**Desafios e Oportunidades com Big Data**

*Um documento comunitário desenvolvido por pesquisadores líderes nos Estados Unidos*

Sumário Executivo

A promessa de tomada de decisão dirigida por dados agora está sendo reconhecida amplamente, e há cada vez mais entusiasmo para a noção de “Big Data”. Embora a promessa de Big Data seja real - Por exemplo, estima-se que o Google sozinho contribuiu com 54 bilhões de dólares para a economia dos EUA em 2009 - não há atualmente uma grande diferença entre o seu potencial e sua realização.

Problemas com o Big Data de heterogeneidade, escala, pontualidade, complexidade e privacidade, impedem o progresso em todas as fases do pipeline que pode criar valor a partir de dados. Os problemas começam logo durante a aquisição de dados, quando o tsunami de dados nos obriga a tomar decisões, atualmente de forma ad hoc, sobre quais dados manter e quais descartar, e como armazenar o que mantemos de forma confiável com os metadados corretos. Muitos dados de hoje não estão nativamente num formato estruturado, por exemplo, tweets e blogs são partes fracamente estruturadas de texto, imagens e vídeo enquanto são estruturados para armazenamento e exibição, mas não por conteúdo semântico e pesquisa: transformar esse conteúdo em um formato estruturado para análise posterior é um grande desafio. O valor dos dados explode quando ele pode ser ligado com outros dados, assim a integração de dados é um criador de grande valor. Como a maioria dos dados é gerado diretamente em formato digital de hoje, temos a oportunidade e o desafio tanto para influenciar a criação de facilitar a ligação e depois vincular automaticamente os dados criados anteriormente. A análise dos dados, organização, recuperação e modelagem são outros desafios fundamentais. A análise dos dados é um gargalo claro em muitas aplicações, tanto devido à falta de escalabilidade dos algoritmos subjacentes e, devido à complexidade dos dados que necessitam de ser analisados. Finalmente, a apresentação dos resultados e sua interpretação por especialistas de domínio não técnicos é fundamental para extrair conhecimento acionável.

Durante os últimos 35 anos, os princípios de gerenciamento de dados, tais como a independência física e lógica, consultas declarativas e otimização baseada em custo tem levado, durante os últimos 35 anos, a uma indústria multibilionária. Mais importante ainda, esses avanços técnicos permitiram a primeira rodada de aplicativos de inteligência de negócios e lançou as bases para a gestão e análise de Big Data hoje. Os novos desafios e muitas oportunidades associadas com Big Data necessitam um repensamento em muitos aspectos destas plataformas de gerenciamento de dados, mantendo outros aspectos desejáveis. Acreditamos que o investimento apropriado em Big Data levará a uma nova onda de fundamentais avanços tecnológicos que serão incorporados nas próximas gerações de gestão de Big Data e plataformas de análise, produtos e sistemas.

Acreditamos que esses problemas de pesquisa não são apenas oportuno, mas também tem o potencial de criar valor econômico enorme na economia dos EUA para os próximos anos. No entanto, eles também são difíceis, obrigando-nos a repensar os sistemas de análise de dados de forma fundamental. Um grande investimento em Big Data, dirigido corretamente, pode resultar não só em grandes avanços científicos, mas também lançar as bases para a próxima geração de avanços na ciência, medicina, e de negócios.

**Desafios e Oportunidades com Big Data**

**1. Introdução**

Estamos nadando em uma inundação de dados de hoje. Em uma ampla gama de áreas de aplicação, os dados estão sendo coletados em escala sem precedentes. Decisões que anteriormente eram baseadas em suposições, ou em modelos meticulosamente construídos da realidade, agora podem ser feita com base nos dados em si. Tal análise de dados Big agora dirige quase todos os aspectos da nossa sociedade moderna, incluindo serviços móveis, fabricação, varejo, serviços financeiros, ciências da vida e ciências físicas.

A pesquisa científica foi revolucionada pelo Big Data [CCC2011a]. O Sloan Digital Sky Survey [SDSS2008] tornou-se hoje um recurso central para os astrônomos de todo o mundo. O campo da Astronomia está sendo transformado de um onde tirar fotos do céu era uma grande parte do trabalho de um astrônomo para um onde as imagens são tudo em um banco de dados já e a tarefa do astrônomo é encontrar objetos interessantes e fenômenos no banco de dados. Nas ciências biológicas, há agora uma tradição bem estabelecida de depósito de dados científicos em um repositório público, e também da criação de bases de dados públicas para uso por outros cientistas. Na verdade, há uma disciplina inteira de bioinformática, que é em grande parte dedicado à curadoria e análise desses dados. Com o avanço da tecnologia, principalmente com o advento do Sequenciamento da Próxima Geração, o tamanho e o número de conjunto de dados experimentais disponíveis está a aumentar exponencialmente.

Big Data tem o potencial de revolucionar não só a pesquisa, mas também a educação [CCC2011b]. Uma comparação detalhada recente quantitativa de diferentes abordagens adotadas por 35 cartas escolares em Nova York descobriu que uma das cinco melhores políticas correlacionadas com eficácia mensurável acadêmico foi o uso de dados para orientar a instrução [DF2011]. Imagine um mundo em que temos acesso a uma enorme base de dados onde coletamos todas as medidas detalhadas do desempenho acadêmico de cada estudante. Estes dados podem ser usados ​​para projetar as abordagens mais eficazes para a educação, a partir de leitura, escrita e matemática, ao avançado, de nível universitário, cursos. Estamos longe de ter acesso a esses dados, mas há tendências poderosas neste sentido. Em particular, há uma forte tendência para a implantação da Web maciça de atividades educacionais, e isso vai gerar uma quantidade cada vez maior de dados detalhados sobre o desempenho dos alunos.

