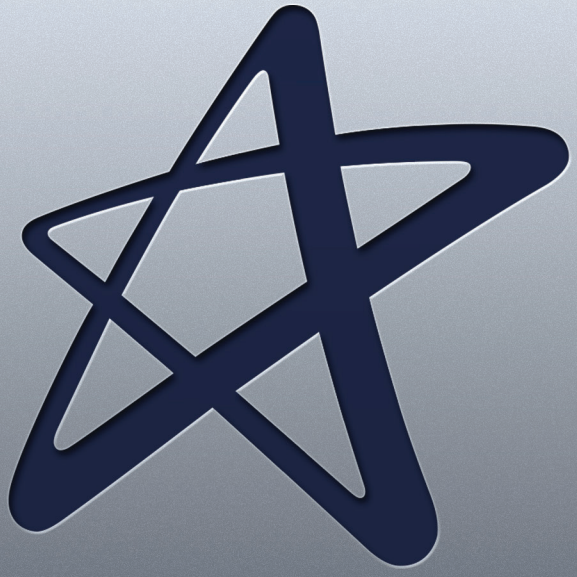
Pós-Graduação

Bancos de Dados Não Estruturados



Cruzeiro do Sul Virtual
Educação a Distância

Introdução a Big Data

Responsável pelo Conteúdo:
Prof. Esp. Milton Roberto y Goya

Revisão Textual:
Profa. Dra. Silvia Albert

Unidade

Introdução a Big Data

Nesta unidade, trabalharemos os seguintes tópicos:

- Introdução ao Tema
- Orientações para leitura Obrigatória
- Material Complementar

Fonte: iStock/Getty Images

Objetivos

- Apresentar conceitos de Big Data
- Apresentar conceitos de Hadoop
- Apresentar conceitos de NoSQL

Caro Aluno(a)!

Normalmente, com a correria do dia a dia, não nos organizamos e deixamos para o último momento o acesso ao estudo, o que implicará o não aprofundamento no material trabalhado ou, ainda, a perda dos prazos para o lançamento das atividades solicitadas.

Assim, organize seus estudos de maneira que entrem na sua rotina. Por exemplo, você poderá escolher um dia ao longo da semana ou um determinado horário todos ou alguns dias e determinar como o seu "momento do estudo".

No material de cada Unidade, há videoaulas e leituras indicadas, assim como sugestões de materiais complementares, elementos didáticos que ampliarão sua interpretação e auxiliarão o pleno entendimento dos temas abordados.

Após o contato com o conteúdo proposto, participe dos debates mediados em fóruns de discussão, pois estes ajudarão a verificar o quanto você absorveu do conteúdo, além de propiciar o contato com seus colegas e tutores, o que se apresenta como rico espaço de troca de ideias e aprendizagem.

Bons Estudos!

Introdução ao Tema

Ao analisarmos as notícias e tendências do mundo da tecnologia da informação podemos perceber que existe um termo que está sendo repetido com muita frequência: "Big Data". Esse é um campo de estudo empolgante, que engloba vários aspectos tecnológicos nem sempre bem compreendidos. Nesta unidade, você será apresentado a conceitos que envolvem esse maravilhoso mundo e que serão importantes para a compreensão do funcionamento dos bancos de dados não-relacionais.

A primeira impressão que o termo Big Data nos passa é que é uma tecnologia para trabalhar com grande volume de dados. Essa ideia é, em parte, verdadeira, mas esse campo abrange muito mais do que isso.

Sistemas comerciais e da área científica vêm trabalhando com grandes volumes de dados há muito tempo. Algumas áreas da ciência, como a astronomia, geologia e meteorologia, produzem um volume enorme de dados que são processados com precisão há muito tempo, muito antes do termo Big Data surgir. Big Data não deve ser definido - apenas - em número de terabytes ou petabytes. A quantidade de bytes pode mudar dependendo do tipo de dados, do setor em que a tecnologia está sendo aplicada ou da própria evolução da tecnologia. É interessante notar, entretanto, que um estudo da empresa de pesquisa IDC aponta que até 2020 teremos produzido 40 zettabytes de bytes, isso equivale a 40 trilhões de gigabytes.

