



# BUSINESS INTELLIGENCE E DATA WAREHOUSE

Big Data  
Conceitos, Técnicas,  
Ferramentas e Arquitetura



# Apresentação - Cetax

- ❑ A Cetax é uma empresa de consultoria e treinamento especializada em sistemas de Business Intelligence e Data Warehouse.
- ❑ Existe desde 2000 trabalhando exclusivamente com BI e DW.
- ❑ Nossos treinamentos são exclusivos sem cursos semelhantes no Brasil
- ❑ Outros cursos são ministrados em parcerias com outras empresas do mercado ou mesmo profissionais que possuem experiência diferenciada

Quem somos

# Apresentação - Instrutor

❑ Marco Antonio Garcia

- ❑ 19 anos de experiência em TI, sendo 14 exclusivamente com Inteligência - Business Intelligence e Data Warehouse.
- ❑ MBA pela FGV, Formado pela FATEC em Processamento de Dados.
- ❑ Certificado pelo Kimball University nos EUA, onde teve aula pessoalmente com Ralph Kimball, um dos principais gurus do data Warehouse, treinamentos realizados no TDWI, maior entidade de pesquisa de Data Warehouses do mundo.
- ❑ Vivência profissional em diversos projetos, passando por Bancos e Financeiras, Construção, Serviços, Varejo, Marketing e outros.

# Objetivos do Curso

- Definir, Conceituar e desmistificar o assunto Big Data
- Demonstrar Requerimentos de Big Data
- Demonstrar Casos de Uso
- Conceituar as Ferramentas necessárias
- Listar algumas aplicações possíveis
- Demonstrar o Ecossistema Hadoop e seus componentes
- Players de Mercado
- Cientista de Dados, perfil e requisitos para a função
- Requisitos para trabalhar com Big Data
- O Data Warehouse para Big Data

Objetivos do Curso de Big Data

## Não são objetivos do curso

- Demonstrar e ensinar uma ferramenta ou software.
- Demonstrar alguma marca de software.

Alinhamento das expectativas dos alunos

# Agenda do Curso

## ❑ Parte 1

- ❑ Conceitos e Fundamentos
- ❑ Data is the New Oil !
- ❑ Alguns números – Brasil e Mundo
- ❑ Ferramentas e Arquitetura

## ❑ Parte 2

- ❑ Nova Geração de Tecnologias
- ❑ Aplicações de Big Data
- ❑ The Big Data – Data Warehouse

## ❑ Parte 3

- ❑ Hadoop e suas ferramentas
- ❑ Map Reduce
- ❑ Players de Mercado

## ❑ Parte 4

- ❑ Cientista de Dados
- ❑ Quero Trabalhar com Big Data
- ❑ Perguntas de Respostas

Itens que serão discutidos e apresentados no treinamento.



A tempos não temos uma palavra tão forte no cenário de informática como Big Data !

O termo está sendo falado em todos os tipos de negócios, cursos, etc.

The screenshot shows the Buscapé website interface. At the top, there is a navigation bar with links for 'Buscapé', 'Bcash', 'SaveMe', 'e-bit', and 'Mais'. Below the navigation bar is a yellow header with the 'buscapé' logo on the left and a search bar containing the text 'big data'. To the right of the search bar are 'BUSCAR' and 'CATEGORIAS' buttons. A green oval highlights the breadcrumb trail 'Livros > big data > Encontramos 81 produtos'. Below the header, the page title 'Literatura Disponível – Brasil' is displayed. The main content area shows a product listing for the book 'Big Data - Como Extrair Volume, Variedade, Velocidade e Valor da Avalanche de Informação Cotidiana - Viktor Mayer-schönberger, Kenneth'. The listing includes the book cover image, the title, author, and price information: 'De: R\$ 26,91' and 'até: R\$ 57,90 em 12x'. There are also filters on the left for 'Filtros selecionados' (big data) and 'Reputação: Todas'.