Acredita-se que o uso da tecnologia da informação pode reduzir o custo dos cuidados de saúde, melhorando sua qualidade [CCC2011c], fazendo dos cuidados preventivos e personalizado, baseando-se na mais extensa (baseado em casa) monitoramento contínuo. Estimativas da McKinsey [McK2011] uma economia de 300 bilhões de dólares a cada ano só nos EUA.

Na mesma linha, já houve casos de persuasão feitas para o valor de Big Data para o planejamento urbano (através da fusão de alta fidelidade de dados geográficos), transporte inteligente (através de análise e visualização de dados detalhados ao vivo de rede de estradas), modelagem ambiental (através de redes de sensores coletando dados ubiquamente) [CCC2011d], de poupança de energia (através de padrões de inauguração de uso), materiais inteligentes (através de novas iniciativas de materiais genoma [MGI2011]), ciências sociais computacionais (Um jejum nova metodologia crescendo em popularidade por causa do custo dramaticamente reduzido de obtenção de dados) [LP 2009], análise de risco financeiro sistêmico (através de análise integrada de uma rede de contratos de encontrar dependências entre entidades financeiras) [FJ 2011], Segurança Interna (através da análise de redes sociais e transações financeiras de possíveis terroristas), a segurança do computador (através de análise de informações registradas e outros eventos, conhecidos como Segurança da Informação e Gestão de Eventos (SIEM)), e assim por diante.

Em 2010, empresas e usuários armazenaram mais de 13 exabytes de novos dados, o que é mais de 50.000 vezes os dados na Biblioteca do Congresso. O valor potencial global de dados de localização pessoal é estimada em US $ 700 bilhões para os usuários finais, e isso pode resultar em uma redução de até 50% no desenvolvimento de produtos e os custos de montagem, de acordo com um relatório recente da McKinsey [McK2011]. McKinsey prevê um efeito igualmente grande de Big Data no emprego, onde 140,000-190,000 trabalhadores com "profunda análise" experiência será necessária nos EUA, além disso, 1,5 milhão de gerentes terão de se tornar dados alfabetizados. Não surpreendentemente, o relatório PCAST ​​recente em Redes e TI I & D [PCAST2010] identificaram Big Data como uma "investigação de fronteira", que pode "acelerar os progressos em uma ampla gama de prioridades." Mesmo mídia popular agora aprecia o valor de Big Data como evidenciado pela cobertura na Economist [Eco2011], o New York Times [NYT2012], e da National Public Radio [NPR2011a, NPR2011b].

Embora os benefícios potenciais de Big Data são reais e significativos, e alguns sucessos iniciais já foram alcançados (como o Sloan Digital Sky Survey), ainda há muitos desafios técnicos que devem ser abordadas para realizar plenamente o seu potencial. O tamanho dos dados, é claro, é um grande desafio, e é aquele que é mais facilmente reconhecida. No entanto, existem outros. Empresas de análise da indústria gostariam de salientar que há desafios não apenas em volume, mas também em variedade e velocidade [Gar2011], e que as empresas não devem focar apenas a primeira delas. Pela Variedade, que geralmente significa heterogeneidade de tipos de dados, representação e interpretação semântica. Pela velocidade, eles significam tanto a taxa em que os dados chegam e o tempo em que deve ser colocada em prática. Enquanto estes três são importantes, esta pequena lista não inclui requisitos adicionais importantes como a privacidade e usabilidade.

A análise do Big Data envolve várias fases distintas, como mostrado na figura a seguir, cada uma das quais apresenta desafios. Muitas pessoas, infelizmente, se concentrar apenas na fase de análise / modelagem: enquanto que a fase é fundamental, é de pouco uso sem as outras fases do pipeline de análise de dados. Mesmo na fase de análise, o que tem recebido muita atenção, as complexidades são mal compreendidas no contexto de clusters de multi locação em que os programas de vários usuários são executados simultaneamente. Muitos desafios significativos estendem além da fase de análise. Por exemplo, Big Data tem de ser gerido no contexto, o que pode ser ruidosa, heterogêneo e não incluir um modelo inicial. A decisão levanta a necessidade de rastrear procedência e para lidar com a incerteza e erro: tópicos que são cruciais para o sucesso, e ainda assim raramente mencionado no mesmo fôlego como Big Data. Da mesma forma, as perguntas para a análise do pipeline de dados não serão normalmente todas definidas com antecedência. Podemos precisar descobrir boas perguntas com base nos dados. Fazendo isso vai exigir mais sistemas inteligentes e também um melhor suporte para interação do usuário com a análise do pipeline. De fato, atualmente temos um gargalo importante no número de pessoas habilitadas a fazer perguntas dos dados e analisá-lo [NYT2012]. Nós podemos aumentar drasticamente este número, apoiando vários níveis de envolvimento com os dados, não exigindo um profundo conhecimento de banco de dados. Soluções para problemas como este não virá de melhorias incrementais para os negócios como de costume, como a indústria pode fazer por conta própria. Ao contrário, eles nos obrigam a repensar fundamentalmente como gerenciar a análise de dados.



Felizmente, as técnicas atuais computacionais podem ser aplicadas, quer tal e qual ou com algumas extensões, para pelo menos alguns aspectos do problema do Big Data. Por exemplo, bancos de dados relacionais dependem da noção de independência de dados lógica: os usuários podem pensar sobre o que eles querem para calcular, enquanto o sistema (com engenheiros qualificados projetar esses sistemas) determina como calcular de forma eficiente. Da mesma forma, o padrão SQL e modelo de dados relacional de uma linguagem uniforme e poderosa para expressar as necessidades de muitas consultas e, em princípio, permite aos clientes escolher entre os fornecedores, aumentando a concorrência. O desafio diante de nós é a de combinar essas características saudáveis ​​de sistemas anteriores como conceber novas soluções para os muitos desafios novos da Big Data.

