para o trabalho, é o valor que será determinado pelo modelo, mas não faz parte do próprio modelo.

3. Agora você pode clicar no botão Preview (ou clique com o botão Marcar Weka passo e selecionar Preview). Existe agora um campo extra disponível com o mesmo nome da classe com uma _predicted sufixo, como mostrado na Figura 16-10.

4. Se você quer ter as probabilidades de saída adicionados a cada linha de saída em vez da classe previsto, assinale a opção correspondente no Weka Scoring guia do arquivo de modelo. Nesse caso, uma coluna extra será adicionada para cada valor de classe distinta, com a probabilidade calculada para cada valor, como mostrado na Figura 16-11.

Note-se que esta última opção só está disponível para os modelos que foram treinados em um problema de classe discreta.

DICA Quando uma transformação é executado, o placar Weka plugin vai carregar o modelo a partir do disco usando o caminho especificado na guia File. Também é possível armazenar as modelo do PDI repositório e usar isso ao invés, quando a transformação é executado. Para fazer isso, primeira carga de um modelo para a etapa de pontuação Weka, como descrito anteriormente. Quando estiver satisfeito que os campos foram mapeados corretamente eo modelo estiver correto, você pode limpar a carga de importação / caixa de texto do modelo e clique em OK. Quando o transformação é salvo, o modelo será armazenado no repositório.

Figura 16-10: resultados de previsão

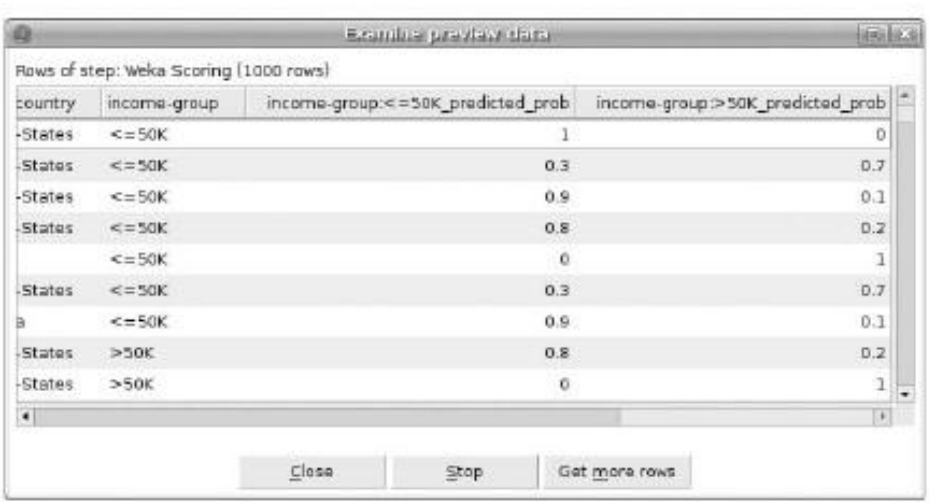


Figura 16-11: probabilidades de saída

Outras dicas e notas referentes à pontuação Weka plugin está disponível em a documentação que faz parte da zipado plugin.

Leitura

A mineração de dados é um campo muito bem documentado de investigação, de modo a Internet tem um
quantidade inimaginável de informações sobre o assunto. Alguns dos mais úteis
fontes de informação são listadas aqui:

- Online Pentaho documentação Weka- http://wiki.pentaho.com/
  display / DataMining / Pentaho Data + + Mining + + Comunidade Documentação
- homepage Weka- www.cs.waikato.ac.nz/ml/weka/
- Pentaho Weka fórum http://forums.pentaho.org/forumdisplay
  . Php? F = 81
- Um completo manual de estatísticas on-line www.statsoft.com/textbook/
- Página inicial dos dados Mineiros Inc., os autores de quatro livros Wiley em dados mineração, Michael Berry e Gordon Linoff- www.data-miners.com
- Dados da comunidade mínero- www.kdnuggets.com

Resumo

Este capítulo foi uma breve introdução ao vasto tema da mineração de dados. Porque da diversidade e da complexidade do tema, nós poderíamos mostrar apenas a ponta do do iceberg, mas tentou tocar nos pontos mais importantes e fornecer
um hands-on exemplo de como os modelos de mineração de dados pode ser usado no Pentaho
BI Suite, especialmente PDI. Futuras versões do Pentaho irá aumentar ainda mais a integração entre o poder analítico da Weka e as capacidades de BI
a plataforma Pentaho. Um próximo passo lógico seria chamar um modelo de Weka uma seqüência de ação (embora isso já pode ser conseguido usando a pontuação plugin e chamando a transformação PDI a partir de uma seqüência de ação). Outro opção para melhorar as capacidades analíticas da Pentaho é a integração de
a biblioteca R estatística, que já está disponível como uma solução de um dos parceiros Pentaho.

Este capítulo abordou os seguintes tópicos:

- Introdução à mineração de dados
- Resumo do Weka Workbench
- adicionais de instalação e opções de configuração para o Weka eo Weka plugins PDI
- Um exemplo completo de como mineração de dados pode ser integrada no plataforma Pentaho

CAPÍTULO

17

Construindo Painéis

No contexto de Business Intelligence, um painel é um aplicativo que está
usadas para apresentar alto nível de conteúdo de BI para usuários finais. Os painéis contêm
apenas alguns indicadores-chave do desempenho de algum aspecto de negócios (vendas"")
ou mesmo o negócio como um todo.

conteúdo do Dashboard é quase invariavelmente gráfica na natureza: em vez de
números, as métricas são simbolizados com fotos, metros, marca, e às vezes
gráficos. O objetivo é fornecer uma visão muito condensada de uma grande área
dos negócios, permitindo que os gerentes de negócios para avaliar o estado de relance.

Normalmente, os indicadores de alto nível gráfico que aparecem nos painéis
fornecer alguma interatividade que permite ao usuário detalhar a mais detalhada
conteúdos de inteligência comercial, como relatórios e cubos OLAP.

O Dashboard Framework Comunidade

A Comunidade Dashboard Framework (CDF) é um conjunto de tecnologias de código aberto,
gias que permite aos desenvolvedores de BI para a construção de dashboards dinâmicos para
o BI Pentaho
Server. painéis CDF são essencialmente as páginas web que usam a tecnologia AJAX
dinamicamente combinam componentes de BI, como relatórios, gráficos, tabelas OLAP,
e mapas. Embora o CDF é, por padrão incluídos no Pentaho BI Server,
que é desenvolvido e mantido pelos membros da Comunidade Pentaho
em vez da Companhia Pentaho.

CDF, a Comunidade, e da Corporação Pentaho

O CDF é um grande exemplo de sinergia entre os objetivos de um comercial
empresa de software de fonte aberta como o Pentaho e sua comunidade.

A Pentaho Corporation inclui o CDF na Comunidade eo
Enterprise Edition do servidor de BI como um plugin. Para os usuários da empresa
Edition, Pentaho fornece também um construtor de painel que pode simplificar a construção
painéis. Você pode encontrar uma seção sobre o CDF com o desenvolvedor do BI
Exemplos de solução, que contém toda a documentação útil, dashboards de exemplo,
e informações básicas (veja Figura 17-1).

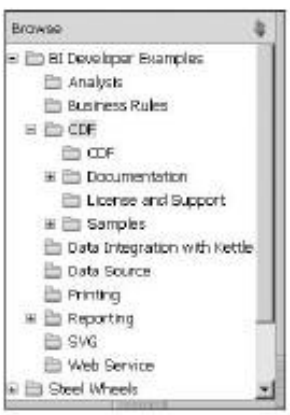


Figura 17-1: O CDF na seção Exemplos BI Developer

Embora a Pentaho Corporation não é formalmente envolvidos no CDF,
um número de desenvolvedores Pentaho são contribuintes activos para o projeto. Em
Além disso, Pentaho oferece hospedagem para a gestão do projeto CDF problema, Wiki
documentação, e fórum.

O CDF é distribuído gratuitamente sob a licença LGPL. O projeto é
código-fonte e os recursos são hospedados em um projeto de código do Google (http://code
.google.com p / cdf-pentaho /).

CDF Projeto História e Quem é Quem

O projeto CDF é levado e mantido por um número de membros proeminentes
da comunidade Pentaho.

As raízes do CDF levar de volta para 2007, quando Ingo Klose apresentou um
solução de dashboard baseados em AJAX como uma alternativa para o Java antes existentes
A tecnologia de painéis baseado em servlet fornecido por Pentaho. Ingo da primária
motivação foi o desejo de uma solução de painéis que era mais fácil de usar e
implantar.

Ingo trabalho foi rapidamente adotado por Pedro Alves. Colaborou em
numa fase inicial com Ingo. Juntos, ele e Ingo fundou o projeto CDF para
centralizar o desenvolvimento e permitir a outros membros da comunidade para contribuir.
Desde então, ele tem muitos recursos e retrabalhadas vários aspectos da
arquitetura.

Além destes principais colaboradores, muitos outros membros da comunidade traço trabalhou para criar dashboards de exemplo, o painel modelos, novos e-componentes da placa.

Emissão de Administração, Documentação e Apoio

CDF problema de gestão está hospedado no jira.pentaho.com site. Pode ser encontradas na seção Comunidade na http://jira.pentaho.com/browse/CDF.

documentação do projeto está atualmente disponível em um número de lugares. Em primeiro lugar,
Wiki Pentaho inclui uma área que é totalmente dedicado ao CDF na http://wiki.pentaho.com/display/COM/Community+Dashboard+Framework.

Além das páginas de documentação da Wikipédia, amostras CDF distribuído na Pentaho BI servidor também contêm uma grande quantidade de documentos valiosos.

Para obter o apoio da comunidade, há um fórum no forums.pentaho.com, Que é dedicado a temas CDF. Para suporte comercial, pode contactar o WebDetails empresa em http://www.webdetails.pt/.

Competências e Tecnologias de Dashboards CDF

Dois tipos diferentes de habilidades são necessárias para construir painéis CDF:

- Geral Pentaho desenvolvimento de competências-Você precisam ser capazes de construir
  seqüências de ação eo conteúdo de BI, como relatórios, gráficos e assim por diante, que você deseja visualizar no painel. Este livro deverá fazê-lo bem
  começou nessa direção.
- habilidades de web-A foco nas habilidades desenvolvimentos web está além do escopo deste livro. Isto é principalmente um livro sobre Business Intelligence, e há uma abundância de livros de alta qualidade e recursos disponíveis na web de desenvolvimento.

Não é obrigatório para unificar essas habilidades bastante diferentes na mesma pessoa. Pelo contrário, pode ser mais produtiva para criar painéis usando pares de desenvolvedores de BI e desenvolvedores web, ou uma equipe de desenvolvedores de BI, os desenvolvedores da Web,
especialistas do domínio de negócio e usuários finais.

O desenvolvimento de competências necessárias web incluem:

- HyperText Markup Language (HTML) em HTML é o padrão
  linguagem utilizada para criar páginas web. No início dos anos noventa, verificou-se e rapidamente se tornou o modelo dominante para páginas da Internet, devido a sua popularidade a uma combinação de simplicidade e de recursos como hiperlinks e imagens. Atualmente HTML é um padrão aberto mantido pela
  W3C. Informação detalhada sobre HTML pode ser encontrada em http://www .w3.org/TR/REC-html40 /.

- Cascading Style Sheets (CSS), CSS emergiu como a lín-
  indicador para definir a apresentação do documento. Assim como o HTML, é um processo aberto
  padrão que é mantido pela W3C. Informações detalhadas sobre o CSS pode ser encontrada em http://www.w3.org/TR/CSS21/.
- JavaScript, o conhecimento do quadro JQuery é recomendado,
  JavaScript é uma linguagem de programação que foi projetada especialmente para adicionar interatividade a páginas da web. Ao longo dos anos, a importância da JavaScript tem aumentado. Lentamente, ele ganhou notoriedade e, durante os últimos alguns anos, a proliferação de uma técnica de programação especial chamado AJAX (às vezes identificada com a chamada Web 2.0) ajudaram a estabelecer sua posição como uma linguagem de programação séria.

  A versão standard do JavaScript chamada ECMAScript é especificado pela ECMA International. Mais informações sobre ECMAScript pode ser encontrada em www.ecma-international.org/publications/standards/Ecma-262.htm.

  JavaScript em si é apenas uma linguagem de programação. Muito do seu valor para criação de páginas web interativas não faz parte da própria linguagem, mas do ambiente de execução (por vezes apelidado de""runtime). O mais normas importantes no que diz respeito à manipulação de documentos HTML é da especificação do DOM, ou melhor, o mapeamento para ECMAScript. O DOM especificação é um padrão aberto mantido pela W3C. Mais informações sobre este assunto podem ser encontradas em www.w3.org/TR/REC-DOM-Level-1/ECMA-script-língua-binding.html.

