Mobile Netw Appl (2014) 19:171–209
DOI 10.1007/s11036-013-0489-0

Big Data: uma pesquisa

Min Chen · **Shiwen Mao** · **Yunhao Liu**

Publicado online: 22 de janeiro de
2014

Resumo Neste artigo, revisamos os antecedentes e o estado da arte do big data. Primeiro, apresentamos o histórico geral de big data e revisamos as tecnologias relacionadas, como computação, Internet das coisas, data centers e Hadoop. Em seguida, focamos nas quatro fases da cadeia de valor de big data, ou seja, geração de dados, aquisição de dados, armazenamento de dados e análise de dados. Para cada fase, apresentamos o histórico geral, discutimos os desafios técnicos e analisamos os avanços mais recentes. Por fim, examinamos as diversas aplicações representativas de big data, incluindo gestão empresarial, Internet das Coisas, redes sociais online, aplicações de mídia, inteligência coletiva e smart grid.

Essas discussões visam fornecer uma visão geral abrangente e um quadro geral aos leitores dessa área empolgante. Esta pesquisa é concluída com uma discussão de problemas em aberto e direções futuras.

Palavras-chave Big data · Computação em nuvem · Internet das coisas · Data center · Hadoop · Smart grid · Análise de big data

M. Chen ()
Escola de Ciência e Tecnologia da Computação,
Universidade de Ciência e Tecnologia Huazhong,
1037 Luoyu Road, Wuhan, 430074, China e-
mail: minchen2012@hust.edu.cn; minchen@ieee.org

S. Mao
Departamento de Engenharia Elétrica e de Computação,
Auburn University, 200 Broun Hall, Auburn, AL
36849-5201, EUA e-
mail: smao@ieee.org

Y. Liu
TNLIST, Escola de Software, Universidade de Tsinghua, Pequim,
China e-mail: yunhao@greenorbs.com

1. Fundo

1.1 Amanhecer da era do big data

Nos últimos 20 anos, os dados aumentaram em larga escala em vários campos. De acordo com um relatório da International Data Corporation (IDC), em 2011, o volume total de dados criados e copiados no mundo foi de 1,8ZB (ÿ 1021B), que aumentou quase nove vezes em cinco anos [1].

Esse número dobrará pelo menos a cada dois anos no futuro próximo.

Sob o aumento explosivo de dados globais, o termo big data é usado principalmente para descrever enormes conjuntos de dados. Em comparação com os conjuntos de dados tradicionais, o big data geralmente inclui massas de dados não estruturados que precisam de mais análise em tempo real. Além disso, big data também traz novas oportunidades para descobrir novos valores, nos ajuda a obter uma compreensão profunda dos valores ocultos e também incorre em novos desafios, por exemplo, como organizar e gerenciar efetivamente esses conjuntos de dados.

Recentemente, as indústrias se interessaram pelo alto potencial do big data, e muitas agências governamentais anunciaram grandes planos para acelerar a pesquisa e as aplicações de big data [2]. Além disso, questões sobre big data são frequentemente abordadas na mídia pública, como *The Economist* [3, 4], *New York Times* [5] e *National Public Radio* [6, 7]. Duas revistas científicas importantes, *Nature* e *Science,* também abriram colunas especiais para discutir os desafios e impactos do big data [8, 9]. A era do big data chegou sem sombra de dúvida [10].

Atualmente, o big data relacionado ao serviço de empresas de Internet cresce rapidamente. Por exemplo, o Google processa dados de centenas de Petabytes (PB), o Facebook gera dados de log de mais de 10 PB por mês, a Baidu, uma empresa chinesa, processa dados de dezenas de PB e a Taobao, uma subsidiária da Alibaba,

gera dados de dezenas de Terabyte (TB) para negociação online por dia. A Figura 1 ilustra o boom do volume global de dados. Embora a quantidade de grandes conjuntos de dados esteja aumentando drasticamente, ela também traz muitos problemas desafiadores que exigem soluções imediatas:

– Os últimos avanços da tecnologia da informação (TI) facilitam a geração de dados. Por exemplo, em média, 72 horas de vídeos são enviados para o YouTube a cada minuto [11]. Portanto, somos confrontados com o principal desafio de coletar e integrar dados massivos de fontes de dados amplamente distribuídas.

– O rápido crescimento da computação em nuvem e da Internet das Coisas (IoT) promove ainda mais o crescimento acentuado dos dados. A computação em nuvem fornece proteção, sites de acesso e canais para ativos de dados. No paradigma da IoT, sensores em todo o mundo estão coletando e transmitindo dados para serem armazenados e processados na nuvem. Tais dados, tanto em quantidade quanto em relações mútuas, ultrapassarão em muito as capacidades das arquiteturas e infra-estrutura de TI das empresas existentes e seus requisitos de tempo real também enfatizarão muito a capacidade de computação disponível. Os dados cada vez mais crescentes causam um problema de como armazenar e gerenciar esses enormes conjuntos de dados heterogêneos com requisitos moderados de infraestrutura de hardware e software.

– Considerando a heterogeneidade, escalabilidade, tempo real, complexidade e privacidade do big data, devemos efetivamente "explorar" os conjuntos de dados em diferentes níveis durante a análise, modelagem, visualização e previsão, de modo a revelar sua propriedade intrínseca e melhorar a tomada de decisão.

1.2 Definição e características de big data

Big data é um conceito abstrato. Além de massas de dados, ele também possui alguns outros recursos, que determinam a diferença entre ele e os "dados massivos" ou "dados muito grandes".

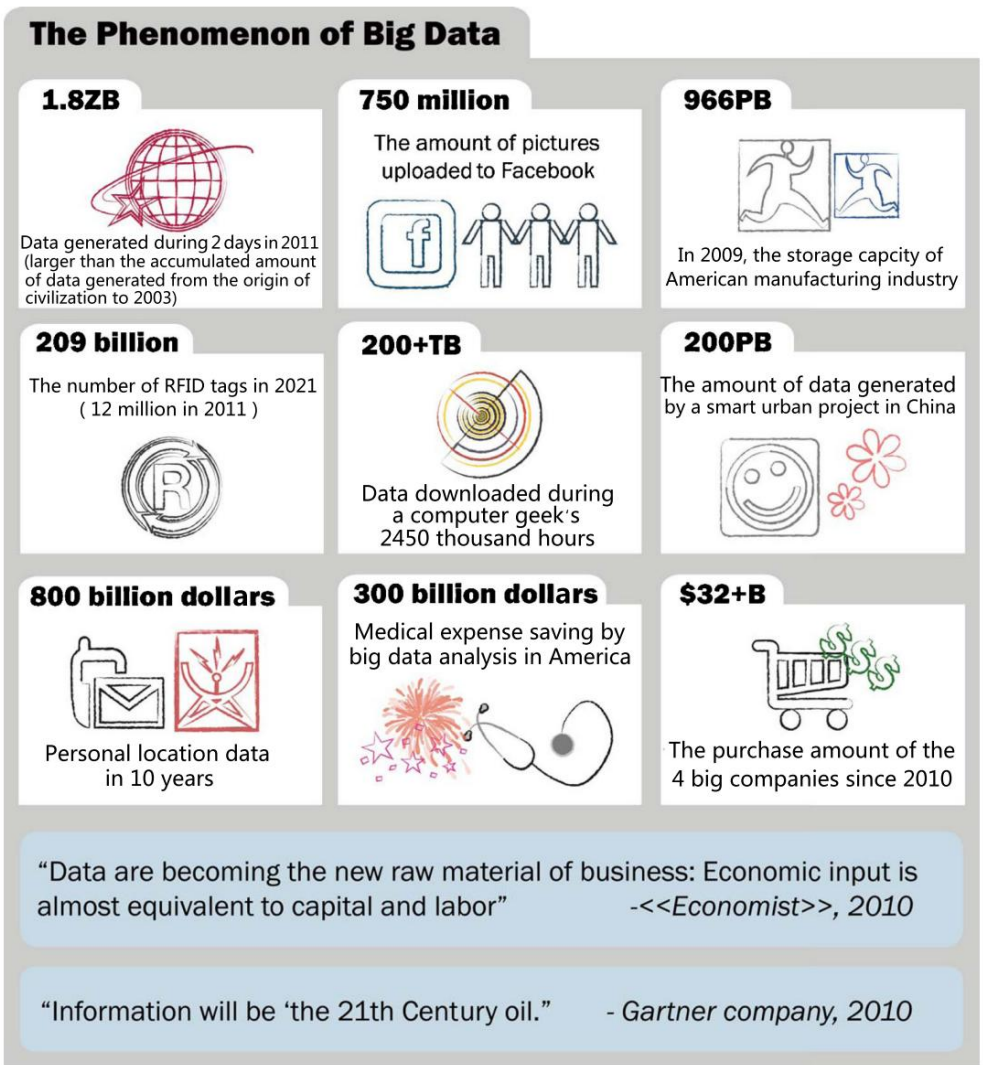

Fig. 1 O big data continuamente crescente

Atualmente, embora a importância do big data tenha sido amplamente reconhecida, as pessoas ainda têm opiniões diferentes sobre sua definição. Em geral, big data significa os conjuntos de dados que não puderam ser percebidos, adquiridos, gerenciados e processados pelas ferramentas tradicionais de TI e software/hardware dentro de um tempo tolerável. Devido a diferentes preocupações, empresas científicas e tecnológicas, acadêmicos de pesquisa, analistas de dados e profissionais técnicos têm diferentes definições de big data. As definições a seguir podem nos ajudar a entender melhor as profundas conotações sociais, econômicas e tecnológicas do big data.

Em 2010, o Apache Hadoop definiu big data como “conjuntos de dados que não podem ser capturados, gerenciados e processados por computadores em geral dentro de um escopo aceitável”. Com base nessa definição, em maio de 2011, a McKinsey & Company, uma agência de consultoria global, anunciou o Big Data como a próxima fronteira para inovação, competição e produtividade. Big data significa conjuntos de dados que não podem ser adquiridos, armazenados e gerenciados pelo software de banco de dados clássico. Essa definição inclui duas conotações: primeiro, os volumes dos conjuntos de dados que estão em conformidade com o padrão de big data estão mudando e podem crescer com o tempo ou com os avanços tecnológicos; Em segundo lugar, os volumes dos conjuntos de dados que estão em conformidade com o padrão de big data em diferentes aplicativos diferem uns dos outros. Atualmente, big data geralmente varia de vários TB a vários PB [10]. A partir da definição da McKinsey & Company, pode-se perceber que o volume de um conjunto de dados não é o único critério para big data. A escala de dados cada vez maior e seu gerenciamento que não poderia ser tratado por tecnologias de banco de dados tradicionais são os próximos dois recursos principais.

Na verdade, big data foi definido já em 2001. Doug Laney, analista da META (atualmente Gartner) definiu os desafios e oportunidades trazidos pelo aumento de dados com um modelo 3Vs, ou seja, o aumento de Volume, Velocidade, e Variedade, em um relatório de pesquisa [12].
Embora tal modelo não tenha sido originalmente usado para definir big data, o Gartner e muitas outras empresas, incluindo a IBM [13] e alguns departamentos de pesquisa da Microsoft [14] ainda usaram o modelo “3Vs” para descrever big data nos dez anos seguintes [15].
No modelo “3Vs”, Volume significa, com a geração e coleta de massas de dados, a escala de dados torna-se cada vez maior; Velocidade significa a pontualidade do big data, especificamente, coleta e análise de dados, etc., deve ser conduzida de forma rápida e oportuna, de modo a utilizar ao máximo o valor comercial do big data; Variedade indica os vários tipos de dados, que incluem dados semiestruturados e não estruturados, como áudio, vídeo, página da Web e texto, bem como dados estruturados tradicionais.

No entanto, outros têm opiniões diferentes, incluindo a IDC, uma das líderes mais influentes em big data e seus campos de pesquisa. Em 2011, um relatório da IDC definiu big data como “tecnologias de big data descrevem uma nova geração de tecnologias e arquiteturas, projetadas para extrair economicamente valor a partir de volumes muito grandes de uma ampla variedade de dados, permitindo a captura, descoberta e/ou análise em alta velocidade”. [1] Com esta definição, as características de big data podem ser resumidas como quatro Vs, ou seja, Volume (grande volume), Variedade (várias modalidades), Velocidade (geração rápida) e Valor (valor enorme, mas densidade muito baixa), como mostrado na Fig. 2. Essa definição de 4Vs foi amplamente reconhecida, pois destaca o significado e a necessidade de big data, ou seja, explorar os enormes valores ocultos. Essa definição indica o problema mais crítico em big data, que é como descobrir valores de conjuntos de dados com escala enorme, vários tipos e geração rápida. Como Jay Parikh, vice-chefe de engenharia do Facebook, disse: “Você só pode possuir um monte de dados que não sejam big data se não utilizar os dados coletados”. [11]

Além disso, o NIST define big data como “Big data deve significar os dados cujo volume de dados, velocidade de aquisição ou representação de dados limita a capacidade de usar métodos relacionais tradicionais para conduzir análises eficazes ou os dados que podem ser efetivamente processados com horizonte importante tecnologias de zoom tal”, que se concentra no aspecto tecnológico de big data. Indica que métodos ou tecnologias eficientes precisam ser desenvolvidos e usados para analisar e processar big data.

Houve discussões consideráveis da indústria e da academia sobre a definição de big data [16, 17].
Além de desenvolver uma definição adequada, a pesquisa de big data também deve se concentrar em como extrair seu valor, como usar dados e como transformar “um monte de dados” em “big data”.

1.3 Valor de big data

A McKinsey & Company observou como o big data criou valor após uma pesquisa aprofundada sobre a saúde dos EUA, a administração do setor público da UE, o varejo dos EUA, a manufatura global e os dados globais de localização pessoal. Por meio de pesquisas sobre as cinco principais indústrias que representam a economia global, o relatório da McKinsey apontou que o big data pode desempenhar plenamente a função econômica, melhorar a produtividade e a competitividade de empresas e setores públicos e criar enormes benefícios para os consumidores. Em [10], a McKinsey resumiu os valores que o big data poderia criar: se o big data pudesse ser utilizado de forma criativa e eficaz para melhorar a eficiência e a qualidade, o valor potencial da indústria médica dos EUA obtido por meio de dados poderia ultrapassar US$ 300 bilhões, reduzindo assim os gastos com saúde nos Estados Unidos em mais de 8%; os varejistas que utilizam totalmente o big data podem aumentar seus lucros em mais de 60%; big data também pode ser utilizado para melhorar a eficiência das operações do governo, de modo que as economias desenvolvidas na Europa possam economizar mais de 100 bilhões de euros (o que exclui o efeito da redução de fraudes, erros e diferenças fiscais).

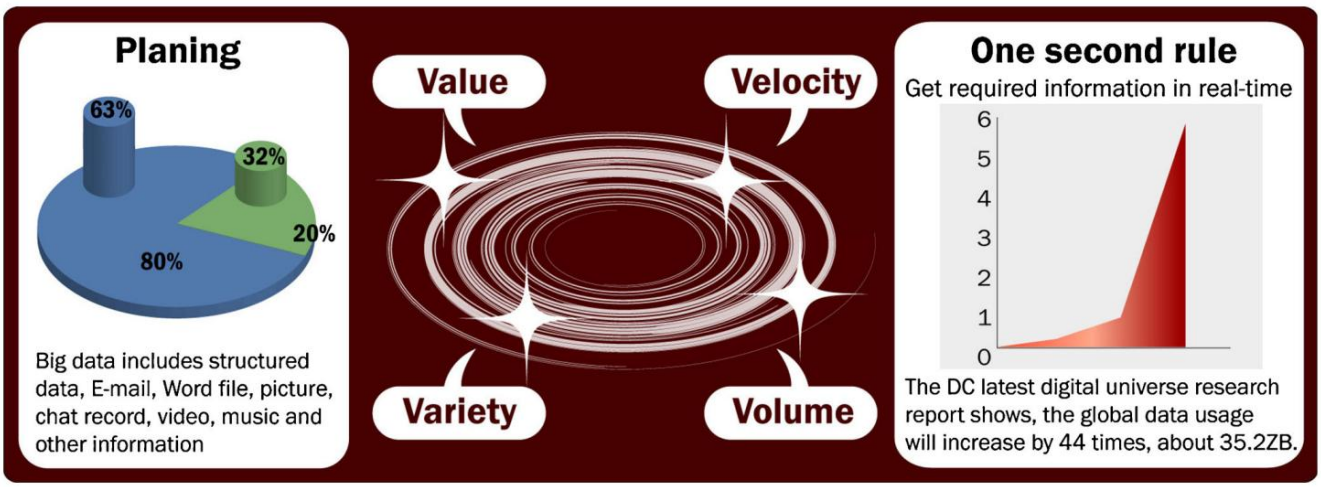

Fig. 2 O recurso 4Vs de big data

O relatório da McKinsey é considerado prospectivo e preditivo, enquanto os seguintes fatos podem validar os valores do big data. Durante a pandemia de gripe de 2009, o Google obteve informações oportunas por meio da análise de big data, que forneceu informações ainda mais valiosas do que as fornecidas pelos centros de prevenção de doenças. Quase todos os países exigiram que os hospitais informassem as agências, como os centros de prevenção de doenças, sobre o novo tipo de casos de influenza. No entanto, os pacientes geralmente não consultavam os médicos imediatamente quando eram infectados.
Também levou algum tempo para enviar informações dos hospitais para os centros de prevenção de doenças, e para os centros de prevenção de doenças analisarem e resumirem essas informações. Portanto, quando o público fica sabendo da pandemia do novo tipo de gripe, a doença já pode ter se espalhado por uma a duas semanas com caráter histerético. O Google descobriu que, durante a propagação da gripe, as entradas freqüentemente procuradas em seus mecanismos de busca seriam diferentes daquelas em horários normais, e as frequências de uso das entradas foram correlacionadas com a propagação da gripe tanto no tempo quanto no local.
O Google encontrou 45 grupos de entrada de pesquisa que eram bastante relevantes para o surto de gripe e os incorporou em modelos matemáticos específicos para prever a propagação da gripe e até mesmo prever locais de onde a gripe se espalhou. Os resultados da pesquisa relacionados foram publicados na Nature [18].

Em 2008, a Microsoft comprou a Farecast, uma empresa de tecnologia científica nos Estados Unidos. A Farecast tem um sistema de previsão de passagens aéreas que prevê as tendências e faixas de aumento/descida do preço das passagens aéreas. O sistema foi incorporado ao buscador Bing da Microsoft. Em 2012, o sistema economizou quase US$ 50 por passagem por passageiro, com uma precisão prevista de até 75%.

Atualmente, os dados se tornaram um importante fator de produção que pode ser comparável aos bens materiais e ao capital humano. À medida que multimídia, mídia social e IoT estão se desenvolvendo, as empresas coletarão mais informações, levando a um crescimento exponencial do volume de dados. Big data terá um potencial enorme e crescente na criação de valores para empresas e consumidores.

1.4 O desenvolvimento de big data

No final da década de 1970, surgiu o conceito de “máquina de banco de dados”, que é uma tecnologia especialmente utilizada para armazenar e analisar dados. Com o aumento do volume de dados, a capacidade de armazenamento e processamento de um único sistema de computador mainframe tornou-se inadequada. Na década de 1980, as pessoas propuseram “share nothing”, um sistema de banco de dados paralelo, para atender à demanda do crescente volume de dados [19]. A arquitetura do sistema sem compartilhamento é baseada no uso de cluster e cada máquina possui seu próprio processador, armazenamento e disco. O sistema Teradata foi o primeiro sistema de banco de dados paralelo comercial bem-sucedido. Esse banco de dados tornou-se muito popular ultimamente. Em 2 de junho de 1986, ocorreu um marco quando a Teradata entregou o primeiro sistema de banco de dados paralelo com capacidade de armazenamento de 1 TB para o Kmart para ajudar a empresa de varejo de grande porte na América do Norte a expandir seu data warehouse [20] .
No final da década de 1990, as vantagens do banco de dados paralelo foram amplamente reconhecidas no campo de banco de dados.

No entanto, surgiram muitos desafios em big data. Com o desenvolvimento dos serviços de Internet, os índices e os conteúdos consultados cresceram rapidamente. Portanto, as empresas de mecanismos de pesquisa tiveram que enfrentar os desafios de lidar com esses grandes dad
O Google criou os modelos de programação GFS [21] e MapReduce [22] para lidar com os desafios trazidos pelo gerenciamento e análise de dados na escala da Internet. Além disso, os conteúdos gerados por usuários, sensores e outras fontes de dados onipresentes também enfrentaram os fluxos de dados avassaladores, o que exigiu uma mudança fundamental na arquitetura de computação e no mecanismo de processamento de dados em larga escala.
Em janeiro de 2007, Jim Gray, um pioneiro do software de banco de dados,

chamou essa transformação de “O Quarto Paradigma” [23]. Ele também achava que a única maneira de lidar com esse paradigma era desenvolver uma nova geração de ferramentas de computação para gerenciar, visualizar e analisar dados massivos. Em junho de 2011, ocorreu outro evento marcante; A EMC/IDC publicou um relatório de pesquisa intitulado *Extracting Values from Chaos* [1], que apresentou o conceito e o potencial de big data pela primeira vez. Este relatório de pesquisa desencadeou um grande interesse tanto na indústria quanto na academia em big data.

Nos últimos anos, quase todas as grandes empresas, incluindo EMC, Oracle, IBM, Microsoft, Google, Amazon e Facebook, etc., iniciaram seus projetos de big data.
Tomando a IBM como exemplo, desde 2005, a IBM investiu US$ 16 bilhões em 30 aquisições relacionadas a big data. Na academia, o big data também estava sob os holofotes. Em 2008, a Nature publicou uma edição especial de big data. Em 2011, a Science também lançou uma edição especial sobre as principais tecnologias de “processamento de dados” em big data. Em 2012, o European Research Consortium for Informatics and Mathematics (ERCIM)
News publicou uma edição especial sobre big data. No início de 2012, um relatório intitulado *Big Data, Big Impact* apresentado no Fórum de Davos, na Suíça, anunciou que o big data se tornou um novo tipo de ativo econômico, assim como moeda ou ouro. A Gartner, uma agência internacional de pesquisa, publicou *Hype Cycles de 2012 a 2013,* que classificou a computação de big data, análise social e análise de dados armazenados em 48 tecnologias emergentes que merecem mais atenção.

Muitos governos nacionais, como os EUA, também prestaram muita atenção ao big data. Em março de 2012, o governo Obama anunciou um investimento de US$ 200 milhões para lançar o “Plano de Pesquisa e Desenvolvimento de Big Data”, que foi a segunda maior iniciativa de desenvolvimento científico e tecnológico após a iniciativa “Information Highway” em 1993. Em julho de 2012, o projeto “Vigorous ICT Japan” lançado pelo Ministério de Assuntos Internos e Comunicações do Japão indicou que o desenvolvimento de big data deve ser uma estratégia nacional e as tecnologias de aplicação devem ser o foco. Em julho de 2012, as Nações Unidas emitiram o relatório Big Data for Development, que resumia como os governos utilizaram big data para melhor servir e proteger seu povo.

1.5 Desafios do big data

O aumento acentuado do dilúvio de dados na era do big data traz enormes desafios na aquisição, armazenamento, gerenciamento e análise de dados. Os sistemas tradicionais de gerenciamento e análise de dados são baseados no sistema de gerenciamento de banco de dados relacional (RDBMS). No entanto, tais RDBMSs se aplicam apenas a dados estruturados, exceto dados semiestruturados ou não estruturados. Além disso, os RDBMSs estão cada vez mais utilizando hardware cada vez mais caro. É aparentemente que os RDBMSs tradicionais não poderiam lidar com o enorme volume e heterogeneidade de big data. A comunidade de pesquisa propôs algumas soluções de diferentes perspectivas. Por exemplo, a computação em nuvem é utilizada para atender aos requisitos de infraestrutura para big data, por exemplo, eficiência de custo, elasticidade e atualização/rebaixamento suave.
Para soluções de armazenamento permanente e gerenciamento de conjuntos de dados desordenados em larga escala, sistemas de arquivos distribuídos [24] e bancos de dados NoSQL [25] são boas escolhas. Tais estruturas de programação obtiveram grande sucesso no processamento de tarefas agrupadas, especialmente para classificação de páginas da web. Vários aplicativos de big data podem ser desenvolvidos com base nessas tecnologias ou plataformas inovadoras. Além disso, não é trivial implantar os sistemas de análise de big data.

Alguma literatura [26-28] discute os obstáculos no desenvolvimento de aplicações de big data. Os principais desafios são listados a seguir:

– *Representação de dados:* muitos conjuntos de dados têm certos níveis de heterogeneidade em tipo, estrutura, semântica, organização, granularidade e acessibilidade. A representação de dados visa tornar os dados mais significativos para análise de computador e interpretação do usuário. No entanto, uma representação de dados imprópria reduzirá o valor dos dados originais e pode até obstruir a análise efetiva dos dados.
A representação eficiente de dados deve refletir a estrutura, classe e tipo de dados, bem como tecnologias integradas, de modo a permitir operações eficientes em diferentes conjuntos de dados.

– *Redução de redundância e compactação de dados:* geralmente, há um alto nível de redundância nos conjuntos de dados.
A redução de redundância e compactação de dados é eficaz para reduzir o custo indireto de todo o sistema na premissa de que os valores potenciais dos dados não são afetados. Por exemplo, a maioria dos dados gerados por redes de sensores são altamente redundantes, que podem ser filtrados e compactados em ordens de grandeza.

– *Gerenciamento do ciclo de vida dos dados:* em comparação com os avanços relativamente lentos dos sistemas de armazenamento, a detecção e a computação generalizadas estão gerando dados em taxas e escalas sem precedentes. Somos confrontados com muitos de desafios prementes, um dos quais é que o sistema de armazenamento atual não suporta dados tão massivos.
De um modo geral, os valores ocultos em big data dependem da atualização dos dados. Portanto, um princípio de importância de dados relacionado ao valor analítico deve ser desenvolvido para decidir quais dados devem ser armazenados e quais dados devem ser armazenados. serão descartados.

– *Mecanismo analítico:* o sistema analítico de big data deve processar massas de dados heterogêneos dentro de um tempo limitado. No entanto, os RDBMSs tradicionais são estritamente projetados com falta de escalabilidade e capacidade de expansão, o que não atende aos requisitos de desempenho.
Bancos de dados não relacionais mostraram suas vantagens únicas no processamento de dados não estruturados e

começou a se tornar mainstream na análise de big data.
Mesmo assim, ainda existem alguns problemas de bancos de dados não relacionais em seu desempenho e aplicações particulares. Encontraremos uma solução de compromisso entre RDBMSs e bancos de dados não relacionais. Por exemplo, algumas empresas utilizaram uma arquitetura de banco de dados mista que integra as vantagens de ambos os tipos de banco de dados (por exemplo, Facebook e Taobao). Mais pesquisas são necessárias no banco de dados na memória e nos dados de amostra com base na análise aproximada.

