Pontifícia Universidade Católica de Minas Gerais
Instituto de Educação Continuada
Professor: Cláudio Lúcio
Introdução e contexto Big Data

Fragmentos extraídas do texto:

Bancos de Dados NoSQL: Conceitos, Ferramentas, Linguagens e Estudos de Casos no Contexto de Big Data

Marcos Rodrigues Vieira[1], Josiel Maimone de Figueiredo[2], Gustavo Liberatti[2], Alvaro Fellipe Mendes Viebrantz[2]

[1]IBM Research Laboratory - Brazil
mvieira@br.ibm.com

[2]Instituto de Computação
Universidade Federal de Mato Grosso (UFMT)
josiel@ic.ufmt.br, {liberatti.gustavo, alvarowolfx}@gmail.com

Um dos grandes desafios atualmente na área de Computação é a manipulação e processamento de grande quantidade de dados no contexto de Big Data. O conceito Big Data pode ser resumidamente definido como uma coleção de bases de dados tão complexa e volumosa que se torna muito difícil (ou impossível) e complexa fazer algumas operações simples (e.g., remoção, ordenação, sumarização) de forma eficiente utilizando Sistemas Gerenciadores de Bases de Dados (SGBD) tradicionais. Por causa desse problema, e outros demais, um novo conjunto de plataformas de ferramentas voltadas para Big Data tem sido propostas, como por exemplo Apache Hadoop [3].

A quantidade de dados gerada diariamente em vários domínios de aplicaçãocomo, por exemplo, da Web, rede sociais, redes de sensores, dados de sensoriamento,entre diversos outros, estão na ordem de algumas dezenas, ou centenas, de Terabytes. Essa imensa quantidade de dados gerados traz novos grandes desafios na forma de manipulação, armazenamento e processamento de consultas em várias áreas de computação,e em especial na área de bases de dados, mineração de dados e recuperação de informação. Nesse contexto, os SGBD tradicionais não são os mais adequados, ou "completos",às necessidades do domínio do problema de Big Data, como por exemplo: execução de consultas com baixa latência, tratamento de grandes volumes de dados, escalabilidade elástica horizontal, suporte a modelos flexíveis de armazenamento de dados, e suporte simples a replicação e distribuição dos dados.

…..

1.2.1. Big Data

O termo Big Data é bem amplo e ainda não existe um consenso comum em sua definição. Porém, Big Data pode ser resumidamente definido como o processamento (eficiente e escalável) analítico de grande volumes de dados complexos produzidos por (várias) aplicações. Exemplos de aplicações no contexto Big Data varia bastante, como aplicações científicas e de engenharias, redes sociais, redes de sensores, dados de Web Click, dados médicos e biológicos, transações de comércio eletrônico e financeiros, entre inúmeras outras. As semelhanças entre os dados desses exemplos de aplicações incluem: grande quantidade de dados distribuídos, características de escalabilidade sob demanda, operações ETL (Extract, Transform, Load) de dados "brutos" (raw) semi- ou não estruturados para dados estruturados e, a necessidade de extrair conhecimento da grande quantidade de dados.

Três fatores influenciaram o grande aumento de volume de dados sendo coletados e

armazenados para posterior análise: difusão dos dispositivos captação de dados, dispositivo com armazenamento na ordem de Terabytes e aumento de velocidade de transmissão nas redes. Os dispositivos de aquisição, bem como os dispositivos de armazenamento de grande escala se difundiram principalmente pelo seu barateamento (e.g., redes de sensores, GPS, smartphones), enquanto que as redes aumentaram sua velocidade e abrangência geográfica. Outro fator importante é a facilidade de geração e aquisição de dados gerados digitalmente, como máquinas fotográficas digitais, smartphones, GPS, etc. Como consequência novas demandas estão surgindo, como a demanda por análise de grande volume de dados em tempo real (data analytics), o aumento do detalhamento das informações, bem como plataformas escaláveis e eficientes de baixo custo.

