CEFET-MG

PÓS-GRADUAÇÃO *LATO SENSU*

Curso: Banco de Dados

Disciplina: *Data Warehouse e Business Intelligence*

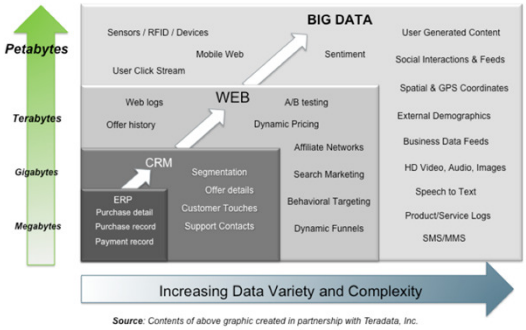
Professor: Fernando Zaidan

Unidade 7 – Big Data

2012

Big Data

Big Data = Transactions + Interactions + Observations



Source: Contents of above graphic created in partnership with Teradata, Inc.

Big Data - Contexto

- **Globalização**
- **Modelo "just in time"**
- **Expansão virtual**

- A partir de 2000 houve uma crescente de dados exponencial que já preocupam os especialistas pela falta de espaço.

Big Data - Contexto

- Em **2008** foram produzidos cerca de **2,5 quintilhões** de bytes **todos os dias** (IBM)

- **90%** dos dados no mundo foram criados **nos últimos dois anos**, decorrente a adesão das grandes empresas à internet, como exemplo as redes sociais, dados dos GPS, dispositivos embutidos e móbil.

Big Data

=
volume
+
variedade
+
velocidade de dados

(+ veracidade
+ valor)

Big Data = volume + variedade + velocidade de dados

- Há **Volume** porque além dos dados gerados pelos sistemas transacionais, temos a imensidão de dados gerados pelos objetos na Internet das Coisas, como sensores e câmeras, e os gerados nas mídias sociais via PCs, smartphones e tablets.

Big Data = volume + variedade + velocidade de dados

- **Variedade** porque estamos tratando tanto de dados textuais estruturados como não estruturados como fotos, vídeos, e-mails e tuites.

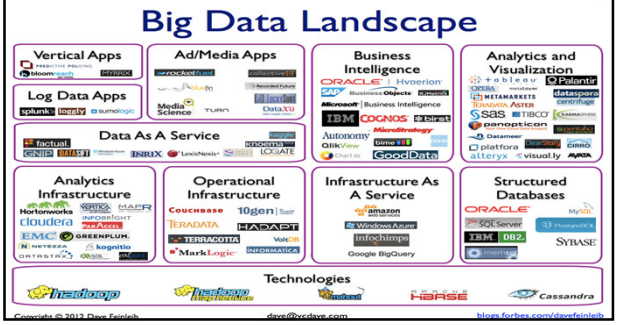
A visualização de dados em forma de cores criado pela IBM.

Big Data = volume + variedade + velocidade de dados

- **Velocidade**, porque muitas vezes precisamos responder aos eventos quase que em tempo real, ou seja, estamos falando de criação e tratamento de dados em volumes massivos.

Big Data = volume + variedade + velocidade de dados

Big Data são dados que testam os limites das tecnologias disponíveis para utilizá-los.



Big Data - Fundamento

- Permitem encontrar padrões e sentido em uma imensa e variada massa amorfa de dados gerados por sistemas transacionais, mídias sociais, sensores, etc.

- É crucial saber tratar os dados na velocidade adequada.

Big Data - Fundamento

- Dados não tratados e analisados em tempo hábil são dados inúteis, pois não geram informação.

- Dados passam a ser ativos corporativos importantes e como tal podem e deverão ser quantificados economicamente.

Big Data - Fundamento

- Portanto, Big Data cria valor para as empresas descobrindo padrões e relacionamentos entre dados que antes estavam perdidos não apenas em data warehouses internos, mas na própria Web, em tuítes, comentários no Facebook e mesmo vídeos no YouTube, assim como RFID.

Big Data - Infraestrutura

As tecnologias que sustentam Big Data podem ser analisadas sob duas óticas:

- As envolvidas com analytics, tendo **Hadoop** e **MapReduce** como nomes principais;
- E as tecnologias de infraestrutura, que armazenam e processam os petabytes (chegando aos zetabyetes) de dados. Neste aspecto, destacam-se os bancos de dados **NoSQL**. Por que estas tecnologias? Por que Big Data é a simples constatação prática que o imenso volume de dados gerados a cada dia excede a capacidade das tecnologias atuais de os tratarem adequadamente.