– *Confidencialidade dos dados:* a maioria dos provedores ou proprietários de serviços de big data no momento não poderia manter e analisar com eficiência conjuntos de dados tão grandes devido à sua capacidade limitada. Eles devem contar com profissionais ou ferramentas para analisar esses dados, o que aumenta os riscos potenciais de segurança. Por exemplo, o conjunto de dados transacionais geralmente inclui um conjunto de dados operacionais completos para conduzir os principais processos de negócios. Esses dados contêm detalhes da granularidade mais baixa e algumas informações confidenciais, como números de cartão de crédito. Portanto, a análise de big data pode ser entregue a terceiros para processamento apenas quando medidas preventivas adequadas são tomadas para proteger esses dados confidenciais, para garantir sua segurança.

– *Gerenciamento de energia:* o consumo de energia dos sistemas de computação do quadro principal tem atraído muita atenção, tanto do ponto de vista econômico quanto do ambiental. Com o aumento do volume de dados e das demandas analíticas, o processamento, armazenamento e transmissão de big data inevitavelmente consumirão cada vez mais energia elétrica.
Portanto, o controle do consumo de energia no nível do sistema e o mecanismo de gerenciamento devem ser estabelecidos para big data, enquanto a capacidade de expansão e a acessibilidade são garantidas.

– *Dispensabilidade e escalabilidade:* o sistema analítico de big data deve suportar conjuntos de dados presentes e futuros. O algoritmo analítico deve ser capaz de processar conjuntos de dados cada vez mais complexos e em expansão.

– *Cooperação:* a análise de big data é uma pesquisa interdisciplinar, que requer que especialistas em diferentes áreas cooperem para colher o potencial de big data.
Uma arquitetura abrangente de rede de big data deve ser estabelecida para ajudar cientistas e engenheiros em vários campos a acessar diferentes tipos de dados e utilizar plenamente seus conhecimentos, de modo a cooperar para completar os objetivos analíticos.

2 tecnologias relacionadas

Para obter uma compreensão profunda de big data, esta seção apresentará várias tecnologias fundamentais que estão intimamente relacionadas a big data, incluindo computação em nuvem, IoT, data center e Hadoop.

2.1 Relação entre computação em nuvem e big data

A computação em nuvem está intimamente relacionada ao big data. Os principais componentes da computação em nuvem são mostrados na Figura 3. Big data é o objeto da operação de computação intensiva e enfatiza a capacidade de armazenamento de um sistema de nuvem. O principal objetivo da computação em nuvem é usar enormes recursos de computação e armazenamento sob gerenciamento concentrado, de modo a fornecer aplicativos de big data com capacidade de computação refinada. O desenvolvimento da computação em nuvem fornece soluções para o armazenamento e processamento de big data. Por outro lado, o surgimento de big data também acelera o desenvolvimento da computação em nuvem. A tecnologia de armazenamento distribuído baseada em computação em nuvem pode efetivamente gerenciar big data; a capacidade de computação paralela em virtude da computação em nuvem pode melhorar a eficiência da aquisição e análise de big data.

Embora existam muitas tecnologias sobrepostas em computação em nuvem e big data, elas diferem nos dois aspectos a seguir. Primeiro, os conceitos são diferentes até certo ponto. A computação em nuvem transforma a arquitetura de TI enquanto o big data influencia a tomada de decisões de negócios.
No entanto, o big data depende da computação em nuvem como infraestrutura fundamental para o bom funcionamento.

Em segundo lugar, big data e computação em nuvem têm clientes-alvo diferentes. A computação em nuvem é uma tecnologia e um produto direcionado aos Chief Information Officers (CIO) como uma solução de TI avançada. Big data é um produto direcionado a Chief Executive Officers (CEO) com foco em operações de negócios.
Como os tomadores de decisão podem sentir diretamente a pressão da concorrência de mercado, eles devem derrotar os oponentes de negócios de maneiras mais competitivas. Com os avanços do big data e da computação em nuvem, essas duas tecnologias estão cada vez mais entrelaçadas. A computação em nuvem, com funções semelhantes às dos computadores e sistemas operacionais, fornece recursos em nível de sistema; grandes dados

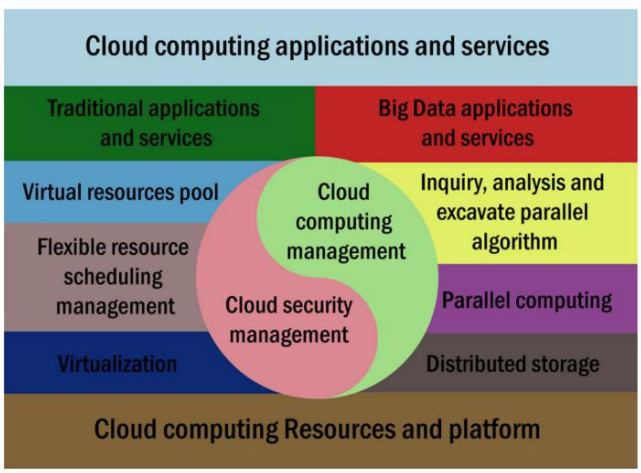

Fig. 3 Principais componentes da computação em nuvem

opera no nível superior suportado por computação em nuvem e fornece funções semelhantes às de banco de dados e capacidade eficiente de processamento de dados. Kissinger, presidente da EMC, indicou que a aplicação de big data deve ser baseada na computação em nuvem.

A evolução do big data foi impulsionada pelo rápido crescimento das demandas de aplicativos e da computação em nuvem desenvolvida a partir de tecnologias virtualizadas. Portanto, a computação em nuvem não apenas fornece computação e processamento para big data, mas também é um modo de serviço. Até certo ponto, os avanços da computação em nuvem também promovem o desenvolvimento de big data, sendo que ambos se complementam.

2.2 Relação entre IoT e big data

No paradigma IoT, uma enorme quantidade de sensores de rede é incorporada em vários dispositivos e máquinas no mundo real. Esses sensores implantados em diferentes campos podem coletar vários tipos de dados, como dados ambientais, dados geográficos, dados astronômicos e dados logísticos.
Equipamentos móveis, instalações de transporte, instalações públicas e eletrodomésticos podem ser equipamentos de aquisição de dados em IoT, conforme ilustrado na Fig. 4.

O big data gerado pela IoT tem características diferentes em comparação com o big data geral devido aos diferentes tipos de dados coletados, dos quais as características mais clássicas incluem heterogeneidade, variedade, característica não estruturada, ruído e alta redundância. Embora os dados IoT atuais não sejam a parte dominante do big data, até 2030, a quantidade de sensores chegarão a um trilhão e então os dados da IoT serão a parte mais importante do big data, de acordo com a previsão da HP. Um relatório da Intel apontou que big data em IoT possui três características que se enquadram no paradigma de big data: (i) terminais abundantes gerando massas de dados; (ii) os dados gerados pela IoT são geralmente semiestruturados ou não estruturados; (iii) dados de IoT são úteis apenas quando são analisados.

Atualmente, a capacidade de processamento de dados da IoT caiu por trás dos dados coletados e é extremamente urgente acelerar a introdução de tecnologias de big data para promover o desenvolvimento da IoT. Muitos operadores de IoT percebem a importância de big data, pois o sucesso de IoT depende da integração efetiva de big data e computação em nuvem. A ampla implantação da IoT também trará muitas cidades para a era do big data.

Há uma necessidade imperiosa de adotar big data para aplicativos de IoT, enquanto o desenvolvimento de big data já está atrasado. É amplamente reconhecido que essas duas tecnologias são interdependentes e devem ser desenvolvidas em conjunto: por um lado, a implantação generalizada de IoT impulsiona o alto crescimento de dados tanto em quantidade quanto em categoria, proporcionando assim a oportunidade para a aplicação e desenvolvimento de big data; por outro lado, a aplicação da tecnologia de big data à IoT também acelera a pesquisa avanços e modelos de negócios da IoT.

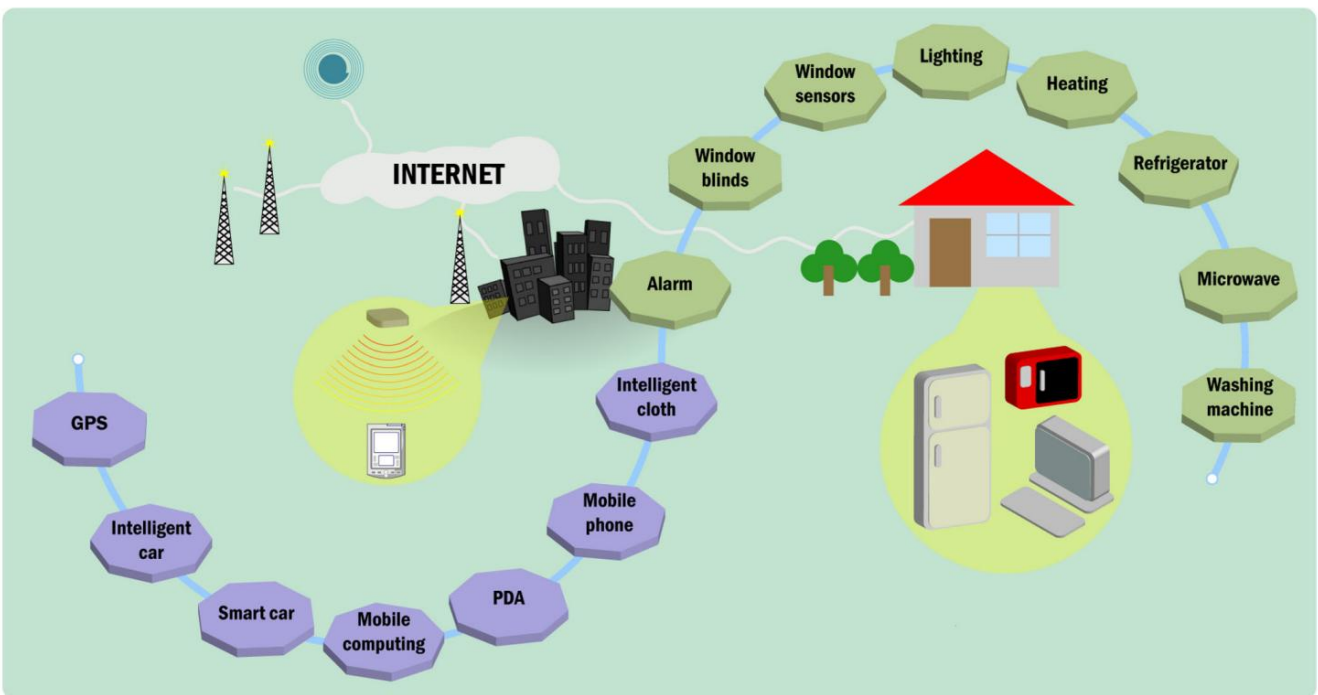

Fig. 4 Ilustração do equipamento de aquisição de dados em IoT

2.3 Centro de dados

No paradigma de big data, o data center não é apenas uma plataforma para armazenamento concentrado de dados, mas também assume mais responsabilidades, como aquisição de dados, gerenciamento de dados, organização de dados e alavancagem dos valores e funções dos dados. Os centros de dados dizem respeito principalmente a “dados” diferentes de “centro”. Possui uma grande quantidade de dados e organiza e gerencia os dados de acordo com seu objetivo principal e caminho de desenvolvimento, o que é mais valioso do que possuir um bom site e recurso. O surgimento de big data traz boas oportunidades de desenvolvimento e grandes desafios para os data centers. Big data é um paradigma emergente, que promoverá o crescimento explosivo da infraestrutura e do software relacionado ao data center. A rede física do centro de dados é o núcleo para suportar big data, mas, atualmente, é a infraestrutura chave que é mais urgentemente necessária [29].

– O big data requer que o data center forneça um poderoso suporte de backstage. O paradigma de big data tem requisitos mais rígidos sobre capacidade de armazenamento e capacidade de processamento, bem como capacidade de transmissão de rede.
As empresas devem levar em consideração o desenvolvimento de data centers para melhorar a capacidade de processamento rápido e eficaz de big data sob uma relação preço/desempenho limitada. O data center deve fornecer a infraestrutura com um grande número de nós, construir uma rede interna de alta velocidade, dissipar o calor de forma eficaz e fazer backup de dados eficaz. Somente quando um centro de dados altamente eficiente em energia, estável, seguro, expansível e redundante é construído, a operação normal de aplicativos de big data pode ser assegurada.

– O crescimento de aplicativos de big data acelera a revolução e a inovação dos data centers. Muitos aplicativos de big data desenvolveram suas arquiteturas exclusivas e promovem diretamente o desenvolvimento de tecnologias de armazenamento, rede e computação relacionadas ao data center. Com o crescimento contínuo dos volumes de dados estruturados e não estruturados e a variedade de fontes de dados analíticos, o processamento de dados e as capacidades de computação do centro de dados devem ser bastante aprimorados. Além disso, como a escala do data center está se expandindo cada vez mais, também é uma questão importante como reduzir o custo operacional para o desenvolvimento de data centers.

– Big data confere mais funções ao data center. No paradigma big data, o data center deve não apenas se preocupar com as instalações de hardware, mas também fortalecer as capacidades soft, ou seja, as capacidades de aquisição, processamento, organização, análise e aplicação de big data. O data center pode ajudar o pessoal de negócios a analisar os dados existentes, descobrir problemas na operação comercial e desenvolver soluções de big data.

2.4 Relação entre hadoop e big data

Atualmente, o Hadoop é amplamente utilizado em aplicativos de big data na indústria, por exemplo, filtragem de spam, pesquisa de rede, análise de fluxo de cliques e recomendação social. Além disso, pesquisas acadêmicas consideráveis agora são baseadas no Hadoop.
Alguns casos representativos são dados abaixo. Conforme declarado em junho de 2012, o Yahoo executa o Hadoop em 42.000 servidores em quatro centros de dados para oferecer suporte a seus produtos e serviços, por exemplo, pesquisa e filtragem de spam, etc. Atualmente, o maior cluster Hadoop tem 4.000 nós, mas o número de nós será aumentou para 10.000 com o lançamento do Hadoop 2.0.
No mesmo mês, o Facebook anunciou que seu cluster Hadoop pode processar 100 PB de dados, que cresceu 0,5 PB por dia em novembro de 2012. Algumas agências conhecidas que usam o Hadoop para realizar computação distribuída estão listadas em [30] . Além disso, muitas empresas fornecem execução e/ou suporte comercial Hadoop, incluindo Cloudera, IBM, MapR, EMC e Oracle.

Entre máquinas e sistemas industriais modernos, os sensores são amplamente utilizados para coletar informações para monitoramento ambiental e previsão de falhas, etc. Bahga e outros em [31] propuseram uma estrutura para organização de dados e infraestrutura de computação em nuvem, denominada CloudView. O Cloud View usa arquiteturas mistas, nós locais e clusters remotos baseados em Hadoop para analisar dados gerados por máquina. Nós locais são usados para a previsão de falhas em tempo real; clusters baseados em Hadoop são usados para análise off-line complexa, por exemplo, análise de dados baseada em casos.

O crescimento exponencial dos dados do genoma e a queda acentuada do custo de sequenciamento transformam a biociência e a biomedicina em ciência orientada por dados. Gunarathne et ai. em [32] utilizou infraestruturas de computação em nuvem, Amazon AWS, Microsoft Azune e estrutura de processamento de dados baseada em MapReduce, Hadoop e Microsoft DryadLINQ para executar dois aplicativos paralelos de biomedicina: (i) montagem de segmentos de genoma; (ii) redução de dimensão na análise da estrutura química. No aplicativo subsequente, os conjuntos de dados 166-D usados incluem 26.000.000 pontos de dados. Os autores compararam o desempenho de todos os frameworks em termos de eficiência, custo e disponibilidade. De acordo com o estudo, os autores concluíram que o baixo acoplamento será cada vez mais aplicado à pesquisa em nuvem de elétrons, e a estrutura da tecnologia de programação paralela (MapReduce) pode fornecer ao usuário uma interface com serviços mais convenientes e reduzir custos desnecessários.

3 Geração e aquisição de big data

Introduzimos várias tecnologias importantes relacionadas a big data, ou seja, computação em nuvem, IoT, data center e Hadoop.
Em seguida, vamos nos concentrar na cadeia de valor de big data, que

pode ser geralmente dividido em quatro fases: geração de dados, aquisição de dados, armazenamento de dados e análise de dados. Se considerarmos os dados como matéria-prima, a geração e aquisição de dados são um processo de exploração, o armazenamento de dados é um processo de armazenamento e a análise de dados é um processo de produção que utiliza a matéria-prima para criar novo valor.

3.1 Geração de dados

A geração de dados é a primeira etapa do big data. Considerando os dados da Internet como exemplo, uma grande quantidade de dados em termos de entradas de pesquisa, postagens em fóruns da Internet, registros de bate-papo e mensagens de microblog são geradas. Esses dados estão intimamente relacionados com a vida cotidiana das pessoas e têm características semelhantes de alto valor e baixa densidade. Esses dados da Internet podem não ter valor individualmente, mas, por meio da exploração de grandes dados acumulados, informações úteis, como hábitos e hobbies dos usuários, podem ser identificadas e até mesmo prever comportamentos e estados emocionais dos usuários.

Além disso, gerados por meio de fontes de dados longitudinais e/ou distribuídas, os conjuntos de dados são mais amplos, altamente diversos e complexos. Essas fontes de dados incluem sensores, vídeos, clickstreams e/ou todas as outras fontes de dados disponíveis.
Atualmente, as principais fontes de big data são as informações de operação e negociação em empresas, informações logísticas e de detecção na IoT, informações de interação humana e informações de posição no mundo da Internet e dados gerados em pesquisas científicas, etc. passa as capacidades das arquiteturas e infraestruturas de TI das empresas existentes, enquanto sua exigência de tempo real também enfatiza muito a capacidade de computação existente.

3.1.1 Dados da empresa

Em 2013, a IBM publicou Analysis: *the Applications of Big Data to the Real World,* que indica que os dados internos das empresas são as principais fontes de big data. Os dados internos das empresas consistem principalmente em dados de negociação on-line e dados de análise on-line, a maioria dos quais são dados historicamente estáticos e gerenciados por RDBMSs de maneira estruturada. Além disso, dados de produção, dados de estoque, dados de vendas e dados financeiros, etc., também constituem dados internos da empresa, que visam capturar atividades informatizadas e orientadas por dados nas empresas, de modo a registrar todas as atividades das empresas na forma de dados internos.

Nas últimas décadas, a TI e os dados digitais contribuíram muito para melhorar a lucratividade dos departamentos de negócios. Estima-se que o volume de dados de negócios de todos os

empresas no mundo pode dobrar a cada 1,2 anos [10], em que o faturamento de negócios pela Internet, empresas para empresas e empresas para consumidores por dia chegará a US$ 450 bilhões [33]. O volume de dados de negócios em constante crescimento requer gerenciamento em tempo real mais eficaz análise, de modo a colher plenamente o seu potencial. Por exemplo, a Amazon processa milhões de operações de terminal e mais de 500.000 consultas de vendedores terceirizados por dia [12].
O Walmart processa um milhão de transações de clientes por hora e esses dados de negociação são importados para um banco de dados com capacidade de mais de 2,5 PB [3]. A Akamai analisa 75 milhões de eventos por dia para seus anúncios alvo [13].

3.1.2 Dados de IoT

Conforme discutido, a IoT é uma importante fonte de big data. Entre as cidades inteligentes construídas com base na IoT, o big data pode vir da indústria, agricultura, tráfego, transporte, assistência médica, departamentos públicos e famílias, etc.

De acordo com os processos de aquisição e transmissão de dados em IoT, sua arquitetura de rede pode ser dividida em três camadas: a camada de detecção, a camada de rede e a camada de aplicação. A camada de detecção é responsável pela aquisição de dados e consiste principalmente em redes de sensores.
A camada de rede é responsável pela transmissão e processamento das informações, onde a transmissão próxima pode depender de redes de sensores e a transmissão remota dependerá da Internet. Finalmente, a camada de aplicação suporta aplicações específicas de IoT.

De acordo com as características da Internet das Coisas, o
os dados gerados da IoT possuem as seguintes características:

– *Dados em larga escala:* em IoT, grandes quantidades de equipamentos de aquisição de dados são implantados de forma distribuída, podendo adquirir dados numéricos simples, por exemplo, localização; ou dados multimídia complexos, por exemplo, vídeo de vigilância. Para atender às demandas de análise e processamento, não apenas os dados adquiridos atualmente, mas também os dados históricos dentro de um determinado período de tempo devem ser armazenados. Portanto, os dados gerados pela IoT são caracterizados por grandes escalas.
– *Heterogeneidade:* devido à variedade de dispositivos de aquisição de dados, os dados adquiridos também são diferentes e tais dados apresentam heterogeneidade.
– *Forte correlação de tempo e espaço:* na IoT, todos os dispositivos de aquisição de dados são colocados em uma localização geográfica específica e todos os dados têm carimbo de data/hora. A correlação de tempo e espaço é uma propriedade importante dos dados da IoT. Durante a análise e processamento de dados, o tempo e o espaço também são dimensões importantes para a análise estatística.

– *Dados efetivos representam apenas uma pequena parcela do big data:* uma grande quantidade de ruídos pode ocorrer durante a aquisição e transmissão de dados em IoT.
Entre os conjuntos de dados adquiridos por dispositivos de aquisição, apenas uma pequena quantidade de dados anormais é valiosa. Para a prova
Por exemplo, durante a aquisição do vídeo de tráfego, os poucos quadros de vídeo que capturam a violação das regras de trânsito

e os acidentes de trânsito são mais valiosos do que aqueles que capturam apenas o fluxo normal do tráfego.

3.1.3 Dados biomédicos

Como uma série de tecnologias de biomedição de alto rendimento são desenvolvidas de forma inovadora no início do século 21, a pesquisa de fronteira no campo da biomedicina também entra na era do big data. Ao construir modelos analíticos inteligentes, eficientes e precisos e sistemas teóricos para aplicações de biomedicina, o mecanismo de governo essencial por trás de fenômenos biológicos complexos pode ser revelado. Não apenas o desenvolvimento futuro da biomedicina pode ser determinado, mas também os papéis principais podem ser assumidos no desenvolvimento de uma série de importantes indústrias estratégicas relacionadas à economia nacional, subsistência das pessoas e segurança nacional, com aplicações importantes como assistência médica, pesquisa e desenvolvimento de novos medicamentos e produção de grãos (por exemplo, cultivos transgênicos).

A conclusão do HGP (Projeto Genoma Humano) e o desenvolvimento contínuo da tecnologia de sequenciamento também levam a aplicações generalizadas de big data no campo.
As massas de dados geradas pelo sequenciamento de genes passam por análises especializadas de acordo com diferentes demandas de aplicação, para combiná-las com o diagnóstico clínico de genes e fornecer informações valiosas para o diagnóstico precoce e tratamento personalizado de doenças. Um sequenciamento de gene humano pode gerar 100 dados brutos de 600 GB. No China National Genebank em Shenzhen, há 1,3 milhão de amostras, incluindo 1,15 milhão de amostras humanas e 150.000 amostras de animais, plantas e microorganismos. Até o final de 2013, 10 milhões de amostras biológicas rastreáveis serão armazenadas e, até o final de 2015, esse número chegará a 30 milhões. É previsível que, com o desenvolvimento das tecnologias de biomedicina, o sequenciamento de genes se torne mais rápido e conveniente, fazendo com que o big data da biomedicina cresça continuamente sem sombra de dúvida.

Além disso, os dados gerados a partir de atendimento médico clínico e P&D médico também aumentam rapidamente. Por exemplo, o University of Pittsburgh Medical Center (UPMC) armazenou 2 TB desses dados. A Explorys, uma empresa americana, fornece plataformas para colocar dados clínicos, dados de operação e manutenção e dados financeiros. Atualmente, cerca de 13 milhões de informações de pessoas foram colocadas, com 44 artigos de dados na escala de cerca de 60 TB, que chegarão a 70 TB em 2013. A Practice Fusion, outra empresa americana, gerencia prontuários eletrônicos de cerca de 200.000 pacientes.

Além dessas pequenas e médias empresas, outras conhecidas empresas de TI, como Google, Microsoft e IBM, têm investido extensivamente em pesquisa e análise computacional de métodos relacionados a big data biológicos de alto rendimento, para ações no enorme mercado como conhecido como a "Próxima Internet". A IBM prevê, na Conferência de Estratégia de 2013, que com o aumento acentuado de imagens médicas e registros médicos eletrônicos, os profissionais médicos podem utilizar big data para extrair informações clínicas úteis de massas de dados para obter um histórico médico e prever os efeitos do tratamento, melhorando assim atendimento ao paciente e redução de custos. Prevê-se que, até 2015, o volume médio de dados de cada hospital aumentará de 167 TB para 665 TB.

3.1.4 Geração de dados de outros campos

À medida que as aplicações científicas estão aumentando, a escala dos conjuntos de dados está se expandindo gradualmente, e o desenvolvimento de algumas disciplinas depende muito da análise de massas de dados. Aqui, examinamos várias dessas aplicações. Apesar de estarem em campos científicos diferentes, as aplicações têm demanda similar e crescente na análise de dados. O primeiro exemplo está relacionado à biologia computacional. GenBank é um banco de dados de sequência de nucleotídeos mantido pelos EUA
Centro Nacional de Inovação em Biotecnologia. Os dados neste banco de dados podem dobrar a cada 10 meses. Em agosto de 2009, o Genbank tinha mais de 250 bilhões de bases de 150.000 organismos diferentes [34]. O segundo exemplo está relacionado à astronomia. Sloan Digital Sky Survey (SDSS), o maior projeto de levantamento do céu em astronomia, registrou dados de 25 TB de 1998 a 2008. À medida que a resolução do telescópio for aprimorada, em 2004, o volume de dados gerados por noite ultrapassará 20 TB. A última aplicação está relacionada à física de alta energia. No início de 2008, o experimento Atlas do Grande Colisor de Hádrons (LHC) da Organização Européia para Pesquisa Nuclear gera dados brutos a 2 PB/s e armazena cerca de 10 TB de dados processados por ano.

Além disso, a detecção generalizada e a computação entre a natureza, comercial, Internet, governo e ambientes sociais estão gerando dados heterogêneos com complexidade sem precedentes. Esses conjuntos de dados têm suas características de dados exclusivas em escala, dimensão de tempo e categoria de dados.
Por exemplo, dados móveis foram registrados com relação a posições, movimentos, graus de aproximação, comunicações, multimídia, uso de aplicativos e ambiente de áudio [108]. De acordo com o ambiente e os requisitos do aplicativo, esses conjuntos de dados são divididos em diferentes categorias, de modo a selecionar as soluções adequadas e viáveis para big data.