Neste trabalho, consideramos cada uma das caixas na figura acima, e discutir tanto o que já foi feito e que desafios permanecem a medida que nós exploráramos o Big Data. Começamos por considerar os cinco estágios no pipeline, em seguida, passar para os cinco desafios horizontais, e termina com uma discussão sobre a arquitetura de todo o sistema que combina todas estas funções.

**2. Fases do processamento do pipeline**

**2.1 Aquisição de Dados e Gravação**

Big Data não surge de um vazio: é gravado a partir de uma fonte de geração de dados. Por exemplo, considere a nossa capacidade de sentir e observar o mundo ao nosso redor, a partir da frequência cardíaca de um cidadão idoso, e presença de toxinas no ar que respiramos, ao telescópio matriz planejada quilômetros quadrados, que irá produzir até 1 milhão de terabytes de dados brutos por dia. Da mesma forma, as experiências científicas e simulações pode facilmente produzir petabytes de dados hoje.

Muitos desses dados não é de interesse, e pode ser filtrado e comprimido por ordens de magnitude. Um desafio é definir esses filtros de tal forma que eles não descartem informação útil. Por exemplo, suponha que uma leitura do sensor difere substancialmente do resto: é provável que seja devido ao sensor falho, mas como podemos ter certeza de que não é um artefato que merece atenção? Além disso, os dados recolhidos por estes sensores na maioria das vezes são espacialmente e temporalmente correlacionada (por exemplo, sensores de tráfego no segmento mesma estrada). Precisamos de pesquisas na ciência da redução de dados que pode inteligentemente processar esses dados brutos para um tamanho que seus usuários podem lidar ao mesmo tempo não perder a agulha no palheiro. Além disso, são necessárias "on-line" de técnicas de análise que podem processar esses dados de streaming em tempo real, já que não podemos dar ao luxo de armazenar primeiro e reduzir depois.

O segundo grande desafio é gerar automaticamente o metadados correto para descrever que dados é gravado e como é registrada sua medida. Por exemplo, em experimentos científicos, detalhes consideráveis ​​sobre as condições específicas e os procedimentos experimentais podem ser necessários para ser capaz de interpretar corretamente os resultados, e é importante que tais metadados sejam gravados com um dado observacional. Sistemas de aquisição de metadados podem minimizar a responsabilidade humana em gravar metadados. Outra questão importante aqui é proveniência de dados. Gravar informações sobre os dados de seu nascimento não é útil a menos que esta informação possa ser interpretada e levada junto com o pipeline de análise de dados. Por exemplo, um erro de processamento num passo pode tornar inútil a análise posterior; com proveniência adequada, pode-se facilmente identificar todo o processamento subsequente que dependem deste passo. Assim, precisamos pesquisar, tanto na geração de metadados adequados e em sistemas de dados que carregam a proveniência de dados e seus metadados através da análise do pipeline de dados.

**2.2 Extração de Informações e Limpeza**

Frequentemente, a informação recolhida não estará num formato pronto para análise. Por exemplo, considere o conjunto de registros eletrônicos de saúde em um hospital, compreendendo ditados transcritos de vários médicos, dados estruturados a partir de sensores e medições (possivelmente com alguma incerteza associada), e dados de imagem, como raios-x. Não podemos deixar os dados desta forma e ainda de forma eficaz analisá-lo. Em vez disso, exigem um processo de extração de informações que puxa as informações necessárias a partir das fontes subjacentes e expressa de uma forma estruturada adequado para análise. Fazendo isso corretamente e completamente é um desafio constante técnico. Note-se que estes dados também inclui imagens e, no futuro, incluir vídeo, extração, muitas vezes é altamente dependente da aplicação (por exemplo, o que você quer sair de uma ressonância magnética é muito diferente do que você poderia sair de uma imagem das estrelas, ou uma foto de vigilância). Além disso, devido à onipresença de câmeras de vigilância e popularidade de GPS habilitados para telefones celulares, câmeras e outros dispositivos portáteis, rico e de alta fidelidade localização e trajetória (isto é, o movimento no espaço) de dados também pode ser extraído.

Estamos acostumados a pensar que o Big Data sempre irá nos dizer a verdade, mas isto está longe da realidade. Por exemplo, os pacientes podem optar por esconder comportamentos de risco e médicos podem, por vezes, mal diagnosticar uma doença, os pacientes podem também imprecisamente lembrar o nome de uma droga ou mesmo que nunca tomou, levando a informação em falta no (a parte história de) sua registro médico. Trabalho existente na limpeza de dados pressupõe restrições bem reconhecidas em dados válidos ou modelos bem compreendidos erro, pois muitos domínios emergentes Big Data estes não existem.

**2.3 Integração de dados, agregação e Representação**

Dada à heterogeneidade do dilúvio de dados, não é suficiente apenas para gravá-lo e jogá-lo em um repositório. Considere, por exemplo, dados de uma série de experimentos científicos. Se só temos um amontoado de conjuntos de dados num repositório, é improvável alguém seja capaz de encontrar, muito menos reutilizar, qualquer destes dados. Com metadados adequados, há alguma esperança, mas, mesmo assim, os desafios permanecem devido a diferenças nos detalhes experimentais e na estrutura do registro de dados.

A análise dos dados é consideravelmente mais desafiadora do que simplesmente localizar, identificar, entender e citar dados. Para uma análise eficaz, em larga escala, tudo isso tem que acontecer de uma forma completamente automatizada. Isso requer diferenças na estrutura de dados e na semântica a ser expressa em formas que são compreensíveis pelo computador, e então "roboticamente" solúvel. Há um forte corpo de trabalho em integração de dados que pode fornecer algumas das respostas. No entanto, um trabalho adicional considerável é necessário para conseguir a resolução diferença automatizado livre de erros.