Se você não estiver familiarizado com estas tecnologias, não deixe que isso tudo assusta você
fora. Domínio não é necessário para construir painéis simples. Mesmo se você é um Desenvolvedor e BI estão trabalhando em conjunto com desenvolvedores web para criar seu dashboards, recomendamos que você vá em frente e trabalhar através deste capítulo. No mínimo, ele vai mostrar quais tecnologias estão envolvidas eo que suas possibilidades e limitações.

Conceitos CDF e Arquitetura

painéis CDF são essencialmente páginas web (HTML documentos) que contêm áreas chamadas de componentes","que são usados para visualizar o conteúdo de BI. Figura 17-2
ilustra o que acontece nos bastidores, quando um painel é aberto por o usuário.

1. O usuário final usa um navegador web para navegar para um painel. Isso faz com que uma requisição HTTP normal para ser enviada ao servidor de BI Pentaho.

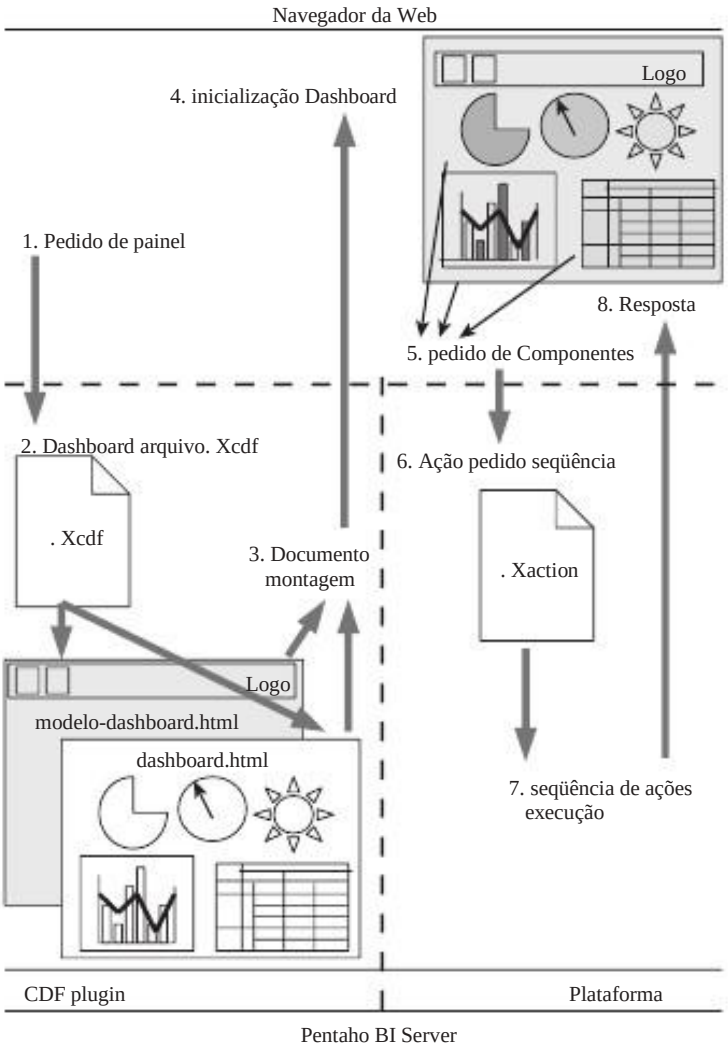


Figura 17-2: Bird's eye view do CDF dashboarding

2. O servidor Pentaho BI recebe o pedido e reconhece que deveria servir a um painel. O pedido contém o nome eo caminho da solução o painel, o que é suficiente para localizar o painel de . Xcdf arquivo.

3. A . Xcdf arquivo especifica os modelo de painel de conteúdo. Esta é uma (parcial) Arquivo HTML que contém espaços reservados para os componentes do painel e instruções JavaScript para enchê-los com os componentes do painel.
O modelo de conteúdo do painel é combinada com uma documento dashboard modelo (Às vezes chamado exterior do modelo) para montar uma página web (um Documento HTML). O modelo externo pode ser especificado explicitamente no . Xcdf arquivo, mas se não é, um padrão será usado.

4. A página web é recebida pelo navegador quando ele é lido e processado para a exposição. Como parte deste processo, o painel é inicializado. Esta é feito como as instruções JavaScript no documento são executadas, levando para a criação efetiva dos componentes do painel. Na página web, os componentes do painel existir como objetos JavaScript que são adicionados a um Dashboards objeto, que também é criado com JavaScript.

5. Após a inicialização do painel, os componentes de responder aos comandos emitido pelo Dashboards objeto. O comando básico é atualização, Que ordens de componentes para recolher o seu conteúdo de BI e colocá-lo em suas marcador correspondente (s) no documento. Normalmente, os componentes reagem ao atualização comando, enviando uma solicitação da Web para o Pentaho Server.

6. O servidor Pentaho recebe pedidos enviados pelos componentes. Normalmente, os componentes solicitar a execução de uma seqüência de ação.

7. O servidor Pentaho executa a seqüência de ação.

8. O conteúdo gerado pela seqüência de ação é enviada de volta para o requerente componente. Lá, a resposta é analisado pelo componente de inclusão na página web. Finalmente, o componente coloca o conteúdo dentro de sua espaço reservado (s) no documento, tornando o conteúdo visível no BI da página.

Vamos agora discutir os elementos-chave na arquitetura CDF em detalhes.

O CDF Plugin

Dashboard pedidos são tratados por um plugin Pentaho. Este plugin é um pedaço de O software Java que sabe como lidar com as solicitações da Web para obter um painel.

O Pentaho sistema de plugins é a maneira preferida para terceiros para ampliar Servidor Pentaho e adicionar funcionalidade personalizada. Uma discussão completa sobre plugins está além do escopo deste livro, mas nesta seção, tocamos alguns conceitos que, esperamos, irá ajudá-lo a entender como um complemento, como o CDF pode ser integrado com a plataforma Pentaho. Você pode encontrar detalhadas informações sobre plugins Pentaho na documentação Wiki em Pentaho wiki.pentaho.com/display/ServerDoc2x/1. + + Desenvolvimento de Plugins.

O Diretório Home CDF

O plugin e todos os arquivos relacionados residem e sob o pentaho-cdf Diretório abaixo do diretório que abriga o sistema Pentaho solução. No restante

deste capítulo, nos referimos a este local como o CDF diretório home, ou simplesmente CDF casa.

Há um par de subdiretórios no diretório home do CDF:

- A lib diretório contém os arquivos java binário que contém o real software que compõe o plugin.
- A js diretório contém os arquivos JavaScript e CSS (bem como alguns outros recursos) que compõem o software de cliente do CDF. Você veremos mais adiante como esses arquivos acabam nas páginas web reais dashboard.
- A recursos diretório contém vários recursos, como imagens e Arquivos CSS stylesheet.

O arquivo plugin.xml

O verdadeiro plugin definição está contida no arquivo plugin.xml. O conteúdo deste arquivo são mostrados na Listagem 17-1.

Listagem 17-1: O arquivo CDF plugin.xml

```
<? Xml version = "1.0" encoding = "UTF-8"?>
<plugin title="Pentaho Comunidade Dashboard Framework">
 <content-types>
  type="xcdf" <content-type mime-type="text/html">
   <title> Dashboard </ title>
   <description> Comunidade Dashboard Arquivo </ description>
   <icon-url> conteúdo / cdf pentaho / resources / cdfFileType.png
   </ Url-icon>
   <operations>
    <operação>
     <id> RUN </ id>
     <comando> conteúdo / cdf-pentaho / RenderXCDF solução? = {} & solução
path = {path} & action = {nome} & template = manto </ command>
    </ Operação>
    <operação>
     NewWindow <id> </ id>
     <comando> conteúdo / cdf-pentaho / RenderXCDF solução? = {} & solução
path = {path} & action = {nome} & template = manto </ command>
    </ Operação>
   </ Operações>
  </ Content-type>
 </ Tipos de conteúdo>
 <content-generator scope="local" id="pentaho-cdf" type="xcdf" url="">
  <classname> org.pentaho.cdf.CdfContentGenerator </ classname>
  <fileinfo-classname> <org.pentaho.cdf.CdfFileInfoGenerator / fileinfo-
```

```
Classname>
  <title> CDF Display Handler </ title>
 </ Gerador de conteúdo>
</ Plugin>
```

Se necessário, você pode alterar este arquivo de configuração para modificar o comportamento
de painéis CDF. Por exemplo, você pode adicionar tipos de conteúdo adicional para permitir que arquivos com uma extensão específica para ser reconhecido como painéis CDF. Para saber mais sobre plugins Pentaho, consulte o documento oficial de Pentaho implementação. As informações pertinentes podem ser encontradas em http://wiki.pentaho .com/display/ServerDoc2x/Developing + Plugins.

ATENÇÃO Você deve garantir que você faça um back-up do plugin.xml arquivo antes de fazer quaisquer alterações.

CDF JavaScript e CSS Recursos

O painel CDF faz amplo uso de JavaScript para criar e controlar componentes e fazer chamadas para o servidor de BI Pentaho. O JavaScript CDF arquivos estão localizados na js subdiretório do diretório Home CDF. Há um número de diferentes arquivos JavaScript no js diretório, bem como no apoio recursos, como arquivos CSS e imagens:

- jquery.js -O CDF é construído em cima do popular quadro AJAX JQuery- de trabalho. Todos os arquivos que têm um nome que começa com jquery são desta quadro. O quadro JQuery oferece funcionalidades para manipular o baixo nível de detalhes dos documentos HTML mudança no tempo de execução (chamada DOM manipulação), bem como programaticamente fazer requisições web e pro- processamento do conteúdo retornado. Combinadas, estas duas funcionalidades são normalmente referido pelo termo AJAX (Sigla para Asynchronous JavaScript and XML).
- jquery. <name>. js e jquery. <name>. CSS arquivos-Além-entregas rando baixo nível de funcionalidade Ajax, jQuery também oferece uma série de extensões que implementam os elementos da interface do usuário (muitas vezes chamado widgets) como a data catadores, textboxes autocomplete, e assim por diante. Esses arquivos JavaScript e suporte a arquivos CSS implementar vários widgets, bem como no apoio funcionalidade para ações como o conteúdo de posicionamento. Há também uma número de subdiretórios no js diretório. Estes contêm ainda mais Widget plugins JQuery.
- Dashboards.js -Este é o núcleo do CDF. Este arquivo instancia um Dashboards Objeto JavaScript, que é a real execução do painel, ção.

- <name> Components.js -Esses arquivos contêm as definições de objeto do vários tipos de componentes que podem ser exibidos em painéis CDF. Na painel de instrumentos reais, as instâncias desses tipos de componentes são criados e adicionado à Dashboards objeto. CoreComponents.js contém definições de componentes para painéis utilizados com frequência. MapComponents.js contém dois componentes para trabalhar com mapas geográficos. Navegação Components.js contém componentes do painel especial que pode ser utilizado para gerar interação de navegação.

- resources.txt Este arquivo-texto lista todos arquivos JavaScript e CSS que estão a ser incluído no documento resultante do painel HTML que será enviado para o cliente.

Durante a montagem do documento, o CDF plugin lê o resources.txt arquivo e injeta referências a arquivos que constam da lista de JavaScript e CSS no painel de instrumentos
Documento HTML na forma de <script> e <link> elementos. Isso garante
Esses recursos serão carregados pelo navegador quando ele recebe o painel de instrumentos, que é necessário para o painel e seus componentes para funcionar corretamente.

O xcdf. Arquivo

A . Xcdf é um arquivo XML contendo pequenas informações que descrevem o painel de instrumentos. Esses arquivos podem ser colocados em qualquer pasta na solução de repositório
permitem aos usuários navegar no painel de instrumentos.

Algumas informações contidas no . Xcdf arquivo, como nome de exibição, descção, eo ícone são utilizados pelo usuário console a oferecer um item que o usuário possa navegue para o painel. A . Xcdf arquivo também se refere a um arquivo HTML parcial que atua como um modelo de conteúdo painel. Opcionalmente, o . Xcdf arquivo também pode especificar um modelo de documento de painel. Esses modelos são discutidos em detalhes na próxima subseção.

Listagem 17-2 mostra o conteúdo de uma amostra . Xcdf arquivo:

Listagem 17-2: Um exemplo de arquivo. Xcdf

```
<? Xml version = "1.0" encoding = "UTF-8"?>
<cdf>
 <title> Pentaho Dashboard Home </ title>
 <author> Webdetails </ author>
 Home <description> Pentaho Dashboard </ description>
 <icon> </ icon>
 <template> dashboard.html </ template>
 <style> manto </ div>
</ Cdf>
```

Conforme mostrado na Lista 17-1, todo o conteúdo do . Xcdf arquivo está contido em <cdf> e </ Cdf> tags. A cdf elemento contém uma lista de marcas que descrevem o várias propriedades do painel.