3.2 Aquisição de Big Data

Como a segunda fase do sistema de big data, a aquisição de big data inclui coleta de dados, transmissão de dados e pré-processamento de dados. Durante a aquisição de big data, uma vez que coletamos os dados brutos, devemos utilizar um mecanismo de transmissão eficiente para enviá-los a um sistema de gerenciamento de armazenamento adequado para suportar diferentes aplicações analíticas. Às vezes, os conjuntos de dados coletados podem incluir muitos dados redundantes ou

dados inúteis, que aumentam desnecessariamente o espaço de armazenamento e afetam a análise de dados subsequente. Por exemplo, alta redundância é muito comum entre conjuntos de dados coletados por sensores para monitoramento do ambiente. A tecnologia de compressão de dados pode ser aplicada para reduzir a redundância. Portanto, as operações de pré-processamento de dados são indispensáveis para garantir armazenamento e exploração de dados eficientes.

3.2.1 Coleta de dados

A coleta de dados consiste em utilizar técnicas especiais de coleta de dados para adquirir dados brutos de um ambiente de geração de dados específico. Quatro métodos comuns de coleta de dados são mostrado a seguir.

– *Arquivos de log:* Como um método de coleta de dados amplamente utilizado, os arquivos de log são arquivos de registro gerados automaticamente pelo sistema de fonte de dados, de modo a registrar atividades em formatos de arquivo designados para análise subsequente. Os arquivos de log são normalmente usados em quase todos os dispositivos digitais. Por exemplo, os servidores da web registram em arquivos de log o número de cliques, taxas de cliques, visitas e outros registros de propriedade dos usuários da web [35]. Para capturar as atividades dos usuários nos sites da Web, os servidores da Web incluem principalmente os três formatos de arquivo de log a seguir: formato de arquivo de log público (NCSA), formato de log expandido (W3C) e formato de log IIS (Microsoft). Todos os três tipos de arquivos de log estão em formato de texto ASCII. Bancos de dados que não sejam arquivos de texto podem, às vezes, ser usados para armazenar informações de log para melhorar a eficiência da consulta do armazenamento massivo de logs [36, 37]. Existem também alguns outros arquivos de log baseados na coleta de dados, incluindo indicadores de estoque em aplicativos financeiros e determinação de estados operacionais em monitoramento de rede e gerenciamento de tráfego.

– *Sensoriamento:* Sensores são comuns na vida diária para medir grandezas físicas e transformar grandezas físicas em sinais digitais legíveis para posterior processamento (e armazenamento). Os dados sensoriais podem ser classificados como ondas sonoras, voz, vibração, automóveis, produtos químicos, correntes, clima, pressão, temperatura, etc. As informações detectadas são transferidas para um ponto de coleta de dados por meio de redes com ou sem fio. Para aplicações que podem ser facilmente implantadas e gerenciadas, por exemplo, sistema de vigilância por vídeo [38], a rede de sensores com fio é uma solução conveniente para adquirir informações relacionadas.

Às vezes, a posição exata de um fenômeno específico é desconhecida e, às vezes, o ambiente monitorado não possui infraestrutura de energia ou comunicação. Então a comunicação sem fio deve

ser usado para permitir a transmissão de dados entre nós sensores sob energia limitada e capacidade de comunicação. Nos últimos anos, as RSSFs têm recebido um interesse considerável e têm sido aplicadas a muitas aplicações, como

como pesquisa ambiental [39, 40], monitoramento da qualidade da água [41], engenharia civil [42, 43] e monitoramento dos hábitos da vida selvagem [44]. Uma RSSF geralmente consiste em um grande número de nós sensores distribuídos geograficamente, cada um sendo um microdispositivo alimentado por bateria.

Esses sensores são implantados em posições designadas conforme exigido pelo aplicativo para coletar dados de sensoriamento remoto. Assim que os sensores forem implantados, a estação base enviará informações de controle para configuração/gerenciamento de rede ou coleta de dados para nós sensores.

Com base nessas informações de controle, os dados sensoriais são montados em diferentes nós sensores e enviados de volta ao

estação base para processamento adicional. Os leitores interessados devem consultar [45] para discussões mais detalhadas.

– *Métodos para aquisição de dados de rede:* Atualmente, a aquisição de dados de rede é realizada usando uma combinação de rastreador da web, sistema de segmentação de palavras, sistema de tarefas e sistema de índice, etc. O rastreador da web é um programa usado pelos mecanismos de pesquisa para baixar e armazenar páginas da web [46]. De um modo geral, o rastreador da Web começa no localizador uniforme de recursos (URL) de uma página da Web inicial para acessar outras páginas da Web vinculadas, durante as quais ele armazena e sequencia todos os URLs recuperados. O rastreador da Web adquire um URL na ordem de precedência por meio de uma fila de URLs e, em seguida, baixa as páginas da Web e identifica todos os URLs nas páginas da Web baixadas e extrai novos URLs para serem colocados na fila. Esse processo é repetido até que o rastreador da Web seja interrompido. A aquisição de dados por meio de um rastreador da Web é amplamente aplicada em aplicativos baseados em páginas da Web, como mecanismos de pesquisa ou cache da Web. As tecnologias tradicionais de extração de páginas da web apresentam várias soluções eficientes e pesquisas consideráveis foram feitas em

este campo. À medida que aplicações de páginas da Web mais avançadas estão surgindo, algumas estratégias de extração são propostas em [47] para lidar com aplicações de Internet ricas.

As atuais tecnologias de aquisição de dados de rede incluem principalmente a tecnologia tradicional de captura de pacotes baseada em Libpcap, tecnologia de captura de pacotes de cópia zero, bem como alguns softwares especializados de monitoramento de rede, como Wireshark, SmartSniff e WinNetCap.

– *Tecnologia de captura de pacotes baseada em Libpcap:* Libpcap (biblioteca de captura de pacotes) é uma biblioteca de funções de captura de pacotes de dados de rede amplamente utilizada. É uma ferramenta geral que não depende de nenhum sistema específico e é usada principalmente para capturar dados na camada de enlace de dados. Ele apresenta simplicidade, facilidade de uso e portabilidade, mas tem uma eficiência relativamente baixa. Portanto, em um ambiente de rede de alta velocidade, perdas consideráveis de pacotes podem ocorrer quando o Libpcap é usado.

– *Tecnologia de captura de pacotes de cópia zero:* A chamada cópia zero (ZC) significa que nenhuma cópia entre quaisquer memórias internas ocorre durante o recebimento e envio de pacotes em um nó. No envio, os pacotes de dados partem diretamente do buffer do usuário dos aplicativos, passam pelas interfaces de rede e chegam a uma rede externa.

  Ao receber, as interfaces de rede enviam pacotes de dados diretamente para o buffer do usuário. A ideia básica da cópia zero é reduzir os tempos de cópia de dados, reduzir as chamadas do sistema e reduzir a carga da CPU enquanto os ddatagramas são passados dos equipamentos de rede para o espaço do programa do usuário. A tecnologia de cópia zero primeiro utiliza a tecnologia de acesso direto à memória (DMA) para transmitir datagramas de rede diretamente para um espaço de endereço pré-alocado pelo kernel do sistema, de modo a evitar a participação da CPU. Enquanto isso, ele mapeia a memória interna dos datagramas no kernel do sistema para a do programa de detecção, ou constrói uma região de cache no espaço do usuário e mapeia para o espaço do kernel. Em seguida, o programa de detecção acessa diretamente a memória interna, de modo a reduzir a cópia da memória interna do kernel do sistema para o espaço do usuário e reduzir a quantidade de chamadas do sistema.

– *Equipamentos móveis:* Atualmente, os dispositivos móveis são mais amplamente utilizados. À medida que as funções dos dispositivos móveis se tornam cada vez mais fortes, eles apresentam meios mais complexos e múltiplos de aquisição de dados, bem como mais variedade de dados. Os dispositivos móveis podem obter informações de localização geográfica por meio de sistemas de posicionamento; adquirir informações de áudio por meio de microfones; adquirir fotos, vídeos, paisagens urbanas, códigos de barras bidimensionais e outras informações multimídia por meio de câmeras; adquira gestos do usuário e outras informações de linguagem corporal por meio de telas sensíveis ao toque e sensores de gravidade. Ao longo dos anos, as operadoras sem fio melhoraram o nível de serviço da Internet móvel adquirindo e analisando essas informações. Por exemplo, o próprio iPhone é um “espião móvel”. Ele pode coletar dados sem fio e informações de localização geográfica e, em seguida, enviar essas informações de volta à Apple Inc. para processamento, do qual o usuário não tem conhecimento. Além da Apple, os sistemas operacionais de smartphones, como o Android do Google e o Windows Phone da Microsoft, também podem coletar informações de maneira semelhante.

Além dos três métodos de aquisição de dados acima mencionados das principais fontes de dados, existem muitos outros métodos ou sistemas de coleta de dados. Por exemplo, em experimentos científicos, muitas ferramentas especiais podem ser usadas para coletar dados experimentais, como espectrômetros magnéticos e radiotelescópios. Podemos classificar os métodos de coleta de dados de diferentes perspectivas. Do ponto de vista das fontes de dados, os métodos de coleta de dados podem ser classificados em duas categorias. gories: registro de métodos de coleta por meio de fontes de dados e registro de métodos de coleta por meio de outras ferramentas auxiliares.

3.2.2 Transporte de dados

Após a conclusão da coleta de dados brutos, os dados serão transferidos para uma infraestrutura de armazenamento de dados para processamento e análise. Conforme discutido na Seção 2.3, big data é armazenado principalmente em um data center. O layout dos dados deve ser ajustado para melhorar a eficiência da computação ou facilitar a manutenção do hardware. Em outras palavras, a transmissão interna de dados pode ocorrer no data center. Portanto, a transmissão de dados consiste em duas fases: transmissões Inter-DCN e transmissões Intra-DCN.

– *Transmissões Inter-DCN:* As transmissões Inter-DCN são da fonte de dados para o centro de dados, o que geralmente é obtido com a infraestrutura de rede física existente. Devido ao rápido crescimento das demandas de tráfego, a infraestrutura de rede física na maioria das regiões do mundo é constituída por sistemas de transmissão de fibra ótica de alto volume, alta taxa e baixo custo.

  Nos últimos 20 anos, equipamentos e tecnologias de gerenciamento avançado foram desenvolvidos, como a arquitetura de rede de multiplexação por divisão de comprimento de onda (WDM) baseada em IP, para conduzir o controle inteligente e o gerenciamento de redes de fibra óptica [48, 49] . WDM é uma tecnologia que multiplexa vários sinais de portadora óptica com diferentes comprimentos de onda e os acopla à mesma fibra óptica do enlace óptico. Nessa tecnologia, lasers com diferentes comprimentos de onda transportam sinais diferentes. De longe, a rede de backbone foi implantada com sistemas de transmissão óptica WDM com taxa de canal único de 40 Gb/s. Atualmente, interfaces comerciais de 100 Gb/s estão disponíveis e sistemas de 100 Gb/s (ou sistemas TB/s) estarão disponíveis em um futuro próximo [50].

  No entanto, as tecnologias tradicionais de transmissão óptica são limitadas pela largura de banda do gargalo eletrônico [51]. Recentemente, a multiplexação ortogonal por divisão de frequência (OFDM), inicialmente projetada para sistemas sem fio, é considerada uma das principais tecnologias candidatas para futuras transmissões ópticas de alta velocidade. OFDM é uma tecnologia de transmissão paralela multiportadora. Ele segmenta um fluxo de dados de alta velocidade para transformá-lo em subfluxos de dados de baixa velocidade a serem transmitidos por múltiplas subportadoras ortogonais [52]. Comparado com o espaçamento de canal fixo do WDM, o OFDM permite que os espectros de frequência de subcanal se sobreponham uns aos outros [53]. Portanto, é uma tecnologia de rede óptica flexível e eficiente.

– *Transmissões Intra-DCN:* as transmissões Intra-DCN são os fluxos de comunicação de dados dentro dos centros de dados.

  As transmissões intra-DCN dependem da comunicação

mecanismo dentro do data center (ou seja, em placas de conexão física, chips, memórias internas de servidores de dados, arquiteturas de rede de data centers e protocolos de comunicação). Um data center consiste em vários racks de servidores integrados interligados com suas redes internas de conexão. Atualmente, as redes de conexão interna da maioria dos data centers são estruturas fat-tree, de duas ou três camadas, baseadas em fluxos de rede multicommodity [51, 54]. Na estrutura topológica de duas camadas, os racks são conectados por switches de rack superior de 1 Gbps (TOR) e, em seguida, esses switches de rack superior são conectados com switches de agregação de 10 Gbps na estrutura topológica. A estrutura topológica de três camadas é uma estrutura aumentada com uma camada no topo da estrutura topológica de duas camadas e tal camada é constituída por switches centrais de 10 Gbps ou 100 Gbps para conectar switches de agregação na estrutura topológica. Existem também outras estruturas topológicas que visam melhorar as redes de data centers [55-58]. Devido à inadequação dos comutadores de pacotes eletrônicos, é difícil aumentar as larguras de banda de comunicação enquanto o consumo de energia é baixo. Ao longo dos anos, devido ao grande sucesso alcançado pelas tecnologias ópticas, a interconexão óptica entre as redes em data centers tem despertado grande interesse. A interconexão óptica é uma solução de alto rendimento, baixo atraso e baixo consumo de energia. Atualmente, as tecnologias ópticas são usadas apenas para links ponto a ponto em data centers. Esses links ópticos fornecem conexão para os switches usando a fibra multimodo (MMF) de baixo custo com taxa de dados de 10 Gbps. A interconexão óptica (switching no domínio óptico) de redes em data centers é uma solução viável, que pode fornecer largura de banda de transmissão em nível de Tbps com baixo consumo de energia. Recentemente, muitos planos de interconexão óptica são propostos para redes de data centers [59]. Alguns planos adicionam caminhos ópticos para atualizar as redes existentes e outros planos substituem completamente os switches atuais [59–64]. Como uma tecnologia de fortalecimento, Zhou et al. em [65] adota links sem fio na banda de frequência de 60 GHz para fortalecer os links com fio. A virtualização da rede também deve ser considerada para melhorar a eficiência e a utilização das redes do data center.

3.2.3 Pré-processamento de dados

Devido à grande variedade de fontes de dados, os conjuntos de dados coletados variam em relação a ruído, redundância e consistência, etc., e sem dúvida é um desperdício armazenar dados sem significado. Além disso, alguns métodos analíticos têm requisitos sérios na qualidade dos dados. Portanto, para permitir uma análise de dados eficaz, devemos pré-processar os dados em muitas circunstâncias para integrar os dados de diferentes fontes, o que pode não apenas reduzir as despesas de armazenamento, mas também melhorar a precisão da análise. Algumas técnicas de pré-processamento de dados relacionais são discutidas a seguir.

– *Integração:* a integração de dados é a pedra angular da informática comercial moderna, que envolve a combinação de dados de diferentes fontes e fornece aos usuários uma visão uniforme dos dados [66]. Este é um campo de pesquisa maduro para banco de dados tradicional. Historicamente, dois métodos têm sido amplamente reconhecidos: data ware house e data federation. O armazenamento de dados inclui um processo denominado ETL (Extrair, Transformar e Carregar).

A extração envolve conectar sistemas de origem, selecionar, coletar, analisar e processar os dados necessários. A transformação é a execução de uma série de regras
para transformar os dados extraídos em formatos padrão.
Carregar significa importar dados extraídos e transformados para a infraestrutura de armazenamento de destino. O carregamento é o procedimento mais complexo entre os três, que inclui operações como transformação, cópia, limpeza, padronização, triagem e organização dos dados.
Um banco de dados virtual pode ser construído para consultar e agregar dados de diferentes fontes de dados, mas esse banco de dados não contém dados. Pelo contrário, inclui informações ou metadados relacionados a dados reais e suas posições.
Essas duas abordagens de “leitura de armazenamento” não satisfazem os requisitos de alto desempenho de fluxos de dados ou programas e aplicativos de pesquisa. Em comparação com as consultas, os dados nessas duas abordagens são mais dinâmicos e devem ser processados durante a transmissão de dados. Geralmente, os métodos de integração de dados são acompanhados por mecanismos de processamento de fluxo e mecanismos de busca [30, 67].

– *Limpeza:* a limpeza de dados é um processo para identificar dados imprecisos, incompletos ou irracionais e, em seguida, modificar ou excluir esses dados para melhorar a qualidade dos dados.
Geralmente, a limpeza de dados inclui cinco procedimentos complementares [68]: definir e determinar tipos de erro, pesquisar e identificar erros, corrigir erros, documentar exemplos de erros e tipos de erros e modificar procedimentos de entrada de dados para reduzir erros futuros.

Durante a limpeza, formatos de dados, integridade, racionalidade e restrição devem ser inspecionados. A limpeza de dados é de vital importância para manter a consistência dos dados, sendo amplamente aplicada em diversas áreas, como bancos, seguros, varejo, telecomunicações e controle de tráfego.

No e-commerce, a maioria dos dados é coletada eletronicamente, o que pode trazer sérios problemas de qualidade de dados. Os problemas clássicos de qualidade de dados vêm principalmente de defeitos de software, erros personalizados ou configuração incorreta do sistema. Os autores em [69] discutiram a limpeza de dados

no comércio eletrônico por rastreadores e recopiando regularmente as informações do cliente e da conta.

Em [70], o problema de limpeza de dados RFID foi examinado. O RFID é amplamente utilizado em muitas aplicações, por exemplo, gerenciamento de estoque e rastreamento de alvos. No entanto, o RFID original apresenta baixa qualidade, o que inclui muitos dados anormais limitados pelo design físico e afetados por ruídos ambientais.

Em [71], um modelo de probabilidade foi desenvolvido para lidar com a perda de dados em ambientes móveis. Khoussainova et al. em [72] propuseram um sistema para corrigir automaticamente erros de dados de entrada definindo restrições de integridade global.

Herbert et ai. [73] propuseram uma estrutura chamada BIO AJAX para padronizar os dados biológicos, de modo a realizar cálculos adicionais e melhorar a qualidade da pesquisa. Com o BIO-AJAX, alguns erros e repetições podem ser eliminados e as tecnologias comuns de mineração de dados podem ser executadas com mais eficiência.

– *Eliminação de redundância:* redundância de dados refere-se a repetições ou excedentes de dados, que geralmente ocorrem em muitos conjuntos de dados. A redundância de dados pode aumentar a despesa desnecessária de transmissão de dados e causar defeitos nos sistemas de armazenamento, por exemplo, desperdício de espaço de armazenamento, levando à inconsistência de dados, redução da confiabilidade dos dados e danos aos dados. Portanto, vários métodos de redução de redundância foram propostos, como detecção de redundância, filtragem de dados e compactação de dados.
Esses métodos podem ser aplicados a diferentes conjuntos de dados ou ambientes de aplicativos. No entanto, a redução da redundância também pode trazer alguns efeitos negativos. Por exemplo, compactação e descompactação de dados causam carga computacional adicional. Portanto, os benefícios da redução de redundância e o custo devem ser cuidadosamente balanceados. Os dados coletados em diferentes campos aparecerão cada vez mais em formatos de imagem ou vídeo.

É bem conhecido que imagens e vídeos contêm redundância considerável, incluindo redundância temporal, redundância espacial, redundância estatística e redundância de detecção. A compactação de vídeo é amplamente usada para reduzir a redundância nos dados de vídeo, conforme especificado em vários padrões de codificação de vídeo (MPEG-2, MPEG-4, H.263 e H.264/AVC). Em [74], os autores investigaram o problema de compressão de vídeo em um sistema de videovigilância com uma rede de sensores de vídeo. Os autores propõem um novo método baseado em MPEG-4, investigando a redundância contextual relacionada ao fundo e ao primeiro plano em uma cena. A baixa complexidade e a baixa taxa de compressão da abordagem proposta foram demonstradas pelos resultados da avaliação.

Na transmissão ou armazenamento generalizado de dados, a exclusão repetida de dados é uma compactação de dados especial tecnologia, que visa eliminar cópias repetidas de dados [75]. Com a exclusão repetida de dados, blocos de dados individuais ou segmentos de dados serão atribuídos com identificadores (por exemplo, usando um algoritmo de hash) e armazenados, com os identificadores adicionados à lista de identificação. Como o anal

O processo de exclusão repetida de dados continua, se um novo bloco de dados tiver um identificador idêntico ao listado

na lista de identificação, o novo bloco de dados será considerado redundante e será substituído pelo bloco de dados armazenado correspondente. A exclusão repetida de dados pode reduzir bastante a necessidade de armazenamento, o que é particularmente importante para um sistema de armazenamento de big data. Além dos métodos de pré-processamento de dados mencionados acima, objetos de dados específicos devem passar por algumas outras operações, como extração de recursos. Tal operação desempenha um papel importante na busca multimídia e análise de DNA [76-78]. Normalmente, vetores de recursos de alta dimensão (ou pontos de recursos de alta dimensão) são usados para descrever tais objetos de dados e o sistema armazena os vetores de recursos dimensionais para recuperação futura. Dados

a transferência é geralmente usada para processar fontes de dados heterogêneas distribuídas, especialmente conjuntos de dados de negócios [79]. De fato, considerando vários conjuntos de dados, não é trivial, ou impossível, construir um procedimento de pré-processamento de dados uniforme e uma tecnologia que seja aplicável a todos os tipos de conjuntos de dados. sobre o recurso específico, problema, requisitos de desempenho e outros fatores dos conjuntos de dados devem ser considerados, de modo a selecionar uma estratégia de pré-processamento de dados adequada.

4 Grande armazenamento de dados

O crescimento explosivo de dados tem requisitos mais rígidos de armazenamento e gerenciamento. Nesta seção, nos concentramos no armazenamento de big data. O armazenamento de big data refere-se ao armazenamento e gerenciamento de conjuntos de dados em grande escala, ao mesmo tempo em que se obtém confiabilidade e disponibilidade de acesso aos dados. Analisaremos questões importantes, incluindo sistemas de armazenamento massivo, sistemas de armazenamento distribuído e mecanismos de armazenamento de big data. Por um lado, a infraestrutura de armazenamento precisa fornecer serviço de armazenamento de informações com espaço de armazenamento confiável; por outro lado, deve fornecer uma poderosa interface de acesso para consulta e análise de uma grande quantidade de dados.

Tradicionalmente, como equipamento auxiliar do servidor, o dispositivo de armazenamento de dados é usado para armazenar, gerenciar, consultar e analisar dados com RDBMSs estruturados. Com o crescimento acentuado de dados, o dispositivo de armazenamento de dados está se tornando cada vez mais importante, e muitas empresas de Internet buscam grande capacidade de armazenamento para serem competitivas. Portanto, há uma necessidade imperiosa de pesquisa sobre armazenamento de dados.

4.1 Sistema de armazenamento para dados massivos

Vários sistemas de armazenamento surgem para atender às demandas de dados massivos. As tecnologias de armazenamento massivo existentes podem ser classificadas como Direct Attached Storage (DAS) e armazenamento de rede, enquanto o armazenamento de rede pode ser ainda classificado em Network Attached Storage (NAS) e Storage Area Network (SAN).

No DAS, vários discos rígidos são conectados diretamente a servidores e o gerenciamento de dados é centrado no servidor, de modo que os dispositivos de armazenamento são equipamentos periféricos, cada um dos quais ocupa uma certa quantidade de recursos de E/S e é gerenciado por um software aplicativo individual. Por esta razão, o DAS é adequado apenas para interconectar servidores de pequena escala. Como

Porém, devido à sua baixa escalabilidade, o DAS apresentará eficiência indesejável quando a capacidade de armazenamento for aumentada, ou seja, a capacidade de atualização e expansão são muito limitadas. Assim, o DAS é usado principalmente em computadores pessoais e computadores de pequeno porte.
servidores.

O armazenamento em rede é utilizado para fornecer aos usuários uma interface de união para acesso e compartilhamento de dados. O equipamento de armazenamento de rede inclui equipamentos especiais de troca de dados, matriz de disco, biblioteca de toques e outras mídias de armazenamento, bem como software de armazenamento especial. É caracterizado com forte capacidade de expansão.

NAS é na verdade um equipamento de armazenamento auxiliar de uma rede. Ele é conectado diretamente a uma rede por meio de um hub ou switch por meio de protocolos TCP/IP. No NAS, os dados são transmitidos na forma de arquivos. Comparado ao DAS, a carga de E/S em um servidor NAS é bastante reduzida, pois o servidor acessa um dispositivo de armazenamento indiretamente por meio de uma rede.

Enquanto o NAS é orientado para a rede, o SAN é especialmente projetado para armazenamento de dados com uma rede escalável e com uso intensivo de largura de banda, por exemplo, uma rede de alta velocidade com conexões de fibra ótica. Em SAN, o gerenciamento de armazenamento de dados é relativamente independente dentro de uma rede de área local de armazenamento, onde a comutação de dados baseada em caminhos múltiplos entre quaisquer nós internos é utilizada para atingir um grau máximo de compartilhamento e gerenciamento de dados.

A partir da organização de um sistema de armazenamento de dados, DAS, NAS e SAN podem ser divididos em três partes: (i) array de discos: é a base de um sistema de armazenamento e a garantia fundamental para o armazenamento de dados; (ii) subsistemas de conexão e rede, que fornecem conexão entre um ou mais conjuntos de discos e servidores; (iii) software de gerenciamento de armazenamento, que lida com compartilhamento de dados, recuperação de desastres e outras tarefas de gerenciamento de armazenamento de vários servidores.

4.2 Sistema de armazenamento distribuído

O primeiro desafio trazido pelo big data é como desenvolver um sistema de armazenamento distribuído em grande escala para processamento e análise de dados com eficiência. Para usar um distribuído sistema para armazenar dados massivos, os seguintes fatores devem ser levados em consideração:

– *Consistência:* um sistema de armazenamento distribuído requer vários servidores para armazenar dados cooperativamente. Como há mais servidores, a probabilidade de falhas do servidor será maior. Normalmente, os dados são divididos em várias partes para serem armazenados em diferentes servidores para garantir a disponibilidade em caso de falha do servidor. No entanto, falhas de servidor e armazenamento paralelo podem causar inconsistência entre diferentes cópias dos mesmos dados. Consistência refere-se a garantir que várias cópias dos mesmos dados sejam idênticas.

– *Disponibilidade:* um sistema de armazenamento distribuído opera em vários conjuntos de servidores. À medida que mais servidores são usados, as falhas do servidor são inevitáveis. Seria desejável se

todo o sistema não é seriamente afetado para satisfazer as solicitações dos clientes em termos de leitura e escrita. Essa propriedade é chamada de disponibilidade.

– *Tolerância à partição:* vários servidores em um sistema de armazenamento distribuído são conectados por uma rede. A rede pode ter falhas de link/nó ou congestionamento temporário. O sistema distribuído deve ter um certo nível de tolerância a problemas causados por falhas de rede. Seria desejável que o armazenamento distribuído ainda funcionasse bem quando a rede fosse particionada.

Eric Brewer propôs uma teoria CAP [80, 81] em 2000, que indicava que um sistema distribuído não poderia atender simultaneamente aos requisitos de consistência, disponibilidade e tolerância de partição; no máximo dois dos três requisitos podem ser satisfeitos simultaneamente. Seth Gilbert e Nancy Lynch, do MIT, provaram a correção da teoria CAP em 2002. Como a consistência, a disponibilidade e a tolerância à partição não podem ser alcançadas simultaneamente, podemos ter um sistema CA ignorando a tolerância à partição, um sistema CP ignorando a disponibilidade e um sistema AP que ignora a consistência, de acordo com diferentes objetivos de projeto. Os três sistemas são discutidos a seguir.