Basicamente, podemos resumir as características do contexto Big Data em quatro propriedades: (1) dados na ordem de dezenas ou centenas de Terabytes (podendo chegar a ordem de Petabytes), (2) poder de crescimento elástico, (3) distribuição do processamento dos dados; e (4) tipos de dados variados, complexos e/ou semiestruturados. A característica de análise dos dados na ordem de Terabytes envolve, entre outros aspectos, o requisito de alto poder computacional de armazenamento e processamento dos dados. A elasticidade esta relacionada ao fato de que a quantidade de dados pode variar de alguns Megabytes a vários Terabytes (e vice-versa) em um espaço de tempo relativamente curto, fazendo com que a estrutura de software/hardware adapta-se sob demanda, i.e. seja alocada/desalocada dinamicamente. A distribuição significa que os dados devem ser distribuídos de forma transparente em vários nós espalhados de processamento, o que demanda armazenamento, processamento e controle de falhas distribuído. Finalmente, a quarta característica esta relacionada a adoção de modelos mais apropriados, flexíveis e eficientes para o armazenamento de tipos de dados variados e semi-estruturados. Vale ressaltar que o modelo relacional não é o mais adequado pois não possui flexibilidade para o armazenamento de dados e evolução no modelo para tais tipos de dados citados acima.

A análise de dados (data analytics) no contexto de Big Data normalmente envolve processamento da ordem de Terabytes em dados de baixo valor (i.e., informação original “bruta”) que são transformados para dados de maior valor (e.g., valores agregados/sumarizados). Mesmo com a grande quantidade de dados Big Data em si não garante a qualidade da informação, pois a análise continua, em grande parte, sendo muito subjetiva. Isso se deve ao fato que os dados em si não são autoexplicativos, onde o processo de limpeza, amostragem, e relacionamento dos dados continua sendo crítico e passível a erros, aproximações e incertezas [14]. Por exemplo, a análise de dados da ordem de Petabytes (ou Exabytes) de cunho científicos (e.g., dados genômicos, física ambiental e simulações numéricas) tem se tornado muito comum hoje em dia, onde é aceitável que o resultado da análise contenham imprecisão (i.e., erro entre certos limites de erros), porém seja computado de forma (relativamente) rápida e/ou em tempo real.

Recentemente, ambientes de computação em nuvem (cloud computing) têm sido utilizados para o gerenciamento de dados em forma de Big Data, enfocando principalmente em duas tecnologias: Bases de Dados Como Serviço (Database as a Service (DaaS)) e Infraestrutura Como Serviço (Infrastructure as a service (IaaS)) (para maiores detalhes). DaaS utiliza um conjunto de ferramentas que permite o gerenciamento remoto dos servidores de dados mantidos por uma infraestrutura externa sob demanda. Essa infraestrutura IaaS fornece elasticidade, pagamento sob demanda, backup automáticos, e rapidez de implantação e entrega.

As principais características que envolvem os ambientes em nuvem são: escalabilidade, elasticidade, tolerância a falhas, auto gerenciamento e a possibilidade de funcionar em hardware commodity (comum). Por outro lado, a maioria dos primeiros SGBD relacionais comerciais foram desenvolvidos e concebidos para execução em ambientes corporativos. Em um ambiente de computação em nuvem traz inúmeros desafios do ponto de vista computacional. Por exemplo, o controle de transação na forma tradicional (i.e., definida pelas propriedades ACID) em um ambiente de nuvem é extremamente complexo.

De uma maneira geral, os ambientes em nuvem precisam ter a capacidade de suportar alta carga de atualizações e processos de análises de dados. Os data centers são uma das bases da computação em nuvem, pois uma grande estrutura como serviço escalável e dinâmica é fornecida para vários clientes. Um ambiente de gerenciamento de dados escalável (scalable data management) pode ser dividido em: (1) uma aplicação complexa com um SGBD gerenciando uma grande quantidade de dados (single tenant); e (2) uma aplicação no qual o ambiente deve suportar um grande número de aplicações com dados possivelmente não muito volumosos [2]. A influência direta dessas duas características é que o SGBD deve fornecer um mesmo esquema genérico para inúmeros clientes com diferentes aplicações, termo denominado bases de dados multitenant.