As marcas mais importantes são <title>,<template>, e <style>:

- <title> é um elemento obrigatório que define o nome do painel de instrumentos como ele aparece para o usuário no console do usuário.
- <template> identifica o modelo de conteúdo painel. O valor pode ser um nome de arquivo no caso de o modelo de conteúdo reside no mesmo local no solução como o repositório . Xcdf arquivo. Se assim o desejar, o modelo de conteúdo pode ser colocado em outro local dentro do repositório, mas depois o caminho precisa ser especificado também. Por exemplo, <template> foo / bar.html </ Template> especifica bar.html localizado na pasta filho do atual local chamado foo como o modelo de conteúdo.
- <style> identifica um modelo de documento particular. Este é um opcional tag. Se estiver presente, ele identifica um arquivo HTML no qual o conteúdo do painel será colocado. Se essa marca não é especificado, o modelo de documento padrão será usado.

Note que a extensão . Xcdf é necessária. Sem essa extensão, o arquivo não será reconhecido como uma definição de painel. A extensão é controlado através da plugin.xml arquivo discutido na subseção anterior. Para ser preciso, Há dois elementos no plugin.xml arquivo que controlam a extensão:

- A type ="xcdf"Atributo na <content-type> elemento
- A type ="xcdf"Atributo na <content-generator> elemento

Modelos

Nós já mencionei que os painéis são, essencialmente, documentos HTML. O CDF gera esses documentos, mesclando um genérico esqueleto HTML documento com outro documento HTML (parcial) que contém o real painel de definição de conteúdo.

Nós nos referimos a estes dois documentos HTML como modelos, e ao final Página web mesclado como o tablier. Chamamos o documento esqueleto genérico do modelo do documento (Às vezes chamado modelo externo). Nós usamos o termo documento modelo de conteúdo para o arquivo que define o painel real. Figura 17-3 ilustra o processo de montagem do documento.

Modelo de Documento (a.k.a. exterior Modelo)

Documento modelos são projetados para permitir que o conteúdo recorrente para ser reutilizado por vários painéis. Exemplos típicos de conteúdo reutilizável incluem:

- <link> e / ou <style> elementos para definir em cascata folhas de estilo para manutenção de uma aparência consistente
- <script> elementos para adicionar interação personalizada ou importação de extensão componentes do painel
- estruturas de navegação, como links, barras de ferramentas e / ou menus
- Os elementos estruturais para obter um layout de documento genérico

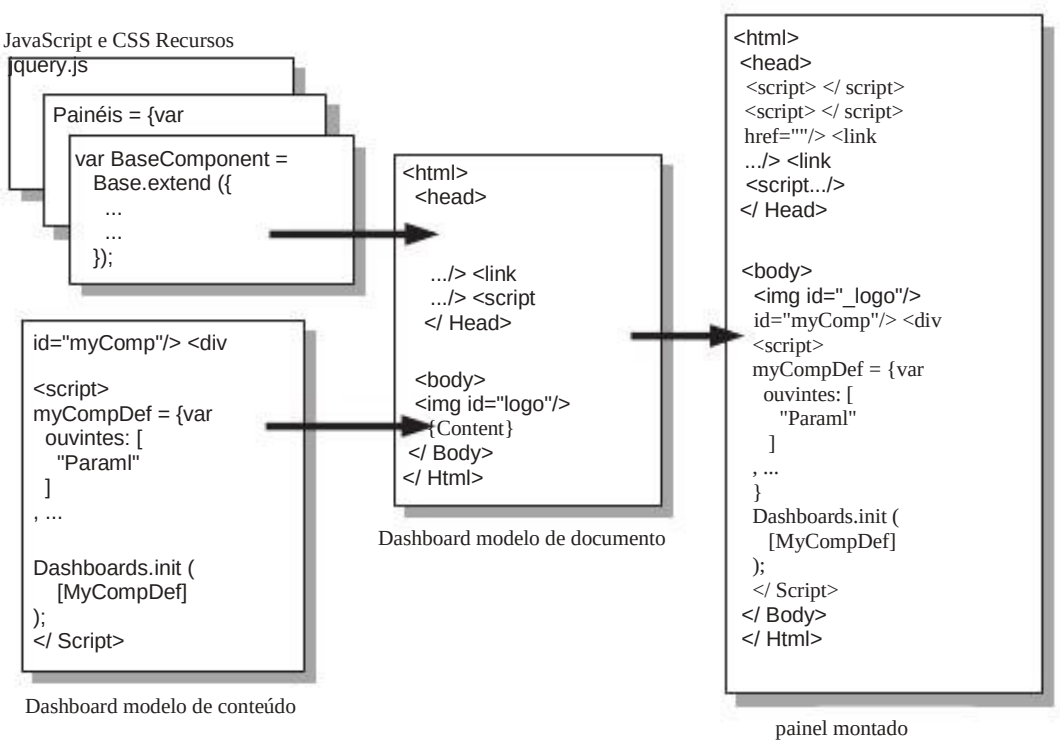


Figura 17-3: Conjunto de documentos

documento CDF modelos residem no diretório home CDF. O padrão Pentaho Community Edition navios servidor com documento de duas dashboard modelos:

- modelo-dashboard.html -Este é o modelo do documento padrão. Se o dashboard . Xcdf arquivo não especificar um <style> elemento, este modelo será usado. Ele inclui alguns elementos de navegação para permitir que os painéis a ser executado fora do ambiente do usuário Pentaho console.
- modelo-painel-mantle.html -Este modelo é um pouco modificada versão do modelo-dashboard.html que não inclui qualquer navegação- ção elementos. Este modelo de documento destina-se painéis que executado dentro de uma página da guia do usuário Pentaho console (que é chamado manto, daí o manto sufixo).

Além dos genéricos, os elementos reutilizáveis de HTML, o modelo de documento também contém espaços reservados, que são substituídos pela CDF plugin durante documento da montagem. Estes marcadores são indicadas utilizando os seguintes sintaxe:

{Nome do espaço reservado-}

Listagem 17-3 mostra alguns dos conteúdos do modelo do documento padrão modelo-dashboard.html.

Listagem 17-3: Alguns dos conteúdos de um modelo de documento da CDF

```
xmlns="http://www.w3.org/1999/xhtml"> <html
  <head>
    <Tipo de link = "text / css" rel = "stylesheet"
      href = "recursos / estilo / template.css" />
  </ Head>
  <body>
    <script type="text/javascript">
      isadmin = "{isadmin}";
      isLoggedIn = "{isLoggedIn}";
    </ Script>
    div id = "content">
      div id = "primaryContentContainer">
        id="primaryContent"> <div
          {Content}
       </ Div>
      </ Div>
    </ Div>
  </ Body>
</ Html>
```

Na Listagem 17-3, você pode ver o número de substituições:

- isadmin -Pode ser substituído por verdade no caso de o usuário atual é o Pentaho administrador.
- isLoggedIn -Pode ser substituído por verdade no caso de o usuário atual está Autenticados.
- conteúdo -Pode ser substituído com o conteúdo do painel de conteúdo modelo. O modelo de conteúdo do painel é discutido em detalhe na próxima subseção.

Os nomes dos arquivos de documento padrão do modelo no CDF início home directório com em modelo de painel e terminam com . Html. Este padrão de nome de arquivo é um requisito para todos os modelos de documentos. Entre os em modelo de painel prefixo eo . Html sufixo não pode aparecer opcionalmente

um nome de modelo, que deve ser precedido por um hífen. Por exemplo, no nome do arquivo do modelo-painel-mantle.html modelo, o texto -Manto aparece entre o prefixo eo sufixo. Neste caso, manto é o real nome do modelo.

Para especificar o modelo do documento, você pode definir o valor da <style> elemento no painel de . Xcdf arquivo para o nome do modelo. Por exemplo, Listagem 17-2 usa <style> manto </ div> para especificar que o manto tem-chapa deve ser usado-map isso para o modelo de painel-manto . Html arquivo.

Se nenhum modelo do documento é especificada no . Xcdf arquivo, o modelo padrão (modelo-dashboard.html) Será usado. Para esses painéis, o documento modelo também pode ser definida dinamicamente, passando o nome do modelo na modelo parâmetro na parte de consulta URI do painel.

Modelo de Conteúdo

O modelo de conteúdo é um arquivo HTML parcial que contém o painel de instrumentos reais definição. Durante a montagem do painel, o {Content} espaço reservado na modelo de documento passa a ter o conteúdo do modelo de conteúdo. A modelo de conteúdo é referenciado a partir do . Xcdf arquivo usando o modelo elemento. Normalmente, você poderia colocar o modelo de conteúdo do painel na solução diretório junto com o . Xcdf arquivo.

O modelo de conteúdo tipicamente contém os seguintes itens:

- JavaScript definições de objeto para os componentes do painel e alguns código para inicializar o Dashboards Objeto JavaScript.
- marcadores HTML para conteúdo componente.
- Opcionalmente, o conteúdo HTML estático para obter um layout de painel componentes.

Listagem 17-4 mostra um exemplo de um modelo de conteúdo muito simples que define um painel com apenas um componente.

Listagem 17-4: Um modelo de conteúdo muito simples CDF

```
id="component_placeholder"> <div
<-! Este espaço reservado componente será preenchido pelo componente ->
</ Div>
<! - Este script define um componente e inicializa o painel ->
language="javascript" type="text/javascript"> <script
/ *
* Definir o componente e configurar suas propriedades.
* /
componente = {var
```

```
    nome: "component_name", tipo: "xaction",
    htmlObject: "component_placeholder",
    solução: "mca", o caminho: "/ painéis / clientes",
    ação: "dashboard_component.xaction",
    parâmetros: []
    executeAtStart: true


}
/ *
* Initalize o painel e adicione o componente.
* /
Dashboards.globalContext = false;
componentes var = [componente];
Dashboards.init (componentes);
</ Script>
```

Listing 17-4 contém apenas dois elementos HTML: a `div` elemento que é usado como componente e um espaço reservado `script` elemento que contém o JavaScript código para configurar o painel e seus componentes. Normalmente, o conteúdo do painel modelos teria mais alguns componentes, assim como algumas HTML para colocar os componentes na página.

Há algumas coisas de nota na listagem 17-4 que se aplicam a todos os painéis modelos de conteúdo:

- espaços reservados de conteúdo são tipicamente `div` ou `span` elementos. Os espaços reservados
  é dado um `id` atributo, que é usado para atribuir um identificador único. Este é necessária para permitir que os componentes para encontrar a área onde eles podem exibir sua saída. No exemplo, usamos um `div` elemento e atribuídas a identificação `component_placeholder`.
- A definição do componente é um objeto literal, que é, pelo menos da a perspectiva do construtor de dashboard mais pequeno do que um conjunto de pares nome / valor. Nós discutiremos os componentes em detalhe no próximo subseção.
- O painel é inicializado com uma chamada para `Dashboards.init ()`. Este inicializa o painel e pode adicionar vários componentes, que são passados como um array.

Exemplo: Clientes e Dashboard Sites

Nesta seção, nós orientá-lo através da criação de um painel de exemplo. Para o exemplo, suponha que você precisa para construir um painel que permite que os gestores da empresa Classe Mundial Filmes para obter rapidamente uma visão geral de seus

clientes e como eles se relacionam com os diversos sites de Classe Mundial Filmes. Figura 17-4 mostra o resultado final deste esforço:

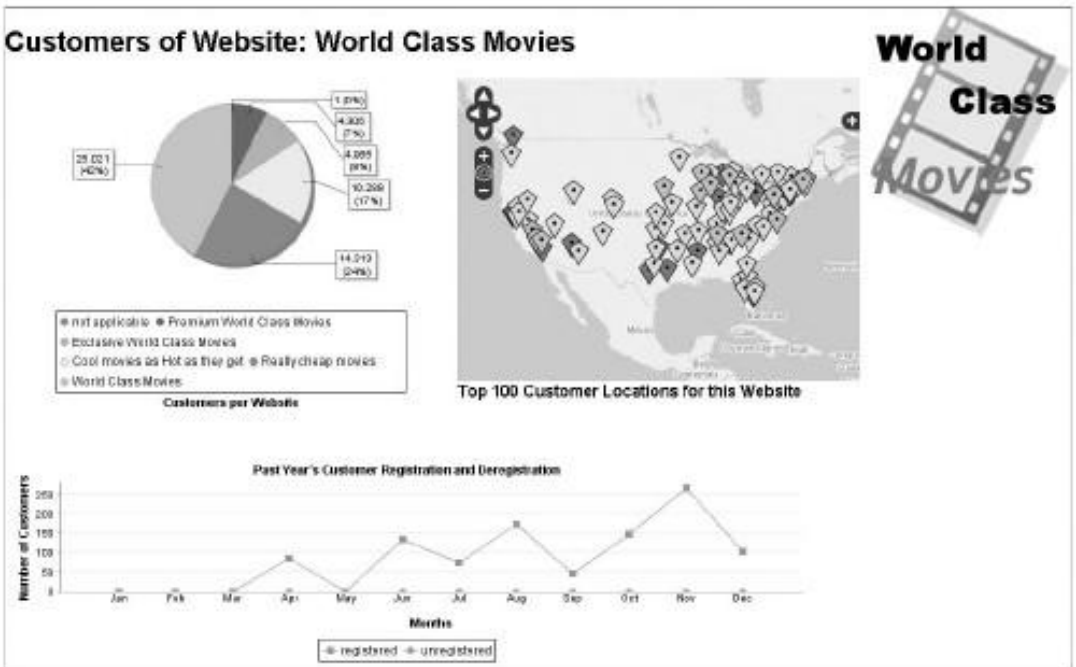


Figura 17-4: Um painel de clientes e websites

O painel contém os seguintes componentes:

- No canto superior esquerdo, um gráfico que mostra o número total de clientes por site. Isso proporciona uma visão rápida e intuitiva do negócio potencial de cada site. Os usuários podem clicar em fatias do bolo para perfurar para baixo e obter informações por site.