Os sistemas CA não têm tolerância de partição, ou seja, eles não poderiam lidar com falhas de rede. Portanto, os sistemas CA são geralmente considerados como sistemas de armazenamento com um único servidor, como os tradicionais bancos de dados relacionais de pequena escala. Esses sistemas apresentam uma cópia única dos dados, de modo que a consistência é facilmente garantida. A disponibilidade é garantida pelo excelente design dos bancos de dados relacionais. No entanto, uma vez que os sistemas CA não conseguiam lidar com falhas de rede, eles não podiam ser expandidos para usar muitos servidores. Portanto, a maioria dos sistemas de armazenamento em grande escala são sistemas CP e sistemas AP.

Em comparação com os sistemas CA, os sistemas CP garantem tolerância à partição. Portanto, os sistemas CP podem ser expandidos para se tornarem sistemas distribuídos. Os sistemas CP geralmente mantêm várias cópias dos mesmos dados para garantir uma

nível de tolerância a falhas. Os sistemas CP também garantem a consistência dos dados, ou seja, várias cópias dos mesmos dados são completamente idênticas. No entanto, o CP não conseguiu garantir uma boa disponibilidade devido ao alto custo de garantia de consistência. Portanto, os sistemas CP são úteis para o cenário ios com carga moderada, mas requisitos rigorosos de precisão de dados (por exemplo, dados comerciais). BigTable e Hbase são dois sistemas CP populares.

Os sistemas AP também garantem a tolerância à partição. No entanto, os sistemas AP são diferentes dos sistemas CP porque os sistemas AP também garantem a disponibilidade. No entanto, os sistemas AP garantem apenas consistência eventual, em vez de consistência forte nos dois sistemas anteriores. Portanto, os sistemas AP se aplicam apenas aos cenários com solicitações frequentes, mas não com requisitos muito altos de precisão. Por exemplo, em sistemas online de Serviços de Redes Sociais (SRS), há muitas visitas simultâneas aos dados, mas uma certa quantidade de erros de dados é tolerável. Além disso, como os sistemas AP garantem consistência eventual, dados precisos ainda podem ser obtidos após um certo atraso. Portanto, os sistemas AP também podem ser usados em circunstâncias sem requisitos rigorosos de tempo real. Dynamo e Cassandra são dois sistemas AP populares.

4.3 Mecanismo de armazenamento para big data

Pesquisas consideráveis sobre big data promovem o desenvolvimento de mecanismos de armazenamento para big data. Os mecanismos existentes de armazenamento de big data podem ser classificados em três níveis de baixo para cima: (i) sistemas de arquivos, (ii) bancos de dados e (iii) modelos de programação.

Os sistemas de arquivos são a base dos aplicativos nos níveis superiores. O GFS do Google é um sistema de arquivos distribuído expansível para suportar aplicativos de grande escala, distribuídos e com uso intensivo de dados [25]. A GFS usa servidores de commodities baratos para obter tolerância a falhas e fornece aos clientes serviços de alto desempenho. O GFS oferece suporte a aplicativos de arquivo em grande escala com leitura mais frequente do que gravação. No entanto, o GFS também possui algumas limitações, como um único ponto de falha e baixo desempenho para arquivos pequenos. Tais limitações foram superadas pelo Colossus [82], o sucessor do GFS.

Além disso, outras empresas e pesquisadores também possuem suas soluções para atender as diferentes demandas de armazenamento de big data. Por exemplo, HDFS e Kosmosfs são derivados de códigos open source de GFS. A Microsoft desenvolveu o Cosmos [83] para dar suporte a seus negócios de busca e publicidade. O Facebook utiliza o Haystack [84] para armazenar a grande quantidade de fotos de tamanho pequeno. Taobao também desenvolveu TFS e FastDFS. Em conclusão, os sistemas de arquivos distribuídos estão relativamente maduros após anos de desenvolvimento e operação comercial. Portanto, vamos nos concentrar nos outros dois níveis no restante desta seção.

4.3.1 Tecnologia de banco de dados

A tecnologia de banco de dados vem evoluindo há mais de 30 anos. Vários sistemas de banco de dados são desenvolvidos para lidar com conjuntos de dados em diferentes escalas e dar suporte a vários aplicativos. Bancos de dados relacionais tradicionais não podem atender ao desafio lenges em categorias e escalas trazidas por big data.
Bancos de dados NoSQL (ou seja, bancos de dados relacionais não tradicionais) estão se tornando mais populares para armazenamento de big data. Os bancos de dados NoSQL apresentam modos flexíveis, suporte para cópia simples e fácil, API simples, consistência eventual e suporte para dados de grande volume. Os bancos de dados NoSQL estão se tornando a principal tecnologia para big data. Nesta seção, examinaremos os três principais bancos de dados NoSQL: bancos de dados chave-valor, bancos de dados orientados a colunas e bancos de dados orientados a documentos, cada um baseado em determinados modelos de dados.

– *Bancos de dados de valores-chave:* Os bancos de dados de valores-chave são constituídos por um modelo de dados simples e os dados são armazenados correspondentes aos valores-chave. Cada chave é única e os clientes podem inserir valores consultados de acordo com as chaves. Esses bancos de dados apresentam uma estrutura simples e os bancos de dados chave-valor modernos são caracterizados por alta capacidade de expansão e tempo de resposta de consulta mais curto do que os bancos de dados relacionais. Nos últimos anos, muitos bancos de dados de valor-chave apareceram motivados pelo sistema Dynamo da Amazon [85]. Apresentaremos o Dynamo e vários outros bancos de dados de chave-valor representativos.

  – *Dynamo:* Dynamo é um armazenamento de dados de valor-chave distribuído altamente disponível e expansível. Ele é usado para o status de alguns armazenar e gerenciar o sistema de idade. serviços principais, que podem ser realizados com acesso por chave, na plataforma de comércio eletrônico da Amazon. O modo público de bancos de dados relacionais pode gerar dados inválidos e limitar a escala e a disponibilidade de dados, enquanto o Dynamo pode resolver esses problemas com uma interface simples de objeto-chave, que é constituída por operações simples de leitura e gravação. O Dynamo alcança elasticidade e disponibilidade por meio dos mecanismos de partição de dados, cópia de dados e edição de objetos. O plano de partição do Dynamo depende do Consistent Hashing [86], que tem a principal vantagem de que a passagem de nó afeta apenas nós diretamente adjacentes e não afeta outros nós, para dividir a carga para várias máquinas de armazenamento principal. O Dynamo copia dados para N conjuntos de servidores, nos quais N é um parâmetro configurável para atingir

alta disponibilidade e durabilidade. O sistema Dynamo também fornece consistência eventual, de modo a realizar atualização assíncrona em todas as cópias.

– *Voldemort:* Voldemort também é um sistema de armazenamento de valor-chave, que foi inicialmente desenvolvido e ainda é usado pelo LinkedIn. Palavras-chave e valores em Voldemort são objetos compostos constituídos por tabelas e imagens. A interface Volde mort inclui três operações simples: leitura, escrita e exclusão, todas confirmadas por palavras-chave. Volde mort fornece atualização assíncrona com controle atual de várias edições, mas não garante consistência de dados. No entanto, o Volde mort oferece suporte ao bloqueio otimista para atualização consistente de vários registros. Quando ocorrer um conflito entre a atualização e quaisquer outras operações, a operação de atualização será encerrada. O mecanismo de cópia de dados do Voldmort é o mesmo do Dynamo. Voldemort não apenas armazena dados na RAM, mas permite que os dados sejam inseridos em um mecanismo de armazenamento. Especialmente, Voldemort suporta dois mecanismos de armazenamento, incluindo Berkeley DB e Random Access Files.

O banco de dados de valor-chave surgiu há alguns anos. Profundamente influenciado pelo Amazon Dynamo DB, outros sistemas de armazenamento de valor-chave incluem Redis, Tokyo Canbinet e Tokyo Tyrant, Memcached e Memcache DB, Riak e Scalaris, todos os quais fornecem capacidade de expansão distribuindo palavras-chave em nós. Voldemort, Riak, Tokyo Cabinet e Memecached podem utilizar dispositivos de armazenamento conectados para armazenar dados em RAM ou discos. Outros sistemas de armazenamento armazenam dados na RAM e fornecem backup em disco ou contam com cópia e recuperação para evitar backup.

– *Banco de dados orientado a colunas:* O banco de dados orientado a colunas bancos de dados armazenam e processam dados de acordo com colunas diferentes de linhas. Tanto as colunas quanto as linhas são segmentadas em vários nós para permitir a capacidade de expansão. Os bancos de dados orientados a colunas são inspirados principalmente no BigTable do Google. Nesta seção, primeiro discutimos o BigTable e depois apresentamos várias ferramentas derivadas.

– *BigTable:* BigTable é um sistema de armazenamento de dados estruturado e distribuído, projetado para processar os dados em grande escala (classe PB) entre milhares de servidores comerciais [87]. A estrutura de dados básica do Bigtable é um mapeamento sequenciado multidimensional com armazenamento esparso, distribuído e persistente. Índices de mapeamento são chave de linha, chave de coluna e timestamps, e cada valor no mapeamento é uma matriz de bytes não analisados. Cada chave de linha no BigTable é uma cadeia de caracteres de 64 KB. Por ordem lexicográfica, as linhas são armazenadas e continuamente segmentadas em Tablets (ou seja, unidades de distribuição) para balanceamento de carga. Assim, a leitura de uma pequena linha de dados pode ser altamente eficaz, pois envolve apenas a comunicação com uma pequena parte das máquinas. As colunas são agrupadas de acordo com os prefixos das chaves, formando assim famílias de colunas. Essas famílias de colunas são as unidades básicas para controle de acesso. Os timestamps são números inteiros de 64 bits para distinguir diferentes edições de valores de células. Os clientes podem determinar com flexibilidade o número de edições de células armazenadas. Essas edições são sequenciadas em ordem decrescente de timestamps, então a última edição sempre será lida.

A API do BigTable também permite a criação e exclusão de tablets e famílias de colunas como modificação de metadados de clusters, tabelas e famílias de colunas. Os aplicativos cliente podem inserir ou excluir valores de BigTable, consultar valores de colunas ou navegar por subconjuntos de dados em uma tabela. O Bigtable também oferece suporte a algumas outras características, como o processamento de transações em uma única linha. Os usuários podem utilizar esses recursos para realizar processamentos de dados mais complexos.

Cada procedimento executado pelo BigTable inclui três componentes principais: servidor mestre, servidor tablet e biblioteca cliente.

O Bigtable permite que apenas um conjunto de servidor Master seja distribuído para ser responsável pela distribuição de tablets para o servidor Tablet, detectando servidores Tablet adicionados ou removidos e conduzindo o balanceamento de carga. Além disso, ele também pode modificar o esquema do BigTable, por exemplo, criando tabelas e famílias de colunas e coletando lixo salvo no GFS, bem como arquivos deletados ou desabilitados, e usando-os em instâncias específicas do BigTable.

Cada servidor tablet gerencia um conjunto Tablet e é responsável pela leitura e escrita de um Tablet carregado. Quando os Tablets são muito grandes, eles serão segmentados pelo servidor. A biblioteca do cliente do aplicativo é usada para se comunicar com as instâncias do BigTable.

BigTable é baseado em muitos componentes fundamentais do Google, incluindo GFS [25], sistema de gerenciamento de cluster, arquivo SSTable para mat e Chubby [88]. O GFS é usado para armazenar dados e arquivos de log. O sistema de gerenciamento de cluster é responsável pelo agendamento de tarefas, compartilhamento de recursos, processamento de falhas de máquinas e monitoramento de status de máquinas. O formato de arquivo SSTable é usado para armazenar dados do BigTable internamente,

e fornece mapeamento entre chaves e valores persistentes, sequenciados e imutáveis como quaisquer cadeias de bytes. BigTable utiliza Chubby para as seguintes tarefas no servidor: 1) garantir que haja no máximo uma cópia Master ativa a qualquer momento; 2) armazenar o local de bootstrap dos dados do BigTable; 3) procure o servidor Tablet; 4) conduzir a recuperação de erros em caso de falha do servidor de Mesa; 5) armazenar informações do esquema BigTable; 6) armazenar a tabela de controle de acesso.

– *Cassandra:* Cassandra é um sistema de armazenamento distribuído para gerenciar a enorme quantidade de dados estruturados distribuídos entre vários servidores comerciais [89]. O sistema foi desenvolvido pelo Facebook e tornou-se uma ferramenta de código aberto em 2008. Adota as ideias e conceitos do Amazon Dynamo e do Google BigTable, integrando especialmente a tecnologia de sistema distribuído do Dynamo com o modelo de dados do BigTable. Tabelas em Cassandra estão na forma de

ping de mapa estruturado quadridimensional distribuído, onde as quatro dimensões, incluindo linha, coluna, família de colunas e supercoluna. Uma linha é distinguida por uma chave de string com comprimento arbitrário. Não importa a quantidade de colunas a serem lidas ou escritas, a operação nas linhas é automática. As colunas podem constituir clusters, chamados de famílias de colunas, e são semelhantes ao modelo de dados do Bigtable. O Cassandra fornece dois tipos de famílias de colunas: famílias de colunas e supercolunas. A supercoluna inclui um número arbitrário de colunas relacionadas aos mesmos nomes. Uma família de colunas inclui colunas e supercolunas, que podem ser continuamente inseridas na família de colunas durante o tempo de execução. Os mecanismos de partição e cópia do Cassandra são muito semelhantes aos do Dynamo, para obter consistência.

– *Ferramentas derivadas do BigTable:* como o código do BigTable não pode ser obtido por meio da licença de código aberto, alguns projetos de código aberto competem para implementar o conceito de BigTable para desenvolver sistemas semelhantes, como HBase e Hypertable.

HBase é uma versão clonada do BigTable programada com Java e faz parte do framework Hadoop of Apache MapReduce [90]. O HBase substitui o GFS pelo HDFS. Ele grava conteúdos atualizados na RAM e os grava regularmente em arquivos em discos. As operações de linha são operações atômicas, equipadas com bloqueio em nível de linha e processamento de transação, que é opcional para grande escala. A partição e a distribuição são operadas de forma transparente e possuem espaço para hash do cliente ou chave fixa.

O HyperTable foi desenvolvido de forma semelhante ao BigTable para obter um conjunto de sistemas de armazenamento e processamento distribuídos, expansíveis e de alto desempenho para dados estruturados e não estruturados [91]. O HyperTable depende de sistemas de arquivos distribuídos, por exemplo, HDFS e gerenciador de bloqueio distribuído. A representação, o processamento e o mecanismo de partição de dados são semelhantes aos do BigTable. A HyperTable tem sua própria linguagem de consulta, chamada HyperTable query language (HQL), e permite que os usuários criem, modifiquem e consultem tabelas subjacentes.

Como os bancos de dados de armazenamento orientados a colunas emulam principalmente o BigTable, seus designs são todos semelhantes, exceto pelo mecanismo de simultaneidade e vários outros recursos. Por exemplo, Cassandra enfatiza consistência fraca de controle simultâneo de várias edições, enquanto HBase e HyperTable focam em consistência forte por meio de bloqueios ou registros de log.

– *Banco de dados de documentos:* Comparado com o armazenamento de valor-chave, o armazenamento de documentos pode suportar formulários de dados mais complexos. Como os documentos não seguem modos estritos, não há necessidade de conduzir a migração de modo. Além disso, os pares chave-valor ainda podem ser salvos. Examinaremos três importantes representantes dos sistemas de armazenamento de documentos, ou seja, MongoDB, SimpleDB e CouchDB.

– *MongoDB:* MongoDB é um banco de dados de código aberto e orientado a documentos [92]. O MongoDB armazena documentos como objetos Binary JSON (BSON) [93], que é semelhante a object. Cada documento tem um campo de ID como chave primária.
A consulta no MongoDB é expressa com syn tax semelhante ao JSON. Um driver de banco de dados envia a consulta como um objeto BSON para o MongoDB.
O sistema permite a consulta de todos os documentos, incluindo objetos embutidos e arrays. Para permitir a consulta rápida, os índices podem ser criados nos campos consultáveis dos documentos. A operação de cópia no MongoDB pode ser executada com arquivos de log nos nós principais que suportam todas as operações de alto nível realizadas no banco de dados. Durante a cópia, os slavers consultam todas as operações de escrita desde a última sincronização ao mestre e executam operações em arquivos de log em bancos de dados locais. O MongoDB oferece suporte à expansão horizontal com compartilhamento automático para distribuir dados entre milhares de nós, equilibrando automaticamente a carga e o failover.

– *SimpleDB:* SimpleDB é um banco de dados distribuído e é um serviço web da Amazon [94]. Os dados no SimpleDB são organizados em vários domínios nos quais os dados podem ser armazenados, adquiridos e consultados. Os domínios incluem diferentes propriedades e conjuntos de pares nome/valor de projetos.
A data é copiada para diferentes máquinas em diferentes centros de dados para garantir a segurança dos dados e melhorar o desempenho. Este sistema não suporta partição automática e, portanto, não pode ser expandido com a alteração do volume de dados. SimpleDB permite aos usuários consultar com SQL. Vale a pena notar que o SimpleDB pode garantir a consistência eventual, mas não oferece suporte ao Muti-Version Concurrency Control (MVCC). Portanto, os conflitos nele contidos não puderam ser detectados do lado do cliente.

– *CouchDB:* Apache CouchDB é um banco de dados orientado a documentos escrito em Erlang [95]. Os dados no CouchDB são organizados em documentos que consistem em campos nomeados por chaves/nomes e valores, que são armazenados e acessados como objetos JSON. Cada documento é fornecido com um identificador único. O CouchDB permite acesso a documentos de banco de dados por meio da API RESTful HTTP. Se um documento precisar ser modificado definido, o cliente deve baixar o documento inteiro para modificá-lo e depois enviá-lo de volta ao banco de dados. Depois que um documento é reescrito uma vez, o identificador será atualizado. O CouchDB utiliza a cópia ideal para obter escalabilidade sem um mecanismo de compartilhamento. Uma vez que vários CouchDBs podem ser executados junto com outras transações simultaneamente, qualquer tipo de topologia de replicação pode ser construída. A consistência do CouchDB depende do mecanismo de cópia. O CouchDB suporta MVCC com registros históricos de Hash.

Big data geralmente são armazenados em centenas e até milhares de servidores comerciais. Assim, os modelos paralelos tradicionais, como Message Passing Interface (MPI) e Open Multi-Processing (OpenMP), podem não ser adequados para suportar tais programas paralelos de larga escala. Recentemente, alguns modelos de programação paralela propostos melhoram efetivamente o desempenho do NoSQL e reduzem a lacuna de desempenho para bancos de dados relacionais. Portanto, esses modelos se tornaram a pedra angular para a análise de dados massivos.

– *MapReduce:* MapReduce [22] é um modelo de programação simples, mas poderoso, para computação em larga escala, usando um grande número de clusters de PCs comerciais para obter processamento e distribuição paralelos automáticos. No MapReduce, o modelo de computação possui apenas dois funções, ou seja, Mapear e Reduzir, ambas programadas pelos usuários. A função Map processa pares chave-valor de entrada e gera pares chave-valor intermediários. Em seguida, o MapReduce irá combinar todos os valores intermediários relacionados à mesma chave e transmiti-los para a função Reduce, que comprime ainda mais o valor definido em um conjunto menor. O MapReduce tem a vantagem de evitar as etapas complicadas para desenvolver aplicativos paralelos, por exemplo, escalonamento de dados, tolerância a falhas e comunicações entre nós. O usuário só precisa programar as duas funções para desenvolver uma aplicação paralela. A estrutura inicial do MapReduce não suportava vários conjuntos de dados em uma tarefa, o que foi mitigado por alguns aprimoramentos recentes [96, 97].

Nas últimas décadas, os programadores estão familiarizados com a linguagem declarativa avançada do SQL, frequentemente usada em um banco de dados relacional, para descrição de tarefas e análise de conjunto de dados. No entanto, a estrutura sucinta do MapReduce fornece apenas duas funções não transparentes, que não podem cobrir todas as operações comuns. Portanto, os programadores precisam gastar tempo programando as funções básicas, que normalmente são difíceis de manter e reutilizar. Para melhorar a eficiência da programação, alguns sistemas avançados de linguagem foram propostos, por exemplo, Sawzall [98] do Google, Pig Latin [99] do Yahoo, Hive [100] do Facebook e Scope [87] da Microsoft.

– *Dryad:* Dryad [101] é um mecanismo de execução distribuída de uso geral para processamento de aplicativos paralelos de dados de baixa granularidade. A estrutura operacional do Dryad é um grafo acíclico direcionado, no qual os vértices representam os programas e as arestas representam os canais de dados.
O Dryad executa operações nos vértices em clusters e transmite dados por meio de canais de dados, incluindo documentos, conexões TCP e FIFO de memória compartilhada.
Durante a operação, os recursos em um gráfico de operação lógica são mapeados automaticamente para recursos físicos.

A estrutura de operação do Dryad é coordenada por um programa central chamado gerenciador de tarefas, que pode ser executado em clusters ou estações de trabalho via rede. Um gerenciador de trabalho consiste em duas partes: 1) códigos de aplicativo que são usados para construir um gráfico de comunicação de trabalho e 2) códigos de biblioteca de programa que são usados para organizar os recursos disponíveis. Todos os tipos de dados são transmitidos diretamente entre os vértices. Portanto, o gerente do trabalho é responsável apenas pela tomada de decisões, o que não impede nenhuma transmissão de dados.

No Dryad, os desenvolvedores de aplicativos podem escolher com flexibilidade qualquer grafo acíclico direcionado para descrever os modos de comunicação do aplicativo e expressar os mecanismos de transmissão de dados. Além disso, o Dryad permite que os vértices usem qualquer quantidade de dados de entrada e saída, enquanto o MapReduce suporta apenas um conjunto de entrada e saída.

DryadLINQ [102] é a linguagem avançada do Dryad e é usada para integrar o já mencionado ambiente de execução de linguagem semelhante a SQL.

– *All-Pairs:* All-Pairs [103] é um sistema especialmente projetado para aplicações de biometria, bioinformática e mineração de dados. Ele se concentra na comparação de pares de elementos em dois conjuntos de dados por uma determinada função. Todos os pares podem ser expressos como três tuplas (Conjunto A, Conjunto B e Função F), em que a Função F é utilizada para comparar todos os elementos no Conjunto A e no Conjunto B. O resultado da comparação é uma matriz de saída M, que é também chamado de produto cartesiano ou junção cruzada do Conjunto A e do Conjunto B.

All-Pairs é implementado em quatro fases: modelagem do sistema, distribuição de dados de entrada, gerenciamento de tarefas em lote e coleta de resultados. Na Fase I, um modelo aproximado de desempenho do sistema será construído para avaliar quanto recurso de CPU é necessário e como

conduzir partição de trabalho. Na Fase II, uma spanning tree é construída para transmissões de dados, o que faz com que a carga de trabalho de cada partição recupere os dados de entrada de forma eficaz. Na Fase III, após o fluxo de dados ser entregue aos nós apropriados, o mecanismo All-Pairs criará uma submissão de processamento em lote para trabalhos em partições, enquanto os sequenciará no sistema de processamento em lote e formulará um comando de execução de nó para adquirir dados. Na última fase, após a conclusão do trabalho do sistema de processamento em lote, o mecanismo de extração irá coletar os resultados e combiná-los em uma estrutura própria, que geralmente é uma única lista de arquivos, na qual todos os resultados são colocados em ordem.

– *Pregel:* O sistema Pregel [104] do Google facilita o processamento de gráficos de grande porte, por exemplo, análise de gráficos de rede e serviços de redes sociais. Uma tarefa computacional é expressa por um gráfico direcionado constituído por vértices e arestas direcionadas. Cada vértice está relacionado a um valor modificável e definido pelo usuário, e cada aresta direcionada relacionada a um vértice de origem é constituída pelo valor definido pelo usuário e o identificador de um vértice de destino. Quando o gráfico é construído, o programa realiza cálculos iterativos, chamados de superetapas, entre as quais os pontos de sincronização globais são definidos até a conclusão do algoritmo e a conclusão da saída.

Em cada superetapa, os cálculos de vértice são paralelos e cada vértice executa a mesma função definida pelo usuário para expressar uma determinada lógica de algoritmo. Cada vértice pode modificar seu estado e de suas arestas de saída, receber uma mensagem enviada do superstep anterior, enviar a mensagem para outros vértices e até mesmo modificar a estrutura topológica de todo o grafo. As arestas não são fornecidas com cálculos correspondentes. As funções de cada vértice podem ser removidas por suspensão. Quando todos os vértices estão em estado inativo sem nenhuma mensagem para transmitir, toda a execução do programa é concluída.

A saída do programa Pregel é um conjunto que consiste nos valores de saída de todos os vértices. De um modo geral, a entrada e a saída do programa Pregel são gráficos direcionados isomórficos.

Inspirados nos modelos de programação acima, outras pesquisas também focaram em modos de programação para tarefas computacionais mais complexas, por exemplo, cálculos iterativos [105, 106], cálculos de memória tolerante a falhas [107], cálculos incrementais [108] e controle de fluxo tomada de decisão relacionada aos dados [109].

A análise de big data envolve principalmente métodos analíticos para dados tradicionais e big data, arquitetura analítica para big data e software usado para mineração e análise de big data. A análise de dados é a fase final e mais importante na cadeia de valor do big data, com o objetivo de extrair valores úteis, fornecer sugestões ou decisões.

Diferentes níveis de valores potenciais podem ser gerados através da análise de conjuntos de dados em diferentes campos [10]. No entanto, a análise de dados é uma área ampla, que muda frequentemente e é extremamente complexa. Nesta seção, apresentamos os métodos, arquiteturas e ferramentas para análise de big data.

4.4 Análise de dados tradicional

A análise de dados tradicional significa usar métodos estatísticos adequados para analisar dados massivos, concentrar, extrair e refinar dados úteis ocultos em um lote de conjuntos de dados caóticos e identificar a lei inerente do assunto, de modo a maximizar o valor dos dados . A análise de dados desempenha um grande papel de orientação na elaboração de planos de desenvolvimento para um país, na compreensão das demandas comerciais dos clientes e na previsão de tendências de mercado para empresas. A análise de big data pode ser considerada como a técnica de análise para um tipo especial de dados.
Portanto, muitos métodos tradicionais de análise de dados ainda podem ser utilizados para análise de big data. Vários métodos tradicionais representativos de análise de dados são examinados a seguir, muitos dos quais são de estatística e ciência da computação.

– *Análise de cluster:* é um método estatístico para agrupar objetos e, especificamente, classificá-los de acordo com algumas características. A análise de cluster é usada para diferenciar objetos com características particulares e dividi-los em algumas categorias (clusters) de acordo com essas características, de modo que objetos da mesma categoria tenham alta homogeneidade enquanto categorias diferentes terão alta heterogeneidade. A análise de agrupamento é um método de estudo não supervisionado sem dados de treinamento.