É importante lembrar que em ambientes multitenant a soma dos tenant pode gerar um grande volume de dados armazenado no SGBD. Este característica é apropriadamente gerenciada pelo ambiente em nuvem, onde novas instâncias de SGBD são criadas em novos nós e/ou servidores do ambiente (os dados de diferentes tenant são independentes entre si).

Em ambientes tradicionais, quando uma aplicação cresce sua complexidade o SGBD atende às requisições e a escalabilidade do sistema no modo que aumenta o seu poder computacional. Por exemplo, o poder computacional do sistema como um todo cresce a medida que mais memória e/ou número de nós do cluster são adicionados ao servidor. No entanto, esta abordagem são caras, dependentes de arquitetura, e normalmente não presentes em SGBD livres (open sources).

A Revolução do Big Data – Jornal O Globo 02/07/2012

Você vai ao mercado com o objetivo de comprar apenas o que falta para o jantar e, ao passar pelo corredor de produtos de higiene, seu celular o surpreende com uma mensagem. O remetente é a própria varejista, que deseja chamar sua atenção para o desodorante em promoção na prateleira ali do lado. O SMS não diz, mas a rede sabe que o seu estoque do produto está mesmo no fim e que, há duas semanas, você escreveu no Facebook o quanto gostava da marca. Se a precisão da mensagem lhe é espantosa, prepare-se: a tecnologia que cruza dados dessa forma já existe, representa um mercado direto estimado em US$ 70 bilhões e está invadindo empresas e governos no Brasil e no mundo — o que deve elevar à enésima potência a possibilidade de ganhos com o uso dessas informações. A promessa é de revolução em várias áreas da economia e até na ciência — além de uma renovada discussão sobre privacidade.

Trata-se do Big Data, termo de mercado para o conjunto de soluções que analisa informações em variedade, volume e velocidade inéditos até hoje. Ferramentas desse tipo surgiram no fim da década passada, mas este ano o conceito extrapolou de vez os limites da academia e dos setores de TI. Isso porque o preço para armazenamento de dados está despencando e diversas ferramentas baratas ou gratuitas para lidar com tamanho volume informações estão surgindo.

O uso dessa nova tecnologia tem vasta abrangência. No último Fórum Econômico Mundial, em Davos, foi publicado um estudo mostrando como o Big Data pode ser uma arma contra problemas socioeconômicos. E até Brad Pitt tem contribuído para sua popularização: o filme "Moneyball (O Homem que mudou o jogo)", que protagoniza, conta a história da mais famosa aplicação do conceito: o gerente de um time de beisebol que usa o Big Data para reunir um elenco de primeira linha sem gastar muito.

"Pré-sal existe por causa do Big Data"

O executivo de operações da EMC, Pat Gelsinger, diz que o mercado global de Big Data já movimenta US$ 70 bilhões por ano — sem contar inestimáveis ganhos nos negócios. A consultoria IDC estima que o segmento crescerá quase 40% ao ano entre 2010 e 2015, mas considera um patamar de US$16,9 bilhões ao fim desse período. A tecnologia envolve tanto dinheiro porque soluciona um problema inadiável para a economia, o da quantidade de dados digitais. O volume deve crescer do atual 1,8 zettabyte para 7,9 zettabytes em 2015, prevê a IDC. Zettabyte equivale a um trilhão de gigabytes.

A centelha dessa revolução é a proliferação de plataformas que geram dados como nunca. São celulares, GPS, redes sociais, câmeras e sensores dos mais diversos tipos. Grande parte da informação gerada é classificada de não-estruturada: ou seja, não é facilmente computável, costuma ser criada pelo ser humano, não por máquina. Até pouco tempo, essa informação só podia ser compreendida por pessoas. Com o Big Data, as máquinas aprendem a lê-la. Essa é, nas palavras de especialistas, a beleza do conceito.

Nos últimos 50 anos, a evolução do mercado de TI se deu apenas no "T" da sigla, a tecnologia. Com o Big Data, é chegada a hora de o "I", de inteligência, guiar o avanço — afirmou Alexandre Kazuki, diretor de marketing da divisão da HP Brasil que cuida de Big Data.

