Integração de Dados, Web e Warehousing

Introdução

Arquiteturas de Integração

Fernando Fonseca
Ana Carolina

Conteúdo

- Introdução
- Integração de Informações
- Consultando a Web

Introdução

- Motivação
- Web e BD
- Arquitetura na Web

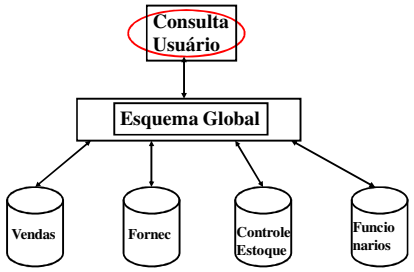
Introdução

- Evolução da Tecnologia de SGBD
- Necessidade por informações mais completas e precisas
- Heterogeneidade
  - Hardware
  - Modelos de Dados
  - Linguagens de Consulta

Introdução

- Manipulação ¨Simultânea¨ de dados heterogêneos
  - Multi-Banco de Dados (MBD)
  - Bancos de Dados Federados
- Processamento de operações que acessam diferentes BD (Operações Globais)

BD Heterogêneos



Introdução

- SGBD
  - Avanços na tecnologia ⇒
    Sistemas economicamente viáveis
    - Uso em grande escala
    - Dados armazenados não em função de uma aplicação (ou mais), porém de modo a representar de forma a mais completa possível um determinado universo

Introdução

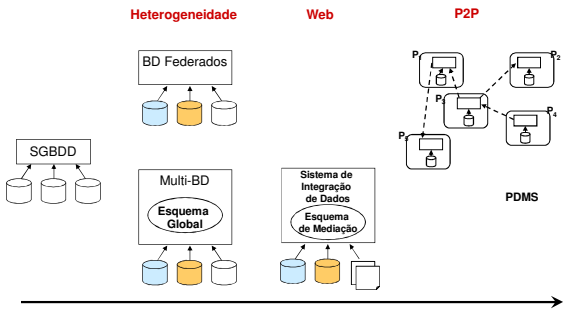
- SGBD (Cont.)
  - Diversos BD armazenando os dados de segmentos distintos e, aparentemente, totalmente independentes da mesma organização
  - Evolução das necessidades por informação mais completa
  - Aplicações precisam consultar diversos BD distintos de forma transparente

Introdução

- SGBD (Cont.)
  - Não é exigido do usuário o conhecimento dos diversos SGBD utilizados
  - Manutenção da consistência volta a ser feita pelos aplicativos
  - Internet
- Interoperabilidade X Integração
  - Interoperabilidade: Sistemas (SGBD)
  - Integração: Dados e Informações

Evolução dos Sistemas de Gerenciamento de Dados Distribuídos

Evolução dos Sistemas de Gerenciamento de Dados Distribuídos



Evolução dos BD Distribuídos

- BD Distribuído
  É uma coleção de múltiplos BD **logicamente inter-relacionados** distribuídos através de uma rede de computadores
- Sistema de BD Distribuídos
  É definido como um sistema que permite o gerenciamento de BD distribuídos deixando a distribuição dos dados **transparente ao usuário**

Evolução dos BD Distribuídos

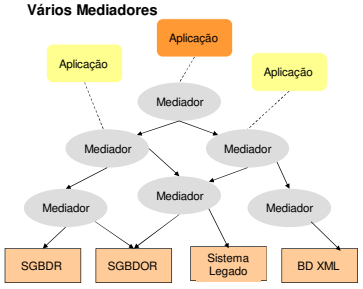
- Bancos de Dados Distribuídos
  - Dados homogêneos
  - Acesso e atualização
- Multi-Bancos de Dados
  - Dados heterogêneos
  - Acesso e atualização usando um esquema global
- Bancos de Dados Federados
  - Dados heterogêneos
  - Importação e exportação de esquemas (não há esquema global)

Evolução dos BD Distribuídos

- Sistemas de Integração de Dados
  - Tipo particular de *middleware*
  - Integração de fontes de dados **distribuídas, autônomas e heterogêneas**
  - **Visão lógica unificada** de dados (esquema global)
    - Evita lidar com as inúmeras fontes, interfaces e representações dos dados diferentes
  - Apenas consultas aos dados
  - Abordagens para integração de dados: **virtual e materializada**
    - Dados atuais x Tempo de resposta

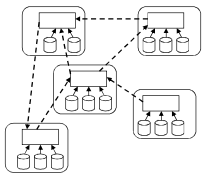
Evolução dos BD Distribuídos

Vários Mediadores



[Wiederhold 92]