– *Análise Fatorial:* visa basicamente descrever a relação entre muitos elementos com apenas alguns fatores, ou seja, agrupar várias variáveis intimamente relacionadas em um fator, e os poucos fatores são então utilizados para revelar o máximo de informações dos dados originais.

– *Análise de Correlação:* é um método analítico para determinar a lei das relações, tais como correlação, dependência correlativa e restrição mútua, entre os fenômenos observados e, conseqüentemente, conduzir a previsão e controle. Tais relações podem ser classificadas em dois tipos: (i) função, refletindo a relação de dependência estrita entre os fenômenos, também chamada de relação de dependência definitiva; (ii) correlação, algumas relações de dependência indeterminadas ou inexatas, e o valor numérico de uma variável pode corresponder a vários valores numéricos da outra variável, e tais valores numéricos apresentam uma flutuação regular em torno de seus valores médios.

– *Análise de Regressão:* é uma ferramenta matemática para revelar correlações entre uma variável e várias outras variáveis. Com base em um conjunto de experimentos ou dados observados, a análise de regressão identifica relações de dependência entre variáveis ocultas pela aleatoriedade. A análise de regressão pode tornar correlações complexas e indeterminadas entre variáveis simples e regulares.

– *Teste A/B:* também chamado de teste de balde. É uma tecnologia para determinar como melhorar as variáveis-alvo comparando o grupo testado. Big data exigirá um grande número de testes para serem executados e analisados.

– *Análise Estatística:* A análise estatística é baseada na teoria estatística, um ramo da matemática aplicada. Na teoria estatística, a aleatoriedade e a incerteza são modeladas com a Teoria da Probabilidade. A análise estatística pode fornecer uma descrição e uma inferência para big data. A análise estatística descritiva pode resumir e descrever conjuntos de dados, enquanto a análise estatística inferencial pode tirar conclusões de dados sujeitos a variações aleatórias. A análise estatística é amplamente aplicada nos campos econômico e de assistência médica [110].

– *Algoritmos de Mineração de Dados:* A mineração de dados é um processo para extrair informações e conhecimentos ocultos, desconhecidos, mas potencialmente úteis, de dados massivos, incompletos, ruidosos, difusos e aleatórios. Em 2006, a IEEE International Conference on Data Mining Series (ICDM) identificou dez algoritmos de mineração de dados mais influentes por meio de um procedimento de seleção rigoroso [111], incluindo C4.5, k-means, SVM, Apriori, EM, Naive Bayes e Cart, etc. Esses dez algoritmos cobrem classificação, agrupamento, regressão, aprendizado estatístico, análise de associação e mineração de links, todos os quais são os problemas mais importantes na pesquisa de mineração de dados.

4.5 Métodos analíticos de big data

No alvorecer da era do big data, as pessoas estão preocupadas em como extrair rapidamente informações importantes de dados massivos, de modo a agregar valor para empresas e indivíduos. Atualmente, os principais métodos de processamento de big data são mostrados a seguir.

– *Bloom Filter:* Bloom Filter consiste em uma série de Hash funções. O princípio do Bloom Filter é armazenar valores Hash de dados que não sejam os próprios dados, utilizando uma matriz de bits, que é, em essência, um índice de bitmap que usa funções Hash para conduzir o armazenamento de compactação com perdas de dados. Ele tem vantagens como alta eficiência de espaço e alta velocidade de consulta, mas também tem algumas desvantagens em reconhecimento e exclusão incorretos.

– *Hashing:* é um método que essencialmente transforma dados em valores numéricos de comprimento fixo mais curtos ou valores de índice. Hashing tem vantagens como leitura rápida, gravação e alta velocidade de consulta, mas é difícil encontrar uma função Hash sólida.

– *Índice:* o índice é sempre um método eficaz para reduzir a despesa de leitura e gravação em disco e melhorar a velocidade de inserção, exclusão, modificação e consulta em bancos de dados relacionais tradicionais que gerenciam dados estruturados e outras tecnologias que gerenciam semiestruturados e não estruturados dados. No entanto, index tem a desvantagem de ter o custo adicional para armazenar arquivos de índice que devem ser mantidos dinamicamente quando os dados são atualizados.

– *Triel:* também chamada de árvore trie, uma variante de Hash Tree. É aplicado principalmente para estatísticas de recuperação rápida e frequência de palavras. A ideia principal do Triel é utilizar prefixos de cadeias de caracteres para reduzir ao máximo a comparação em cadeias de caracteres, de modo a melhorar a eficiência da consulta.

– *Computação Paralela:* em comparação com a computação serial tradicional, a computação paralela refere-se à utilização simultânea de vários recursos de computação para concluir uma tarefa de computação. Sua ideia básica é decompor um problema e atribuí-lo a vários processos separados para serem concluídos independentemente, de modo a alcançar o coprocessamento. Atualmente, alguns modelos clássicos de computação paralela incluem MPI (Message Passing Interface), MapReduce e Dryad (veja uma comparação na Tabela 1).

Embora os sistemas ou ferramentas de computação paralela, como MapReduce ou Dryad, sejam úteis para análise de big data, eles são ferramentas de baixo nível difíceis de aprender e usar. Portanto, algumas ferramentas ou linguagens de programação paralela de alto nível estão sendo desenvolvidas com base nesses sistemas. Essas linguagens de alto nível incluem Sawzall, Pig e Hive usadas para MapReduce, bem como Scope e DryadLINQ usadas para Dryad.

4.6 Arquitetura para análise de big data

Por causa dos 4Vs de big data, diferentes arquiteturas analíticas devem ser consideradas para diferentes requisitos de aplicação.

Tabela 1 Comparação entre MPI, MapReduce e Dryad

	MPI	MapReduce	Dríade
Implantação	Nó de computação e dados armazenamento organizado separadamente (Os dados devem ser movidos do nó de computação)	Computação e armazenamento de dados dispostos no mesmo nó (A computação deve estar próxima dos dados)	Computação e armazenamento de dados dispostos no mesmo nó (A computação deve estar próxima dos dados)
Gestão de recursos/ agendamento	–	Workqueue(google) HOD (Yahoo)	Não está claro
Programação de baixo nível	API MPI	API MapReduce	API Dryad
programação de alto nível	–	Porco, Colmeia, Jaql, ⋯	Escopo, DryadLINQ
Armazenamento de dados	O sistema de arquivos local, NFS, ⋯	GFS(google) HDFS (Hadoop), KFS Amazon S3, ⋯	NTFS, Cosmos DFS
Particionamento de tarefas	Partição manual do usuário as tarefas	Automação	Automação
Comunicação	Mensagens, Remoto acesso à memória	Arquivos (FS local, DFS)	Arquivos, TCP Pipes, FIFOs de memória compartilhada
Tolerante a falhas	ponto de verificação	Tarefa reexecutada	Tarefa reexecutada

4.6.1 Análise em tempo real vs. offline

De acordo com os requisitos de pontualidade, a análise de big data pode ser classificada em análise em tempo real e análise off-line.

– *Análise em tempo real:* é usada principalmente em comércio eletrônico e finanças. Uma vez que os dados mudam constantemente, é necessária uma análise rápida dos dados e os resultados analíticos devem ser devolvidos com um atraso muito curto. As principais arquiteturas existentes de análise em tempo real incluem (i) clusters de processamento paralelo usando bancos de dados relacionais tradicionais e (ii) plataformas de computação baseadas em memória. Por exemplo, Greenplum da EMC e HANA da SAP são arquiteturas de análise em tempo real.

– *Análise off-line:* geralmente é usada para aplicações com requisitos elevados de tempo de resposta, por exemplo, aprendizado de máquina, análise estatística e algoritmos de recomendação. A análise off-line geralmente realiza análises importando logs para uma plataforma especial por meio de ferramentas de aquisição de dados. Sob a configuração de big data, muitas empresas de Internet utilizam a arquitetura de análise off-line baseada em Hadoop para reduzir o custo de conversão de formato de dados e melhorar a eficiência da aquisição de dados. Exemplos incluem a ferramenta de código aberto do Facebook Scribe, a ferramenta de código aberto do LinkedIn Kafka, a ferramenta de código aberto Taobao Timetunnel e Chukwa do Hadoop, etc. Essas ferramentas podem atender às demandas de aquisição e transmissão de dados com centenas de MB por segundo.

4.6.2 Análise em diferentes níveis

A análise de big data também pode ser classificada em análise de nível de memória, análise de nível de Business Intelligence (BI) e análise de nível massivo, que são examinadas a seguir.

– *Análise de nível de memória:* é para o caso em que o volume total de dados é menor que a memória máxima de um cluster. Hoje em dia, a memória do cluster de servidor ultrapassa centenas de GB enquanto até mesmo o nível de TB é comum. Portanto, uma tecnologia de banco de dados interno pode ser usada e os dados quentes devem residir na memória para melhorar a eficiência analítica. A análise de nível de memória é extremamente adequada para análise em tempo real.
O MongoDB é uma arquitetura analítica de nível de memória representativa. Com o desenvolvimento do SSD (Solid-State Drive), a capacidade e o desempenho da análise de dados no nível da memória foram aprimorados e amplamente aplicados.

– *Análise de BI:* é para o caso em que a escala de dados ultrapassa o nível de memória, mas pode ser importada para o ambiente de análise de BI. Atualmente, os principais produtos de BI são fornecidos com planos de análise de dados para suportar o nível de TB.

– *Análise massiva:* é para o caso quando a escala de dados ultrapassou completamente as capacidades de produtos de BI e bancos de dados relacionais tradicionais. Atualmente, a maioria das análises massivas utiliza HDFS do Hadoop para armazenar dados e usa MapReduce para análise de dados. A maioria das análises massivas pertence à categoria de análise offline.

4.6.3 Análise com diferentes complexidades

A complexidade de tempo e espaço dos algoritmos de análise de dados difere muito entre si de acordo com diferentes tipos de dados e demandas de aplicativos. Por exemplo, para aplicações que são passíveis de processamento paralelo, um algoritmo distribuído pode ser projetado e um modelo de processamento paralelo pode ser usado para análise de dados.

4.7 Ferramentas para mineração e análise de big data

Muitas ferramentas para mineração e análise de big data estão disponíveis, incluindo software profissional e amador, software comercial caro e software de código aberto. Nesta seção, revisamos brevemente os cinco softwares mais usados, de acordo com uma pesquisa de “Qual software de análise, mineração de dados e big data você usou nos últimos 12 meses para um projeto real?” de 798 profissionais feita pela KDNuggets em 2012 [112].

– *R* (30,7%): R, uma linguagem de programação e ambiente de software de código aberto, é projetado para mineração/análise e visualização de dados. Enquanto tarefas intensivas de computação são executadas, o código programado com C, C++ e Fortran pode ser chamado no ambiente R. Além disso, usuários qualificados podem chamar objetos R diretamente em C.
Na verdade, R é uma realização da linguagem S, que é uma linguagem interpretada desenvolvida pela AT&T Bell Labs e usada para exploração de dados, análise estatística e desenho de gráficos. Comparado ao S, o R é mais popular por ser de código aberto. R ocupa o primeiro lugar na pesquisa KDNuggets 2012. Além disso, em uma pesquisa de “Linguagens de design que você usou para mineração/análise de dados no ano passado” em 2012, R também ficou em primeiro lugar, derrotando SQL e Java. Devido à popularidade do R, fabricantes de banco de dados, como Teradata e Oracle, lançaram produtos que suportam R.

– *Excel* (29,8%): o Excel, um componente central do Microsoft Office, fornece recursos avançados de processamento de dados e análise estatística. Quando o Excel é instalado, alguns plug-ins avançados, como Analysis ToolPak e Solver Add-in, com funções poderosas para análise de dados são integrados inicialmente, mas esses plug-ins podem ser usados apenas se os usuários os habilitarem. O Excel também é o único software comercial entre os cinco primeiros.

– *Rapid-I Rapidminer* (26,7%): Rapidminer é um software de código aberto usado para mineração de dados, aprendizado de máquina e análise preditiva. Em uma investigação do KDnuggets em 2011, ele foi usado com mais frequência do que o R (classificado como Top 1). Os programas de mineração de dados e aprendizado de máquina fornecidos pelo RapidMiner incluem Extrair, Transformar e Carregar (ETL), pré-processamento e visualização de dados, modelagem, avaliação e implantação.
O fluxo de mineração de dados é descrito em XML e exibido por meio de uma interface gráfica do usuário (GUI). Rapid Miner é escrito em Java. Ele integra o método de aprendizado e avaliação do Weka e funciona com o R. As funções do Rapidminer são implementadas com conexão de processos incluindo vários operadores. Todo o fluxo pode ser considerado como uma linha de produção de uma fábrica, com entrada de dados originais e saída de resultados do modelo. Os operadores podem ser considerados como algumas funções específicas com diferentes características de entrada e saída.

– *KNMINE* (21,8%): KNIME (Konstanz Information Miner) é uma plataforma de integração de dados, processamento de dados, análise de dados e mineração de dados rica em código aberto, inteligente e fácil de usar [113] . Ele permite que os usuários criem fluxos de dados ou canais de dados de maneira visualizada, executem seletivamente alguns ou todos os procedimentos analíticos e forneçam resultados analíticos, modelos e visualizações interativas. O KNIME foi escrito em Java e, baseado no Eclipse, fornece mais funções como plug-ins. Por meio de arquivos plug-in, os usuários podem inserir módulos de processamento para arquivos, imagens e séries temporais e integrá-los em vários projetos de código aberto, por exemplo, R e Weka. O KNIME controla a integração de dados, limpeza, conversão, filtragem, estatísticas, mineração e, finalmente, visualização de dados.

Todo o processo de desenvolvimento é conduzido sob um ambiente visualizado. O KNIME foi projetado como uma estrutura expansível e baseada em módulos. Não há dependência entre suas unidades de processamento e recipientes de dados, tornando-o adaptável ao ambiente distribuído e desenvolvimento independente. Além disso, é fácil expandir o KNIME. Os desenvolvedores podem expandir facilmente vários nós e visualizações do KNIME.

– *Weka/ Pentaho* (14,8%): Weka, abreviado de Waikato Environment for Knowledge Analysis, é um software de aprendizado de máquina e mineração de dados gratuito e de código aberto escrito em Java. Weka fornece funções como processamento de dados, seleção de recursos, classificação, regressão, agrupamento, regra de associação e visualização, etc.
Pentaho é um dos softwares de BI de código aberto mais populares. Inclui uma plataforma de servidor web e várias ferramentas para dar suporte a relatórios, análises, gráficos, integração de dados e mineração de dados, etc., todos os aspectos do BI. Os algoritmos de processamento de dados da Weka também estão integrados no Pentaho e podem ser chamados diretamente.

5 aplicações de big data

Na seção anterior, examinamos a análise de big data, que é a fase final e mais importante da cadeia de valor de big data. A análise de big data pode fornecer valores úteis por meio de julgamentos, sugestões, suportes ou decisões.

No entanto, a análise de dados envolve uma ampla gama de aplicações, que mudam frequentemente e são extremamente complexas. Nesta seção, primeiro revisamos a evolução das fontes de dados. Em seguida, examinamos seis dos campos de análise de dados mais importantes, incluindo análise de dados estruturados, análise de texto, análise de sites, análise de multimídia, análise de rede e análise móvel. Por fim, apresentamos vários campos de aplicação importantes de big data.

5.1 Principais aplicações de big data

5.1.1 Evoluções do aplicativo

Recentemente, a análise de big data foi proposta como uma tecnologia analítica avançada, que normalmente inclui programas complexos e de grande escala sob métodos analíticos específicos. Na verdade, aplicativos baseados em dados surgiram nas últimas décadas. Por exemplo, já na década de 1990, o BI tornou-se uma tecnologia predominante para aplicativos de negócios e os mecanismos de pesquisa de rede baseados no processamento massivo de mineração de dados surgiram no início do século XXI. Algumas aplicações potenciais e influentes de diferentes campos e seus dados e características de análise são discutidos a seguir.

– *Evolução dos Aplicativos Comerciais:* Os primeiros dados de negócios geralmente eram dados estruturados, coletados por empresas de sistemas legados e armazenados em RDBMSs. As técnicas analíticas usadas em tais sistemas prevaleciam na década de 1990 e eram intuitivas e simples, por exemplo, nas formas de relatórios, painel de controle, consultas com condição, inteligência de negócios baseada em pesquisa, processamento de transações on-line, visualização interativa, cartões de pontuação, preditiva modelagem e mineração de dados [114]. Desde o início do século XXI, as redes e a World Wide Web (WWW) têm proporcionado uma oportunidade única para que as organizações tenham uma exibição online e interajam diretamente com os clientes. Produtos abundantes e informações de clientes, como registros de dados de clickstream e comportamento do usuário, podem ser adquiridos na WWW. A otimização do layout do produto, a análise comercial do cliente, as sugestões de produtos e a análise da estrutura do mercado podem ser realizadas por meio de técnicas de análise de texto e mineração de sites. Conforme relatado em [115], a quantidade de telefones celulares e tablet PC ultrapassou pela primeira vez a de laptops e PCs em 2011. Os telefones celulares e a Internet das Coisas baseados em sensores estão abrindo uma nova geração de aplicações de inovação e exigindo uma capacidade consideravelmente maior de apoiar a detecção de localização, orientação para as pessoas e operação sensível ao contexto.
– *Evolução das Aplicações de Rede:* A primeira geração da Internet fornecia principalmente e-mail e serviços WWW. Análise de texto, mineração de dados e a análise de páginas da Web foi aplicada à mineração de conteúdo de e-mail e à construção de mecanismos de pesquisa. Hoje em dia, a maioria das aplicações são baseadas na web, independentemente de seu campo e objetivos de design. Os dados de rede respondem por uma grande porcentagem do volume global de dados. A Web tornou-se uma plataforma comum para páginas interconectadas, cheias de vários tipos de dados, como texto, imagens, áudio, vídeos e conteúdos interativos, etc. Portanto, surgiu uma abundância de tecnologias avançadas usadas para dados semiestruturados ou não estruturados no momento certo. Por exemplo, a análise de imagens pode extrair informações úteis das imagens (por exemplo, reconhecimento facial). As tecnologias de análise multimídia podem ser aplicadas a sistemas automatizados de vigilância por vídeo para aplicações comerciais, policiais e militares. Desde 2004, mídias sociais online, como fóruns na Internet, comunidades online, blogs, serviços de redes sociais e sites sociais multimídia, oferecem aos usuários grandes oportunidades para criar, carregar e compartilhar conteúdos.
– *Evolução das Aplicações Científicas:* A pesquisa científica em muitos campos está adquirindo dados massivos com sensores e instrumentos de alto rendimento, como astrofísica, oceanologia, genômica e pesquisa ambiental. A US National Science Foundation (NSF) anunciou recentemente o programa BIGDATA para promover esforços para extrair conhecimento e insights de grandes e complexas coleções de dados digitais. Algumas disciplinas de pesquisa científica desenvolveram plataformas de big data e obtiveram resultados úteis. Por exemplo, em biologia, o iPlant [116] aplica infraestrutura de rede, recursos físicos de computação, ambiente de coordenação, recursos de máquinas virtuais, software de análise interoperacional e serviço de dados para auxiliar pesquisadores, educadores e estudantes no enriquecimento das ciências vegetais. Os conjuntos de dados do iPlant têm grandes variedades na forma, incluindo especificações ou dados de referência, dados experimentais, dados de log ou modelo, dados de observação e outros dados derivados.

Conforme discutido, podemos dividir a pesquisa de análise de dados em seis campos técnicos principais, ou seja, análise de dados estruturados, análise de dados de texto, análise de dados da web, análise de dados multimídia, análise de dados de rede e análise de dados móveis. Essa classificação visa enfatizar as características dos dados, mas alguns dos campos podem utilizar tecnologias básicas semelhantes. Como a análise de dados tem um escopo amplo e não é fácil ter uma cobertura abrangente, vamos nos concentrar nos principais problemas e tecnologias na análise de dados nas discussões a seguir.

5.1.2 Análise de dados estruturados

Aplicações de negócios e pesquisas científicas podem gerar dados estruturados massivos, cuja gestão e análise dependem de tecnologias maduras comercializadas, como RDBMS, data warehouse, OLAP e BPM (Business Process Management) [28] . A análise de dados é baseada principalmente em mineração de dados e análise estatística, ambas bem estudadas nos últimos 30 anos.

No entanto, a análise de dados ainda é um campo de pesquisa muito ativo e novas demandas de aplicação impulsionam o desenvolvimento de novos métodos. Por exemplo, aprendizado de máquina estatístico baseado em modelos matemáticos exatos e algoritmos poderosos foram aplicados à detecção de anomalias [117] e controle de energia [118]. Explorando as características dos dados, a mineração de tempo e espaço pode extrair estruturas de conhecimento escondidas em fluxos de dados e sensores de alta velocidade [119]. Impulsionado pela proteção de privacidade em aplicativos de comércio eletrônico, governo eletrônico e assistência médica, a mineração de dados de proteção de privacidade é um campo de pesquisa emergente [120]. Na última década, a mineração de processos está se tornando um novo campo de pesquisa, especialmente na análise de processos com dados de eventos [121].

5.1.3 Análise de dados de texto

O formato mais comum de armazenamento de informações é o texto, por exemplo, e-mails, documentos comerciais, páginas da web e mídias sociais. Portanto, considera-se que a análise de texto apresenta mais potencial baseado em negócios do que dados estruturados. Geralmente, a análise de texto é um processo para extrair informações e conhecimentos úteis de um texto não estruturado. A mineração de texto é interdisciplinar, envolvendo recuperação de informações, aprendizado de máquina, estatística, linguística computacional e mineração de dados em particular. A maioria dos sistemas de mineração de texto são baseados em expressões de texto e processamento de linguagem natural (NLP), com mais ênfase no último. O NLP permite que os computadores analisem, interpretem e até gerem texto. Alguns métodos comuns de PNL incluem aquisição lexical, desambiguação de sentido de palavra, marcação de parte da fala e gramática livre de contexto probabilístico [122]. Algumas técnicas baseadas em NLP foram aplicadas à mineração de texto, incluindo extração de informações, modelos de tópicos, resumo de texto, classificação, agrupamento, resposta a perguntas e mineração de opinião.

5.1.4 Análise de dados da Web

A análise de dados da Web emergiu como um campo de pesquisa ativo. Ele visa recuperar, extrair e avaliar automaticamente informações de documentos e serviços da Web para cobrir conhecimentos úteis. A análise da Web está relacionada a vários campos de pesquisa, incluindo banco de dados, recuperação de informações, NLP e mineração de texto. De acordo com as diferentes partes ser minerados, classificamos a análise de dados da Web em três campos relacionados: mineração de conteúdo da Web, mineração de estrutura da Web e mineração de uso da Web [123].

A mineração de conteúdo da Web é o processo para descobrir conhecimento útil em páginas da Web, que geralmente envolvem vários tipos de dados, como texto, imagem, áudio, vídeo, código, metadados e hiperlink. A pesquisa em mineração de imagem, áudio e vídeo foi recentemente chamada de análise multimídia, que será discutida na Seção 6.1.5. Já que a maioria

Os dados de conteúdo da Web são dados de texto não estruturados, a pesquisa sobre análise de dados da Web concentra-se principalmente em texto e hipertexto. A mineração de texto é discutida na Seção 6.1.3, enquanto a mineração de hipertexto envolve a mineração de arquivos HTML semiestruturados que contêm hiperlinks. O aprendizado supervisionado e a classificação desempenham papéis importantes na mineração de hiperlinks, por exemplo, e-mail, gerenciamento de grupos de notícias e manutenção de catálogos da Web [124]. A mineração de conteúdo da Web pode ser realizada com dois métodos: o método de recuperação de informações e o banco de dados método. A recuperação de informações principalmente auxilia ou melhora a pesquisa de informações ou filtra as informações do usuário de acordo com deduções ou documentos de configuração. O método de banco de dados visa simular e integrar dados na Web, de forma a realizar consultas mais complexas do que pesquisas baseadas em palavras-chave.

A mineração de estruturas da Web envolve modelos para descobrir estruturas de links da Web. Aqui, a estrutura refere-se aos diagramas esquemáticos vinculados em um site ou entre vários sites. Os modelos são construídos com base em estruturas topológicas providas de hiperlinks com ou sem descrição de link. Tais modelos revelam as semelhanças e correlações entre diferentes sites e são usados para classificar as páginas do site. Page Rank [125] e CLEVER [126] fazem pleno uso dos modelos para procurar páginas de sites relevantes. Rastreador orientado a tópicos é outro caso de sucesso ao utilizar os modelos [127].

A mineração de uso da Web visa extrair dados auxiliares gerados por diálogos ou atividades na Web. A mineração de conteúdo da Web e a mineração de estrutura da Web usam os dados principais da Web. Os dados de uso da Web incluem logs de acesso a servidores Web e servidores proxy, registros históricos de navegadores, perfis de usuário, dados de registro, sessões ou negociações de usuários, cache, consultas de usuários, dados de favoritos, cliques e rolagens do mouse e quaisquer outros tipos de dados gerados por meio de interação com a Web. À medida que os serviços da Web e a Web 2.0 estão se tornando maduros e populares, os dados de uso da Web terão uma variedade cada vez maior. A mineração de uso da Web desempenha papéis importantes em espaço personalizado, comércio eletrônico, privacidade/segurança de rede e outros campos emergentes. Por exemplo, sistemas de recomendação colaborativos podem personalizar o comércio eletrônico utilizando as diferentes preferências de Usuários.

5.1.5 Análise de dados multimídia

Os dados multimédia (incluindo principalmente imagens, áudio e vídeos) têm vindo a crescer a uma velocidade espantosa, onde se extraem conhecimentos úteis e se compreendem os semantemas por análise. Como os dados multimídia são heterogêneos e a maioria desses dados contém informações mais ricas do que sim

Plenos dados estruturados ou dados textuais, a extração de informação confronta-se com o enorme desafio das diferenças semânticas. A pesquisa sobre análise de multimídia abrange muitas disciplinas. Algumas prioridades de pesquisa recentes incluem resumo de multimídia, anotação de multimídia, índice e recuperação de multimídia, sugestão de multimídia e detecção de eventos de multimídia, etc.

A sumarização de áudio pode ser realizada extraindo as palavras ou frases proeminentes dos metadados ou sintetizando uma nova representação. A sumarização de vídeo serve para interpretar a sequência de conteúdo de vídeo mais importante ou representativa, podendo ser estática ou dinâmica. Os métodos de resumo de vídeo estático utilizam uma sequência de quadros-chave ou quadros-chave sensíveis ao contexto para representar um vídeo. Esses métodos são simples e têm sido aplicados a muitos aplicativos de negócios (por exemplo, pelo Yahoo, AltaVista e Google), mas seu desempenho é ruim. Os métodos de resumo dinâmico usam uma série de quadros de vídeo para representar um vídeo e tomam outras medidas suaves para fazer o resumo final

aparência mais natural. Em [128], os autores propõem um sistema de sumarização multimídia orientado a tópicos (TOMS) que pode resumir automaticamente as informações importantes em um vídeo pertencente a uma determinada área de tópicos, com base em um determinado conjunto de recursos extraídos do vídeo.