Mesmo para análises mais simples, que dependem apenas um conjunto de dados, ainda há uma questão importante de design de banco de dados adequado. Normalmente, haverá muitas formas alternativas para armazenar a mesma informação. Certos modelos terá vantagens sobre os outros para efeitos de certos inconvenientes e, possivelmente, para outros fins. Testemunha, por exemplo, a variedade enorme na estrutura de bases de dados com informações relacionadas sobre entidades substancialmente semelhantes, tais como genes. Projeto de banco de dados é hoje uma arte, e é cuidadosamente executado no contexto empresarial por profissionais altamente remunerados. Devemos permitir que outros profissionais, tais como cientistas de domínio, para criar projetos de banco de dados eficaz, seja por meio de ferramentas, concebendo, para auxiliá-los no processo de concepção ou através de renúncia ao processo de design e desenvolvimento de técnicas completamente para que os bancos de dados possam ser utilizados de forma eficaz na ausência de bom projetista de banco de dados.

**2.4 processamento de consultas, Modelagem de Dados e Análise**

Métodos para consultas e mineração do Big Data são fundamentalmente diferentes da tradicional análise estatística em pequenas amostras. Big Data é muitas vezes ruidoso, dinâmico, heterogêneo, inter-relacionados e não confiável. No entanto, mesmo ruidoso, Big Data pode ser mais valioso do que pequenas amostras porque a estatística geral obtida a partir de padrões frequentes e análise de correlação geralmente dominam flutuações individuais e muitas vezes divulgam ​​padrões ocultos e conhecimentos mais confiáveis. Além disso, Big Data quando interligados, formam grandes redes de informação heterogêneas, com redundância de informação que pode ser explorada para compensar a falta de dados, a casos CrossCheck conflitantes, para validar relações de confiança, para divulgar grupos inerentes, e para descobrir relações ocultas e modelos.

Mineração requer dados acessíveis integrados, limpos, confiáveis e eficientes, consulta declarativa e interfaces de mineração, algoritmos de mineração escaláveis ​​e ambientes de computação de Big Data. Ao mesmo tempo, mineração de dados também pode ser usada para ajudar a melhorar a qualidade e confiabilidade dos dados, entender a sua semântica, e fornecer funções inteligentes de consulta. Como observado anteriormente, na vida real os registros médicos têm erros, são heterogêneos e, frequentemente, são distribuídos em vários sistemas. O valor da análise do Big Data na área da saúde, para dar apenas um exemplo de domínio de aplicação, só pode ser realizado se ele pode ser aplicado de forma robusta sob estas condições difíceis. Por outro lado, o conhecimento desenvolvido a partir de dados pode ajudar a corrigir erros e remoção de ambiguidade. Por exemplo, um médico pode escrever "DVT", como o diagnóstico de um paciente. Esta sigla é comumente usada para tanto "trombose venosa profunda" e "diverticulite", dois muito diferentes condições médicas. A base de conhecimentos construídos a partir de dados relacionados pode usar sintomas associados ou medicamentos para determinar qual dos dois o médico quis dizer.

Big Data também está permitindo que a próxima geração de análise de dados interativos com respostas em tempo real. No futuro, as consultas para com Big Data serão geradas automaticamente para criação de conteúdo em sites, para preencher favoritos ou recomendações, e fornecer uma análise ad hoc do valor de um conjunto de dados para decidir se armazenar ou descartá-lo. Dimensionar técnicas de processamento de consultas complexas para terabytes e permitir tempos de resposta interativa é um problema importante de investigação aberto hoje.

Um problema com a atual análise Big Data é a falta de coordenação entre os sistemas de banco de dados, que hospedam os dados e fornecem consultas SQL, com análise de pacotes que executam várias formas de processamento não-SQL, tais como mineração de dados e análises estatísticas. Analistas de hoje são impedidos por um tedioso processo de exportação de dados do banco de dados, realizando um processo de não-SQL e trazer os dados de volta. Este é um obstáculo ao transporte de mais interativa a elegância da primeira geração de sistemas OLAP SQL dirigidos para o tipo de mineração de dados de análise que está na demanda crescente. A união entre linguagens de consulta declarativas e as funções de tais pacotes irá beneficiar tanto a expressividade e desempenho da análise.

**2.5 Interpretação**

Ter a capacidade de analisar Big Data é de valor limitado se os usuários não podem entender a análise. Em última análise, um tomador de decisões, provido com o resultado da análise, tem de interpretar esses resultados. Esta interpretação não pode acontecer em um vácuo. Normalmente, trata-se de examinar todas as hipóteses formuladas e refazendo a análise. Além disso, como vimos acima, existem muitas fontes possíveis de erro: sistemas de computadores podem ter bugs, modelos quase sempre têm pressupostos, e os resultados podem basear-se em dados errados. Por todas estas razões, nenhum usuário responsável vai ceder autoridade ao sistema de computador. Em vez disso, ela vai tentar entender, e verificar, os resultados produzidos pelo computador. O sistema de computador deve tornar mais fácil para ela fazer isso. Isto é particularmente um desafio com dados grandes, devido à sua complexidade. Há frequentemente suposições cruciais por trás dos dados gravados. A análise de pipeline muitas vezes podem envolver várias etapas, mais uma vez com os pressupostos construídos. O recente choque relacionado à hipoteca no sistema financeiro ressaltou drasticamente a necessidade de tal tomador de decisão - em vez de aceitar a estabelecida solubilidade de uma instituição financeira no valor nominal, um tomador de decisão tem que examinar criticamente os pressupostos muitos em vários estágios de análise.