- O título do painel exibe o valor da seleção atual feita em o gráfico de pizza. Isso é mostrado no topo do tablier.

- Um mapa que mostra onde os clientes de um site particular, são localizado.

- Um gráfico de linha que mostra registros de clientes e de-inscrições. Nós não incluem os dados reais para adicionar isso, mas deixo isto como um exercício para o leitor. Você deve ser capaz de adicioná-lo em seu próprio país, se você conseguiu a adição dos componentes anteriores.

O restante desta seção descreve passo a passo como criar esse painel de instrumentos.

NOTA Note que não são necessárias ferramentas especiais para criar o painel. Tudo o que você precisa é Pentaho Design Studio para criar seqüências de ação, e um editor de texto para criar os arquivos de painel específico. Também não há instrumentos envolvidos para implementar o

painel de instrumentos. É simplesmente uma questão de colocar os arquivos para a solução de repositório de arquivos
do sistema. Para este exemplo, nós assumimos que você está trabalhando como Administrador com uma
sendo executado localmente servidor Pentaho BI. Dessa forma, você pode simplesmente colocar os arquivos que você
criar para o painel diretamente na solução Pentaho adequado
subdiretório.

Instalação

Antes que você pode construir o painel, é uma boa idéia para criar uma pasta separada manter todos os itens relacionados ao painel. Este é provavelmente o melhor feito a partir do usuário do console.

Nós recomendamos que você criar uma partição dashboards pasta na wcm solução, e abaixo disso, uma Clientes pasta para este painel exemplo específico.

Criar um separado dashboards pasta é uma questão de conveniência. Ele fornece um espaço para manter todos os tipos de recursos compartilhados por vários painéis, e também torna mais fácil criar um novo painel, simplesmente copiando e colando um já existente. É de nenhuma maneira a obrigação de fazê-lo.

Criando o. Xcdf Arquivo

O primeiro passo na criação de um painel é para configurar o . Xcdf arquivo.

1. Criar um novo arquivo de texto chamado customers.xcdf diretamente no sistema de arquivos
   diretório que corresponde à Clientes pasta da solução Pentaho
   repositório.
2. Abra o customers.xcdf arquivo, e adicionar o conteúdo mostrado na
   Listagem 17-5.

3. Salve as alterações para o arquivo.

Listagem 17-5: O arquivo customers.xcdf

```
<? Xml version = "1.0" encoding = "UTF-8"?>
<cdf>
  <title> WCM clientes Painel </ title>
  <author> World Class Filmes - Randy Coon </ author>
  <description>
    Os clientes que WCM Dashboard fornece
    uma visão geral da distribuição de
    - Os clientes WCM sobre websites
    - Cliente local
    - Cancelamento de registro de clientes /
  </ Description>
  <icon> </ icon>
  <template> clientes-template.html </ template>
</ Cdf>
```

Criando o arquivo HTML Dashboard

Na subseção anterior, você criou o customers.xcdf arquivo. Este arquivo inclui o <template> tag e usa-lo para especificar que cliente -Template.html é o seu modelo de conteúdo do painel:

```
<template> clientes-template.html </ template>
```

A partir de nossa discussão anterior sobre o . Xcdf arquivo, você pode se lembrar que o modelo elemento especifica o caminho eo nome do conteúdo do painel modelo. Neste caso, nenhum caminho for especificado, eo CDF plugin irá procurar um arquivo chamado cliente template.html no diretório atual. Assim, você pode basta criar o cliente template.html arquivo no mesmo diretório que o customer.xcdf arquivo.

1. Abra o cliente template.html arquivo e adicionar o conteúdo mostrado na Listing 17-6.

2. Salve as alterações quando tiver terminado.

Listagem 17-6: O arquivo cliente template.html

```
<h1> O WCM Dashboard Clientes </ h1>
language="javascript" type="text/javascript"> <script
  / / Obter a solução eo caminho da localização actual
  solução var = Dashboards.getQueryParameter ("solução");
  var caminho = Dashboards.getQueryParameter ("caminho");



    / / Parâmetros Dashboard
    Dashboards.globalcontext = false;
    / / Definições de parâmetros clique aqui



    / / Definições de componentes
    componentes var = [];


  / Painel de inicialização /
  Dashboards.init (componentes);
</ Script>
```

Código clichê: Como a solução e caminho

O conteúdo inicial da cliente template.html arquivo são apenas uma código clichê esqueleto que agora você pode expandir gradualmente a construir a painel de instrumentos. Nós queremos mencionar algumas coisas sobre esse código.

As duas primeiras declarações definem os solução e caminho variáveis globais. Estes são usados para manter o nome da solução Pentaho corrente e repos-solução caminho itory respectivamente. Esta é uma questão de conveniência: você precisa configurar

uma série de componentes que recorrer a uma seqüência de ação. Além da nome da . Xaction arquivo, esses componentes precisam de saber o nome da Pentaho solução eo caminho de solução de repositório onde o . Xaction arquivo reside.

Você pode obter os valores para as variáveis de solução eo caminho da URI da página web atual. Isso funciona porque o painel em si é servido a partir da solução de repositório, eo nome da solução eo caminho são na parte de consulta da URI. A Dashboards objeto fornece a getQueryParameter método, que foi criado especialmente para analisar a atual URI e pegar o valor do parâmetro de consulta especificado pelo argumento.

Por exemplo, a parte de consulta da URI do nosso painel exemplo é:

```
? Wcm solução = & = caminho dashboard / clientes & action = customers2.xcdf
```

Chamando Dashboards.getQueryParameter ("caminho"), Obtemos a valor / Dashboard / cliente, E chamando Dashboards.getQueryParameter ("Solução") resulta no valor wcm.

Código clichê: Parâmetros Dashboard

A declaração no código clichê Dashboards.globalcontext = true; assinala o início de uma seção que você irá usar para definir os parâmetros do painel e seus valores iniciais.

Dashboard parâmetros são usados para conduzir atualizações de componentes do painel. Não existem definições de parâmetro no código clichê ainda, e você irá adicioná-los como necessários, mas é sempre uma boa idéia de reservar uma seção em antecedência. Uma vez que você achar que você precisa fazer parâmetros, isto irá ajudar a manter los em um só lugar. Nós colocamos a seção de parâmetros antes que os componentes, porque isso irá tornar mais fácil para configurar os componentes que precisam referem-se a todos os parâmetros do painel.

Por enquanto, a única linha desta seção define a globalcontext propriedade da Dashboards objeto a falsa. Por padrão, essa propriedade é verdade, Que permite o painel para usar todas as variáveis globais como parâmetros. Por várias razões, recomendamos que você sempre este conjunto de falsa.

Código clichê: Componentes Dashboard

A seção final do código clichê estão lá para fazer além de componentes um pouco mais conveniente. O padrão típico é a criação de todos os seus componentes definições e adicioná-los todos de uma vez para o Dashboards objeto com uma chamada para o seu init método.

Listagem 17-6 criada uma variável global chamada componentes, E recebe um novo matriz vazia para ele. Como você criar definições novo componente, um por um, você adicioná-los a essa matriz. Finalmente, na última linha do código, você passa a componentes da matriz para Dashboards.init.

A vantagem dessa abordagem é que permite que você mantenha o código que cria cada definição de componente, juntamente com o código que faz com que seja adicionada ao painel. Uma vantagem adicional é que a linha atual que adiciona os componentes não tem a lista de nomes de variáveis do indivíduo componentes. Em nossa experiência, isso torna mais fácil de erro e menos propenso a adicionar
e remover componentes.

Teste

Embora o painel em si ainda não contém nenhum conteúdo real, você pode já que o acesso do usuário do console. É, de fato, sábio para testar isso agora, só para ter certeza que temos os primeiros passos à direita.

1. Porque o customers.xcdf e cliente template.html arquivos residem em a solução de repositório, você deve atualizá-lo para testar o painel. Você pode fazer isso a partir do console ou Administração Pentaho diretamente do usuário consola a última opção é provavelmente a mais conveniente. (Lembre-se você pode atualizar o repositório do usuário console menu, escolha Ferramentas Refresh Repositório de cache, ou você pode clicar no ícone de atualização pouco no canto superior direito do Procurar painel).

2. Depois de atualizar o repositório, você deve ser capaz de invocar o dashboard da consola de utilizador. Figura 17-5 mostra um screenshot do que o painel pode ser semelhante a este ponto.

Figura 17-5: Um painel de clientes vazia

Agora que você tem um painel vazio, você pode iniciar a adição de componentes. Por enquanto, a melhor maneira é um passo de cada vez. Testando o painel de instrumentos após cada
Além torna mais fácil detectar quaisquer problemas logo no início.

Clientes por gráfico de pizza Website

Você vai primeiro adicionar os clientes por gráfico de pizza site. Existem duas maneiras de incluir gráficos em painéis:

- Construir uma seqüência de ação que proporciona a carta, e que incluem no painel usando CDF XactionComponent.
- Criar um CDF JFreeChartComponent e configurá-lo para carregar os dados diretamente a partir de uma consulta SQL ou MDX ou transformação PDI.

Para este exemplo, você vai ficar com a primeira opção. Há uma série de razões para esta escolha.

Primeiro de tudo, uma seqüência de ação em separado serão reutilizáveis ao longo do plataforma. Isto pode não ser a melhor razão para todos, mas se o gráfico é susceptíveis de serem reutilizadas em outras seqüências de ação, ou mesmo outros painéis, em seguida, um seqüência de ação contribuirá para separar a manutenção de uma solução de BI como um inteiro.

A segunda razão para preferir uma seqüência de ação é que não são importantes questões de segurança envolvendo todos os componentes que permitem que dados sejam obtidos directamente a partir de uma consulta. Atualmente, essas consultas são executadas pelo servidor, mas a consulta texto é controlado a partir do lado do cliente. Embora a definição de componentes decorre de tudo que estava especificado no modelo de conteúdo no lado do servidor, o texto da consulta podem ser manipulados pelo cliente, levando a SQL e / ou MDX injeção. (Isto pode ser facilmente alcançado através de um depurador de JavaScript do navegador como o Firebug ou scripting plugin como o Greasemonkey).

Para consultas MDX, as ramificações talvez não sejam tão graves como para SQL, porque os cubos de Mondrian que são alvo de MDX pode ser fixada com autorização baseada em função. Para consultas SQL, isso é praticamente impossível. Na estado atual das coisas, nós recomendamos fortemente contra o uso dela. O CDF comunidade está bem ciente desses problemas e está trabalhando atualmente em uma forma de fornecer uma solução segura para este problema.

Clientes / Website: Seqüência de Ação Pie Chart

O próximo passo é criar a seqüência de ações para entregar a pizza.

1. Para entregar a carta, criar uma seqüência de ação denominada customers_per _website_piechart.xaction e armazená-lo na wcm / painéis / Clientes diretório juntamente com os arquivos do painel outra. A ação design seqüência é mostrado na Figura 17-6.

2. O primeiro passo na seqüência de ação é chamado Obter Contagem cliente por Website. Esta é uma ação do processo relacional. (Você pode encontrar este tipo de ação no submenu Obter dados de.) A consulta SQL que é usada para recuperar os dados do gráfico de pizza é mostrado na Lista 17-7.