A anotação multimídia insere rótulos para descrever o conteúdo de imagens e vídeos nos níveis de sintaxe e semântica. Com tais rótulos, o gerenciamento, resumo e recuperação de dados multimídia podem ser facilmente implementados.
Como a anotação manual é demorada e trabalhosa, a anotação automática sem qualquer intervenção humana torna-se altamente atraente. O principal desafio para a anotação multimídia automática é a diferença semântica.

Embora muito progresso tenha sido feito, o desempenho dos métodos de anotação automática existentes ainda precisa ser melhorado. Atualmente, muitos esforços estão sendo feitos para explorar de forma sincronizada a anotação multimídia manual e automática [129].

A indexação e recuperação de multimídia envolvem a descrição, armazenamento e organização de informações multimídia e assistência aos usuários para procurar recursos multimídia de maneira conveniente e rápida [130]. Geralmente, indexação e recuperação multimídia incluem cinco procedimentos: análise estrutural, extração de características, mineração de dados, classificação e anotação, consulta e recuperação [131]. A análise estrutural visa segmentar um vídeo em vários elementos estruturais semânticos, incluindo detecção de limite de lente, extração de quadro-chave e

segmentação, etc. De acordo com o resultado da análise estrutural, o segundo procedimento é a extração de recursos, que inclui principalmente a mineração adicional dos recursos de quadros-chave, objetos, textos e movimentos, que são a base da indexação e recuperação de vídeo. A mineração, classificação e anotação de dados devem utilizar os recursos extraídos para encontrar os modos de conteúdo de vídeo e colocar os vídeos em categorias agendadas para gerar índices de vídeo. Ao receber uma consulta, o sistema usará um método de medição de similaridade para procurar um vídeo candidato. O resultado da recuperação otimiza o feedback relacionado.

A recomendação multimédia consiste em recomendar conteúdos multimédia específicos de acordo com as preferências dos utilizadores. Está provado ser uma abordagem eficaz para fornecer serviços personalizados. A maioria dos sistemas de recomendação existentes pode ser classificada em sistemas baseados em conteúdo e sistemas baseados em filtragem colaborativa. Os métodos baseados em conteúdo identificam características gerais dos usuários ou seus interesses, e recomendam usuários para outros conteúdos com características semelhantes. Esses métodos dependem amplamente da medição de similaridade de conteúdo, mas a maioria deles é prejudicada por limitações de análise e especificações excessivas. Os métodos baseados em filtragem colaborativa identificam grupos com interesses semelhantes e recomendam conteúdos para os membros do grupo de acordo com seu comportamento [132]. Atualmente, é introduzido um método misto, que integra as vantagens dos dois tipos de métodos acima mencionados para melhorar a qualidade da recome

O Instituto Nacional de Padrões e Tecnologia dos EUA ogy (NIST) iniciou o TREC Video Retrieval Evaluation para detectar a ocorrência de um evento em videoclipes baseados no Event Kit, que contém alguma descrição de texto relacionada a conceitos e exemplos de vídeo [134]. Em [135], o autor propôs um novo algoritmo para detecção de eventos multimídia especiais usando alguns exemplos de treinamento positivo. A pesquisa sobre detecção de eventos de vídeo ainda está em sua infância e se concentra principalmente em esportes ou eventos de notícias, corrida ou eventos anormais em vídeos de monitoramento e outros eventos semelhantes com padrões repetitivos.

5.1.6 Análise de dados de rede

A análise de dados de rede evoluiu da análise quantitativa inicial [136] e análise de rede sociológica [137] para a emergente análise de rede social online no início do século XXI. Muitos serviços de redes sociais online, incluindo Twitter, Facebook e LinkedIn, etc., tornaram-se cada vez mais populares ao longo dos anos. Esses serviços de rede social on-line geralmente incluem dados vinculados maciços e dados de conteúdo. Os dados vinculados estão principalmente na forma de estruturas gráficas, descrevendo as comunicações entre duas entidades. Os dados de conteúdo contêm texto, imagem e outros dados multimídia de rede. O rico conteúdo em

tais redes trazem desafios sem precedentes

e oportunidades para análise de dados. De acordo com a perspectiva centrada em dados, a pesquisa existente sobre contextos de serviços de redes sociais pode ser classificada em duas categorias: análise estrutural baseada em links e análise baseada em conteúdo [138].

A pesquisa sobre análise estrutural baseada em links sempre foi comprometida com previsão de links, descoberta de comunidades, evolução de redes sociais e análise de influência social, etc. Os SRS podem ser visualizados como grafos, nos quais cada vértice corresponde a um usuário e as arestas correspondem às correlações entre os usuários. Como os SNS são redes dinâmicas, novos vértices e arestas são continuamente adicionados aos grafos.
A previsão de link é prever a possibilidade de conexão futura entre dois vértices. Muitas técnicas podem ser usadas para previsão de link, por exemplo, classificação baseada em recursos, métodos probabilísticos e álgebra linear. A classificação baseada em recursos é selecionar um grupo de recursos para um vértice e utilizar as informações do link existente para gerar classificadores binários para prever o link futuro [139]. Métodos probabilísticos visam construir modelos para probabilidades de conexão entre vértices no SNS [140]. A Álgebra Linear calcula a similaridade entre dois vértices de acordo com a matriz singular de similaridades [141]. Uma comunidade é representada por uma matriz subgráfica, na qual as arestas que conectam vértices no subgrafo apresentam alta densidade, enquanto as arestas entre dois subgrafos apresentam densidade muito menor [142].

Muitos métodos para detecção de comunidade foram propostos e estudados, a maioria dos quais são funções de destino baseadas em topologia, contando com o conceito de capturar a estrutura da comunidade. Du et al. utilizou a propriedade de comunidades sobrepostas na vida real para propor um método eficaz de detecção de comunidade SNS em larga escala [143]. A pesquisa em SRS visa buscar um modelo de lei e dedução para interpretar a evolução da rede. Alguns estudos empíricos descobriram que o viés de proximidade, limitações geográficas e outros fatores desempenham papéis importantes na evolução do SNS [144-146], e alguns métodos de geração são propostos para auxiliar no projeto de redes e sistemas [147].

A influência social refere-se ao caso em que os indivíduos mudar seu comportamento sob a influência de outras pessoas. A força da influência social depende da relação entre indivíduos, distâncias de rede, efeito de tempo e características de redes e indivíduos, etc. Marketing, propaganda, recomendação e outras aplicações podem se beneficiar da influência social medindo qualitativa e quantitativamente a influência de indivíduos em outros [148, 149].

Geralmente, se a proliferação de conteúdos no SNS for considerada, o desempenho da análise estrutural baseada em links pode ser melhorado.

A análise baseada em conteúdo em SRS também é conhecida como análise de mídia social. As mídias sociais incluem texto, multimídia, posicionamento e comentários. No entanto, a análise de mídia social enfrenta desafios sem precedentes. Em primeiro lugar, os dados de mídia social massivos e em constante crescimento devem ser analisados automaticamente dentro de uma janela de tempo razoável. Em segundo lugar, os dados de mídia social contêm muito ruído. Por exemplo, a blogosfera contém um grande número de blogs de spam, assim como Tweets triviais no Twitter. Em terceiro lugar, os SNS são redes dinâmicas, que variam e são atualizadas com frequência e rapidez. A pesquisa existente sobre análise de mídia social ainda está em sua infância. Considerando que o SRS contém informações massivas, o aprendizado por transferência em redes heterogêneas visa transferir informações de conhecimento entre diferentes mídias [150].

5.1.7 Análise de dados móveis

Em abril de 2013, o Android Apps forneceu mais de 650.000 aplicativos, abrangendo quase todas as categorias. Até o final de 2012, o fluxo mensal de dados móveis atingiu 885 PB [151]. Os dados massivos e os aplicativos abundantes exigem análise móvel, mas também trazem alguns desafios. Como um todo, os dados móveis têm características únicas, por exemplo, detecção móvel, flexibilidade de movimento, ruído e uma grande quantidade de redundância. Recentemente, novas pesquisas sobre análise móvel foram iniciadas em diferentes campos. Como a pesquisa sobre análise móvel está apenas começando, apresentaremos apenas algumas aplicações de análise recentes e representativas nesta seção.

Com o crescimento do número de usuários móveis e melhor desempenho, os telefones celulares agora são úteis para construir e manter comunidades, como comunidades com localizações geográficas e comunidades baseadas em diferentes origens e interesses culturais (por exemplo, o último Webchat). As comunidades de rede tradicionais ou comunidades de SNS carecem de interação on-line entre os membros, e as comunidades são ativas apenas quando os membros estão sentados diante dos computadores. Pelo contrário, os telefones celulares podem oferecer suporte a interações ricas a qualquer hora e em qualquer lugar. As comunidades móveis são definidas como um grupo de indivíduos com os mesmos hobbies (ou seja, saúde, segurança e entretenimento, etc.) se reúnem em redes, se reúnem para fazer um objetivo comum, decidem medidas por meio de consulta para atingir o objetivo e começar a implementar seu plano [152]. Em [153], os autores propuseram um modelo qualitativo de uma comunidade móvel. Atualmente, acredita-se amplamente que os aplicativos da comunidade móvel promoverão muito o desenvolvimento da indústria móvel.

Recentemente, o progresso em sensor sem fio, tecnologia de comunicação móvel e processamento de fluxo permite que as pessoas construam uma rede de área corporal para monitorar em tempo real a saúde das pessoas. Geralmente, os dados médicos de vários sensores têm características diferentes em termos de atributos, relações de tempo e espaço, bem como relações fisiológicas, etc.

Além disso, tais conjuntos de dados envolvem privacidade e proteção de segurança. Em [154], Garg et al. introduzir um mecanismo de análise de transporte multimodal de dados brutos para monitoramento em tempo real da saúde. Sob a circunstância de que apenas características altamente abrangentes relacionadas à saúde estão disponíveis, Park et al. em [155] examinou abordagens para melhor utilização.

Pesquisadores da Gjovik University College na Noruega e Derawi Biometrics colaboraram para desenvolver um aplicativo para smartphones, que analisa os passos quando as pessoas andam e usa as informações de ritmo para desbloquear o sistema de segurança [11]. Enquanto isso, Robert Delano e Brian Parise, do Georgia Institute of Technology, desenvolveram um aplicativo chamado iTrem, que monitora o tremor do corpo humano com um sismógrafo embutido em um telefone celular, de modo a lidar com Parkinson e outras doenças do sistema nervoso [11].

5.2 Principais aplicações de big data

5.2.1 Aplicação de big data nas empresas

Atualmente, big data vem principalmente e é usado principalmente em empresas, enquanto BI e OLAP podem ser considerados os predecessores da aplicação de big data. A aplicação de big data nas empresas pode melhorar sua eficiência de produção e competitividade em muitos aspectos. Em particular, em marketing, com análise de correlação de big data, as empresas podem prever com mais precisão o comportamento do consumidor e encontrar novos modos de negócios. No planejamento de vendas, após a comparação de dados massivos, as empresas podem otimizar seus preços de commodities. Na operação, as empresas podem melhorar sua eficiência operacional e satisfação, otimizar a força de trabalho, prever com precisão os requisitos de alocação de pessoal, evitar o excesso de capacidade de produção e reduzir o custo do trabalho. Na cadeia de suprimentos, usando big data, as empresas podem realizar otimização de estoque, otimização logística e coordenação de fornecedores, etc., para mitigar a lacuna entre oferta e demanda, controlar orçamentos e melhorar serviços.

Em finanças, a aplicação de big data nas empresas tem se desenvolvido rapidamente. Por exemplo, o China Merchants Bank (CMB) utiliza a análise de dados para reconhecer que atividades como "acumulação de pontuação múltipla" e "troca de pontuação em lojas" são eficazes para atrair clientes de qualidade. Ao construir um modelo de aviso de desistência do cliente, o banco pode vender produtos financeiros de alto rendimento para os 20% de clientes com maior probabilidade de desistir para retê-los. Como resultado, os índices de evasão dos clientes com Cartão Gold e Cartão Girassol foram reduzidos em 15% e 7%, respectivamente. Ao analisar os registros de transações dos clientes, os clientes de pequenas empresas em potencial podem ser identificados com eficiência. Ao utilizar serviços bancários remotos e a plataforma de referência em nuvem para implementar vendas cruzadas, ganhos consideráveis de desempenho foram alcançados.

Obviamente, a aplicação mais clássica é no comércio eletrônico. Dezenas de milhares de transações são realizadas no Taobao e o tempo de transação correspondente, preços de commodities e quantidades de compra são registrados todos os dias, e mais importante, junto com idade, sexo, endereço e até hobbies e interesses de compradores e vendedores. O Data Cube of Taobao é um aplicativo de big data na plataforma Taobao, por meio do qual os comerciantes podem conhecer o status industrial macroscópico da plataforma Taobao, as condições de mercado de suas marcas e os comportamentos dos consumidores etc. decisões de produção e estoque. Enquanto isso, mais consumidores podem comprar suas commodities favoritas com preços mais favoráveis. O empréstimo de crédito do Alibaba analisa e julga automaticamente o tempo para emprestar empréstimos a empresas por meio dos dados de transações corporativas adquiridas em virtude da tecnologia de big data, enquanto a intervenção manual não ocorre em todo o processo. É divulgado que, até agora, o Alibaba emprestou mais de RMB 30 bilhões de Yuan com apenas cerca de 0,3% de empréstimos ruins, o que é muito menor do que os de outros bancos comerciais.

5.2.2 Aplicação de big data baseada em IoT

A IoT não é apenas uma fonte importante de big data, mas também um dos principais mercados de aplicativos de big data. Devido à grande variedade de objetos, as aplicações de IoT também evoluem infinitamente.

As empresas de logística podem ter uma experiência profunda com a aplicação de big data de IoT. Por exemplo, os caminhões da UPS são equipados com sensores, adaptadores sem fio e GPS, para que a sede possa rastrear as posições dos caminhões e evitar falhas no motor. Enquanto isso, este sistema também ajuda a UPS a supervisionar e gerenciar seus funcionários e otimizar as rotas de entrega. As rotas de entrega ideais especificadas para caminhões UPS são derivadas de sua experiência de condução anterior. Em 2011, os motoristas da UPS dirigiram cerca de 48,28 milhões de km a menos.

Cidade inteligente é uma área de pesquisa quente baseada na aplicação de dados IoT. Por exemplo, a cooperação do projeto de cidade inteligente entre o Condado de Miami-Dade, na Flórida, e a IBM conecta 35 tipos de departamentos governamentais importantes do condado e a cidade de Miami e ajuda os líderes do governo a obter melhor suporte de informações na tomada de decisões para o gerenciamento de recursos hídricos, reduzindo engarrafamentos e melhorar a segurança pública. A aplicação da cidade inteligente traz benefícios em muitos aspectos para o Condado de Dade. Por exemplo, o Departamento de Gestão de Parques do Condado de Dade economizou um milhão de dólares em contas de água devido à identificação e conserto oportunos de canos de água que estavam funcionando e vazar

5.2.3 Aplicação de big data orientado a redes sociais online

SRS online é uma estrutura social constituída por indivíduos sociais e conexões entre indivíduos com base em um

rede de informações. O big data do SNS online vem principalmente de mensagens instantâneas, social online, microblog e espaço compartilhado, etc., que representam várias atividades do usuário. A análise de big data de SRS online usa método analítico computacional fornecido para entender as relações na sociedade humana em virtude de teorias e métodos, que envolvem matemática, informática, sociologia e ciência da administração, etc., de três dimensões, incluindo estrutura de rede, interação em grupo e divulgação de informações. O aplicativo inclui análise de opinião pública de rede, coleta e análise de inteligência de rede, marketing socializado, suporte à tomada de decisão do governo e educação online, etc. A Fig. 5 ilustra a estrutura técnica da aplicação de big data de SNS online.

Aplicações clássicas de big data de SNS online são apresentadas a seguir, que principalmente extraem e analisam informações de conteúdo e informações estruturais para adquirir valores.

– *Aplicativos baseados em conteúdo:* Linguagem e texto são as duas formas mais importantes de apresentação em SRS. Por meio da análise da linguagem e do texto, podem ser reveladas as preferências, emoções, interesses, demandas etc. do usuário.

– *Aplicações baseadas em estrutura:* No SRS, os usuários são representados como nós, enquanto relações sociais, interesses e hobbies, etc. agregam relações entre os usuários em uma estrutura agrupada. Tal estrutura com relações estreitas entre indivíduos internos, mas relações externas soltas também é chamado de comunidade. A análise baseada na comunidade é de vital importância para melhorar a propagação da informação e para a análise das relações interpessoais.

O Departamento de Polícia de Santa Cruz dos Estados Unidos experimentou a aplicação de dados para análise preditiva. Ao analisar o SNS, o departamento de polícia pode descobrir tendências e modos de crime e até mesmo prever as taxas de criminalidade nas principais regiões [11].

Em abril de 2013, a Wolfram Alpha, uma empresa de computação e mecanismos de busca, estudou a lei do comportamento social analisando dados sociais de mais de um milhão de usuários americanos do Facebook. De acordo com a análise, descobriu-se que a maioria dos usuários do Facebook se apaixona por volta dos 20 anos, fica noiva por volta dos 27 anos e depois se casa por volta dos 30 anos. Finalmente, seus relacionamentos conjugais apresentam mudanças lentas entre 30 e 60 anos. Tais resultados de pesquisa são altamente consistentes com os dados do censo demográfico dos EUA. Além disso, a Global Pulse realizou uma pesquisa que revelou algumas leis em atividades sociais e econômicas usando dados do SNS. Este projeto utilizou mensagens do Twitter publicamente disponíveis em inglês, japonês e indonésio de julho de 2010 a outubro de 2011, para analisar tópicos relacionados a alimentos, combustível, habitação e empréstimos. O objetivo é entender melhor o comportamento e as preocupações do público. Este projeto analisou o big data do SNS sob vários aspectos: 1) prever a ocorrência de eventos anormais, detectando o crescimento acentuad

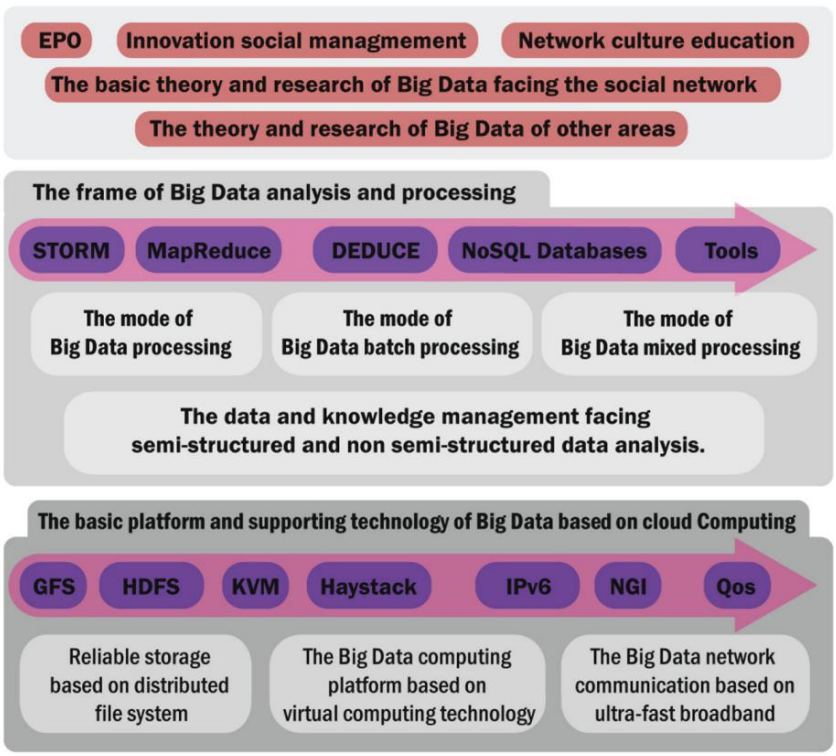

Fig. 5 Habilitando tecnologias para big data orientado a redes sociais online

ou diminuição da quantidade de tópicos, 2) observar as tendências semanais e mensais de diálogos no Twitter; desenvolvendo modelos para a variação do nível de atenção em

tópicos específicos ao longo do tempo, 3) entender as tendências de transformação do comportamento ou interesse do usuário comparando proporções de diferentes subtópicos e 4) prever tendências com indicadores externos envolvidos nos diálogos do Twitter.

Como exemplo clássico, o projeto descobriu que a variação da inflação dos preços dos alimentos nas estatísticas oficiais da Indonésia corresponde ao número de Tweets ao preço do arroz no Twitter, conforme mostrado na Fig. 6 .

De um modo geral, a aplicação de big data de SRS online pode ajudar a entender melhor o comportamento do usuário e dominar as leis de atividades sociais e econômicas de

os três seguintes aspectos:

– *Early Warning:* para lidar rapidamente com a crise, se houver, detectando anormalidades no uso de dispositivos e serviços eletrônicos.

– *Monitoramento em tempo real:* fornecer informações precisas para a formulação de políticas e planos, monitorando o comportamento atual, a emoção e a preferência dos usuários.

– *Feedback em tempo real:* obtenha feedbacks dos grupos em relação a algumas atividades sociais com base no monitoramento em tempo real.

5.2.4 Aplicações de big data médico e de saúde

Os dados médicos e de saúde são dados complexos em crescimento contínuo e rápido, contendo valores de informações abundantes e diversos. Big data tem potencial ilimitado para

armazenando, processando, consultando e analisando dados médicos de forma eficaz. A aplicação de big data médica influenciará profundamente o negócio de assistência médica.

Por exemplo, a Aetna Life Insurance Company selecionou 102 pacientes de um grupo de mil pacientes para completar um experimento a fim de ajudar a prever a recuperação de pacientes com síndrome metabólica. Em um experimento independente, escaneou 600.000 resultados de exames laboratoriais e 180.000 reclamações por meio de uma série de resultados de exames de detecção de síndrome metabólica de pacientes em três anos consecutivos.

Além disso, resumiu o resultado final em um plano de tratamento extremamente personalizado para avaliar os fatores perigosos e os principais planos de tratamento dos pacientes. Então, os médicos podem reduzir a morbidade em 50% nos próximos 10 anos, prescrevendo estatinas e ajudando os pacientes a perder peso em 2,5 quilos, ou sugerindo aos pacientes que reduzam o total de triglicerídeos em seus corpos se o teor de açúcar em seus corpos for superior a 20 .

O Mount Sinai Medical Center, nos EUA, utiliza tecnologias da Ayasdi, uma empresa de big data, para analisar todas as sequências genéticas de Escherichia Coli, incluindo mais de um milhão de variantes de DNA, para investigar por que as cepas bacterianas resistem aos antibióticos. A Ayasdi usa análise de dados topológicos, um novo método de pesquisa matemática, para entender as características dos dados.

O HealthVault da Microsoft, lançado em 2007, é um excelente aplicativo de big data médico lançado em 2007. Seu objetivo é gerenciar informações de saúde individual em dispositivos médicos individuais e familiares. Atualmente, as informações de saúde podem ser inseridas e carregadas com dispositivos móveis inteligentes e

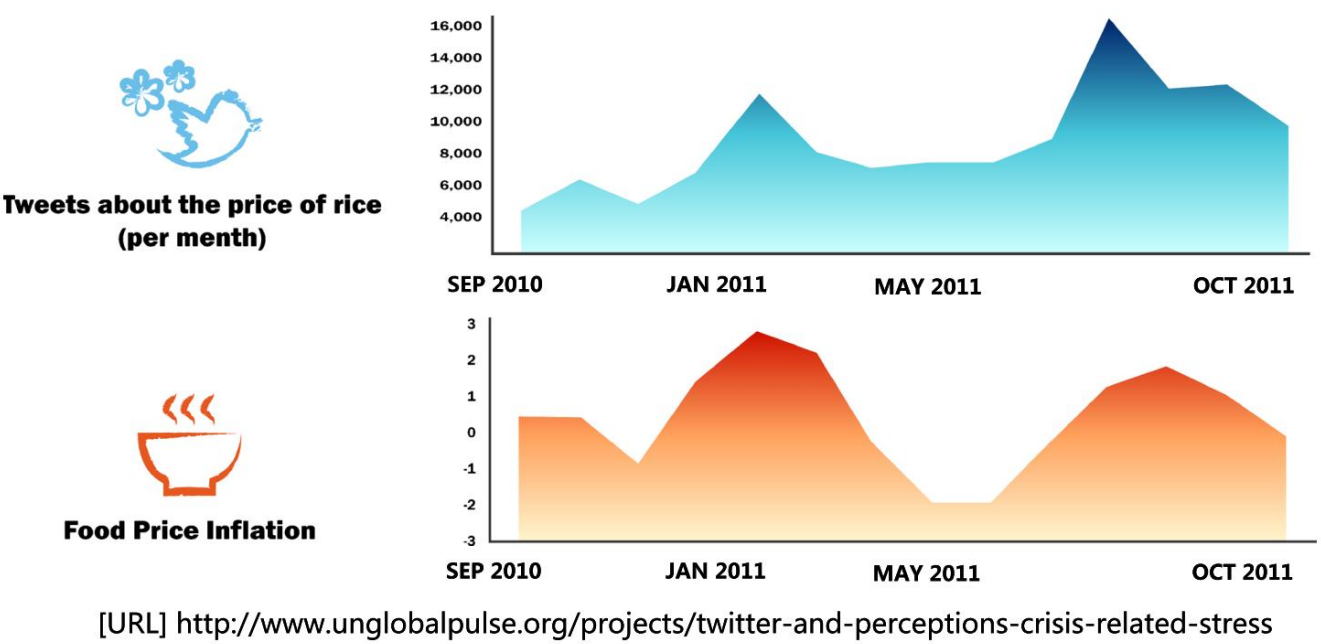

Fig. 6 A correlação entre os Tweets sobre o preço do arroz e a inflação dos preços dos alimentos

importados de prontuários médicos individuais por uma agência terceirizada. Além disso, pode ser integrado a um aplicativo de terceiros com o kit de desenvolvimento de software (SDK) e interface aberta.

5.2.5 Inteligência coletiva

Com o rápido desenvolvimento da comunicação sem fio e das tecnologias de sensores, os telefones celulares e tablets têm capacidades de computação e detecção cada vez mais fortes. Como resultado, a detecção de multidão está se tornando uma questão-chave da computação móvel. Na detecção de multidão, um grande número de usuários em geral utiliza dispositivos móveis como unidades básicas de detecção para conduzir a coordenação com redes móveis para distribuição de tarefas detectadas.
e coleta e utilização de dados detectados. Ele pode nos ajudar a concluir tarefas de detecção social complexas e de larga escala. No sensoriamento de multidões, os participantes que realizam tarefas complexas de sensoriamento não precisam ter habilidades profissionais. A detecção de multidão na forma de Crowdsourcing foi aplicada com sucesso a fotografia georreferenciada, posicionamento e navegação, detecção de tráfego rodoviário urbano, previsão de mercado, mineração de opinião e outras aplicações de trabalho intensivo.