Em suma, ela raramente é suficiente para fornecer apenas os resultados. Em vez disso, deve-se fornecer informações suplementares que explicam como cada resultado foi obtido, e com base precisamente nas entradas. Essas informações suplementares são chamadas a proveniência dos dados (resultado). Ao estudar a melhor forma de capturar, armazenar e proveniência de consulta, em conjunto com técnicas de capturar metadados adequada, pode-se criar uma infraestrutura para oferecer aos usuários a capacidade tanto para interpretar os resultados analíticos obtidos e repetir a análise com pressupostos diferentes, parâmetros, ou conjuntos de dados.

Sistemas com uma rica paleta de visualizações se tornam importantes na transmissão aos usuários os resultados das consultas de uma maneira que é melhor compreendida no domínio particular. Considerando que os usuários de sistemas de inteligência de negócios iniciais eram de conteúdo com apresentações tabulares, analistas de hoje precisam para empacotar e apresentar os resultados de visualizações poderosas que ajudam a interpretação, e colaboração de suporte ao usuário, como discutido na Seção. 3.5.

Além disso, com apenas alguns cliques o usuário deve ser capaz de detalhar cada pedaço de dados que ela vê e entender a sua origem, que é um elemento-chave para a compreensão dos dados. Ou seja, os usuários precisam ser capazes de ver não apenas os resultados, mas também entender por que eles estão vendo os resultados. No entanto, proveniência crus, em especial em relação às fases no pipeline de análise, é provável que seja muito técnico para muitos usuários a compreender completamente. Uma alternativa é permitir que os usuários "brincar" com os passos da análise - fazer pequenas alterações no pipeline, por exemplo, ou modificar os valores para alguns parâmetros. Os usuários podem ver os resultados dessas mudanças incrementais. Por esses meios, os usuários podem desenvolver um sentimento intuitivo para a análise e verificar também se ele funciona como esperado em casos de canto. Conseguir isso requer que o sistema para fornecer instalações convenientes para o usuário especificar análises. Especificação declarativa, discutida no Sec. 4, é um componente de um sistema deste tipo.

**3. Desafios na Análise do Big Data**

Depois de descrever as várias fases da análise do pipeline do Big Data, nos voltamos agora para alguns desafios comuns que sustentam muitos, e às vezes tudo, uma dessas fases. Elas são mostradas como cinco caixas na segunda fila da FIG. 1.

**3.1 Heterogeneidade e incompletude**

Quando os seres humanos consomem informação, uma grande heterogeneidade é confortavelmente tolerada. Na verdade, a nuance e a riqueza da linguagem natural pode proporcionar uma profundidade valiosa. No entanto, algoritmos de análise de máquinas esperam que os dados homogêneos, e não consegue entender nuances. Em consequência, os dados devem ser cuidadosamente estruturados como um primeiro passo na (ou antes) de análise de dados. Considere-se, por exemplo, um paciente que tenha vários procedimentos médicos de um hospital. Podemos criar um registro por procedimento médico ou exame de laboratório, um registro para o período de internação inteiro, ou um registro para todas as interações da vida hospitalar do paciente. Com qualquer coisa que não seja o primeiro design, o número de procedimentos médicos e testes de laboratório por registro seria diferente para cada paciente. As três escolhas de design listados têm sucessivamente menos estrutura e, por outro lado, sucessivamente maior variedade. Maior estrutura é susceptível de ser necessário por muitos (tradicional) de sistemas de análise de dados. No entanto, o design menos estruturado é provável que seja mais eficaz para muitas finalidades - por exemplo a questões relacionadas com a progressão da doença ao longo do tempo irá requerer uma operação dispendiosa juntar com os dois primeiros designs, mas pode ser evitado com o último. No entanto, os sistemas de computador funcionam mais eficientemente se eles podem armazenar vários itens que são todas idênticas em tamanho e estrutura. Representação eficiente, acesso e análise de dados semi-estruturados requerem mais trabalho.

Considere um registro de um modelo de banco de dados eletrônico de saúde que tem campos para data de nascimento, ocupação e tipo de sangue de cada paciente. O que vamos fazer se uma ou mais dessas peças de informação não é fornecida por um paciente? Obviamente, o registro de saúde ainda é colocado no banco de dados, mas com os valores de atributos correspondentes a ser definido como NULL. A análise de dados que olha para classificar os pacientes por, digamos, ocupação, deve levar em conta os pacientes para as quais esta informação não é conhecida. Pior, esses pacientes com ocupações desconhecidos pode ser ignorado na análise apenas se temos razão para acreditar que eles são de outra maneira estatisticamente semelhante ao dos pacientes com ocupação conhecida para a análise realizada. Por exemplo, se os pacientes desempregados são mais propensos a esconder sua situação de trabalho, os resultados da análise podem ser distorcidos em que considera uma mistura da população mais empregada do que existe, e, portanto, potencialmente, um que tem diferenças no perfil de ocupação relacionadas à saúde.

Mesmo após a correção de dados limpeza e erro, algumas incompletude e alguns erros nos dados deverão permanecer. Esta incompletude e esses erros devem ser gerenciados durante a análise de dados. Fazer isso corretamente é um desafio. Trabalhos recentes sobre o gerenciamento de dados probabilísticos sugerem uma maneira de fazer progresso.

**3.2 Tamanho**

Naturalmente, a primeira coisa que alguém pensa com Big Data é o seu tamanho. Afinal, a palavra "grande" está lá no próprio nome. Gerenciar grandes volumes e o rápido crescente volume de dados tem sido um problema desafiador para muitas décadas. No passado, este desafio foi atenuado por processadores cada vez mais rápidos, seguindo a lei de Moore, para nos fornecer os recursos necessários para lidar com volumes cada vez maiores de dados. Mas, há uma mudança fundamental em andamento agora: volume de dados está aumentando mais rápido do que recursos de computação, e as velocidades de CPU são estáticos.