Outro desafio com que o Big Data lida é a variedade de fontes de dados. Apenas 10% (dez por cento) dos dados disponíveis são estruturados, isto é, estão no formato tradicional de tuplas em um banco de dados, os outros 90% (noventa por cento) dos dados estão em formato não-estruturados, sendo que esses últimos são provenientes de várias fontes como: contratos, formulários, imagens, manuais, raios-x, e-mails, PDFs, mensagens instantâneas, documentos, páginas da WEB, áudio e vídeo.

Tratar grandes volumes de dados e grande variedade de dados nos leva a perguntar: qual é o tempo de resposta desse ambiente? Quanto tempo estamos dispostos a esperar por uma resposta do sistema? É comum que as respostas a essas perguntas seja "o mais rápido possível, de preferência instantaneamente".

Segundo o McKinsey Global Institute (2011), Big Data refere-se:

> "ao conjunto de dados cujo tamanho está além da capacidade de análise, captura, armazenamento e gerenciamento de uma ferramenta desenvolvida por um software típico".

McKINSEY GLOBAL INSTITUTE, 2011, p.27.

Podemos dizer, então, que Big Data trata de grandes volumes de dados (Volume), com informações vindas de diversas fontes (Variedade) e produzem respostas rápidas (Velocidade). Nesta unidade, iremos considerar apenas estes três requisitos.

Para abordar os três requisitos que definimos para esta unidade, precisamos nos colocar questões que pertencem a dois aspectos que envolvem Big Data:

- **Declaração de problema:** Como processar Big Data usando o estado-da-arte da tecnologia atual sem "estourar" o limite de tempo e o orçamento?

Onde está o "gargalo" do processamento? A velocidade das CPUs está cada vez maior, mas a velocidade de acesso a disco ou a volumes de discos convencionais, ainda é lenta. O aumento da velocidade de CPU não beneficia muito os programas que têm necessidade de acessar grandes volumes de dados.

Em 2000, Eric Brewer, em sua palestra na Universidade de Berkley, em São Francisco, nos Estados Unidos,(checar ortografia) propôs o teorema CAP. Essencialmente o teorema afirma que em qualquer sistema distribuído *stateful* é preciso escolher entre

- ***Consistency*** (consistência forte). Todos os nós veem os mesmos dados ao mesmo tempo.
- ***Availability*** (alta disponibilidade). Toda solicitação recebe uma resposta, seja ela bem-sucedida ou não.
- ***Network Partition Tolerance*** (tolerância a particionamento dos dados na rede). O sistema continua funcionando mesmo que mensagens sejam perdidas ou parte do sistema falhe.

Entre essas três propriedades, somente duas podem ser garantidas ao mesmo tempo.

Lembramos que sistemas de banco de dados tradicionais possuem consistência forte e alta disponibilidade, mas não trabalham muito bem com tolerância a particionamento dos dados na rede. Já o sistema de bancos de dados não-relacionais são sistemas altamente tolerantes a falhas, desde que exista um grande número de servidores suportando o sistema, fornecem disponibilidade, mas apenas consistência eventual.

Veremos a seguir alguns sistemas que auxiliam a vencer essa limitação imposta tanto pelo banco de dados estruturados quanto pelo banco de dados não-relacionais.

Em relação à tolerância a particionamento, veremos como ela pode ser fornecida pelo **Hadoop Distributed File System** (HDFS). O **Hadoop Distributed File System** (HDFS) é um sistema de arquivos altamente tolerante a falhas, projetado para executar em hardware, padrão de baixo custo. O HDFS disponibiliza acesso de alto rendimento para os dados do aplicativo e é adequado para aplicativos com grandes conjuntos de dados.

Segundo a Hadoop, "Hadoop é um storage confiável e um sistema analítico" [2014], composto por duas partes essenciais:

- Hadoop Distributed Filesystem (HDFS), sistema de arquivos distribuído e confiável, responsável pelo armazenamento dos dados
- Hadoop MapReduce, responsável pela análise e processamento dos dados.