Se o Big Data está dando os primeiros passos no mundo, a tecnologia ainda engatinha no Brasil, na avaliação de Kátia Vaskys, diretora de Bussiness Analytics da IBM. Ela cita a forma como a maioria das empresas brasileiras monitora suas marcas nas redes sociais:

- Aqui costuma-se contratar um time de estagiários para isso. Isso é basear a estratégia de marketing na intuição, mas não há intuição que resista a tanta informação! Há uma ferramenta tecnológica para fazer isso com muito mais precisão e em tempo real.
- A aplicação por aqui está restrita a setores como varejo, financeiro (análise de risco),

telecomunicações e petrolífero, mas começa a chegar à mídia.

- A Renner usa o Big Data para monitorar em tempo real o fluxo de mercadorias da loja ao cruzar dados de localização GPS dos caminhões dos fornecedores com os níveis dos estoques, contou o diretor de TI Leandro Balbinot. A rede também acompanha a aceitação dos produtos nas redes sociais. E já imagina outros usos, como a possibilidade de reorganizar uma loja com base em dados meteorológicos. Exemplo: se, nas últimas chuvas, os clientes compraram mais calças ou acessórios, a rede pode dar destaque a esses itens quando os primeiros pingos caírem.

Apesar de o uso no Brasil ainda ser pouco maduro, a expectativa é enorme. Temos um dos principais mercados de internet no mundo, sobretudo de redes sociais, o que é crucial para a adesão ao Big Data — observou Maurício Prado, gerente geral de servidores da Microsoft Brasil.

Só sabemos que o pré-sal existe por causa do Big Data e da economia da nuvem — resumiu Patrícia Florissi, CTO da EMC para as Américas.

Isso porque a tecnologia agiliza o processamento de dados sísmicos captados pelas sondas que procuram petróleo no fundo do mar. Como são milhões as variáveis, o trabalho exige intermináveis simulações de imagens, e só o Big Data é capaz de dar conta do trabalho em um tempo razoável.

Visando a esse mercado, a gigante EMC está construindo no Parque Tecnológico do Fundão um centro de pesquisas totalmente dedicado ao uso de Big Data para a indústria do petróleo. Ele ficará pronto em no máximo dois anos e empregará 35 pesquisadores. A companhia vai investir R$ 100 milhões no país nos próximos quatro anos.

Polícia chega antes do crime

Há também iniciativas brasileiras na seara governamental, aceleradas pela proximidade da Lei de Acesso à Informação, que entra em vigor em maio. Uma parceria do Ministério do Planejamento, do Serviço Federal de Processamento de Dados (Serpro) e da PUC Rio disponibilizou na web dados abertos dos mandatos do governo Lula.

A massificação do Big Data, porém, enfrenta obstáculos. O maior deles é com a privacidade. Mas, para Karin Breitman, da PUC-Rio, os cientistas não devem "censurar" pesquisas:

É uma questão ética. Cabe à sociedade impor limites à aplicação da ciência e da tecnologia, mas os pesquisadores precisam trabalhar na ponta.

Outro problema é a escassez de profissionais com habilidades em matemática, estatística e computação. O Big Data levou as empresas a uma disputa frenética por esse perfil e tornou a IBM a maior empregadora de matemáticos PhDs no mundo. O instituto McKinsey prevê que faltarão até 190 mil desses profissionais em 2018 nos EUA.

Já há carência desse profissional no Brasil. Se houver uma explosão do Big Data, teremos problemas — advertiu Kazuki, da HP.

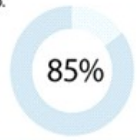
Apesar dos desafios, a expectativa é enorme. A "Economist" escreveu que o Big Data pode transformar modelos de negócio de empresas centenárias. A RollsRoyce, cita, deixaria de vender turbinas para alugá-las, cobrando pelo uso. Sensores e o histórico do cliente dariam o preço.