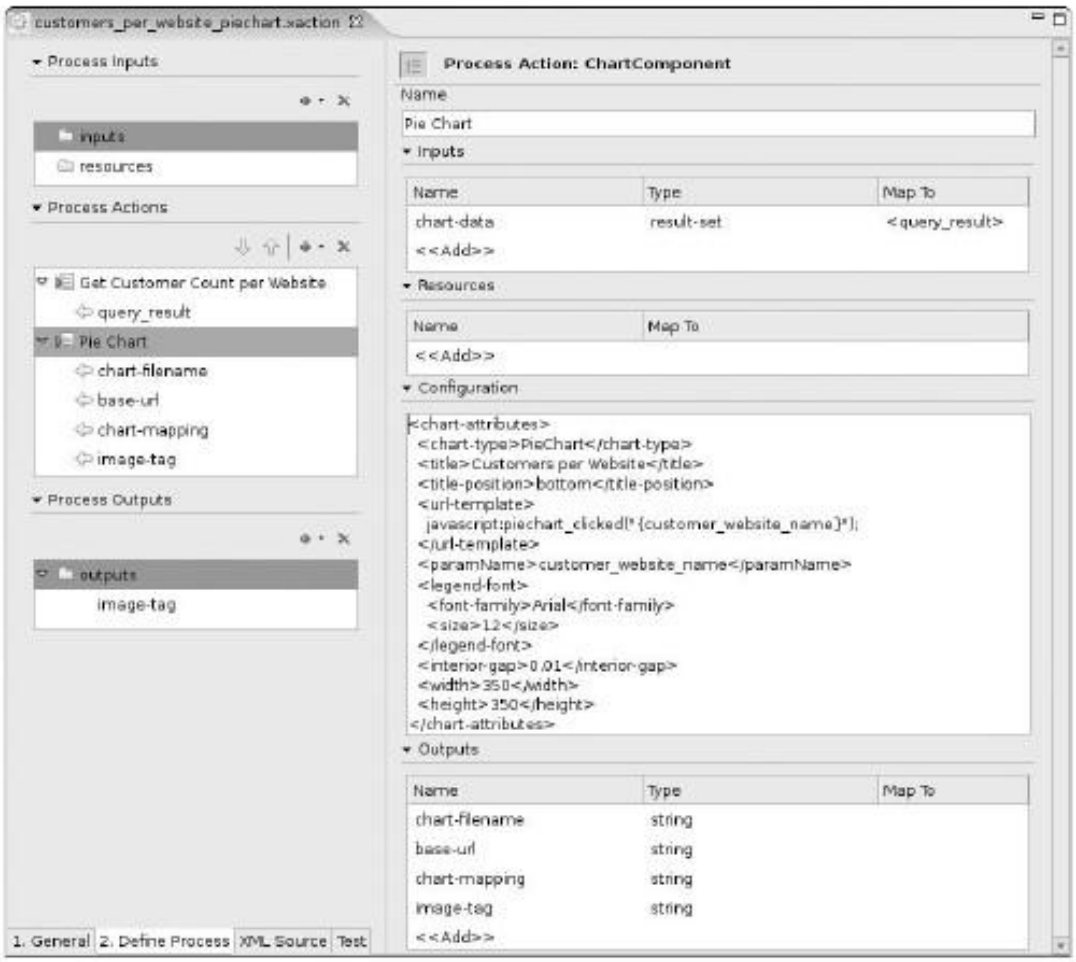


Figura 17-6: O projeto seqüência de ação para os clientes por gráfico de pizza Website

Listagem 17-7: A consulta SQL para os clientes por gráfico de pizza Website

```
SELECTCustomer_website_name,
       COUNT (*) AS customer_count
FROMdim_customer
WHEREcurrent_record 1 =
ANDcustomer_date_unregistered CURRENT_DATE> ()
GRUPO PELA customer_website_name
ORDER BY 2
```

Esta é uma pergunta bastante simples que busca a linha atual do cada cliente registrado, e agrupa as linhas para contar o número de clientes pelo site. Finalmente, o resultado é ordenado pelo customer_count. A ordem é importante e vai afetar a ordem das fatias do bolo gráfico. Na seqüência de ação, o resultado da consulta é acessível para o o restante da seqüência de ação no query_result parâmetro de saída.

3. O segundo passo na seqüência de ação é simplesmente chamado Gráfico de pizza. Você Pode encontrar este processo de ação no menu no submenu Gráfico. A Torta Gráfico ação do processo aceita a query_result da etapa anterior e usa-lo para definir o gráfico de dados parâmetro de entrada. Além disso, alguns dos o Gráfico de pizza'S opções são definidas especificando os atributos diretamente no gráfico a seqüência de ação. A configuração é mostrado na Lista 17-8.

Listing 17-8: Configurando a ação do processo gráfico de pizza

```
<chart-attributes>
  <chart-type> PieChart </ tipo de gráfico->
  Clientes <title> por site </ title>
  <title-position> fundo </ a posição title->
  <url-template>
    javascript: piechart_clicked ("{customer_website_name}");
  </ Template-url>
  <paramname> customer_website_name </ paramName>
  <legend-font>
    <font-family> Arial </ font-family>
    <size> 12 </ size>
  </ Font legenda->
  <interior-gap> 0,01 </ lacuna interior->
  <width> 350 </ largura>
  <height> 350 </ altura>
</ Carta-atributos>
```

NOTA O XML de configuração usados na Listagem 17-08 maio parece um pouco assustador no em primeiro lugar. No entanto, parece mais difícil do que realmente é. O Wiki Pentaho contém excelentes informações sobre a criação e configuração de gráficos. A documentação pode ser encontrado em:

http://wiki.pentaho.com/display/ServerDoc2x/Chart+Reference

As linhas em negrito da Listagem 17-8 são de interesse particular. Estas linhas permitir que o documento HTML que contém o gráfico para reagir no caso de uma fatia de o gráfico é clicado. Uma discussão completa sobre como essa configuração afeta o HTML gerado está além do escopo deste livro. Por agora é suficiente para perceber que o url-modelo elemento provoca uma função JavaScript chamada piechart_clicked () para ser chamado sempre que uma fatia do bolo é clicado. A paramName elemento eo {} Customer_website_name espaço reservado para assegurar que o nome do site que corresponde à fatia de pizza é passado como um argumento para a função. Observe que o nome customer_website_name é

idêntico ao nome da primeira coluna do conjunto de resultados emitido pelo primeiro passo na seqüência de ação.

A tag de imagem parâmetro de saída é usado para entregar o resultado da pizza Gráfico ação do processo para o chamador da seqüência de ação. Isto irá gerar o HTML necessário para tornar um gráfico de pizza no seu painel.

Depois de criar a seqüência de ação, é uma boa idéia testá-lo. Se você acidentalmente cometi um erro na seqüência, mas não testá-lo de antemão, você pode perder muito tempo depurando seu painel apenas para descobrir que ocorreu o problema a montante.

Então, atualize a solução de repositório. Você deverá ver uma janela pop up novo item na usuário do console. Abri-lo para verificar se o gráfico de pizza é exibida. Deve olhar como o gráfico mostrado anteriormente na Figura 17-4.

Clientes / Website: XactionComponent

Depois de ter confirmado a seqüência de ação, você precisará modificar o cliente template.html arquivo para incluí-lo em seu painel. Você precisa:

- Adicionar um elemento de espaço reservado HTML que atua como recipiente onde a saída a partir da seqüência de ação pode ser exibida.
- Criar um objeto JavaScript que sabe como chamar a seqüência de ação. Este objeto é o componente do painel de instrumentos reais.
- Adicione o componente do painel.

1. Para isso, abra o clientes-template.html arquivo e adicionar um div elemento para atuar como um espaço reservado. Inseri-lo logo após a tag de fechamento da inicial h1 elemento, mas antes da script elemento, como mostrado aqui:

```
<h1> O WCM Dashboard Clientes </ h1>
id="customersPerWebsite"> </ div>
language="javascript" type="text/javascript"> <script
... Resto do ficheiro ...
```

Observe o id atributo no div elemento. Como você vai ver, você precisa passar o valor para a definição do componente para que ele saiba onde colocar sua saída.

2. Agora você pode adicionar uma nova seção dentro do elemento script para definir seu componente. Primeiro, encontre a linha que lê:

```
componentes var = [];
```

Imediatamente após essa linha, adicione o código listado na listagem 17-9 e depois salvar suas alterações.

Listagem 17-9: O gráfico circular definição de componentes

```
/ / Clientes / site Pie Chart Component
componentes [components.length] = {
  nome: "customersPerWebsite", tipo: "XactionComponent",
  : solução, o caminho: caminho,
  ação: "customers_per_website_piechart.xaction",
  parâmetros: []
  htmlObject: "customersPerWebsite",
  executeAtStart: true
};
```

O código na listagem 17-9 é uma declaração de atribuição. Para o direito do sinal de igual, você tem um objeto literal (esta é a peça a partir da abertura chaveta de imediatamente após o sinal de igual à chave de fechamento em ao final do anúncio), que será usado pelo Dashboards objeto para criar e configurar o componente real. Para a esquerda do sinal de igual, você tem componentes [components.length]. Como um todo, a declaração tem o efeito de acrescentando um novo componente de configuração para o final da matriz de componentes (Que é armazenado no componentes variável). Nesta fase, o objeto literal em si é pouco mais do que um saco de pares nome / valor. Aqui vai uma breve explicação de suas finalidades:

- nome Um valor de cadeia. Isso deve identificar o componente entre todos os outros componentes na página.
- tipo Um valor de cadeia. Internamente, isso é usado para fazer o instanciados real- ção de uma das classes de componentes. Neste caso particular, você quer para obter um resultado de uma seqüência de ação, que é porque você usou o string XactionComponent. Alternativamente, você também pode usar xaction ou Xaction -Todos estes são mapeados para o XactionComponent componente em o CoreComponents.js arquivo.
- solução Um valor de cadeia. Este é o nome da solução Pentaho onde a seqüência de ação, queremos chamar reside. Basta atribuir o solução variável, que nós vimos em nossa discussão sobre o código clichê.
- caminho Um valor de cadeia. Este é o caminho da solução de repositório onde o seqüência de ação, queremos chamar reside. Basta atribuir o caminho variável, que nós vimos em nossa discussão sobre o código clichê.
- ação -O nome do . Xaction arquivo.
- parâmetros -Esta é uma matriz para especificar quais parâmetros devem ser passou ao invocar a seqüência de ação. A Gráfico de pizza ação seqüência não exige (nem esperar) quaisquer parâmetros, razão pela qual a matriz está vazia. Quando você configurar outros componentes, você vai aprender como configurar o parâmetros membro.

- htmlObject Um valor de cadeia. Aqui, você pode atribuir o valor do atributo ID do titular do componente local. Isto estabelece uma ligação entre a objeto componente JavaScript e HTML do documento e permite a componente para mostrar a sua saída no painel.

- executeAtStart -Um valor booleano. Configurando-o para verdade faz com que o com- nente a ser prestado, logo que todos os componentes são adicionados à painel de instrumentos. Caso contrário, o componente será atualizado somente em resposta a uma mudança de parâmetros. Isso é discutido na próxima subseção.

Agora que você mudou o dashboard para incluir um componente, você deve Refresca a solução de repositório e testar o painel. Quando você invocar o painel você verá que o gráfico de pizza foi adicionada ao painel.

Alterar dinamicamente o título Dashboard

Nesta seção, vamos explicar como permitir que componentes do painel de interagir com uns aos outros. Isto envolve três adições distintas para cliente template.html:

- A piechart_clicked () função, que será chamado sempre que o usuário clica em uma fatia do gráfico de pizza
- A site parâmetro do painel, que é utilizado para comunicar a nome da fatia clicado gráfico de pizza para o Dashboards objeto
- O marcador de código HTML e JavaScript para uma TextComponent que reage a uma mudança na site parâmetro, mostrando o seu valor

Adicionando o parâmetro website_name Dashboard

Porque tanto a piechart_clicked () função e os novos TextComponent necessidade de se referir ao site parâmetro, é melhor criar esse primeiro.

Acrescente o seguinte código para a parâmetro seção que ele lê:

```
/ Nome / parâmetro
param_website var = "website";
/ / Valor inicial do parâmetro
Dashboards.setParameter (param_website, "Classe Mundial Filmes");
```

Este novo trecho de código faz duas coisas:

- Atribui o nome do parâmetro site para a variável global param_website
- Na verdade, cria o parâmetro na Dashboards objeto e fornece- com um valor padrão, neste caso "Classe Mundial Filmes"

Como você pode ver, o nome do parâmetro é armazenado em uma variável global para você pode usar isso para se referir ao nome do parâmetro. A vantagem dessa abordagem é que você não precisará digitar repetidamente o nome do parâmetro literal todo o código. Se você tivesse que usar a seqüência literal "Site", Há o risco de

acidentalmente apertou o nome. Você pode, naturalmente, também cometeu um erro quando digitando o nome da variável param_website, Mas isso vai resultar em uma execução clara de erro sempre que você se refere à variável não-existentes. Problemas causados por erros de digitação do nome do parâmetro literal são muito difíceis de solucionar: qualquer erros de execução podem surgir numa fase muito mais tarde, depois de dashboard inicialização, e é também provável que os componentes parecem falhar silenciosamente.