O que existe disponível sobre o Big Data no Brasil

The screenshot shows the Amazon website interface. At the top, the Amazon logo and 'Try Prime' button are visible. Below the logo, there are links for 'MARCO's Amazon.com', 'Today's Deals', 'Gift Cards', 'Sell', and 'Help'. A search bar contains the text 'big data'. To the left of the search bar, a dropdown menu says 'Shop by Department' with 'Books' selected. Other categories like 'Advanced Search', 'New Releases', 'Best Sellers', 'The New York Times® Best Sellers', 'Children's Books', and 'Te' are also listed. A green oval highlights the search results summary: '1-12 of 3,608 results for Books : Computers & Technology : "big data"'.

On the left sidebar, under 'Show results for Books', there are links for 'Any Category', 'Books', 'Computers & Technology', 'Data Mining (425)', 'Databases (825)', 'Information Theory (271)', 'Modeling & Simulation (224)', 'Information Visualization', and 'Graphic Design (144)'. The main content area displays a book cover for 'Big Data, Big Analytics' with a 33% discount offer. Below the book cover, 'Related Searches' include 'big data analytics', 'data science', and 'hadoop'. Under 'Book Format', there are buttons for 'Paperback', 'Hardcover', and 'Kindle Edition'.

O que existe disponível sobre o Big Data nos Eua.

## Big Data = Grandes Dados ?

- Muitas definições podem cercar o assunto :
  - Alto Volume.
  - Alta Velocidade.
  - Diversas Fontes.
- Uma combinação de tudo isso e muito mais.
- Assim como BI, é um termo “guarda-chuva”.

O termo é recente, muitas possibilidades, muitas definições ( algumas ainda vagas )

O mercado está em formação muitas coisas estão ainda acontecendo e muitas ainda estão por vir.

O que é certo : Big Data é uma tendência que vai mudar a maneira em que analisamos os dados em qualquer tipo de negócio !

## Muitos Dados Gerados

- Além dos sistemas utilizados em empresas de todos os portes, temos milhares de outros dispositivos que geram dados diariamente :
  - Em 2010 existiam 5 bilhões de celulares no mundo.
  - Um avião Boeing pode gerar até 20 TB/hora para seus engenheiros examinar em tempo real.
  - Em pouco tempo teremos muito mais equipamentos ligados a internet gerando informações para análise “internet das coisas”

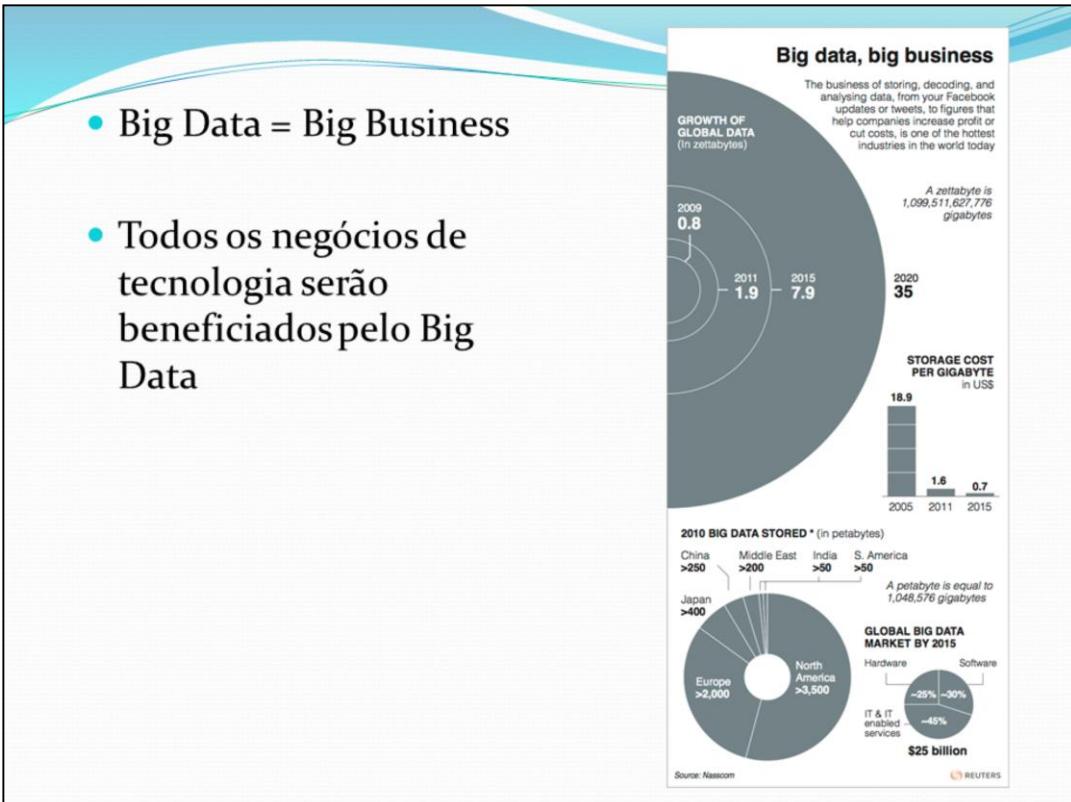
O **Facebook** armazena, acessa e analisa mais de 50 **petabytes** de informações geradas pelos usuários, a cada mês são gerados mais de 700 milhões de minutos por mês.