Fonte: IBM, 2012.

Big Data - Infraestrutura

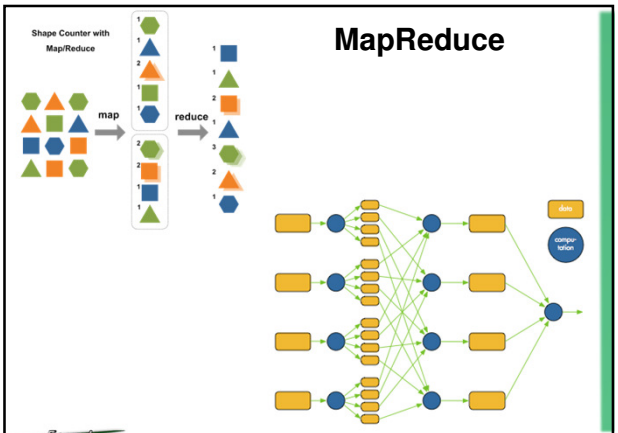
MapReduce é um modelo de programação para o processamento de grandes conjuntos de dados, bem como o nome de uma implementação do modelo pelo Google.

MapReduce é normalmente usado para fazer a computação distribuída em clusters de computadores.

O modelo é inspirado no mapa e visa reduzir as funções comumente usadas na programação funcional.

Fonte: Wikipedia, 2012.

MapReduce



Big Data - Infraestrutura

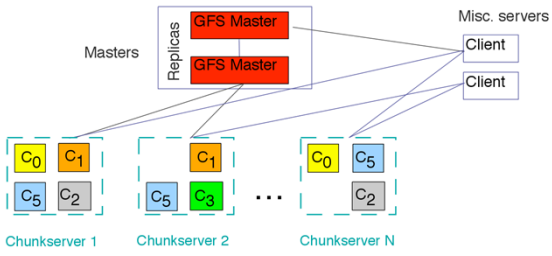
Google File System (GFS ou GoogleFS) é um sistema de arquivos distribuídos proprietária desenvolvida pela Google para seu próprio uso.

Ele é projetado para acesso eficiente e confiável de dados através de grandes conjuntode hardware.

Uma nova versão do Sistema de Arquivos do Google tem o codinome Colossus.

Fonte: Wikipedia, 2012.

Google File System



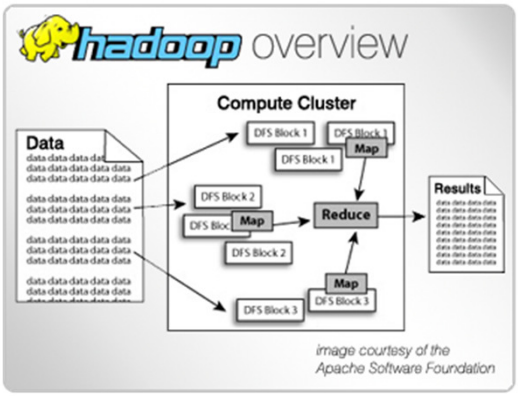
Big Data - Infraestrutura

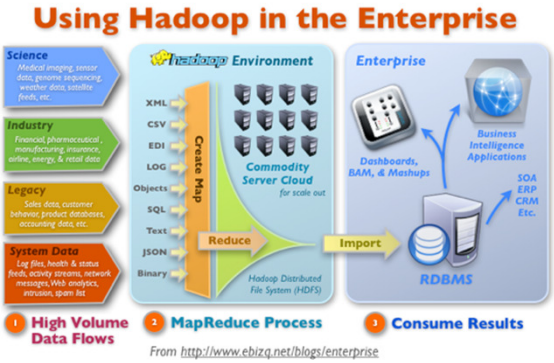
Hadoop é uma plataforma de software em Java de computação distribuída voltada para clusters e processamento de grandes massas de dados. Foi inspirado pelo MapReduce e Google File System.

Trata-se de um projeto de alto-nível da Apache que vai sendo construído por uma comunidade utilizando a linguagem JAVA.