Crowdsourcing, uma nova abordagem para resolução de problemas, tem como base um grande número de usuários em geral e distribui tarefas de forma livre e voluntária. Na verdade, o Crowdsourcing foi aplicado por muitas empresas antes do surgimento do big data. Por exemplo, P&G, BMW e Audi melhoraram suas capacidades de P&D e design em virtude do Crowdsourcing. A ideia principal do Crowdsourcing é distribuir tarefas para usuários em geral e completar tarefas que usuários individuais não podem ou não querem realizar. Sem a necessidade de implantar intencionalmente módulos de sensoriamento e empregar profissionais, o Crowdsourcing pode ampliar o escopo de um sistema de sensoriamento para atingir a escala da cidade e escalas ainda maiores.

Na era do big data, o Spatial Crowdsourcing se torna um tema quente. A estrutura de operação do Spatial Crowdsourcing é apresentada a seguir. Um usuário pode solicitar o serviço e os recursos relacionados a um local especificado. Em seguida, os usuários móveis que desejam participar da tarefa irão para o local especificado para adquirir dados relacionados (como vídeo, áudio ou fotos). Por fim, os dados adquiridos serão enviados ao solicitante do serviço. Com o rápido crescimento dos dispositivos móveis e as funções cada vez mais poderosas fornecidas pelos dispositivos móveis, pode-se prever que o Crowdsourcing Espacial prevalecerá mais do que o Crowdsourcing tradicional, por exemplo, Amazon Turk e Crowdflower.

5.2.6 Rede inteligente

Smart Grid é a rede elétrica de próxima geração constituída por redes de energia tradicionais integradas com computação, comunicação e controle para geração otimizada, fornecimento e consumo de energia elétrica. Os big data relacionados à rede inteligente são gerados a partir de várias fontes, como (i) hábitos de utilização de energia dos usuários, (ii) dados de medição fasorial, que são medidos pela unidade de medição fasorial (PMU) implantada em todo o país, (iii) consumo de energia dados de medição medidos pelos medidores inteligentes na Infraestrutura de Medição Avançada (AMI), (iv) dados de precificação e licitação do mercado de energia, (v) dados de gerenciamento, controle e manutenção de dispositivos e equipamentos nas redes de geração, transmissão e distribuição de energia (como como monitores de disjuntores e transformadores). Smart Grid traz os seguintes desafios na exploração de big data.

– *Planejamento da rede:* por meio da análise dos dados do Smart Grid, podem ser identificadas as regiões que apresentam alta carga elétrica excessiva ou altas frequências de falta de energia. Mesmo as linhas de transmissão com alta probabilidade de falha podem ser identificadas. Tais resultados analíticos podem contribuir para a atualização, transformação e manutenção da rede, etc. Por exemplo, pesquisadores da Universidade da Califórnia, em Los Angeles, projetaram um “mapa elétrico” de acordo com a teoria de big data e fizeram um mapa da Califórnia integrando informações do censo e informações de utilização de energia em tempo real fornecidas pelas empresas de energia elétrica. O mapa considera um bloco como uma unidade para demonstrar o consumo de energia de cada bloco no momento. Ele pode até comparar o consumo de energia do quarteirão com a renda média per capita e os tipos de construção, de modo a revelar hábitos de uso de energia mais precisos de todos os tipos de grupos na comunidade. Este mapa fornece previsão de carga efetiva e visual para o planejamento da rede elétrica em uma cidade. A transformação preferencial nas instalações da rede elétrica em blocos com altas frequências de interrupção e sobrecargas graves pode ser realizada, conforme exibido no mapa.
– *Interação entre geração e consumo de energia:* Uma rede elétrica ideal deve equilibrar geração e consumo de energia. No entanto, a rede elétrica tradicional é construída com base na abordagem unidirecional de transmissão-transformação-distribuição de consumo, o que não permite ajustar a capacidade de geração de acordo com a demanda de consumo de energia, levando a redundância e desperdício de energia elétrica. Portanto, medidores elétricos inteligentes são desenvolvidos para melhorar a eficiência do fornecimento de energia. A TXU Energy tem várias implantações bem-sucedidas de medidores elétricos inteligentes, que podem ajudar o fornecedor a ler os dados de utilização de energia a cada 15 minutos, exceto a cada mês no passado. O custo de mão-de-obra para leitura do medidor é bastante reduzido, porque os dados de utilização de energia (uma fonte de big data) são adquiridos e analisados com frequência e rapidez, as empresas de fornecimento de energia podem ajustar o preço da eletricidade

de acordo com os períodos de pico e baixo consumo de energia. A TXU Energy utilizou esse nível de preço para estabilizar as flutuações de pico e baixa do consumo de energia. De fato, a aplicação de big data na rede inteligente pode ajudar na realização de preços dinâmicos de compartilhamento de tempo, que é uma situação ganha-ganha para fornecedores e usuários de energia.

– *O acesso à energia renovável intermitente:* atualmente, muitos novos recursos energéticos, como a eólica e a solar, podem ser conectados às redes elétricas. No entanto, como as capacidades de geração de energia de novos recursos energéticos estão intimamente relacionadas às condições climáticas que apresentam aleatoriedade e intermitência, é um desafio conectá-los às redes elétricas. Se o big data das redes de energia for efetivamente analisado, essas novas fontes de energia renováveis intermitentes podem ser gerenciadas com eficiência: a eletricidade gerada pelos novos recursos energéticos pode ser alocada para regiões com escassez de eletricidade.
Tais recursos energéticos podem complementar as tradicionais gerações hidrelétricas e termelétricas.

6 Conclusão, questões em aberto e perspectivas

Neste artigo, revisamos os antecedentes e o estado da arte do big data. Em primeiro lugar, apresentamos o histórico geral de big data e revisamos as tecnologias relacionadas, como computação, IoT, data centers e Hadoop. Em seguida, nos concentramos nas quatro fases da cadeia de valor de big data, ou seja, geração de dados, aquisição de dados, armazenamento de dados e análise de dados. Para cada fase, apresentamos o histórico geral, discutimos os desafios técnicos e analisamos os avanços mais recentes.
Finalmente revisamos as diversas aplicações representativas de big data, incluindo gestão empresarial, IoT, redes sociais, aplicações médicas, inteligência coletiva e smart grid. Essas discussões visam fornecer uma visão geral abrangente e um quadro geral para os leitores deste empolgante
área.

No restante desta seção, resumimos os pontos críticos de pesquisa e sugerimos possíveis direções de pesquisa de big data. Também discutimos possíveis tendências de desenvolvimento nesta ampla área de pesquisa e aplicação.

6.1 Problemas em aberto

A análise de big data enfrenta muitos desafios, mas a pesquisa atual ainda está em estágio inicial. Esforços consideráveis de pesquisa são necessários para melhorar a eficiência da exibição, armazenamento e análise de big data.

6.1.1 Pesquisa teórica

Embora big data seja uma área de pesquisa quente com grande potencial tanto na academia quanto na indústria, ainda há muitos problemas importantes a serem resolvidos, que serão discutidos a seguir.

– *Problemas fundamentais de big data:* há uma necessidade imperiosa de uma definição rigorosa e holística de big data, um modelo estrutural de big data, uma descrição formal de big data e um sistema teórico de ciência de dados. Atualmente, muitas discussões sobre big data parecem mais especulação comercial do que pesquisa científica. Isso ocorre porque o big data não é definido formal e estruturalmente e os modelos existentes não são rigorosamente verificados.
– *Padronização de big data:* Deve ser desenvolvido um sistema de avaliação da qualidade dos dados e um padrão/benchmark de avaliação da eficiência da computação de dados.
Muitas soluções de aplicativos de big data afirmam que podem melhorar as capacidades de processamento e análise de dados em todos os aspectos, mas ainda não há um padrão de avaliação unificado e benchmark para equilibrar a eficiência computacional de big data com métodos matemáticos rigorosos. O desempenho só pode ser avaliado quando o sistema é implementado e implantado, o que não permite comparar horizontalmente as vantagens e desvantagens de várias soluções alternativas antes e depois da implementação do big data. Além disso, uma vez que a qualidade dos dados é uma base importante do pré-processamento, simplificação e triagem dos dados, também é um problema urgente avaliar a qualidade dos dados com eficácia e rigor.

– *Evolução dos modos de computação de big data:* Isso inclui modo de memória, modo de fluxo de dados, modo PRAM e modo MR, etc. uma abordagem intensiva em dados. A transferência de dados tem sido o principal gargalo da computação de big data. Portanto, surgiram muitos novos modelos de computação adaptados para big data, e mais modelos desse tipo estão no horizonte.

6.1.2 Desenvolvimento de tecnologia

A tecnologia de big data ainda está em sua infância. Muitos problemas técnicos importantes, como computação em nuvem, computação em grade, computação em fluxo, computação paralela, arquitetura de big data, modelo de big data e sistemas de software que suportam big data, etc., devem ser totalmente investigados.

– *Conversão de formato de big data:* devido a fontes de dados amplas e diversas, a heterogeneidade é sempre uma característica de big data, bem como um fator chave que restringe a eficiência da conversão de formato de dados. Se essa conversão de formato puder ser mais eficiente, a aplicação de big data poderá criar mais valores.

– *Transferência de big data:* a transferência de big data envolve geração, aquisição, transmissão, armazenamento e outras transformações de big data no domínio espacial. Conforme discutido, a transferência de big data geralmente incorre em altos custos, que é o gargalo para a computação de big data. No entanto, a transferência de dados é inevitável em aplicativos de big data.
Melhorar a eficiência de transferência de big data é um fator chave para melhorar a computação de big data.

– *Desempenho em tempo real de big data:* o desempenho em tempo real de big data também é um problema fundamental em muitos cenários de aplicativos. Meios eficazes para definir o ciclo de vida dos dados, calcular a taxa de depreciação dos dados e construir modelos de computação de aplicativos online e em tempo real influenciarão os resultados da análise de big data.

– *Processamento de big data:* À medida que a pesquisa de big data avança, novos problemas no processamento de big data surgem da análise de dados tradicional, incluindo (i) reutilização de dados, com o aumento da escala de dados, mais valores podem ser extraídos de re- utilização de dados existentes; (ii) reorganização de dados, conjuntos de dados em diferentes negócios podem ser reorganizados, o que pode gerar mais valor; (iii) esgotamento de dados, o que significa dados errados durante a aquisição. Em big data, não apenas os dados corretos, mas também os dados errados devem ser utilizados para gerar mais valor.

6.1.3 Implicações práticas

Embora já existam muitas aplicações bem-sucedidas de big data, muitos problemas práticos devem ser resolvidos:

– *Gerenciamento de big data:* o surgimento de big data traz novos desafios para o gerenciamento de dados tradicional. Atualmente, muitos esforços de pesquisa estão sendo feitos em bancos de dados orientados a big data e tecnologias de Internet, modelos de armazenamento e bancos de dados adequados para novo hardware, integração de dados heterogêneos e multiestruturados, gerenciamento de dados de computação móvel e pervasiva, gerenciamento de dados de SNS e gerenciamento de dados distribuídos.

– *Pesquisa, mineração e análise de big data:* o processamento de dados é sempre um ponto crítico de pesquisa em big data. Problemas relacionados incluem pesquisa e mineração de modelos SNS, algoritmos de pesquisa de big data, pesquisa distribuída, pesquisa P2P, análise visualizada de big data, sistemas de recomendação massivos, sistemas de mídia social, mineração de big data em tempo real, mineração de imagem, mineração de texto, mineração semântica, mineração de dados multiestruturados e aprendizado de máquina, etc.

– *Integração e proveniência de big data:* Conforme discutido, o valor adquirido da utilização abrangente de vários conjuntos de dados é muito maior do que o valor da soma de conjunto de dados individual. Portanto, a integração de diferentes fontes de dados é um problema atual. A integração de dados enfrenta muitos desafios, como diferentes padrões de dados e uma grande quantidade de dados redundantes.
Proveniência de dados é o processo de geração e evolução de dados ao longo do tempo e é usado principalmente para investigar vários conjuntos de dados diferentes de um único conjunto de dados. Portanto, vale a pena estudar como integrar informações de proveniência de dados com diferentes padrões e de diferentes conjuntos de dados.

– *Aplicação de big data:* Atualmente, a aplicação de big data está apenas começando e devemos explorar maneiras mais eficientes de utilizar totalmente o big data. Portanto, aplicações de big data em ciência, engenharia, medicina, assistência médica, finanças, negócios, aplicação da lei, educação, transporte, varejo e telecomunicações, aplicações de big data em pequenas e médias empresas, aplicações de big data em departamentos governamentais, serviços de big data e interação humano-computador de big data, etc. são todos problemas de pesquisa importantes.

6.1.4 Segurança de dados

Em TI, segurança e privacidade são sempre duas preocupações principais. Na era do big data, como o volume de dados está crescendo rapidamente, há riscos de segurança mais graves, enquanto os métodos tradicionais de proteção de dados já demonstraram não ser aplicáveis ao big data. Em particular, a segurança de big data é confrontada com os seguintes desafios relacionados à segurança.

– Privacidade de big data: a privacidade de big data inclui dois aspectos: (i) Proteção da privacidade pessoal durante a aquisição de dados: interesses pessoais, hábitos e propriedades corporais, etc. dos usuários podem ser adquiridos mais facilmente e os usuários podem não estar cientes. (ii) Os dados de privacidade pessoal também podem vazar durante o armazenamento, transmissão e uso, mesmo se adquiridos com a permissão dos usuários. Por exemplo, o Facebook é considerado uma empresa de big data com a maioria dos dados SNS atualmente. De acordo com um relatório [156], Ron Bowes, pesquisador da Skull Security, adquiriu dados nas páginas públicas de usuários do Facebook que falharam em modificar sua configuração de privacidade por meio de uma ferramenta de aquisição de informações. Ron Bowes empacotou esses dados em um pacote de 2,8 GB e criou uma semente de BitTorrent (BT) para download por outros. A capacidade de análise de big data pode levar à mineração de privacidade a partir de informações aparentemente simples. Portanto, a proteção da privacidade se tornará um problema novo e desafiador.

– Qualidade dos dados: a qualidade dos dados influencia a utilização de big data. Dados de baixa qualidade desperdiçam recursos de transmissão e armazenamento com baixa usabilidade. Existem muitos fatores que podem restringir a qualidade dos dados, por exemplo, geração, aquisição e transmissão podem influenciar os dados

qualidade. A qualidade dos dados se manifesta principalmente em sua precisão, integridade, redundância e consistência. Embora muitas medidas tenham sido tomadas para melhorar a qualidade dos dados, os problemas relacionados ainda não foram bem resolvidos. Portanto, métodos eficazes para detectar automaticamente a qualidade dos dados e reparar alguns dados danificados precisam ser investigados.

– Mecanismo de segurança de big data: Big data traz desafios para a criptografia de dados devido à sua grande escala e alta diversidade. O desempenho dos métodos de criptografia anteriores em dados de pequena e média escala pode

não atendem às demandas de big data, portanto, abordagens eficientes de criptografia de big data devem ser desenvolvidas. Esquemas eficazes para gerenciamento de segurança, controle de acesso e comunicações de segurança devem ser investigados para dados estruturados, semiestruturados e não estruturados. Além disso, no modo multilocatário, o isolamento, a confidencialidade, a integridade, a disponibilidade, a capacidade de controle e a rastreabilidade dos dados dos locatários devem ser ativados com base na garantia de eficiência.

– Aplicação de big data na segurança da informação: Big data não só traz desafios para a segurança da informação, mas também oferece novas oportunidades para o desenvolvimento de mecanismos de segurança da informação. Por exemplo, podemos descobrir potenciais brechas de segurança e APT (Ameaça Persistente Avançada) após a análise de big data na forma de arquivos de log de um Sistema de Detecção de Intrusão. Além disso, características de vírus, brechas e características de ataque, etc. também podem ser identificadas mais facilmente por meio da análise de big data.

A segurança do big data tem atraído grande atenção dos pesquisadores. No entanto, há apenas pesquisas limitadas sobre a representação de big data heterogêneo de várias fontes, métodos de medição e compreensão semântica, teorias de modelagem e modelos de computação, armazenamento distribuído de otimização de eficiência energética e arquiteturas de sistemas de hardware e software processados, etc. Particularmente, a segurança de big data, incluindo credibilidade, backup e recuperação, manutenção de integridade e segurança, deve ser mais investigada.

6.2 Perspectivas

O surgimento de big data abre grandes oportunidades. Na era da TI, o "T" (Tecnologia) era a principal preocupação, enquanto a tecnologia impulsiona o desenvolvimento dos dados. Na era do big data, com destaque para o valor dos dados e avanços no "I" (Informações), os dados conduzirão o progresso das tecnologias em um futuro próximo. Big data não só terá impacto social e econômico, mas também influenciará a maneira de viver e pensar de todos, o que está acontecendo. Poderíamos não prever o futuro, mas pode tomar precauções para possíveis eventos que possam ocorrer no futuro.

– *Dados com maior escala, maior diversidade e estruturas mais complexas:* embora as tecnologias representadas pelo Hadoop tenham obtido grande sucesso, espera-se que tais tecnologias fiquem para trás e sejam substituídas devido ao rápido desenvolvimento do big data. A base teórica do Hadoop surgiu já em 2006. Muitos pesquisadores se preocuparam com melhores maneiras de lidar com dados estruturados de maior escala, maior diversidade e complexidade. Esses esforços são representados pelo Spanner (banco de dados distribuído globalmente) do Google e banco de dados relacional distribuído, expansível e tolerante a falhas F1. No futuro, a tecnologia de armazenamento de big data empregará bancos de dados distribuídos, suportará mecanismos de transação semelhantes aos bancos de dados relacionais e tratará os dados de maneira eficaz por meio de gramáticas semelhantes ao SQL.

– *Desempenho de recursos de dados:* como big data contém valores enormes, dominar big data significa dominar recursos. Por meio da análise da cadeia de valor do big data, pode-se perceber que seu valor vem dos próprios dados, tecnologias e ideias, e o núcleo são os recursos de dados. A reorganização e integração de diferentes conjuntos de dados podem criar mais valores. A partir de agora, as empresas que dominam os recursos de big data podem obter enormes benefícios alugando e cedendo os direitos de uso de seus dados.

– *Big data promove a fusão cruzada da ciência:* Big data não apenas promove a fusão abrangente de computação em nuvem, IoT, data center e redes móveis, etc., mas também força a fusão cruzada de muitas disciplinas.
O desenvolvimento de big data deve explorar tecnologias e métodos inovadores em termos de aquisição de dados, armazenamento, processamento, análise e segurança da informação, etc. ser examinada para as empresas modernas a partir da perspectiva de gestão. Além disso, a aplicação de big data a campos específicos requer a participação de talentos interdisciplinares.

– *Visualização:* Em muitos cenários de interação humano-computador, o princípio do que você vê é o que você obtém é seguido, por exemplo, como em editores de texto e imagem. Em aplicações de big data, os dados mistos são muito úteis para a tomada de decisões. Somente quando os resultados analíticos são exibidos de forma amigável, eles podem ser efetivamente utilizados pelos usuários. Relatórios, histogramas, gráficos de pizza e curvas de regressão, etc., são frequentemente usados para visualizar os resultados da análise de dados. Novas formas de apresentação ocorrerão no futuro, por exemplo, Microsoft Renlifang, um mecanismo de busca s

utiliza diagramas relacionais para expressar o relacionamento interpessoal.

– *Orientado a dados:* é bem conhecido que os programas consistem em estruturas de dados e algoritmos, e estruturas de dados são usadas para armazenar dados. Na história do design de programas, observa-se que o papel dos dados está se tornando cada vez mais significativo. Na era dos dados de pequena escala, em que a lógica é mais complexa do que os dados, o design do programa é principalmente orientado para o processo. Como os dados de negócios estão se tornando mais complexos, métodos de design orientados a objetos são desenvolvidos. Atualmente, a complexidade dos dados de negócios ultrapassou em muito a lógica de negócios. Conseqüentemente, os programas são gradualmente transformados de intensivos em algoritmos para intensivos em dados. Prevê-se que certamente surgirão métodos de design de programas orientados a dados, que terão uma influência de longo alcance no desenvolvimento de TI em engenharia de software, arquitetura e design de modelos, entre outros.

– *Big data desencadeia a revolução do pensamento:* Gradualmente, big data e sua análise influenciarão profundamente nossas formas de pensar. Em [11], os autores resumem a revolução do pensamento desencadeada pelo big data da seguinte forma:

  – Durante a análise de dados, tentaremos utilizar todos os dados além de apenas analisar um pequeno conjunto de dados de amostra.
  – Em comparação com dados precisos, gostaríamos de aceitar dados numerosos e complicados.
  – Devemos prestar mais atenção às correlações entre as coisas além de explorar a relação causal.

  – Os algoritmos simples de big data são mais eficazes do que algoritmos complexos de small data.

  – Resultados analíticos de big data reduzirão fatores precipitados e subjetivos durante a tomada de decisão, e cientistas de dados substituirão “especialistas”.

Ao longo da história da sociedade humana, as demandas e vontades dos seres humanos são sempre a fonte de energia para promover o progresso científico e tecnológico. O big data pode fornecer respostas de referência para os seres humanos tomarem decisões por meio de mineração e processamento analítico, mas não pode substituir o pensamento humano. É o pensamento humano que promove as utilizações generalizadas de big data.
Big data é mais como um cérebro humano extensível e expansível do que um substituto do cérebro humano. Com o

Com o surgimento da IoT, o desenvolvimento da tecnologia de detecção móvel e o progresso da tecnologia de aquisição de dados, as pessoas não são apenas usuários e consumidores de big data, mas também seus produtores e participantes. Sensoriamento de relações sociais, crowd sourcing, análise de big data em SNS e outras aplicações intimamente relacionadas a atividades humanas baseadas em big data serão cada vez mais preocupados e certamente provocarão enormes transformações nas atividades sociais da sociedade futura.

Agradecimentos Este trabalho foi financiado pela China National Natural Science Foundation (No. 61300224), o Ministério da Ciência e Tecnologia (MOST), China, o Programa Internacional de Colaboração em Ciência e Tecnologia (Project No.: S2014GAT014) e a Hubei Provincial Key Projeto (nº 2013CFA051). A pesquisa de Shiwen Mao é apoiada em parte pela US NSF sob as subvenções CNS 1320664, CNS-1247955 e CNS-0953513, e através do site NSF Broadband Wireless Access & Applications Center (BWAC) na Auburn University.

Referências

1. Gantz J, Reinsel D (2011) Extraindo valor do caos. IDC iView, pp 1–12

2. Ficha informativa: Big data no governo federal (2012). http://www.whitehouse.gov/sites/default/files/microsites/ostp/big data fact sheet 3 29 2012.pdf

3. Cukier K (2010) Dados, dados em todos os lugares: um relatório especial sobre gerenciamento de informações. Jornal Economist 4.
Afogando-se em números - os dados digitais inundarão o planeta- e nos ajudarão a entendê-lo melhor (2011). http://www.economist.com/blogs/ dailychart/2011/11/bigdata-0 5. Lohr S (2012)
A era do big data. New York Times, pp 11 6. Yuki N (2011) Seguindo a trilha digital para o ouro do big data. http://www.npr.org/2011/11/29/142521910/thedigitalbreadcrumbs that-lead-to-big-data

7. Yuki N A busca por analistas para dar sentido ao big data (2011). http://www.npr.org/2011/11/30/142893065/the-searchforanalysts to-make-sense-of-big-data 8. Grandes dados
(2008). http://www.nature.com/news/specials/bigdata/index.html

9. Coleta online especial: lidando com big data (2011). http://www.sciencemag.org/site/special/data/ 10.
Manyika J, McKinsey Global Institute, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: a próxima fronteira para inovação, competição e produtividade.
McKinsey Global Institute 11.
Mayer-Schönberger V, Cukier K (2013) Big data: uma revolução que transformará a forma como vivemos, trabalhamos e pensamos. Eamon Dolan/Houghton Mifflin Harcourt
12. Laney D (2001) Gerenciamento de dados 3D: controlando o volume, a velocidade e a variedade dos dados. META Group Research Note, 6 de fevereiro de 13. Zikopoulos P, Eaton C et al (2011) Compreendendo big data: análise para hadoop de classe empresarial e dados de streaming. Mídia McGraw Hill Osborne
14. Meijer E (2011) O mundo de acordo com o linq. Communications of the ACM 54(10):45–51 15. Beyer M
(2011) O Gartner diz que resolver o desafio de big data envolve mais do que apenas gerenciar volumes de dados. Gartner. http://www. gartner.com/it/page.jsp
16. OR Team (2011) Big data agora: perspectivas atuais do OReilly Radar. OReilly Media 17. Grobelnik M (2012) Tutorial
de big data. http://videolectures.net/eswc2012grobelnikbigdata/ _ 18. Ginsberg J, Mohebbi MH, Patel RS, Brammer
L, Smolinski MS, Brilliant L (2008) Detectando epidemias de influenza usando dados de consulta de mecanismos de pesquisa. Nature 457(7232):1012–1014 19. DeWitt D, Gray J (1992) Sistemas de banco de dados
paralelos: o futuro dos sistemas de banco de dados de alto desempenho. Comum ACM 35(6):85–98

20. Walter T (2009) Teradata passado, presente e futuro. UCI ISG
série de palestras sobre gerenciamento de dados
escaláveis 21. Ghemawat S, Gobioff H, Leung ST (2003) The google file system.
In: ACM SIGOPS Operating Systems Review, vol 37. ACM, pp 29–43

22. Dean J, Ghemawat S (2008) Mapreduce: processamento de dados
simplificado em grandes clusters. Commun ACM 51(1):107–
113 23. Hey AJG, Tansley S, Tolle KM et al (2009) O quarto paradigma:
descoberta científica intensiva em dados
24. Howard JH, Kazar ML, Menees SG, Nichols DA, Satyanarayanan M,
Sidebotham RN, West MJ (1988) Escala e desempenho em um sistema de
arquivos distribuído. ACM Trans Comput Syst (TOCS) 6(1):51–81 25. Cattell
R (2011) Scalable sql e nosql
data stores. ACM SIG MOD Registro 39(4):12–27 26. Labrinidis A, Jagadish HV
(2012) Desafios e oportunidades
com big data. Proc VLDB Endowment 5(12):2032–2033 27. Chaudhuri S, Dayal
U, Narasayya V (2011) Uma visão geral da tecnologia de inteligência
de negócios. Comum ACM 54(8): 88–98