PDMS

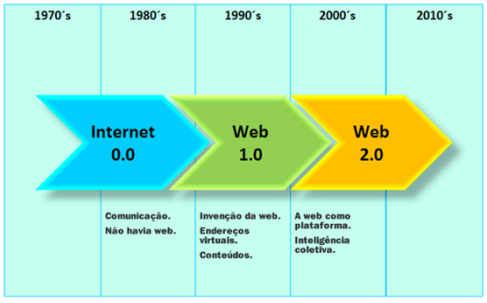


Evolução dos BD Distribuídos

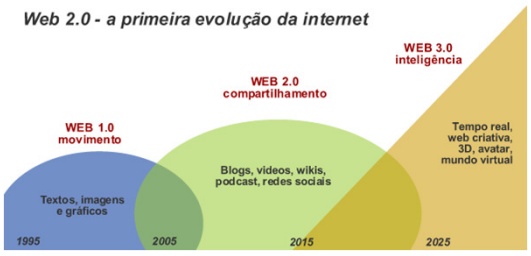
- Sistemas de Gerenciamento de Dados *Peer-to-Peer* (PDMS)
  - Compartilhamento de dados descentralizado
  - Processamento e armazenamento de dados distribuídos em pontos autônomos
  - Comportamento dinâmico dos pontos: podem entrar e sair do sistema a qualquer momento
  - Escalabilidade sem servidores poderosos
  - Mapeamentos semânticos dos dados armazenados nos pontos

Evolução da Web

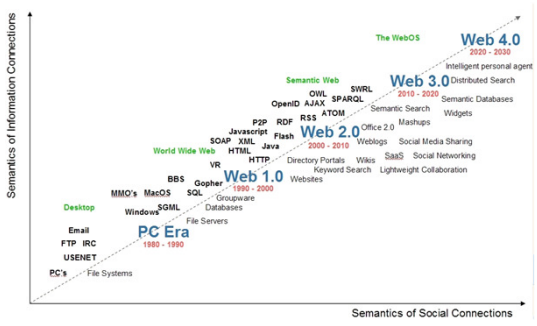
Evolução da Web



Evolução da Web



Evolução da Web



Integração de Dados na Web

Motivação

- Web: enorme banco de dados
- Documentos são gerados para serem disponibilizados para leitura
- Alguns destes documentos foram gerados a partir de consultas a BD
- Dados podem ser extraídos das páginas web para serem utilizados por outros programas

Web X Banco de Dados

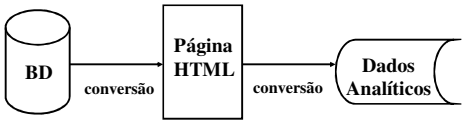
Web

- Padrão simples e universal para troca de informações
- Informações decompostas como unidades que possam ter nome (URL) e ser transmitidas (HTTP)
- Estrutura da informação (HTML)

Banco de Dados

- Esquemas (relacional) e diagramas (E-R) para descrever a estrutura
- Linguagem de consulta, controle de concorrência, recuperação e integridade
- Separa a visão lógica da implementação física

Conversão de Dados



- **Fragilidade (mudança de formato dos dados fonte)**
- **Pode ser necessário todo o BD para análise dos dados**
  - **Proposta de solução: XML (eXtensible Markup Language)**

WEB

- A Web proporcionou
  - Uma infraestrutura global e um conjunto de padrões para dar suporte à troca de documentos
  - Um formato de apresentação para hipertexto (HTML)
  - Interfaces com o usuário bem construídas para recuperação de documentos
  - Um novo formato, XML, para troca de dados com estrutura

Tecnologia de Banco de Dados

- A tecnologia de BD oferece
  - Técnicas de armazenamento e linguagens de consulta a dados altamente estruturados
  - Modelos e métodos para estruturar dados
  - Mecanismos para a manutenção da integridade e consistência de dados
  - Um novo modelo, de dados semiestruturados, que abranda os rigores dos sistemas de BD altamente estruturados

Web X Banco de Dados

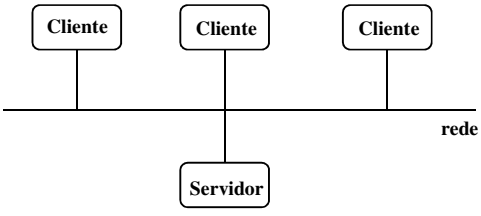
- Onde a Web é diferente de BD
  - Não tem estrutura uniforme
  - Não tem restrição de integridade
  - Não tem transações
  - Não tem uma linguagem de consulta ou um modelo de dados padrão