A Yahoo tem sido o maior patrocinador do projeto, utilizando-o intensivamente no seu negócio.

Fonte: Wikipedia, 2012.





Big Data - Infraestrutura

NoSQL

Termo genérico para uma classe definida de banco de dados não-relacionais que rompe uma longa história de banco de dados relacionais com propriedades ACID.

Outros termos equivalentes para esta categoria de bancos é *NF²*, *N1NF* (*non first normal form*), *nested relational*, *dimensional*, *multivalue*, *free-form*, *schemaless*, *document database* e MRNN (Modelo Relacional Não Normalizado). Os banco de dados que estão sob estes rótulos não podem exigir esquemas de tabela fixa e, geralmente, não suportam instruções e operações de junção SQL.

Fonte: Wikipedia, 2012.

Big Data - Infraestrutura

Tendências em arquiteturas de computadores, como a computação na nuvem, e a necessidade crescente de prover serviços escaláveis, fazem surgir novas tecnologias.

Há alguns exemplos de softwares de código fechado que atendem estes requisitos, sendo alguns deles Google BigTable e Amazon DynamoDB. E alguns exemplos de sofware open-source como Apache Cassandra (originalmente desenvolvido para o Facebook), Apache HBase, Linkedins Project Voldemort e dentre outros.

É importante entender que o intuito não é eliminar bancos de dados relacionais, mas oferecer uma alternativa.

Fonte: Wikipedia, 2012.

Big Data - Analytics

Depois da infraestrutura é necessário atenção aos componentes de analytics, pois estes é que transformam os dados em algo de valor para o negócio.

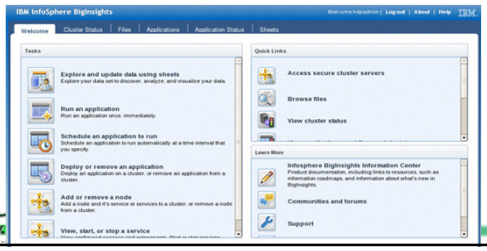
Big Data Analytics não signfica eliminar os tradicionais sistemas de BI que existem hoje, mas pelo contrário, devem coexistir.

Um bom exemplo de uso de Hadoop para analytics é o BigInsights da IBM.

Fonte: IBM, 2012.

Big Data - Analytics

BigInsights IBM InfoSphere traz o poder do Hadoop para a empresa. Permite que empresas de todos os tamanhos custo efetivamente gerenciar e analisar o enorme volume, variedade e velocidade de dados que os consumidores e as empresas a criar todos os dias. Fonte: IBM, 2012.

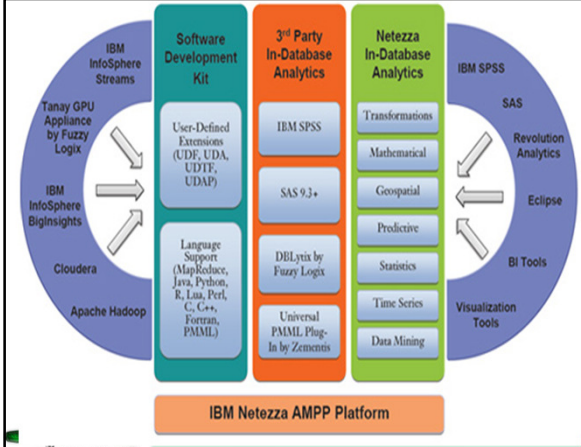


Big Data - Analytics

Netezza IBM - Família de Ferramentas (appliance) de Data Warehouse para análise de negócios, com banco de dados, servidores e storages integrados. De fácil implementação, otimizado e pronto para o uso, com manutenção contínua e nenhuma necessidade de tuning.

Com servidor, storage e banco de dados em um mesmo equipamento desenvolvido exclusivamente para a função, equipes geram dados confiáveis para a tomada de decisões em segundos.

Fonte: IBM, 2012.



Big Data - Analytics

Stream computing: um novo paradigma. No modelo de data mining tradicional uma empresa filtra dados dos seus vários sistemas e após criar um Data Warehouse, dispara "queries".

Na prática faz-se garimpagem em cima de dados estáticos, que não refletem o momento, mas sim o contexto de horas, dias ou mesmo semanas atrás.