Primeiro, ao longo dos últimos cinco anos, a tecnologia de processador fez uma mudança dramática - em vez de processadores dobrar sua frequência ciclo de relógio a cada 18-24 meses, agora, devido a limitações de potência, velocidades de clock em grande parte estagnado e processadores estão sendo construídos com um número crescente de núcleos. No passado, os grandes sistemas de processamento de dados tiveram de se preocupar com o paralelismo entre os nós de um cluster, agora, tem que lidar com paralelismo dentro de um único nó. Infelizmente, as técnicas de processamento de dados paralelos que foram aplicadas no passado para o processamento de dados através de nós não se aplicam diretamente para o paralelismo intra-nó, já que a arquitetura é muito diferente, por exemplo, há muitos mais recursos de hardware, como caches do processador e processador canais de memória que são compartilhados entre os núcleos em um único nó. Além disso, o movimento em direção a embalagem múltiplos soquetes (cada um com 10s de núcleos) acrescenta outro nível de complexidade para o paralelismo intra-nó. Finalmente, com previsões de "silício negro", ou seja, que a análise de energia, provavelmente, no futuro, nos proíbem de usar todo o hardware no sistema de forma contínua, sistemas de processamento de dados provavelmente vai ter que gerenciar ativamente o consumo de energia do processador. Essas mudanças sem precedentes nos obrigam a repensar a forma como projetar, construir e operar os componentes de processamento de dados.

A segunda mudança dramática que está em andamento é a transição para a computação em nuvem, que agora agrega várias cargas de trabalho diferentes, com metas de desempenho variáveis ​​(por exemplo, a demanda de serviços interativos que o motor de processamento de dados retornar uma resposta dentro de um limite fixo de tempo de resposta) em clusters muito grandes. Este nível de partilha de recursos em clusters caros e de grande porte exigem novas formas de determinar a forma de executar e executar trabalhos de processamento de dados, para que possamos cumprir as metas de cada carga de trabalho de forma rentável, e para lidar com falhas do sistema, que ocorrem mais frequentemente como operamos em clusters maiores e maiores (que são obrigados a lidar com o rápido crescimento do volume de dados). Isto coloca um prêmio sobre abordagens declarativos para programas expressando, mesmo aqueles que fazem as tarefas de aprendizagem complexas máquinas, uma vez que a otimização global em programas de vários usuários é necessário para um bom desempenho global. Dependência de otimizações usuário orientadas programa é susceptível de conduzir a utilização do cluster pobres, já que os usuários não têm conhecimento de programas de outros usuários. Sistema de otimização voltada holística requer programas que seja suficientemente transparente, por exemplo, como em sistemas de banco de dados relacional, onde linguagens de consulta declarativas são concebidos com isto em mente.

A terceira mudança dramática que está em andamento é a mudança transformadora do subsistema de I / O tradicional. Por muitas décadas, discos rígidos (HDDs) foram usados ​​para armazenar dados persistentes. HDDs teve muito mais lento desempenho IO IO aleatória do que o desempenho seqüencial, e motores de processamento de dados formatados seus dados e projetaram seus métodos de processamento de consulta para "contornar" esta limitação. Mas, HDDs estão cada vez mais sendo substituídos por drives de estado sólido, hoje e outras tecnologias, como memória de mudança de fase ao virar da esquina. Essas novas tecnologias de armazenamento não tem a mesma propagação grande de desempenho entre o seqüencial e aleatória desempenho de I / O, o que exige um repensar da forma como subsistemas de armazenamento de dados de projeto para sistemas de processamento. Implicações deste subsistema de armazenamento mudando potencialmente tocar todos os aspectos de processamento de dados, incluindo algoritmos de processamento de consulta, agendamento de consulta, o projeto de banco de dados, métodos de controle de concorrência e métodos de recuperação.

**3.3 Velocidade**

O outro lado de tamanho é a velocidade. Quanto maior o conjunto de dados a serem processados, o que levará mais tempo para analisar. A concepção de um sistema que trata eficazmente o tamanho é igualmente susceptível de resultar em um sistema que possa processar um dado tamanho de conjunto de dados mais rápidos. No entanto, não é apenas a velocidade esta que é normalmente significava quando se fala de Velocidade no contexto de Big Data. Em vez disso, existe uma taxa de aquisição de desafio, conforme descrito no cap. 2.1, e um desafio oportunidade descrito a seguir.

Há muitas situações em que o resultado da análise é necessária imediatamente. Por exemplo, se uma transação de cartão de crédito fraudulento é suspeito, deve idealmente ser marcada antes da transação ser concluída - potencialmente impedindo a operação de tomar lugar em tudo. Obviamente, uma análise completa da história de um usuário compra não é provável que seja viável em tempo real. Em vez disso, é necessário desenvolver resultados parciais de antemão, de modo que uma pequena quantidade de computação incremental com novos dados podem ser utilizadas para se chegar a uma determinação rápida.

Dado um conjunto grande de dados, muitas vezes é necessário encontrar elementos nele que atendam a um critério especificado. No decurso da análise dos dados, este tipo de pesquisa é susceptível de ocorrer repetidamente. Digitalizar os dados de todo o conjunto para encontrar elementos adequados é obviamente impraticável. Em vez disso, estruturas de índice são criados com antecedência, para permitir encontrar elementos de qualificação rapidamente. O problema é que cada estrutura do índice é projetado para suportar apenas algumas classes de critérios. Com novas análises desejadas usando Big Data, há novos tipos de critérios especificados, e uma necessidade de desenvolver novas estruturas de índice para suportar tais critérios. Por exemplo, considere um sistema de gestão de tráfego com informações sobre milhares de veículos e locais hot spots nas estradas. O sistema pode precisar de prever potenciais pontos de congestionamento ao longo de uma rota escolhida por um usuário, e sugerir alternativas. Isso requer avaliar várias consultas de proximidade espacial de trabalho com as trajetórias de objetos móveis. Estruturas de índice novos são necessários para suportar essas consultas. Projetando tais estruturas torna-se particularmente difícil quando o volume de dados está crescendo rapidamente, e as consultas têm prazos apertados de resposta.