Uma curiosidade: o nome "Hadoop" veio do elefante de pelúcia que pertencia ao filho do criador, Doug Cutting.

Vamos entender melhor uma dessas duas partes essenciais, o MapReduce?

Segundo a IBM [2009], MapReduce é um "Modelo de programação que permite o processamento de dados massivos em um algorítmo paralelo e distribuído, geralmente em um cluster de computadores" (IBM, 2009, p. 2). MapReduce é baseado nas operações de Map e Reduce de linguagens funcionais como o LISP. Ele trata os dados como um conjunto de pares <Key, Value> (chave/valor). As operações de entrada leem os dados e geram os pares <Key, Value> e o usuário fornece duas funções Map e Reduce, que são chamadas em tempo de execução.

Em uma operação típica de Map, o nó principal recebe os dados, divide em partes menores e envia aos outros nós para serem processados. Ao final do processamento estes nós devolvem o resultado ao nó principal. O nó principal o obtêm uma lista de pares <Key, Value>, processa os pares e gera um conjunto de pares <Key, Value> intermediário. O nó principal, então, repassa o valor intermediário para a função Reduce. Cada par é processado em paralelo.

Em uma operação típica de Reduce, o nó principal combina as respostas obtidas pelos outros nós gerando o resultado final do processamento. O nó principal processa todos os valores associados com a mesma <Key> e o nó principal mescla os valores para formar um conjunto de valores possivelmente menor. Geralmente, apenas um valor de saída de 0 ou 1 é produzido a cada chamada Reduce. Os valores intermediários são fornecidos à função Reduce do usuário por um iterador, o que permite identificar listas de valores que são grandes demais para a memória.

Um banco de dados não-relacional também é conhecido como banco NoSQL. e pode trabalhar com o modelo <Chave, Valor> ou <Key, Value>. Veremos mais à frente que existem vários modelos de armazenamento no banco de dados NoSQL, sendo que um dos mais populares é o modelo de <Chave>, <Valor>.

O termo NoSQL é de 1998 e é abreviação para "**not-only SQL**", indicando que não processaria apenas instruções SQL. O site https://goo.gl/Z3Rh fornece uma lista atualizada dos principais bancos NoSQL disponíveis no mercado. Alguns dos principais bancos NoSQL são:

- Oracle NoSQL
- IBM Cloudant
- Cassandra
- Voldemort
- MongoDB
- BigTable
- DynamoDB

Uma de suas características marcantes é que ele é "schema free", ou seja, não segue o modelo tradicional dos bancos de dados relacional, permitindo que cada registro armazenado em seu banco tenha uma estrutura diferente do registro anterior.

É um banco que permite escalabilidade horizontal, ou seja, permite aumentar o número de máquinas disponíveis. A escalabilidade horizontal em modelos relacionais seria inviável devido a concorrência. Como nos modelos NoSQL não existem bloqueios, esse tipo de escalabilidade é a mais viável.

O NoSQL possui suporte à replicação de uma forma nativa o que provê uma escalabilidade maior e também uma diminuição do tempo gasto para a recuperação de informações. Uma API simples é fornecida para que o acesso às informações seja feito da forma mais rápida possível.

Um banco de dados NoSQL pode usar um dos seguintes modelos de dados:

- Chave/Valor (Key/value)
- Colunas (Columnar)
- Documento (Document)
- Grafos (Graph)

O modelo Chave/Valor é o modelo mais simples. Permite a visualização do banco como uma grande tabela. Todo o banco é composto por um conjunto de chaves que estão associadas a um único valor. O valor é armazenado, sem preocupação com o que representa e a aplicação faz o tratamento e se preocupa com o entendimento do valor. É muito escalar devido ao acesso somente pela chave, podendo armazenar tudo em um *bucket* ou criar *buckets* de domínio. Limitação: consulta só pela chave, retornando o valor. Valor não pode ser consultado pelo atributo.