  ⇒O poder de abstração desenvolvido pela comunidade de banco de dados pode prover a chave para diminuir a complexidade da Web

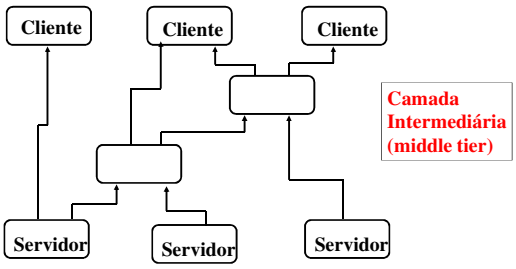
Web X Banco de Dados

- Onde BD pode beneficiar a Web
  - XML
  - Metadados sobre as fontes Web
  - Combinação de consultas a dados estruturados e semiestruturados

Arquitetura Tradicional Cliente/Servidor



Arquitetura de aplicação baseada na Web



Middleware

- Camada Intermediária (*middle tier*)
  - Os sistemas que a gerenciam: *Middleware*
- Duas abordagens
  - *Data Warehouse*: dados importados de fontes de dados diversas e armazenados em um BD intermediário especialmente construído
  - Sistema Mediador: consultas do cliente são transformadas e decompostas diretamente em consultas junto à fonte de dados

Sistema Virtual de Integração de Dados

- Distinção do Sistema Mediador para um sistema tradicional
  - O sistema não se comunica diretamente com a fonte de dados local. Esta tarefa é feita por um programa (*wrapper* ou tradutor), específico de cada site, que traduz os dados locais em uma forma que pode ser processada pelo sistema de integração de dados
    - O que o tradutor provê depende do que está disponível na fonte de dados: seu esquema exportado

Sistema Virtual de Integração de Dados

- Distinção do Sistema Mediador para um sistema tradicional (Cont.)
  - O usuário não faz consultas diretamente sobre o esquema no qual os dados estão armazenados mas sobre um Esquema do Mediador que contém relações virtuais relativas a uma dada aplicação de integração
    - O sistema de integração deve reformular a consulta do usuário em consultas que se referem diretamente aos esquemas nas fontes de dados

Integração de Informação

- Integração de informação
  - **Sistemas**
  - **Arquiteturas**
- Construindo Sistemas de Integração de Dados

Integração de Informação

- Problema: A Web contém um número crescente de fontes de informação que podem ser vistas como um grande repositório de dados
- Tarefa: Responder consultas que podem requerer extração e combinação de dados de várias fontes de dados na WEB, ou Fontes Web

Integração de Informação

- Exemplo: Considere um Banco de Dados de Cinema na Internet contendo dados sobre elenco, gênero e diretores. Informações sobre críticas de filmes podem ser encontradas em várias outras fontes web (ex: revistas) e muitas outras fontes provêm horários dos cinemas. Combinando dados de todas estas fontes, podemos responder a
  - **"Quais os filmes com Julia Roberts, suas críticas e horários de exibição, em cartaz hoje à noite em Paris"**

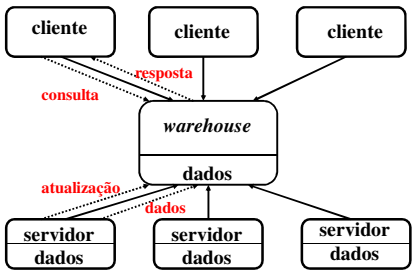
Sistemas para Integração de Informação

- Vários sistemas já foram construídos
- Problemas similares aos apontados para integração de sistemas de banco de dados heterogêneos
- E mais:
  - Fontes Web **variadas e evolutivas**
  - Poucos **metadados** sobre características das fontes
  - Alto grau de **autonomia** das fontes Web

Sistemas para Integração de Informação

- Duas abordagens
  - *Warehousing*: os dados gerados das diversas fontes Web são carregados (materializados) em um repositório (*warehouse*) e as consultas são aplicadas a estes dados
    - Vantagem: desempenho garantido no momento da consulta
    - Desvantagem: atualização do repositório sempre que houver mudança nos dados