Com stream computing esta garimpagem é efetuada em tempo real. Em vez de disparar queries em cima de uma base de dados estática, coloca-se uma corrente contínua de dados (streaming data) atravessando um conjunto de queries. Fonte: IBM, 2012.

Big Data - Analytics

Stream computing

Podemos pensar em inúmeras aplicações, sejam estas em finanças, saúde e mesmo manufatura.

Vamos ver este último exemplo: um projeto em desenvolvimento com uma empresa de fabricação de semicondutores pode monitorar em tempo real o processo de detecção e classificação de falhas. Com stream computing as falhas nos chips sendo fabricados são detectados em minutos e não horas ou mesmo semanas. Os wafers defeituosos podem ser reprocessados e, mais importante ainda, pode-se fazer ajustes em tempo real nos próprios processos de fabricação. Fonte: IBM, 2012.

Big Data - na prática

- Uma companhia que tira fotos de satélites e vende aos seus clientes informações em tempo real sobre a disponibilidade de vagas de estacionamento livres em uma cidade numa determinada hora.

- Uma varejista americana controla as combinações de produtos que seus clientes põem no carrinho, ou seja, ganhou eficácia e ainda descobriu várias curiosidades que podem ajudar.

Big Data – case Pixomondo

Criar, analisar, armazenar e acessar grandes quantidades de dados digitais é uma questão urgente para muitas organizações, e sua solução pode colocar algumas delas no tapete vermelho. Uma estrela de criação e análise de dados é a Pixomondo, um estúdio de efeitos visuais indicado para o Prêmio da Academia de efeitos visuais pelo filme "A Invenção de Hugo Cabret".

PIXOMONDO
VISUAL EFFECTS

Big Data – case Pixomondo

Todos os componentes de cada frame nas 854 cenas em que a Pixomondo trabalhou, desde o relógio de Hugo até um menu ou um guarda-chuva, representam um conjunto de dados que devem ser codificados, armazenados e manipulados.

Big Data – case Pixomondo

A Pixomondo tem uma rede global com 7 estúdios,
Alemanha,
Los Angeles,
Burbank,
Londres,
Xangai,
Pequim,
Nova York
e Toronto
para trabalhar em um ciclo de 24 horas que leva os dados de um escritório a outro, seguindo o sol.

Big Data – case Pixomondo

A Pixomondo não poderia perder um dia de produtividade; o cronograma de edição do Hugo estava 20 semanas mais curto que o normal.

Para tal empregou o Big Data na análise de dados.

As tecnologias usadas por um estúdio global de efeitos visuais que nunca interrompe suas operações estão disponíveis para qualquer negócio sem nenhuma mágica.

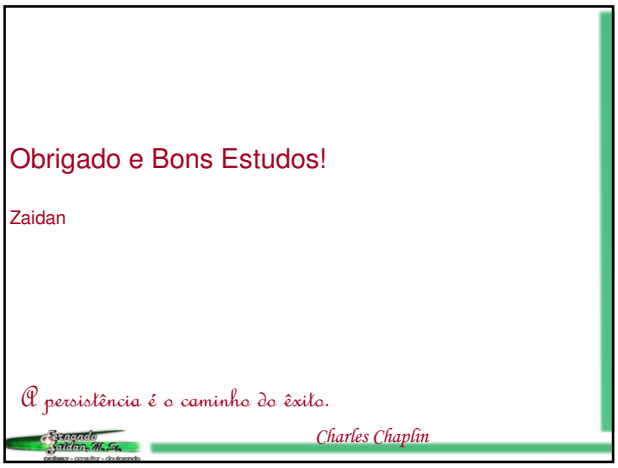
Big Data – Novo profissional

A Gartner prevê que, até 2015, a procura por recursos humanos relacionada coma Big Data levará à criação de 4,4 milhões de empregos globalmente. Mas apenas um terço dos postos de trabalho será preenchido.

Fonte: Computerworld, 2012.

Leiam o texto: **7 new types of jobs created by Big Data**
http://www.smartplanet.com/blog/bulletin/7-new-types-of-jobs-created-by-big-data/682

Big Data – Leitura complementar sobre nteligência dos Negócios – Thomas Davenport


Obrigado e Bons Estudos!
Zaidan
A persistência é o caminho do êxito.
Charles Chaplin