Bom Para:

- Armazenamento de informações de sessão
- Perfil de usuário e preferências
- Dados de carrinho de compra

Ruim para:

- Relacionamento entre dados
- Transações com múltiplas operações
- Consultas por dados e atributos
- Operações por conjuntos

Iremos trabalhar com um banco de dados orientado para documentos, chamado MongoDB. O banco MongoDB usa a estrutura <Chave><Valor>, uma IDE *on-line*: https://goo.gl/5Ymclu O software para instalação gratuita pode ser encontrado em https://goo.gl/0gZsq7

Trataremos da instalação, configuração e operação do banco na próxima unidade.

Orientações para leitura Obrigatória

Considerando que a ideia central desta Unidade é lhe mostrar os principais conceitos que envolvem o banco de dados não-relacional, é de suma importância que primeiro conheçamos os princípios de banco de dados relacional.

O livro "Sistemas de banco de dados - 7ª ed.", de Elmasri e Navathe, oferece uma introdução bastante completa ao amplo campo de sistemas de bancos de dados.

O livro apresenta uma base sólida sobre os alicerces da tecnologia de bancos de dados, ao mesmo tempo em que esclarece como o campo deve se desenvolver no futuro.

Sistemas de banco de dados - 7ª ed. - Ramez Elmasri, Shamkant B. Navathe

Após entrar em sua "área do aluno", disponível em: https://goo.gl/h8Ur9g, no menu à esquerda da tela, clique em "Serviços", depois em "Biblioteca" e, no centro da tela, clique em "E-books - Bib. Virtual Universitária". No topo da tela que abrirá, haverá um campo de busca para autor, título, assunto etc., nele digite "Sistemas de banco de dados" (sem as aspas) e clique na capa que aparecer como resultado.

Note a seta ao lado direito da tela para avançar página a página, assim como perceba que os ícones no rodapé da tela correspondem a determinadas funções, entre as quais ampliar a visualização (zoom), marcar a obra como favorita, imprimir trechos que escolher e pular para um número específico de página.

Material Complementar

Indicações para saber mais sobre os assuntos abordados nesta Unidade:

Vídeos

Big Data Storymap

História do Big Data, como surgiu e evoluiu ao longo dos anos.

https://youtu.be/iFyGuvyesw4

Big Data Architecture Patterns

Vários modelos de implementação de Big Data. Altamente recomendado para os inciantes.

https://youtu.be/-N9i-YXoQBE

Explaining Big Data

Aprofunda alguns conceitos vistos no vídeo anterior (Big Data Architecture Patterns). É interessante assistir este vídeo agora e revê-lo após ter estudado o módulo 2.

https://youtu.be/7D1CQ_L0izA

Introduction to Hadoop

Introdução gráfica ao Hadoop. É interessante assisti-lo para consolidar os conceitos de Map/Reduce

https://youtu.be/Pq30yQ0-I3E

Referências

DATE, C.J.. **Introdução a Sistemas de Bancos de Dados**. Trad. Daniel Vieira. Rio de Janeiro: Elsiever Editora, 2015.

Elmasri Rames, Navathe B. Shamkant. **Sistemas de banco de dados**. Trad. Daniel Vieira; revisão técnica Enzo Seraphim, Thatyana de Faria Piola Seraphim - 7. ed. - São Paulo: Pearson Education do Brasil, 2018.

Portal Educação, Google Analytics. Disponível em: <http://www.portaleducacao.com.br/informatica/artigos/48358/google-analytics>. Acesso em 3 de dezembro de 2016.

IBM, Big Data para Impacientes <http://www.ibm.com/developerworks/br/data/library/techarticle/dm-1209hadoopbigdata/>. Acesso em 3 de janeiro de 2017

IBM, Processamento de Dados Distribuído com Hadoop <https://www.ibm.com/developerworks/br/cloud/library/l-hadoop-1/index.html>. Acesso em 3 de janeiro de 2017

Cruzeiro do Sul Virtual
Educação a Distância

www.cruzeirodosulvirtual.com.br
Campus Liberdade
Rua Galvão Bueno, 868
CEP 01506-000
São Paulo - SP - Brasil
Tel: (55 11) 3385-3000


Cruzeiro do Sul
Educacional