28. Agrawal D, Bernstein P, Bertino E, Davidson S, Dayal U, Franklin M, Gehrke
J, Haas L, Halevy A, Han J et al (2012) Desafios e oportunidades com big
data. Um white paper da comunidade desenvolvido pelas principais pesquisas
nos Estados Unidos 29. Sun Y, Chen M, Liu B, Mao S (2013) Far:
um método de roteamento que evita falhas para redes de data center com
topologia regular.
In: Anais do simpósio ACM/IEEE sobre arquiteturas para redes e sistemas
de comunicação (ANCS'13). ACM 30. Wiki (2013). Aplicativos e
organizações que usam hadoop.
http://wiki.apache.org/hadoop/PoweredBy 31.
Bahga A, Madisetti VK (2012) Analisando dados massivos de manutenção de
máquinas em uma nuvem de computação. IEEE Transac Parallel Distrib Syst
23(10):1831–1843 32. Gunarathne
T, Wu TL, Choi JY, Bae SH, Qiu J (2011)
Paradigmas de computação em nuvem para aplicativos biomédicos
agradavelmente paralelos. Concurr Comput Prac Experience 23(17):2338–
2354
33. Gantz J, Reinsel D (2010) A década do universo digital - você está pronto.
Publicação externa de informações e dados da IDC (Analyse the Future), pp
1–16 34. Bryant RE (2011)
Computação escalável com uso intensivo de dados para aplicações científicas.
Comput Sci Eng 13(6):25–33 35. Wahab MHA, Mohd
MNH, Hanafi HF, Mohsin MFM (2008)
Pré-processamento de dados em logs de servidor web para algoritmo de
mineração de regras de associação generalizada. World Acad Sci Eng
Technol 48:2008
36. Nanopoulos A, Manolopoulos Y, Zakrzewicz M, Morzy T (2002) Indexação de
logs de acesso à web para consultas de padrões. In: Anais do 4º workshop
internacional sobre informação na web e gerenciamento de dados. ACM, pp
63–68
37. Joshi KP, Joshi A, Yesha Y (2003) Sobre o uso de um warehouse para analisar
logs da web. Distrib Parallel Databases 13(2):161–180 38.
Chandramohan V, Christensen K (2002) Uma primeira olhada em redes de
sensores com fio para sistemas de vigilância por vídeo. In: Proceedings
LCN 2002, 27ª conferência anual do IEEE sobre redes locais de computadores.
IEEE, pp 728–729 39. Selavo L,
Wood A, Cao Q, Sookoor T, Liu H, Srinivasan A, Wu Y, Kang W, Stankovic J,
Young D et al (2007) Luster: rede de sensores sem fio para pesquisa
ambiental . In: Proceedings of the 5th International Conference on Embedded
Networked Sensor Systems. ACM, pp 103–116 40. Barrenetxea G, Ingelrest
F, Schaefer G, Vetterli M, Couach
O, Parlange M (2008) Sensorscope: monitoramento ambiental pronto para uso. In:
Processamento de informação em redes de sensores, 2008, conferência
internacional IPSN'08. IEEE, páginas 332–343

41. Kim Y, Schmid T, Charbiwala ZM, Friedman J, Srivastava MB (2008) Nawms:
sistema autônomo não intrusivo de monitoramento de água. In: Proceedings
of the 6th ACM Conference on Embedded Network Sensor Systems. ACM,
pp 309–322
42. Kim S, Pakzad S, Culler D, Demmel J, Fenves G, Glaser S, Turon M (2007)
Monitoramento da integridade de infraestruturas civis usando redes de
sensores sem fio. In Information Processing in Sensor Networks 2007, 6th
International Symposium on IPSN 2007.
IEEE, pp 254–263
43. Ceriotti M, Mottola L, Picco GP, Murphy AL, Guna S, Corra M, Pozzi M, Zonta
D, Zanon P (2009) Monitoramento de edifícios históricos com redes de
sensores sem fio: a implantação da torre aquila .
In: Proceedings of the 2009 International Conference on Information
Processing in Sensor Networks. IEEE Computer Society, pp 277–288 44.
Tolle G,
Polastre J, Szewczyk R, Culler D, Turner N, Tu K, Burgess S, Dawson T,
Buonadonna P, Gay D et al (2005) A macroscope in the redwoods. In:
Proceedings of the 3rd International Conference on Embedded Networked
Sensor Systems. ACM, pp 51–63

45. Wang F, Liu J (2011) Coleta de dados de sensores sem fio em rede:
problemas, desafios e abordagens. IEEE Commun Surv Tutor 13(4):673–687

46. Cho J, Garcia-Molina H (2002) Rastreadores paralelos. In: Anais da 11ª
Conferência Internacional sobre World Wide Web. ACM, pp 124–135 47.
Choudhary S,
Dincturk ME, Mirtaheri SM, Moosavi A, von Bochmann G, Jourdan GV, Onut IV
(2012) Crawling rich internet applications: the state of the art. Em: CASCON.
páginas 146–160

48. Ghani N, Dixit S, Wang TS (2000) Sobre a integração ip-over-wdm.
IEEE Comum Mag 38(3):72–84
49. Manchester J, Anderson J, Doshi B, Dravida S, Ip over sonet
(1998) IEEE Commun Mag 36(5):136–142 50. Jinno
M, Takara H, Kozicki B (2009) Dynamic Optical Mesh Net Works: Drivers,
Challenges and Solutions for the Future. In: Comunicação óptica, 2009, 35ª
conferência europeia sobre a ECOC'09.
IEEE, pp 1–4
51. Barroso LA, Hölzle U (2009) O datacenter como um computador: uma
introdução ao design de máquinas em escala de armazém. Synt Lect Comput
Archit 4(1):1–108 52. Armstrong J
(2009) Ofdm para comunicações ópticas. J Light Technol 27(3):189–204 53. Shieh
W (2011) Ofdm para redes
ópticas flexíveis de alta velocidade.
J Light Technol 29(10):1560–1577
54. Guia de design e implantação de interconexão de data center da Cisco
(2010)
55. Greenberg A, Hamilton JR, Jain N, Kandula S, Kim C, Lahiri P, Maltz DA, Patel
P, Sengupta S (2009) Vl2: uma rede de data center escalável e flexível. In
ACM SIGCOMM computer communication review, vol 39. ACM, pp 51–62
56. Guo C, Lu G, Li D, Wu H, Zhang X, Shi Y, Tian C,
Zhang Y, Lu S (2009) Bcube: a arquitetura de rede centrada no servidor de alto
desempenho para data centers modulares. ACM SIGCOMM Comput Commun
Rev 39(4):63–74 57. Farrington N, Porter G, Radhakrishnan S, Bazzaz HH,
Subramanya V, Fainman Y,
Papen G, Vahdat A (2011) Helios: uma arquitetura de comutação elétrica/óptica
híbrida para centros de dados modulares. ACM SIGCOMM Comput Commun
Rev 41(4):339–350

58. Abu-Libdeh H, Costa P, Rowstron A, O'Shea G, Donnelly A (2010) Roteamento
simbiótico em data centers futuros. ACM SIG COMM Comput Commun Rev
40(4):51–62 59. Lam C, Liu H, Koley B, Zhao X,
Kamalov V, Gill V, Tecnologias de comunicação de fibra óptica: o que é necessário
para operações de rede de datacenter (2010) IEEE Commun Mag 48(7):32–
39

60. Wang G, Andersen DG, Kaminsky M, Papagiannaki K, Ng TS, Kozuch M, Ryan M (2010) c-through: Part-time optics in data centers. In: ACM SIGCOMM Computer Communication Review, vol 40. ACM, pp 327–338

61. Ye X, Yin Y, Yoo SJB, Mejia P, Proietti R, Akella V (2010) Dos: um comutador óptico escalável para datacenters. In Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems. ACM, p 24 62. Singla A, Singh A, Ramachandran K,
Xu L, Zhang Y (2010) Pro teus: a topology maleable data center network. In Proceedings of the 9th ACM SIGCOMM workshop on hot topics em redes.

ACM, pág. 8
63. Liboiron-Ladouceur O, Cerutti I, Raponi PG, Andriolli N, Castoldi P (2011) Projeto energeticamente eficiente de uma arquitetura de interconexão óptica multiplano escalável. IEEE J Sel Top Quantum Electron 17(2):377–383

64. Kodi AK, Louri A (2011) Redes fotônicas reconfiguráveis com eficiência energética e largura de banda para sistemas de computação de alto desempenho (hpc). IEEE J Sel Top Quantum Electron 17(2):384–395

65. Zhou X, Zhang Z, Zhu Y, Li Y, Kumar S, Vahdat A, Zhao BY, Zheng H (2012) Espelho espelho no teto: links sem fio flexíveis para data centers. ACM SIGCOMM Comput Commun Rev 42(4):443–454 66. Lenzerini M (2002) Integração de dados: uma
perspectiva teórica.
In: Anais do vigésimo primeiro simpósio ACM SIGMOD-SIGACT SIGART sobre princípios de sistemas de banco de dados. ACM, pp 233–246 67. Cafarella MJ, Halevy A,
Khoussainova N (2009) Integração de dados para a web relacional. Proc VLDB Endowment 2(1):1090–1101

68. Maletic JI, Marcus A (2000) Limpeza de dados: além da integridade
análise. Em: QI. Citeseer, pp 200–209
69. Kohavi R, Mason L, Parekh R, Zheng Z (2004) Lições e desafios da mineração de dados de comércio eletrônico de varejo. Mach Learn 57(1-2):83–113 70. Chen H, Ku WS, Wang H,
Sun MT (2010) Aproveitando a redundância temporal espacial para limpeza de dados RFID. In: Anais da conferência internacional ACM SIGMOD 2010 sobre gerenciamento de dados. ACM, pp 51–62 71. Zhao Z, Ng W (2012) Uma abordagem baseada em modelo para
limpeza de fluxo de dados RFID. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM, pp 862–871 72. Khoussainova N, Balazinska M, Suciu D (2008) Extração probabilística de eventos de dados rfid.
In: Engenharia de Dados, 2008.

IEEE 24ª conferência internacional sobre ICDE 2008. IEEE, pp 1480–1482
73. Herbert KG, Wang JTL (2007) Limpeza de dados biológicos: um caso
estudar. Int J Inf Qual 1(1):60–82
74. Tsai TH, Lin CY (2012) Explorando a redundância contextual na melhoria da codificação de vídeo baseada em objetos para vigilância de redes de sensores de vídeo. IEEE Transac Multmed 14(3):669–682 75. Sarawagi S,
Bhamidipaty A (2002) Desduplicação interativa usando aprendizado ativo. Em Anais da oitava conferência internacional ACM SIGKDD sobre descoberta de conhecimento e mineração de dados. ACM, pp 269–278

76. Kamath U, Compton J, Dogan RI, Jong KD, Shehu A (2012)
Uma abordagem de algoritmo evolutivo para geração de recursos a partir de dados de sequência e sua aplicação para predição do local de junção do DNA. IEEE/ACM Transac Comput Biol Bioinforma (TCBB) 9(5):1387–1398 77. Leung KS, Lee KH, Wang
JF, Ng EYT, Chan HLY, Tsui SKW, Mok TSK, Tse PC-H, Sung JJ-Y (2011) Mineração de dados em sequências de DNA do vírus da hepatite b. IEEE/ACM Transac Comput Biol Bioinforma 8(2):428–440

78. Huang Z, Shen H, Liu J, Zhou X (2011) Redução efetiva de dados para busca de similaridade multimídia. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, pp 1021–1032 79. Bleiholder J, Naumann F (2008) Fusão de dados.
ACM Comput Surv (CSUR) 41(1):1 80. Brewer EA (2000) Rumo a sistemas distribuídos robustos. Em:

PODC. p 7
81. Gilbert S, Lynch N (2002) A conjectura de Brewer e a viabilidade de serviços da Web consistentes, disponíveis e tolerantes a partições. ACM SIGACT News 33(2):51–59 82. McKusick MK,
Quinlan S (2009) Gfs: eqvolution on fast forward. ACM Queue 7(7):10 83. Chaiken R, Jenkins B, Larson PA, Ramsey B,
Shakib D, Weaver S, Zhou J (2008) Escopo: processamento paralelo fácil e eficiente de conjuntos de dados massivos. Proc VLDB Endowment 1(2):1265–

1276
84. Beaver D, Kumar S, Li HC, Sobel J, Vajgel P et al (2010) Encontrando uma agulha no palheiro: armazenamento de fotos do facebook. Em OSDI, vol 10. pp 1–8

85. DeCandia G, Hastorun D, Jampani M, Kakulapati G, Lakshman A, Pilchin A, Sivasubramanian S, Vosshall P, Vogels W (2007)
Dynamo: armazenamento de valor-chave altamente disponível da amazon. In: SOSP, vol 7. pp 205–
220 86. Karger D, Lehman E, Leighton T, Panigrahy R, Levine M, Lewin D (1997) Hashing consistente e árvores aleatórias: protocolos de cache distribuídos para aliviar pontos quentes em todo o mundo rede.
In: Anais do vigésimo nono simpósio anual da ACM sobre teoria da computação. ACM, pp 654–663 87. Chang F, Dean J, Ghemawat
S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE (2008) Bigtable: um sistema de armazenamento distribuído para dados estruturados. ACM Trans Comput Syst (TOCS) 26(2):4 88. Burrows M (2006) O serviço chubby lock para sistemas distribuídos
fracamente acoplados. In: Anais do 7º Simpósio de Projeto e Implementação de Sistemas Operacionais. USENIX Association, pp 335–350 89. Lakshman A, Malik P (2009) Cassandra: sistema de armazenamento estruturado em uma rede p2p. In: Anais do 28º
Simpósio ACM sobre Princípios de Computação Distribuída. ACM, pp 5–5

90. George L (2011) HBase: o guia definitivo. O'Reilly Media Inc 91. Judd D (2008)
hypertable-0.9. 0.4-alpha 92. Chodorow K (2013)
MongoDB: o guia definitivo. O'Reilly
Media Inc
93. Crockford D (2006) O tipo de mídia application/json para notação de objeto
javascript (json)
94. Murty J (2009) Programação de serviços da web amazon: S3, EC2, SQS, FPS e SimpleDB. O'Reilly Media Inc 95. Anderson JC, Lehnardt
J, Slater N (2010) CouchDB: o guia definitivo. O'Reilly Media Inc 96. Blanas S, Patel JM, Ercegovac V, Rao J, Shekita EJ,
Tian Y (2010) Uma comparação de algoritmos de junção para processamento de log em mapreduce. In: Anais da conferência internacional ACM SIGMOD 2010 sobre gerenciamento de dados. ACM, pp 975–986

97. Yang HC, Parker DS (2009) Traverse: indexação simplificada em grandes clusters map-reduce-merge. In: Sistemas de banco de dados para aplicações avançadas. Springer, pp 308–322
98. Pike R, Dorward S, Griesemer R, Quinlan S (2005) Interpretando os dados: análise paralela com sawzall. Programa de Ciências 13(4):277–298

99. Gates AF, Natkovich O, Chopra S, Kamath P, Narayanamurthy SM, Olston C, Reed B, Srinivasan S, Srivastava U (2009) Building a high-level dataflow system on top of map-reduce: the pig experience. Processo VLDB Endowment 2(2):1414–1425

100. Thusoo A, Sarma JS, Jain N, Shao Z, Chakka P, Anthony S, Liu H, Wyckoff P, Murthy R (2009) Hive: a warehousing solution over a map-reduce framework. Proc VLDB Endowment 2(2):1626–1629 101. Isard M, Budiu M, Yu Y, Birrell A, Fetterly D (2007)
Dryad: programas paralelos de dados distribuídos a partir de blocos de construção sequenciais.
ACM SIGOPS Oper Syst Rev 41(3):59–72 102. Yu ´
Y, Isard M, Fetterly D, Budiu M, Erlingsson U, Gunda PK, Currey J (2008) Dryadlinq: a system for general-purpose dis tributed data- computação paralela usando uma linguagem de alto nível. In: OSDI, vol 8. pp 1–14 103. Moretti C, Bulosan J, Thain D, Flynn PJ
(2008) All-pairs: an abstracttion for data-intensive cloud computing. In: Processamento paralelo e distribuído, 2008. Simpósio internacional IEEE sobre IPDPS 2008. IEEE, pp 1–11

104. Malewicz G, Austern MH, Bik AJC, Dehnert JC, Horn I, Leiser N, Czajkowski G (2010) Pregel: um sistema para processamento de gráficos em grande escala. In: Anais da conferência internacional ACM SIGMOD 2010 sobre gerenciamento de dados. ACM, pp 135–146 105. Bu Y, Bill H, Balazinska
M, Ernst MD (2010) Haloop: processamento de dados iterativo eficiente em grandes clusters. Proc VLDB Endowment 3(1-2):285–296 106. Ekanayake J, Li H, Zhang B, Gunarathne T, Bae SH,
Qiu J, Fox G (2010) Twister: um tempo de execução para mapreduce iterativo. In Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. ACM, pp 810–818 107. Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, Franklin M, Shenker S, Stoica I
(2012) Conjuntos de dados distribuídos resilientes: uma abstração tolerante a falhas para memória computação em cluster. In: Anais da 9ª conferência USENIX sobre projeto e implementação de sistemas em rede. Associação USENIX, pp 2–2 108. Bhatotia P, Wieder A, Rodrigues R, Acar UA, Pasquin R (2011)
Incoop: mapreduce para cálculos incrementais. In: Anais do 2º Simpósio ACM sobre Computação em Nuvem. ACM, pág. 7

109. Murray DG, Schwarzkopf M, Smowton C, Smith S, Madhavapeddy A, Hand S (2011) Ciel: um mecanismo de execução universal para computação de fluxo de dados distribuídos. In: Anais da 8ª Conferência USENIX sobre Projeto e Implementação de Sistemas em Rede. p 9

110. Anderson TW (1958) Uma introdução à estatística multivariada análise, vol 2. Wiley, Nova York
111. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Philip SY et al (2008) Os 10 principais algoritmos em mineração de dados. Knowl Inf Syst 14(1):1–37 112. Qual software analítico de
mineração de dados e big data você usou nos últimos 12 meses para um projeto real? (2012) http://www.kdnuggets. com/polls/2012/analytics-data-mining-big-data-software.html 113. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter
T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2008) KNIME: o minerador de informações Konstanz. Springer 114. Sallam RL, Richardson J, Hagerty J, Hostmann B (2011) Quadrante mágico para
plataformas de inteligência de negócios. CT, Grupo Gartner, Stamford

115. Além do PC. Relatório Especial sobre Tecnologia Pessoal (2011)
116. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A et al (2011) O iplant colaborativo: ciberinfraestrutura para biologia vegetal. Front Plant Sci 34(2):1–16. doi:10.3389/fpls.2011.00034 117. Baah GK, Gray A, Harrold MJ (2006)
Detecção de anomalias on-line de software implantado: uma abordagem estatística de aprendizado de máquina.
In: Anais do III Workshop Internacional de Garantia de Qualidade de Software. ACM, pp 70–77
118. Moeng M, Melhem R (2010) Aplicando aprendizagem de máquina estatística para tensão multicore e escalonamento de frequência. Em: Processos de a 7ª conferência internacional da ACM sobre fronteiras da computação. ACM, págs. 277–286
119. Gaber MM, Zaslavsky A, Krishnaswamy S (2005) Fluxos de dados de mineração: uma revisão. ACM Sigmod Record 34(2):18–26 120. Verykios
VS, Bertino E, Fovino IN, Provenza LP, Saygin Y, Theodoridis Y (2004) Estado da arte em mineração de dados de preservação de privacidade. ACM Sigmod Record 33(1):50–57 121. van der Aalst W (2012)
Mineração de processos: visão geral e oportunidades. ACM Transac Manag Inform Syst (TMIS) 3(2):7 122. Manning CD, Sch¨utze H (1999)
Fundamentos do processamento estatístico de linguagem natural, vol 999. MIT Press 123. Pal SK, Talwar V, Mitra P (2002) Mineração
da Web em estrutura de computação leve, relevância, estado da arte e direções futuras. IEEE Transac Neural Netw 13(5):1163–1177 124. Chakrabarti S (2000) Mineração de dados para hipertexto:
uma pesquisa tutorial.
ACM SIGKDD Explor Newsl 1(2):1–11 125. Brin
S, Page L (1998) A anatomia de um mecanismo de busca da web hipertextual em grande escala. Comput Netw ISDN Syst 30(1):107–117 126. Konopnicki
D, Shmueli O (1995) W3qs: um sistema de consulta para a rede mundial. In: VLDB, vol 95. pp 54–65 127. Chakrabarti S, Van den Berg M,
Dom B (1999) Rastreamento focado: uma nova abordagem para a descoberta de recursos da Web específicos de tópicos.
Rede de computação 31(11):1623–1640
128. Ding D, Metze F, Rawat S, Schulam PF, Burger S, Younessian E, Bao L, Christel MG, Hauptmann A (2012) Além da recuperação de áudio e vídeo: rumo à sumarização multimídia. In: Anais da 2ª Conferência Internacional da ACM sobre Recuperação Multimídia. ACM, pp 2 129. Wang M, Ni B, Hua XS, Chua TS (2012) Assistive
tagging: a survey of multimedia tagging with human-computer joint exploit. ACM Comput Surv (CSUR) 44(4):25 130. Lew MS, Sebe N, Djeraba C, Jain R (2006) Recuperação de informações multimídia baseada em
conteúdo: estado da arte e desafios. ACM Trans Multimed Comput Commun Appl (TOMCCAP) 2(1):1–19 131. Hu W, Xie N, Li L, Zeng X, Maybank S (2011) Uma pesquisa sobre indexação e recuperação de vídeo baseada em conteúdo
visual. IEEE Trans Syst Man Cybern Parte C Appl Rev 41(6):797–819 132. Park YJ, Chang KN (2009) Modelo de perfil de cliente baseado em comportamento individual e em grupo para recomendação personalizada de
produtos. Expert Syst Appl 36(2):1932–1939 133. Barrag´ans-Mart´ınez AB, Costa-Montenegro E, Burguillo JC, Rey-L´opez M, Mikic-Fonte FA, Peleteiro A (2010) Um conteúdo híbrido Abordagem de filtragem
colaborativa baseada em itens e baseada em itens para recomendar programas de TV aprimorados com decomposição de valor singular. Inf Sci 180(22):4290–4311

134. Naphade M, Smith JR, Tesic J, Chang SF, Hsu W, Kennedy L, Hauptmann A, Curtis J (2006) Ontologia conceitual em larga escala para multimídia. IEEE Multimedia 13(3):86–91 135. Ma Z, Yang Y, Cai Y,
Sebe N, Hauptmann AG (2012) Adaptação de conhecimento para detecção de eventos multimídia ad hoc com poucos exemplares. In: Anais do 20º Congresso Internacional de Multimídia da ACM. ACM, pp 469–478 136. Hirsch JE (2005) Um índice para quantificar a produção de pesquisa
científica de um indivíduo. Proc Natl Acad Sci USA 102(46):16569 137. Watts DJ (2004) Seis graus: a ciência de uma era conectada.
WW Norton & Company 138.
Aggarwal CC (2011) Uma introdução à análise de dados de redes sociais. saltador

139. Scellato S, Noulas A, Mascolo C (2011) Explorando recursos de lugar na previsão de links em redes sociais baseadas em localização. In: Anais da 17ª Conferência Internacional ACM SIGKDD sobre Descoberta de Conhecimento e Mineração de Dados. ACM, pp 1046–1054 140. Ninagawa A, Eguchi K
(2010) Previsão de link usando modelos de grupo probabilísticos de estrutura de rede. In: Anais do Simpósio ACM 2010 sobre Computação Aplicada. ACM, pp 1115–1116

141. Dunlavy DM, Kolda TG, Acar E (2011) Previsão de vínculo temporal usando fatorações de matriz e tensor. ACM Transac Knowl Discov Data (TKDD) 5(2):10 142. Leskovec J,
Lang KJ, Mahoney M (2010) Comparação empírica de algoritmos para detecção de comunidade de rede. In: Anais da 19ª Conferência Internacional sobre World Wide Web. ACM, pp 631–640 143. Du N, Wu B, Pei X, Wang B, Xu L (2007)
Detecção de comunidade em redes sociais de larga escala. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, pp 16–25 144. Garg S, Gupta T, Carlsson N, Mahanti A (2009) Evolução de uma rede
de agregação social online: um estudo empírico. In: Pro ceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference. ACM, pp 315–321 145. Allamanis M, Scellato S, Mascolo C (2012) Evolução de uma rede social online baseada em
localização: análise e modelos. In: Proceedings of the 2012 ACM Conference on Internet Measurement Conference. ACM, pp 145–158

146. Gong NZ, Xu W, Huang L, Mittal P, Stefanov E, Sekar V, Song D (2012) Evolução de redes de atributos sociais: medições, modelagem e implicações usando o Google+. Em: Processos de a conferência ACM 2012 sobre medição da Internet. ACM, pp 131–144
147. Zheleva E, Sharara H, Getoor L (2009) Co-evolução de redes sociais e de afiliação. In: Anais da 15ª Conferência Internacional ACM SIGKDD sobre Descoberta de Conhecimento e Mineração de Dados. ACM, pp 1007–1016
148. Tang J, Sun J, Wang C, Yang Z (2009) Análise de influência social em redes de larga escala. In: Anais da 15ª conferência internacional ACM SIGKDD sobre descoberta de conhecimento e mineração de dados. ACM, pp 807–816
149. Li Y, Chen W, Wang Y, Zhang ZL (2013) Dinâmica de difusão de influência e maximização de influência em redes sociais com relacionamentos de amigos e inimigos. In Proceedings of the six ACM international conference on Web search and data mining. ACM, pp 657–666 150. Dai W, Chen Y, Xue GR,
Yang Q, Yu Y (2008) Aprendizagem traduzida: transferência de aprendizagem através de diferentes espaços de recursos: In: Avanços em sistemas de processamento de informações neurais. pp 353–360 151. Cisco Visual
Networking Index (2013) Atualização global de previsão de tráfego de dados móveis, 2012–2017 http://www.cisco.com/en.US/solutions/collateral/ ns341/ns525/ns537/ns705/ ns827/white paper c11-520862.html (Son eris¸im: 5 de maio de 2013)
152. Rhee Y, Lee J (2009) Sobre a modelagem de um modelo de comunidade móvel: projetando interfaces de usuário para apoiar a interação do grupo. Interactions 16(6):46–51
153. Han J, Lee JG, Gonzalez H, Li X (2008) Mineração massiva de conjuntos de dados rfid, trajetória e tráfego. In: Anais da 14ª conferência internacional ACM SIGKDD sobre descoberta de conhecimento e mineração de dados. ACM, p 2 154. Garg MK,
Kim DJ, Turaga DS, Prabhakaran B (2010) Análise multimodal de fluxos de dados de rede de sensores corporais para cuidados de saúde em tempo real. In: Proceedings of the International Conference on Multimedia Information Retrieval. ACM, pp 469–478
155. Park Y, Ghosh J (2012) Uma estrutura de imputação probabilística para análise preditiva usando dados de saúde de várias fontes agregados de forma variável. In: Anais do 2º Simpósio Internacional de Informática em Saúde ACM SIGHIT. ACM, pp 445–454

156. Tasevski P (2011) Ataques de senha e estratégias de geração. Tartu University: Faculdade de Matemática e Informática ciências