Note que são necessários para criar o parâmetro usando setParameter porque você defina explicitamente a globalcontext membro da Dashboards objeto para falsa. Se você deixou globalcontext ativada, você ainda seria necessária para ""Criar o parâmetro, mas teria que se resumia a criação de uma global site variável, assim:

```
/ / Inicializa o parâmetro website
var site = "Classe Mundial Filmes";
```

Embora isso possa parecer simples à primeira vista, o uso de variáveis globais faz mais difícil manter o código de dashboard. Por um lado, é mais difícil distinguir entre as variáveis globais que são"apenas"variáveis globais e as os que são usados como parâmetros. É muito fácil de substituir acidentalmente o valor de uma variável global, o que torna difícil para depurar problemas. Ao desativar globalcontext, Você é forçado a usar explicitamente as chamadas para Dashboards.setParameter () e Dashboards.getParameterValue (), Que torna claro que a intenção é.

Reagindo aos cliques do mouse sobre o gráfico de pizza

O segundo passo é criar o piechart_clicked () função. Lembre-se que, quando você configurou o url-modelo propriedade da tabela (veja a listação 17-8), só fez com que o piechart_clicked () função a ser chamada. Você nunca definiu a função em si. Agora você irá adicionar o código para o piechart_clicked () função.

Listagem 17-10 mostra uma possível implementação do piechart_clicked () função. Adicione este código para o cliente template.html arquivo, diretamente abaixo a definição do componente gráfico de pizza.

Listagem 17-10: A função piechart_clicked

```
/ / Função que lida com os cliques do mouse sobre as fatias de pizza
função piechart_clicked (website) {
  curr_param_website var = Dashboards.getParameterValue (param_website);
  if (curr_param_website! site =) {
     Dashboards.fireChange (site param_website);
  }
}
```

Quando piechart_clicked ()é chamado, o nome da categoria que pertence a fatia correspondente é passado através da sua função site argumento. Este é um conseqüência imediata da maneira que você configurou o url-modelo propriedade da Gráfico de pizza seqüência de ação. Este código é mostrado na íntegra na Lista 17-8.

No corpo da função, você primeiro determinar o valor atual da o site painel de parâmetros. Isso é feito chamando o getParameter Valor () método da Dashboards objeto, passando o param_website variável, que detém o nome do parâmetro.

Em seguida, verifique se o valor atual da site parâmetro difere de o novo valor passado como argumento. Isto é usado por um se declaração para decidir se deve chamar o fireChange () método da Dashboards objeto. A fireChange () método da Dashboards objeto de espera para ser passados dois parâmetros:

- O nome de um painel existente parâmetro Mais uma vez, você pode usar o variável global param_website.
- O novo valor para o parâmetro especificado-Este é o valor do argumento você recebe do gráfico de pizza.

O objetivo da fireChange () método consiste em comunicar a alteração ao Dashboards objeto, que então notificar os outros componentes do mudar.

NOTA Note-se que fireChange""é um pouco enganador como o fireChange () método não significa necessariamente uma mudança de fogo. Se você realmente quiser garantir que você está apenas a sinalização mudar, você tem que certificar-se explicitamente a si mesmo. É por isso que o nosso execução do piechart_clicked () método utiliza um se declaração evitar falsas sinalização de mudanças que não ocorreu.

Adicionando um TextComponent

O passo final é adicionar um componente do painel que pode exibir o valor de o site parâmetro. A TextComponent destina-se precisamente este tipo da tarefa.

1. Para incluir esse componente do painel, você precisa adicionar um elemento HTML espaço reservado onde o componente pode tornar seu texto. Você vai usar um span elemento para esse fim, que você vai colocar na já existente h1 elemento:

   ```
   Os clientes da <h1> Website: id="websiteName"> <span </ span> </ h1>
   ```

   Você usa um span elemento para garantir o texto processado pelo componente irá aparecer"inline"com o texto da contendo h1 elemento. Nota

que incluiu uma id atributo e atribuído o valor WebSiteName. A id atributo será usado pelo TextComponent para manipular o documento HTML, neste caso, para processar um texto. Esta é semelhante ao nosso uso prévio do id atributo no espaço reservado correspondente ao XactionComponent nós usado para exibir o gráfico de pizza.

2. Depois de adicionar o espaço reservado, você pode adicionar o código JavaScript para criar a componente. O código para a TextComponent é mostrado na Listagem 17-11. Adicione este código logo abaixo do piechart_clicked () função:

Listagem 17-11: Código para o TextComponent para alterar dinamicamente o nome do parâmetro website no título do painel

```
/ / Textcomponent nome Website
componentes [components.length] = {
  nome: "website", tipo: "TextComponent",
  ouvintes: [param_website]
  expressão:
   function () {
    retorno Dashboards.getParameterValue (param_website);
   }
  htmlObject: "site da Web",
  executeAtStart: true
};
```

O código na Listagem 17-11 revela uma série de semelhanças com o código para o XactionComponent usado para exibir o gráfico de pizza (que é mostrado na Listagem 17-9). A nome,tipo,htmlObject e executeAtStart membros foram todos os mostrados aqui, e tem um significado semelhante para este tipo de componente. Ao contrário da XactionComponent mostrado na Lista 17-9, o TextComponent não não invocar uma seqüência de ação e, portanto, as variáveis de membro solução, caminhoE ação e os parâmetros não são aplicáveis. Há dois membros variáveis que não encontrou antes:

- expressão Função de uma seqüência. Para o TextComponent, Esta função ser chamado para entregar o valor de texto que será colocado no elemento identificados pelo htmlObject membro. Na Listagem 17-11, optamos por anexar uma função inline chamada anónima directamente para o membro. Se não houver uma função chamada já está disponível, você pode simplesmente atribuir a função nome próprio. Se você usar uma função chamada, note que você não deve anexar o parênteses após o nome da função a que tem o efeito de chamar a função, o que fará com que o valor de retorno da função para ser atribuído, e não a própria função.

- ouvintes -Uma matriz de nomes de parâmetro. Esse membro pode ser configurado para todos os tipos de componentes. Listagem de um nome de parâmetro dashboard

aqui faz com que o componente de ouvir esse parâmetro. Na prática, isso significa que o componente será chamado para atualizar-se sempre que o fireChange () método da Dashboards objeto é chamado com o respec- nome do parâmetro tiva como seu primeiro argumento. Na Listagem 17-11, nós configuramos o componente para ouvir o site parâmetro, incluindo o param_website variável em sua matriz ouvintes.

Após concluir essas etapas, você deve testar para ver se tudo funciona. Assim Refresca a solução de repositório e em seguida, abra o painel. Clique sobre a torta para garantir que os cliques são capturados. Também verifique se o título do painel é atualizado automaticamente quando clicar no gráfico de pizza.

Mostrando a localização do cliente

Agora que você tem um parâmetro do painel, que é controlado pelo mouse clique no gráfico de pizza, você pode adicionar mais componentes e configurar ouvintes para isso.

Nesta subseção, você vai aprender como adicionar um CDF MapComponent para mostrar distribuição geográfica dos clientes. Você vai usar isso para mostrar um mapa de Estados Unidos, que marca os 100 primeiros (por número de clientes) locais dos clientes da empresa de classe mundial de filmes para o site atual.

NOTA O CDF MapComponent é baseado na biblioteca JavaScript OpenLayers. Ao contrário de outras soluções de web popular mapa da página, OpenLayers é open source e disponível sob a licença BSD. É enviado junto com o CDF. Você pode encontrar mais informações sobre o OpenLayers na http://openlayers.org.

Além da biblioteca OpenLayers, o CDF MapComponent Também usa web pedidos de www.openstreetmap.org, que fornece os dados para tirar da rua sobreposições para um mapa OpenLayers.

O CDF MapComponent também podem fazer solicitações da Web para http://www.geonames . Org / que é utilizado para fazer pesquisas de longitude / latitude de nomes local.

MapComponent CDF formato de dados

O CDF MapComponent pode ser usado para indicar as localizações no mapa mundial. O componente usa uma seqüência de ação para obter os dados que aparecem no no mapa. O conjunto de dados tem o seguinte formato:

- id -Um identificador exclusivo para um local.
- latitude -A latitude geográfica do local. Isso define o distância de um local a partir do equador, expresso em número de graus. Latitude varia de +90 (No Pólo Norte) a 90 (centro de Antártica). Se a latitude não estiver disponível, você pode fornecer um vazio string.

- longitude De longitude geográfica do local. Isso define o distância de um local a partir do primeiro meridiano, expressa como um número de graus. Longitude varia de +180 a -180. Se a longitude não é disponível, você pode fornecer uma seqüência vazia.
- nome -A seqüência que representa um nome legível do local.
- valor -O valor da métrica que você deseja mapear.
- título -Esse valor opcional pode ser utilizado pelo MapBubbleComponent. (Este é um componente que pode ser chamado a partir de uma localização no mapa para mostrar os detalhes relativos a esse local específico.)

Se você olhar para o formato do conjunto de dados, você pode perceber uma redundância: Localização pode ser especificado por nome, mas também por longitude / latitude. Se o latitude e longitude não estão presentes, o nome do local é utilizado para pesquisa um serviço web fornecido por www.geonames.org. Isso funciona bem se você estiver interessados em mapear alguns locais, mas pode se tornar bastante lento quando centenas de mapeamento de locais. Além disso, um único nome pode ser mapeado para múltiplas
locais, por isso é melhor sempre explicitamente fornecer dados de longitude e latitude.

Mais tarde nesta seção descrevem como criar a seqüência de ação para fornecer os dados para o MapComponent.

Acrescentando uma dimensão Geografia

Antes que você possa criar uma seqüência de ação para entregar os dados para o MapComponent, Você precisará obter a latitude e longitude para todas as localidades você deseja mapear. Para manter as coisas simples, optamos por não incluir uma geografia dimensão no nosso armazém de dados Classe Mundial Filmes. No entanto, para o finalidade de demonstrar a MapComponent neste painel, precisamos fornecer a longitude ea latitude de dados para a localização do cliente.

NOTA A dimensão geográfica ou local é uma característica comum em muitos dados projetos armazém. Na Classe Filmes do Mundo de data warehouse, uma geografia tabela de dimensão pode ser usado para navegar na fact_customer, fact_order, e fact_inventory tabelas de fatos. Além disso, pode servir para o floco de neve dim_customer e dim_warehouse tabelas de dimensão. A única razão para não incluí-la foi manter a simplicidade. No entanto, incorporando uma geografia tabela de dimensão na wcm_dwh banco de dados (e modificar o processo de ETL nesse sentido) é um excelente exercício para a esquerda para o leitor.

Listagem 17-12 mostra o layout da tabela de dimensão geográfica.

Listagem 17-12: O layout da tabela de dimensão dim_geography

```
CREATE TABLE dim_geography (
  geography_keyINTEGER                    NOT NULL,
  Geography_country_code CHAR (2)         NOT NULL,
```

```
  Geography_country_name VARCHAR (50) NOT NULL,
  geography_region_code CHAR (2) NOT NULL,
  Geography_region_name VARCHAR (50) NOT NULL,
  geography_longitudeDOUBLENOT NULL,
  geography_latitudeDOUBLENOT NULL,
  geography_city_nameVARCHAR (50) NOT NULL,
  PRIMARY KEY ("geography_key '),
  ÍNDICE
(Geography_city_name, geography_region_code geography_country_code)
)
```

Os dados para a dim_demography tabela pode ser obtido de diversas fontes. Nós usado o''a população mundial das cidades''dataset fornecido por MaxMind. Você pode baixá-lo como worldcitiespop.txt.gz a partir de http://geolite.maxmind.com/download / worldcities. Alternativamente, você pode usar os dados da geonames . Org. Já mencionamos que este serviço web é usada pelo CDF MapComponent para obter dados de latitude e longitude online, mas o mesmo site também fornece arquivos comprimidos que estão disponíveis para download http://download.geonames.org/export/dump.

Nós não vamos discutir o carregamento do dim_geography dimensão em detalhes, como o processo é bastante simples. Nós carregamos os dados com PDI. Primeiro nós usou o texto de entrada e saída de mesa etapas para carregar os dados em uma tabela na estadiamento área de banco de dados. Em seguida, esta tabela indexada para permitir uma eficiente pesquisa no país, o estado (região), cidade e. Em uma transformação individual, foi utilizado um Tabela etapa de entrada para carregar todas as linhas atuais do cliente tabela na wcm banco de dados. Nós adicionamos Database Lookup etapas para localizar as linhas correspondentes na o região e país tabelas, e isso permitiu-nos olhar para cima e longitude latitude pelo nome da cidade, o código de região (estado), eo código do país. O resultado foi despejados no dim_geography tabela.

Localização Seqüência de Ação de Dados

Agora que você tem os dados necessários, você pode criar uma seqüência de ação que pode entregar os dados para um MapComponent. Neste caso, você pode usar um muito seqüência de ações simples, contendo apenas um Obter dados de / Relacional etapa. O projeto da seqüência de ação é mostrado na Figura 17-7.