**3.4 de Privacidade**

A privacidade dos dados é outra grande preocupação, e que aumenta no contexto de Big Data. Para os registros eletrônicos de saúde, existem leis rígidas que regem o que pode e não pode ser feito. Para outros dados, regulamentos, especialmente nos EUA, são menos contundentes. No entanto, existe um grande medo público sobre o uso inadequado de dados pessoais, nomeadamente através da ligação de dados de múltiplas fontes. Gerenciar a privacidade é efetivamente um problema tanto técnico e sociológico, que deve ser tratado em conjunto de ambas as perspectivas para cumprir a promessa do Big Data.

Considere, por exemplo, os dados recolhidos a partir de serviços baseados em localização. Essas novas arquiteturas requerem um usuário pode compartilhar seu / sua posição com o prestador de serviços, resultando em problemas de privacidade óbvias. Note que esconder a identidade do usuário sozinho, sem esconder sua localização não tratam corretamente essas preocupações com a privacidade. Um atacante ou um (potencialmente malicioso) servidor baseado em localização pode-se inferir a identidade da fonte de consulta a partir de sua informação de localização (subsequente). Por exemplo, a informação de um usuário local pode ser rastreado através de vários pontos de conexão fixas (por exemplo, torres de celular). Depois de um tempo, o usuário deixa "um rastro de migalhas de pacotes", que podem ser associados a uma residência certa ou localização do escritório e, assim, usado para determinar a identidade do usuário. Vários outros tipos de informações surpreendentemente privadas, tais como problemas de saúde (por exemplo, a presença em um centro de tratamento de câncer) ou preferências religiosas (por exemplo, a presença de uma igreja) também pode ser revelado por apenas observando o movimento de usuários anônimos e padrão de utilização ao longo do tempo. Em geral, Barabási et al. mostrou que há uma correlação estreita entre as identidades das pessoas e seus padrões de movimento [Gon2008]. Observe que ocultar a localização de um usuário é muito mais desafiador do que esconder sua identidade. Isto é porque com serviços baseados na localização, a localização do usuário é necessária para um acesso aos dados com sucesso ou uma coleção de dados, enquanto que a identidade do utilizador que não seja necessário.

Existem muitos outros problemas de pesquisas desafiadoras. Por exemplo, nós não sabemos ainda como compartilhar dados privados, limitando a divulgação e garantir a utilidade de dados suficiente nos dados compartilhados. O paradigma existente de privacidade diferencial é um passo muito importante na direção certa, mas infelizmente ele reduz grandemente o conteúdo de informações, a fim de ser útil em casos mais práticos. Além disso, os dados reais não são estáticos, mas aumentam e mudam ao longo do tempo, nenhuma das técnicas predominantes resultam em qualquer conteúdo útil sendo lançado neste cenário. No entanto, outro sentido muito importante é repensar em segurança para o compartilhamento de informações nos casos de uso do Big Data. Muitos serviços online hoje exigem de nós compartilharmos informações privadas (pense em aplicativos do Facebook), mas, além do nível de registro de controle de acesso, não entendemos o que significa compartilhar dados, como os dados compartilhados podem ser ligados, e como dar aos usuários fino controle refinado sobre este compartilhamento.

**3.5 colaboração humana**

Apesar dos grandes avanços feitos na análise computacional, ainda há muitos padrões que os humanos podem facilmente detectar, mas algoritmos de computador têm um tempo difícil encontrar. Na verdade, CAPTCHAs exploram precisamente este fato para informar aos usuários da web humanos para além de programas de computador. Idealmente, a análise de Big Data não vai ser tudo computacional - em vez disso, será projetado explicitamente para ter um humano no ciclo. O novo sub-campo de análise visual está tentando fazer isso, pelo menos no que diz respeito à fase de modelagem e análise do pipeline. Há um valor semelhante para a contribuição humana em todas as fases de análise do pipeline.

No complexo mundo de hoje, que muitas vezes leva vários especialistas de diferentes domínios para realmente entender o que está acontecendo. Um sistema de análise de Big Data deve apoiar a entrada de vários especialistas humanos e exploração compartilhada de resultados. Estes peritos múltiplos podem ser separados no espaço e no tempo em que é muito caro para montar uma equipe inteira juntos em um quarto. O sistema de dados tem de aceitar esta entrada especialista distribuída, e apoiar a sua colaboração.

Um método popular novo do engenho humano aproveitando para resolver os problemas é através de crowd-sourcing. Wikipedia, a enciclopédia on-line, é talvez o exemplo mais conhecido de dados multidão origem. Estamos confiando em informações fornecidas por estranhos não controlados. Na maioria das vezes, o que dizem é correto. No entanto, devemos esperar que haja indivíduos que têm outros motivos e habilidades - alguns podem ter um motivo para fornecer informações falsas em uma tentativa intencional de enganar. Enquanto a maioria tais erros serão detectados e corrigidos por outras pessoas na multidão, precisamos de tecnologias para facilitar isso. Também precisamos de uma estrutura para usar na análise desses dados multidão de origem com declarações contraditórias. Como seres humanos, nós podemos olhar para comentários de um restaurante, alguns dos quais são positivos e outros críticos, e chegar a uma avaliação sumária com base no qual podemos decidir se tentar comer lá. Precisamos de computadores para ser capaz de fazer o equivalente. As questões de incerteza e erro tornar-se ainda mais pronunciada em um tipo específico de crowd-sourcing, denominado senso participativo. Neste caso, cada pessoa com um telefone celular pode agir como um sensor multi-modal coleta de vários tipos de dados instantaneamente (por exemplo, imagem, vídeo, áudio, localização, tempo, velocidade, direção, aceleração). O desafio extra aqui é a incerteza inerente dos dispositivos de coleta de dados. O fato de que os dados coletados são, provavelmente, espacialmente e temporalmente correlacionados pode ser explorado para avaliar melhor a sua correção. Quando multidão de origem os dados são obtidos de aluguel, por exemplo, com "turcos Mecânica," grande parte dos dados podem ser criados com o objetivo principal de conseguir o feito rapidamente e não corretamente. Este é mais um modelo de erro, que deve ser planejado para explicitamente quando se aplica.