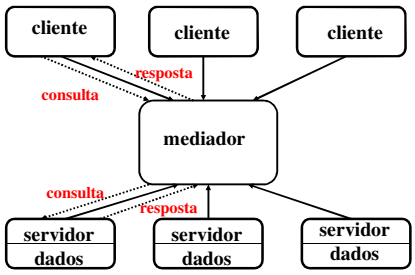
Arquitetura com *Data Warehouse*



Sistemas para Integração de Informação

- Duas abordagens (Cont.)
  - Enfoque Virtual: os dados são mantidos nas fontes Web e as consultas são decompostas em tempo real e submetidas às diversas fontes
    - Vantagem: os dados não são replicados e tem-se a garantia de estarem atualizados no momento da consulta
    - Desvantagem: como as fontes de dados são autônomas, são necessários métodos para otimização de consultas para garantir um desempenho adequado

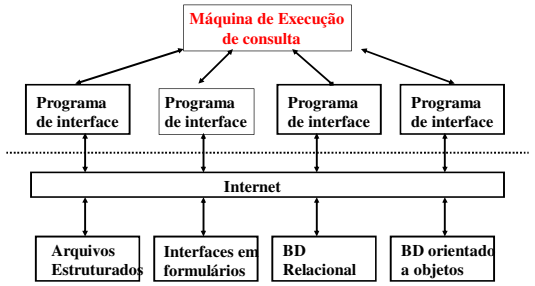
Arquitetura com Mediador



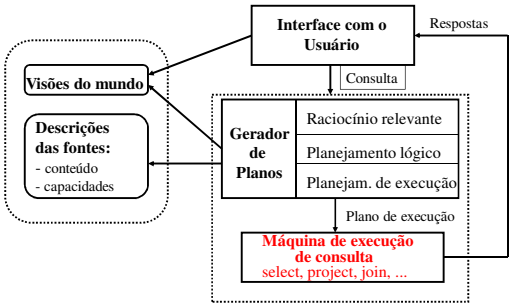
Sistemas para Integração de Informação

- O Enfoque Virtual é mais apropriado pelas seguintes razões
  - O número de fontes Web é grande
  - Os dados são atualizados frequentemente
  - Existe pouco controle sobre as fontes de dados

Arquitetura de um Sistema de Integração de Dados



Arquitetura de um Sistema de Integração de Dados (cont.)



Construindo Sistemas de Integração de Dados na Web

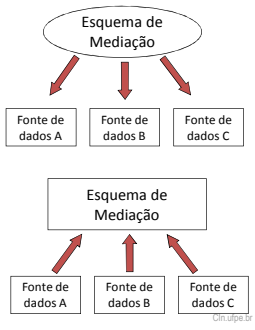
- Especificação do esquema de mediadores
  - O esquema é o conjunto de nomes de atributos e coleções usados para formular consultas
  - Para avaliar uma consulta, o sistema tem que traduzi-la em consultas aos esquemas locais das fontes de dados
  - Problemas
    - Como especificar as descrições das fontes de dados
    - Como usá-las na reformulação de consultas

Construindo Sistemas de Integração de Dados na Web

- Especificação do esquema de mediadores (Cont.)
  - Para especificar a descrição das fontes de dados
    - Visão Global: cada entidade no esquema do mediador tem uma correspondência com o esquema da fonte de dado. Neste caso a reformulação de consultas torna-se mais simples.
    - Visão Local: cada informação na fonte tem uma correspondência com uma entidade no esquema do mediador. É mais fácil a manutenção das fontes de dados.

Mapeamentos entre Esquemas

- Mapeamentos em Sistemas de Integração de Dados
  - Global-as-view
    - Relações no esquema global são descritas como visões sobre a coleção de relações locais
  - Local-as-view
    - Cada relação em uma visão local é descrita como visão sobre o esquema de mediação



Construindo Sistemas de Integração de Dados na Web

- Completude dos dados nas fontes Web
  - As fontes Web, em geral, não são completas com relação ao domínio associado
  - Uma informação negativa pode ser útil ao sistema de integração que redirecionará o acesso às outras fontes
  - Como determinar se a fonte é completa
    - Formalismos probabilísticos para descrever o conteúdo e sobreposições entre as fontes
    - Algoritmos para escolha "ótima" entre as diversas fontes de informação

Construindo Sistemas de Integração de Dados na Web

- Divergência na capacidade de processamento de consultas
  - Duas razões para a divergência
    - Os dados podem estar armazenados em arquivos estruturados ou sistemas legados com interface de acesso limitada
    - Mesmo que os dados estejam armazenados em BD, a fonte de dados pode prover acesso limitado por razões de segurança ou desempenho

Construindo Sistemas de Integração de Dados na Web

- Divergência na capacidade de processamento de consultas (Cont.)
  - Para construir um sistema de integração de dados efetivo, estas capacidades devem estar explicitamente descritas, combinadas e exploradas para garantir desempenho
  - Dois tipos de capacidade
    - Negativa: limita o acesso aos dados
    - Positiva: executa operações algébricas adicionais, facilitando o acesso