Note que a seqüência de ação tem um parâmetro de entrada chamado cliente _website_name. Este será usado para obter os locais para os clientes do cur- atualmente selecionado site. O resultado da ação do processo é mapeado para o consulta _result parâmetro de saída. Isso garante que os dados podem ser interpretados a partir do resposta depois de fazer uma solicitação da web para a seqüência da ação.

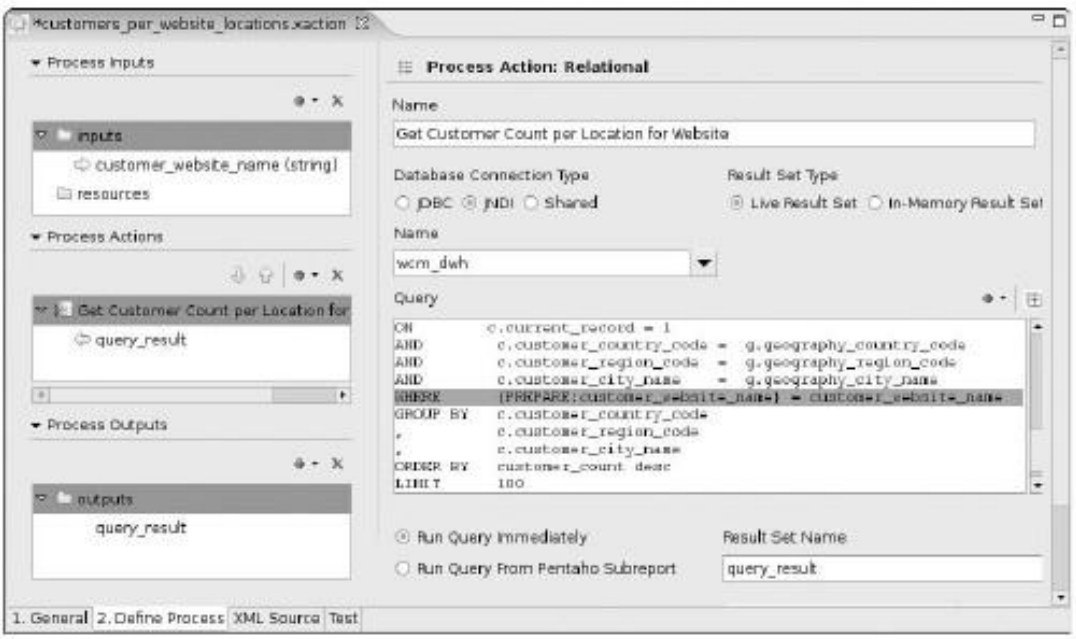


Figura 17-7: Uma seqüência de ações para entregar os dados de localização

O SQL é mostrado na Listagem 17-13.

Listagem 17-13: Uma consulta SQL para obter os locais dos clientes

```
SELECT      g.geography_key
,           CASO QUANDO g.geography_key IS NULL THEN"
              g.geography_latitude ELSE latitude ENDAS
            CASO QUANDO g.geography_key IS NULL THEN"
,             g.geography_longitude ELSE longitude ENDAS
            CONCAT (c.customer_city_name
            , '', C.customer_region_code
,           , '(', C.customer_country_code, ')') AS LOCATION_NAME
            COUNT (*) AS customer_count
            CONCAT (c.customer_city_name) LOCATION_TITLE AS
            dim_customerc
,           g dim_geography
,           c.current_record = 1
DA          c.customer_country_code = g.geography_country_code
LEFT JOIN   c.customer_region_code = g.geography_region_code
ON          c.customer_city_name = g.geography_city_name
E           {PREPARO: customer_website_name customer_website_name} =
E           c.customer_country_code
E           c.customer_region_code
ONDE        c.customer_city_name
GROUP BY    customer_count desc
,           100
,
ORDER BY
LIMITE
```

NOTA Se o wcm_dwh projeto teve destaque o dim_geography tabela, o dim_customer tabela provavelmente teria sido floco de neve, permitindo uma simples INNER JOIN usando o geography_key.

Colocando no Mapa

Você pode agora adicionar o código para incluir o MapComponent em seu painel.

1. Abra o cliente template.html arquivo e adicionar a linha a seguir após o espaço reservado para o gráfico de pizza.

   id="map" div style="width:400px; height:300px;"> </ div>

   Este será o espaço reservado para o MapComponent. Há duas coisas de uma nota, esse elemento de espaço reservado HTML:

   - Atualmente, o CDF suporta exatamente um mapa por painel, e deve estar em um espaço reservado com o id = "map".
   - A largura ea altura são explicitamente especificados usando um estilo atributo. Normalmente, é aconselhável que se abstenha de utilizar os atributos de estilo inline porque torna mais difícil de painéis tema mais tarde. No entanto, neste caso, você precisa especificar explicitamente as dimensões do placeholder '', ou o mapa não será visível.

2. Agora você pode adicionar o MapComponent O código JavaScript para o componentes seção. O código é mostrado na Listagem 17-14.

Listagem 17-14: O componente de mapa

```
/ / Posições do cliente componente do mapa
componentes [components.length] = {
  nome: "customerLocations", digite: "mapa",
  : solução, o caminho: caminho,
  ação: "customers_per_website_locations.xaction",
  ouvintes: [param_website]
  parâmetros: [["customer_website_name", param_website]]
  htmlObject: "mapa",
  executeAtStart: true,
  initPosLon: -37,370157, initPosLat: -100,458984,
  initZoom: 1,
  expressão:
    function () {
      return "" js / Openmap / OpenLayers / img / marker.png ";
    }
};
```

Um número de membros que configurada na Lista 17-14 foram introduzidas na listagem 17-9 (o código do componente gráfico de pizza) e 17-11 (o título do painel

componente de código). As configurações seguintes membros merecem um pouco mais discussão:

- tipo,nomeE htmlObject Seqüência de valores. Atualmente, o MapComponent depende de algumas variáveis globais que restringem todas estas propriedades de ter um valor diferente do "Mapa". Note que isto também se aplica a o valor da id atributo do elemento HTML espaço reservado.

- ouvintes -Note-se que o membro está configurado exatamente como você viu na Listagem 17-11. Ele garante que o mapa será atualizado em resposta a um mudança na site parâmetro.

- parâmetros -Uma matriz de matrizes. Você já encontrou o parâmetros membro na Lista 17-9. Desta vez, você passa um componente definição de parâmetro para vincular o parâmetro do painel à entrada parâmetro da seqüência de ação. Como mostrado na Listagem 17-13, um parâmetro definição de um componente é também uma matriz com duas entradas. O primeiro entrada é uma expressão que avalia o nome da seqüência de ação parâmetro. Na Listagem 17-13, esta é a seqüência de caracteres literal "Customer_website _name ". Note que isso é exatamente o nome do parâmetro de entrada do customers_per_website_locations.xaction. A segunda entrada no definição de parâmetros de componentes é a variável param_website, Que contém o nome do site painel de parâmetros. Juntos, esse garante que sempre que o componente de mapa é chamado para atualizar a si mesmo, irá utilizar o valor da site parâmetro painel para parametrizar o customer_website_name da seqüência de ação.

- initPosLon e initPosLat -dupla valores. Estes membros são específicos ao MapComponent. Eles controlam o local deve ser inicialmente o centro do mapa. Neste caso, o mapa irá utilizar o centro dos EUA como o ponto focal.

- initZoom -inteiro. Especifica o fator de zoom inicial para o mapa.

- expressão -Uma função string retornando de uma expressão. Para cada ponto o conjunto de dados, essa função será chamada. A função deve retornar uma expressão que avalia para um URI (como um valor de cadeia). A URI é usado para criar uma imagem que está a ser colocado no mapa, o correspondente Local. Na Listagem 17-14, sempre voltamos a URL de um marcador pouco Imagem que é fornecido com OpenLayer.

Isso é tudo que existe para ela. Agora você pode atualizar o cache de repositório e de teste o painel novamente. Clique no gráfico de pizza, e repare que o título eo o mapa são atualizados.

Usando marcadores diferentes dependendo dos dados

Na subseção anterior, explicamos como adicionar um MapComponent ao painel de visualização da distribuição geográfica dos clientes. No entanto,

você realmente não fazer nada com a métrica atual, o número de cus-
entes. Nesta subseção, você adicionará alguns códigos para mostrar diferentes marcadores dependendo da quantidade de clientes que residem naquela localidade.

Para o nosso exemplo, vamos supor que você está interessado em distinguir entre locais que têm poucas quantidades, moderada ou grande quantidade de clientes. Por enquanto
Vamos resolver por algumas regras simples:

- Localidades com 100 ou menos clientes são considerados pequenos, e vai ficar um marcador verde.
- Localidades com mais de 250 clientes são considerados importantes e obter um marcador vermelho.
- posições restantes (com mais de 100, mas não mais de 250 clientes) são considerados moderados e obter um marcador amarelo.

Vamos primeiro tentar e tentar chegar a algo que funciona mais ou menos.
Você pode, então, gradualmente melhorar o código, se quiser. Uma execução muito simples, implementação do novo requisito envolve a escrita de uma expressão alterada de o expressão membro o direito de escolher marcador da imagem URI.

Se você manter as coisas simples e se contentar com as imagens junto com OpenLayer, o código na Listagem 17-15 implementa suas exigências.

Listagem 17-15: Código modificado para o membro de expressão

```
expressão:
 function () {
  return "valor <= 100? 'Js / Openmap / OpenLayers img / marcador green.png "
  + ": Valor> 250? 'Js / Openmap / OpenLayers / img / marker.png "
  + ":" Js / Openmap / OpenLayers img / marcador gold.png "
  ;
 }
```

No entanto, esse código deixa muito a desejar. Primeiro de tudo, é muito difícil para ler, devido ao fato de que você tem que colocar o seu código em uma seqüência literal. Se
fazer o exercício mental de chamar a função atribuída à expressão membros e, em seguida, imagine o código que avalia, você acaba com algo equivalente ao trecho a seguir:

```
    valor <= 100? 'Js / Openmap / OpenLayers img / marcador green.png'
  : Valor> 250? 'Js / Openmap / OpenLayers / img / marker.png'
  : 'Js / Openmap / OpenLayers img / marcador gold.png'
```

Agora que abriu o código, você pode ver que há uma variável chamada valor que contém o valor da métrica. Esta variável é colocado à disposição os componentes internos. O código desembrulhou permite detectar alguns problemas:

- O código refere-se às imagens usando literais string. Não há nenhuma maneira de configurar o que as imagens serão mostradas. Isso é ruim porque é difícil uma fios

aspecto da apresentação (a imagem do marcador) para a lógica que é escolhido para um apresentação especial.

- A lógica em si, com os valores-limite de 100 e 250, também é cabeadas. Se o conceito de"grande"e"locais"pequena é provável que mudar ao longo do tempo, este tipo de código impede que o painel de adaptação para esse tipo de mudança.

A solução para ambos os problemas é fazer com que os valores literais configurável. Felizmente, o MapComponent foi projetado para ter uma instalação para configurar o imagens do marcador. Isso pode ser configurado através do marcadores membros, que pode ser definido como um array de strings URI. Assim como o valor variável, o marcadores membro é magicamente colocados à disposição da expressão. Isto significa que você pode utilizar o seguinte código para extrair as strings URI literal da expressão código:

```
componentes [components.length] = {
  nome: "customerLocations", digite: "mapa",

    ... Os outros membros ...,

    Marcadores: [
      "Js / Openmap / OpenLayers img / marcador green.png",
      "Js / Openmap / OpenLayers / img / marker.png",
      "Js / Openmap / OpenLayers img / marcador gold.png"
    ]
    expressão:
      function () {
        return "valor <= 100? marcadores [0] "
        + ": Valor> 250? marcadores [1] "
        + ": Marcadores" [2]
        ;
      }



}
```

Este é definitivamente melhor, como nós podemos agora alterar as imagens sem marcador alterar o código que executa a lógica. Dito isto, há ainda alguns espaço para melhorias. Agora, os limiares de 100 e 250 ainda estão hard-wired no código. Agora, esses valores não são imutáveis, nossa idéia de que um grande número de clientes é pode mudar ao longo do tempo, ou neste caso, nós podemos desenvolver a opinião de que os valores-limite deve ser realmente depende do site. Outra razão pela qual não estamos tão felizes com o hard-wired constantes é que não podemos simplesmente aplicar a mesma lógica de outros painéis que não é, sem duplicar o código. Combine isso com a obrigação de alterar os valores em algum ponto no futuro, e você está olhando para o inferno manutenção.