Patrícia Florissi, da EMC, diz que ainda é incipiente o uso da presciência da tecnologia. Por exemplo: como são capazes de entender imagens, softwares de Big Data poderiam monitorar as câmeras de uma cidade e acionar a polícia antes de um crime acontecer com base em padrões que antecedem assaltos e assassinatos. Sairíamos de "Moneyball" para cair em "Minority Report" — com os prós e contras disso.

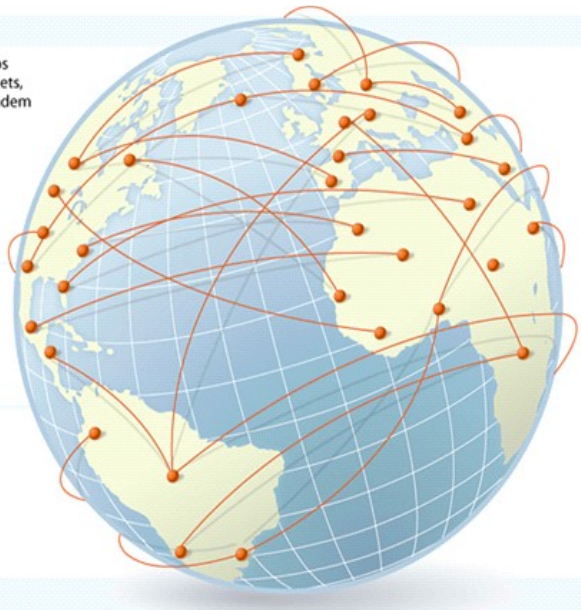
A NOVIDADE

A grande novidade das soluções de Big Data é lidar também com os chamados dados não-estruturados, que até então só podiam ser compreendidos por pessoas. São tweets, posts no Facebook, vídeos, geolocalização e comportamentos de clientes que dependem de contexto para ter sentido.

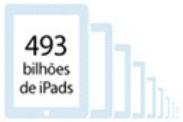
Esses dados não-estruturados representam 85% das informações com as quais as empresas lidam hoje



O mercado de Big Data crescerá quase 40% ao ano até 2015



A quantidade global de dados digitais deve crescer de 1,8 zettabyte, hoje, para 7,9 zettabytes em 2015. Daqui a três anos, toda a informação do mundo poderá ser armazenada em:



COMPARE

1 Zettabyte é igual
1.000.000.000.000.000.000.000 bytes

1 Gigabyte é igual
1.000.000.000 bytes

Alguns exemplos de como a solução tem sido usada

A companhia Skybox tira fotos de satélite e vende a seus clientes informações em tempo real sobre a disponibilidades de vagas de estacionamento livres numa cidade em determinada hora ou quantos navios estão ancorados no mundo neste momento

O projeto Global Pulse, das Nações Unidas, vai utilizar um programa que decifra a linguagem humana na análise de mensagens de texto e posts em redes sociais para prever o aumento do desemprego, o esfriamento econômico e epidemias de doenças

A varejista americana Dollar General monitora as combinações de produtos que seus clientes põem nos carrinhos. Ganhou eficácia e ainda descobriu curiosidades: quem bebe Gatorade tem mais chances de comprar também laxante

A Sprint Nextel saltou da última para a primeira posição no ranking de satisfação dos usuários de celular nos EUA ao integrar os dados de todos os seus canais de relacionamento. De quebra, cortou pela metade os gastos com call center

No terremoto do Haiti, pesquisadores americanos perceberam antes de todo mundo a diáspora de Porto Príncipe por meio dos dados de geolocalização de 2 milhões de chips SIM de celulares, facilitando a atuação da ajuda humanitária

Um hospital no Canadá usou tecnologia da IBM e da Universidade de Ontário para monitorar em tempo real dezenas de indicadores de saúde de bebês prematuros. O cruzamento permitiu aos médicos antecipar ameaças às vidas das crianças

Em busca dos melhores lugares para instalar turbinas eólicas, a dinamarquesa Vestas Wind analisou petabytes de dados climáticos, de nível das marés, mapas de desmatamento etc. O que costumava levar semanas durou algumas horas