**4. Arquitetura do Sistema**

As empresas hoje já usa, e apreciam o valor da inteligência de negócios. Os dados de negócios são analisados ​​para muitas finalidades: a empresa pode realizar análises do sistema de registro e análise de mídias sociais para avaliação de risco, retenção de clientes, gestão da marca, e assim por diante. Normalmente, tais tarefas variadas foram tratadas por sistemas separados, mesmo que cada sistema inclui medidas comuns de extração de informações, limpeza de dados, relacional como processamento (junta, grupo-by, agregação), modelagem estatística e preditiva, e exploração adequada e visualização ferramentas, como mostrado na Fig. 1.

Com Big Data, a utilização de sistemas separados desta maneira torna-se proibitivamente dispendioso devido ao grande tamanho dos conjuntos de dados. A despesa é devida não só ao custo dos próprios sistemas, mas também o tempo necessário para carregar os dados em vários sistemas. Em consequência, Big Data tornou necessário para executar cargas de trabalho heterogêneas em uma única infraestrutura é suficientemente flexível para lidar com todas essas cargas de trabalho. O desafio aqui não é construir um sistema que é ideal para todas as tarefas de processamento. Em vez disso, a necessidade é para a arquitetura subjacente do sistema para ser flexível o suficiente para que os componentes construídos em cima dela para expressar os vários tipos de tarefas de processamento pode ajustá-lo para executar com eficiência essas cargas de trabalho diferentes. Os efeitos de escala na arquitetura física foram considerados na Seção 3.2. Nesta seção, vamos nos concentrar nas necessidades de programação.

Se os usuários estão a compor e construir complexos pipelines mais analíticos Big Data, é essencial que eles tenham adequados de alto nível primitivas para especificar suas necessidades em tais sistemas flexíveis. O quadro Map-Reduce foi tremendamente valioso, mas é apenas um primeiro passo. Mesmo linguagens declarativas que explorá-la, como porco Latina, estão em um nível bastante baixo quando se trata de tarefas de análise complexas. Semelhantes especificações declarativas são necessárias em níveis mais altos para atender as necessidades programação e composição destes dutos análise. Além da necessidade técnica básica, há um imperativo de negócio forte também. Empresas normalmente vai terceirizar o processamento de dados grande, ou muitos aspectos dele. Especificações declarativas são obrigados a permitir que os acordos de nível de serviços tecnicamente significativos, desde o ponto do out-sourcing é especificar exatamente qual tarefa será executada sem entrar em detalhes de como fazê-lo.

Especificação declarativa é necessária não só para a composição gasoduto, mas também para as operações individuais próprios. Cada operação (limpeza, extração de modelagem, etc) potencialmente executado em um conjunto de dados muito grande. Além disso, cada operação em si é suficientemente complexo que há muitas opções e otimizações possíveis em como ela é implementada. Em bancos de dados, há um trabalho considerável na otimização de operações individuais, tais como associações. É bem conhecido que pode haver várias ordens de magnitude diferença no custo de duas maneiras diferentes para executar a mesma consulta. Felizmente, o usuário não tem que fazer essa escolha - o sistema de banco de dados faz isso por ela. No caso do Big Data, essas otimizações pode ser mais complexa, porque nem todas as operações de I / O intensivo como em bancos de dados. Algumas operações podem ser, mas outros podem ser intensivo da CPU, ou uma mistura. Técnicas de otimização de modo padrão de banco de dados não pode ser usado diretamente. No entanto, deve ser possível desenvolver novas técnicas para operações de dados Big inspirados por técnicas de banco de dados.

O próprio fato de que a análise de Big Data tipicamente envolve várias fases destaca um desafio que se coloca rotineiramente na prática: sistemas de produção deve executar complexas pipelines analíticas, ou fluxos de trabalho, em intervalos de rotina, por exemplo, horário ou diário. Novos dados devem ser incrementalmente representaram, tendo em conta os resultados da análise prévia e dados pré-existentes. E, claro, proveniência deve ser preservado, e deve incluir as fases do gasoduto analítica. Os sistemas atuais oferecem pouco ou nenhum suporte para tais condutas de dados grande, e isso é em si um objetivo desafiador.

**5. Conclusão**

Entramos em uma era de Big Data. Através de uma melhor análise de grandes volumes de dados que estão se tornando disponíveis, há o potencial para fazer avanços mais rápidos em muitas disciplinas científicas e melhorar a rentabilidade e sucesso de muitas empresas. No entanto, muitos desafios técnicos descritos neste documento devem ser abordadas antes que este potencial possa ser realizado plenamente. Os desafios não incluem apenas as questões óbvias de escala, mas também de heterogeneidade, a falta de estrutura, privacidade, tratamento de erros, pontualidade, visualização, proveniência e, em todas as fases do pipeline de análise de aquisição de dados para a interpretação dos resultados. Estes desafios técnicos são comuns em uma grande variedade de domínios de aplicação e, portanto, não o custo-benefício para resolver no contexto de um domínio só. Além disso, estes desafios exigem soluções transformadoras, e não será abordada naturalmente pela próxima geração de produtos industriais. Devemos apoiar e incentivar a pesquisa fundamental para enfrentar estes desafios técnicos, se quisermos alcançar os benefícios prometidos de Big Data.