Construindo Sistemas de Integração de Dados na Web

- Otimização de Consultas
  - Seleção de um conjunto mínimo de fontes de dados a consultar e determinação da consulta mínima a ser enviada para cada fonte
  - Poucas estatísticas sobre os dados nas fontes e, portanto, pouca informação para avaliar o custo de execução de consultas
  - Fusão de consultas, permitindo acessar vários atributos de um dado objeto em fontes diversas

Construindo Sistemas de Integração de Dados na Web

- Máquinas para execução de consultas
  - Construir um mecanismo de execução de consultas específico para integração de dados na Web (decomposição/reformulação)
  - Os desafios são a autonomia das fontes de dados e a imprevisibilidade do desempenho da rede

Construindo Sistemas de Integração de Dados na Web

- Construção de Tradutores para páginas HTML
  - Dificuldade: as páginas HTML foram criadas para serem lidas e não para se extrair dados, além de mudarem frequentemente
  - Ferramentas baseadas em
    - Gramáticas especializadas em como os dados aparecem em páginas HTML
    - Técnicas de aprendizado indutivo para automaticamente manter o tradutor

Construindo Sistemas de Integração de Dados na Web

- Comparando objetos nas fontes Web (*semântica!*)
  - Um dos problemas mais difíceis é decidir que dois objetos de duas fontes distintas se referem à mesma entidade do mundo real
  - Cada fonte usa sua própria convenção para nomear objetos
  - Alguns sistemas resolvem usando
    - Heurísticas específicas do domínio
    - Técnicas de recuperação de informação (IR)

Considerações

- Integração e navegação
- Representação de Dados para Web/DB

WWW

- Por que desejamos tratar a Web como um BD?
  - Para manter integridade
  - Para fazer consultas de acordo com alguma estrutura (oposto a consultas baseadas em conteúdo)
  - Para introduzir alguma organização
- A Web não tem estrutura. O melhor que podemos afirmar é que a Web é um enorme grafo

Formato de Dados

- A maioria dos dados do mundo real está representada em algum formato de dados
- Os formatos de dados são definidos para a troca e armazenamento de dados
- Formatos científicos tendem a ter “esquemas fixos”
- A representação textual oferecida algumas vezes pode não ser imediatamente traduzida para uma representação padrão relacional /objeto-relacional

Integração de Dados

- O objetivo é integrar todos os tipos de informação, incluindo informação não estruturada
  - Informação irregular ou ausente
  - Informação com estrutura não conhecida completamente
  - Esquemas que evoluem dinamicamente

Integração de Dados

- Modelos de dados tradicionais e linguagens inadequadas
  - Incapazes de acomodar conjuntos de dados heterogêneos (diferentes tipos e estruturas)
  - Dificuldade em desenvolver mecanismos capazes de fazer conversões entre dois modelos de dados distintos

Navegação

- Para consultar um BD é preciso entender seu esquema
- Esquemas não têm uma terminologia muito clara.
- Na Web, o usuário pode querer consultar os dados com pouco ou nenhum conhecimento do esquema
  - Existem números inteiros no BD maiores do que 2^{16}?
  - Quais objetos no BD têm um nome de atributo que começa com “at”?
- Extensões de linguagens de consulta relacionais têm sido propostas, mas não existe uma técnica genérica para interpretá-las

Modelo de dados

- Representação dos dados com algum tipo de modelo baseado em grafo ou em árvore
  - Ciclos são permitidos, mas geralmente os referenciamos como árvores
  - Diversas abordagens com pequenas diferenças (fácil conversão): dados nos *labels* ou arcos, nós que carregam informação ou não
- Codificação direta para BD relacional ou objeto-relacional
  - Identidade de objetos

Representação de Dados para Web/BD

- Necessidades de modelar
  - A própria Web
  - A estrutura de Web sites
  - A estrutura interna de páginas da Web
  - O conteúdo do Web site em menor granularidade

Modelos de Dados para Web/BD

- Grafos rotulados: a maneira natural de representar páginas (nós) e *links* entre elas (arcos). Os rótulos representam os nomes de atributos
- Dados semiestruturados: baseados no modelo de grafos rotulados. Neste modelo não existem restrições sobre o conjunto de arcos que partem de um dado nó ou sobre os tipos dos valores dos atributos
- Outras características: presença de construtores específicos da Web na representação de dados