Felizmente, existe uma maneira simples de lidar com isso. Tudo o que você realmente tem que fazer
é substituir a ocorrência dos valores constantes, com referências ao painel parâmetros. Então, primeiro adicionar dois novos parâmetros para os valores-limite na seção de parâmetros:

```
low_customer_count var = "low_customer_count";
Dashboards.setParameter (low_customer_count, "100");
high_customer_count var = "high_customer_count";
Dashboards.setParameter (high_customer_count, "250");
```

Note-se que atualmente, o CDF suporta apenas parâmetros de valor da cadeia. A membro da expressão modificada é mostrado aqui:

```
expressão:
  function () {
  <voltar valor "= parseInt (Dashboards.getParameterValue
    (Low_customer_count))? Marcadores [0] "
+ ": ParseInt> valor (Dashboards.getParameterValue
    (High_customer_count))? Marcadores "[2]
+ ":
    marcadores [1] "
  ;
  }
```

Como você pode ver, as variáveis limite são parametrizadas. Observe como o built-in parseInt () função é usada para converter os valores de parâmetro para o tipo inteiro. Isso é necessário para uma comparação (numérica) adequada.

Note que nós não (ainda) adicionar o low_customer_count e high_customer _count parâmetros para o ouvintes matriz do MapComponent. Tornaria bom senso para fazê-lo se os valores-limite foram dependentes do site, ou Se o painel permitiu que o usuário final para inserir um novo valor limite para permitir uma análise what-if. Isso é deixado como um exercício para o leitor.

Styling e Personalização

Nas seções anteriores, você criou um painel, componentes e interação. Até agora, nós deliberadamente suprimida a tentação de ter de fazer a painel de boa aparência. Isto permitiu que você se concentre em detalhes técnicos de tornar o trabalho componentes do painel, sem ter que lidar com o distração dos problemas de layout e afins. Outra consideração é que o habilidades necessárias para criar painéis e componentes do painel de construção são realmente muito diferente do que você precisa para fazer um layout e criar um esteticamente agradável todo. As tarefas são ex-melhor feito por um desenvolvedor de BI apenas com o desenvolvimento de competências poucos web, enquanto o segundo é melhor feito por um
desenvolvedor web, talvez em conjunto com um especialista em usabilidade.

Nesta seção, mostramos algumas técnicas que você pode usar para adicionar um layout para
seu painel. Uma discussão detalhada sobre este tema está fora do escopo deste
livro, mas pelo menos podemos mostrar os primeiros passos do processo. Estamos muito confiante de que você deve ser capaz de começar por aqui e siga o seu próprio percurso, desde que você tenha algumas habilidades em páginas da Web de estilo, ou se você pode confiar em uma
desenvolvedor web para ajudá-lo.

Denominando o Dashboard

Há duas coisas a considerar quando se denominar um painel:

- Organizando-Layout colocação dos componentes. Isso vale principalmente para painéis individuais.
- Theming-Garantia uma aparência consistente é usado em painéis, ou um aplicativo inteiro.

Por definição, o componente layout será em grande parte específica para um indivíduo painel de instrumentos. Por esta razão, o estilo ea estrutura do documento para controlar o layout
pertence principalmente no modelo de conteúdo. Theming, por outro lado, envolve recorrência de elementos e atributos, tais como cores e fontes.
Porque o modelo do documento pode ser reutilizada por vários painéis, faz
bom senso de controle theming a esse nível. (Usando documento modelos personalizados é o tema da próxima subseção).

Você pode usar dois diferentes dispositivos para controlar o layout:

- estrutura de documento usando elementos HTML específicos, eo uso de elemento de assentamento
- Cascading Style Sheets (CSS)

Uma discussão completa sobre esses métodos e as técnicas envolvidas seria
bem fora do escopo deste livro. Vamos resolver para o ponto de vista que tanto métodos são válidos e úteis, e muitas vezes o efeito desejado pode ser melhor alcançado através da combinação de ambos os métodos.

as tabelas HTML são ainda a maneira mais fácil de alcançar rapidamente um esquema robusto. A
layout simples que coloca o título do painel na parte superior do painel, e
Abaixo dele, o gráfico do lado esquerdo eo mapa na mão direita
lado, é mostrado na Listagem 17-16.

Listagem 17-16: Usando uma tabela HTML para colocar os componentes

```
<table>
  <- Linha 1:! Cabeçalho ->
  <tr>
    <-! Vãos de largura da célula da tabela ->
    colspan="100%"> <td
      Os clientes da <h1> Website: id="websiteName"> <span </ span> </ h1>
```

```
      <br/>
    </ Td>
  </ Tr>
  <! - Linha 2: componentes ->
  <tr>
    <-! Coluna esquerda ->
    <td>
      id="customersPerWebsite"> </ div>
    </ Td>
    <-! Coluna da direita ->
    <td>
      id="map" div style="width:400px; height:300px;"> </ div>
    </ Td>
  </ Tr>
</ Table>
```

Este método usa o tabela,tr (Para a linha da tabela) e td (Por dados da tabela) elementos para criar um layout tabular. Embora este método seja fácil de configurar e trabalha em uma maneira razoavelmente semelhantes entre os navegadores, é geralmente desaprovado
por desenvolvedores web moderna porque o layout é basicamente um efeito colateral da estrutura do documento (ou seja, os elementos). Isso torna mais difícil de entender documentos HTML, porque não é possível determinar se o documento
estrutura existe para seu próprio benefício (para processar dados tabulares) ou para alcançar um
layout efeito. Portanto, boas práticas de desenvolvimento exigem que os layouts devem ser controlada utilizando CSS ao invés de estrutura do documento.

Na Listagem 17-17, você encontrará o código HTML que permite atingir um layout similar ao que resulta o código na Listagem 17-16.

Listagem 17-17: Usando CSS para colocar os componentes

```
<style>
  # {CustomersPerWebsite
      position: absolute;
      top: 75px;
      left: 20px;
      width: 250px;
      height: 250px;
  }
  # {Mapa
      position: absolute;
      top: 75px;
      esquerda: 400px;
      width: 400px;
      height: 300px;
  }
</ Style>
Os clientes da <h1> Website: id="websiteName"/> <span </ span> </ h1>
id="customersPerWebsite"> </ div>
id="map"> </ div>
```

NOTA O layout da Listagem 17-17 não funciona particularmente bem com o padrão CDF modelos, uma vez que estes não antecipar o posicionamento absoluto de elementos. Na próxima seção, nós desenvolvemos um modelo de conteúdo que é mais adequado para este tipo de layout.

Criando um modelo de documento personalizado

Até agora, você tem usado o modelo do documento padrão para seus clientes painel de instrumentos. Isso significa que o painel sempre contou com o cabeçalho padrão com o logotipo da CDF, eo rodapé com o link para o site WebDetails.

Se você gosta, você pode usar o seu próprio modelo de documento para personalizar o aparência de seu painel adicional. As possibilidades para adicionar sua próprios fundos, layouts, e de navegação são bastante ilimitado. Um detalhado discussão sobre todas as coisas e técnicas que você poderia aplicar está além do escopo deste livro. No entanto, podemos, pelo menos, mostrar-lhe como criar os seus próprios modelo de documento.

Para adicionar o seu modelo de documento próprio, criar um novo arquivo chamado em modelo painel-wcm.html diretamente no diretório home do CDF. Lembre-se que o nome do arquivo é importante, mas deve começar com painel de modelo- e ter a . Html extensão. Adicione o conteúdo mostrado na Listagem 17-18 para o arquivo.

Listagem 17-18: O arquivo de modelo-dashboard-wcm.html

```
<DOCTYPE html PUBLIC! "- / / W3C / / DTD XHTML 1.0 Strict / / EN"
"Http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
xmlns="http://www.w3.org/1999/xhtml"> <html
 <head>
  <META HTTP-EQUIV="Content-Type" content="text/html; charset=utf-8" />
  <title> Comunidade Dashboard Framework para o Word Class Filmes </ title>
  <script language="javascript">
    isadmin = "{isadmin}";
    isLoggedIn = "{isLoggedIn}";
  </ Script>
 </ Head>
 <body>
 id="header"> </ div>
 id="content"> <div
  div id = "primaryContentContainer">
   id="primaryContent"> <div
   {Content}
   </ Div>
   </ Div>
  </ Div>
  id="footer"> </ div>
 </ Body>
</ Html>
```

Lembre-se, a {Content} espaço reservado será substituído pelo painel modelo de conteúdo durante a montagem do documento.

Você deve modificar o . Xcdf arquivo para usar o documento personalizado modelo. Basta adicionar uma linha como a seguinte, entre os cdf abertura e tags de fechamento:

```
<style> wcm </ style>
```

Observe que o valor, wcm, Corresponde diretamente ao wcm postfix no nome do arquivo do modelo do documento exemplo, modelo-painel-wcm.html. Consulte a Lista de 17-2 para um exemplo de uma . Xcdf arquivo que especifica um modelo de documento.

Resumo

Neste capítulo, abordamos a criação de dashboards baseados na Comunidade Dashboard Framework. Neste capítulo, você aprendeu o seguinte:

- O CDF é criado e mantido pela comunidade Pentaho.
- A Pentaho Corporation inclui o CDF no servidor.
- Os painéis são, na verdade páginas web baseadas em tecnologias como o HTML, CSS e JavaScript, e usar AJAX construída sobre JQuery.
- pedidos painel CDF são manipulados por um plugin que reúne os painel de instrumentos a partir de um modelo de documento e um modelo de conteúdo.
- A . Xcdf arquivo contém as informações sobre o documento e conteúdo modelo a ser usado.
- modelos de documentos e conteúdos são na verdade arquivos HTML.
- Dashboard componentes são adicionados com JavaScript e muitas vezes exigem uma espaço reservado HTML para exibir seu conteúdo.
- componentes Dashboard pode ouvir as mudanças nos parâmetros do painel, que atua como uma sugestão para automaticamente atualizar-los.
- Os componentes podem sinalizar uma mudança de parâmetro com o fireChange método.
- componentes Dashboard pode confiar em seqüências de ação de entregar o real BI conteúdo.
- A XactionComponent pode ser usado para exibir o conteúdo de um arbitrário seqüência de ação.
- O texto dinâmico pode ser implementado usando a TextComponent.

Também discutimos o seguinte:

- Como capturar cliques do mouse em um JFreeChart
- Como usar o MapComponent para mostrar a distribuição geográfica da data
- Como mostrar diferentes marcadores em um MapComponent dependendo da valor das métricas
- Como o seu estilo de painéis
- Como criar e utilizar um modelo de documento personalizado

Índice

B

C

D

E

F

G

H

J

K

L

M

N

O

Q

S

T

U

V

W

X

Y

Z

Seu recurso one-stop para o código aberto BI e soluções de data warehousing

Pentaho é uma full-featured, Open Source Business Intelligence Suite, que permite que você crie dados depósitos e aplicações ricas e poderosas de BI em uma fração do custo de uma solução proprietária.

Este livro leva você para cima e correndo com Pentaho em poucos minutos: desde o começo você vai ser

relatórios de exemplo em execução, painéis e tabelas dinâmicas de OLAP, enquanto você aprende sobre Pentaho conceitos e arquitetura. Usando um estudo de caso prático, você vai aprender o que a modelagem dimensional

é e como aplicá-lo para projetar um data warehouse. Você vai criar e preencher seus dados

armazém com ferramentas de integração de dados Pentaho. Finalmente, você vai aprender como construir seu próprio BI aplicações em cima de seu data warehouse com Pentaho relatórios, análise, dashboards,

e ferramentas de mineração de dados.

- Compreender conceitos importantes Pentaho, incluindo seqüências de ação ea solução repositório
- Aplicar os conceitos-chave da dimensão modelagem e construção de um data warehouse usando esquemas estrela
- Use Pentaho ferramentas de integração de dados para construir aplicações ETL
- Explore avançada PDI recursos, incluindo execução remota e clustering
- Design e relatórios e gráficos utilizando implantar Pentaho Report Designer
- Aproveite OLAP e criar interativo tabelas dinâmicas com drill up / drill down utilizando Pentaho Analysis Services
- Concentração e conteúdo de BI compacto para usuários corporativos com ampla dashboards
- Descobrir e explorar padrões em seus dados Pentaho usando mineração de dados

Roland Bouman é um desenvolvedor de aplicações com foco em tecnologia open source da Web, bancos de dados e Business Inteligência. Ele é membro ativo das comunidades MySQL e Pentaho, e você pode acompanhar seu blog em http://rpbouman.blogspot.com/.

Jos van Dongen é um autor experiente profissional de Business Intelligence e bem conhecido e apresentador. Ele fala regularmente em conferências e seminários. Você pode encontrar mais informações sobre a Jos http://www.tholis.com.

Visite www.wiley.com / go / pentahosolutions para exemplos de código e dados da amostra.

ISBN: 978-0-470-48432-6

WILEY
wiley.com

$ 50,00 EUA / CAN $ 60,00

Visite nosso site em www.wiley.com / compbooks

Database Management / Geral