A cada minuto são feitos uploads de 48 horas de vídeos no **Youtube**, ou seja, nunca ninguém conseguirá assistir todos os vídeos do **Youtube**.

Diariamente mais de 500 milhões de mensagens são enviadas pelo **Twitter**, com uma média de 5700 TPS (Twittes per Second ou Mensagens por Segundo), o recorde é de 143.199 TPS.

O **Google** processa diariamente mais de 3 bilhões de pesquisas em todo o mundo, sendo desse total 15% totalmente inéditas. Seu "motor" de pesquisa rastreia 20 bilhões de sites diariamente, armazenando 100 **petabytes** de informação.

Sem contar todas as informações que as companhias geram diariamente, sejam elas estruturadas ou não.



Fonte Imagem – Nasscom e Reuters

## 3 Vs – uma definição

- **Volume** – o volume crescente de dados em todas as áreas e empresas.
- **Velocidade** – o tempo necessário para disponibilizar os dados para análise é cada vez menor
- **Variedade** – a variedade de dados é cada vez maior, sensores, imagens, dados não estruturados ou semi estruturados.

**Volume – Volume dos Dados:** Passamos a falar muito rápido de Gigabytes para Terabytes e agora estamos falando de Petabytes e outros volumes que não vou saber colocar aqui de cabeça para vocês.

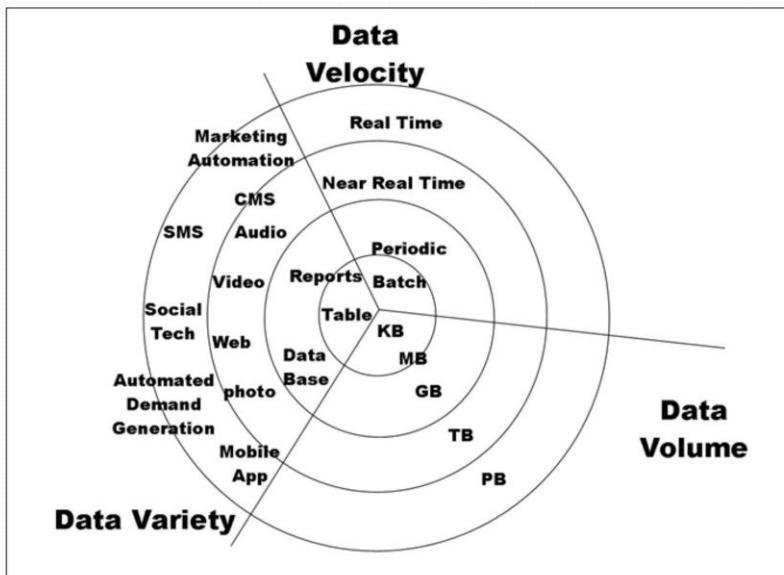
Hoje são contabilizados em média 12 Terabytes de Tweets diariamente, em 2012 foram gerados cerca de 2.834 Exabytes (que são milhões de Gigabytes) a previsão é que em 2020 se gerem anualmente 40.026 Exabytes de informações.

**Velocity – Velocidade:** Hoje para alguns negócios, 1 minuto pode ser muito tempo, detecção de fraudes, liberações de pagamentos, análises de dados médicos ou qualquer outra informação sensível a tempo.

A maior parte dos projetos de DW/BI (Data Warehouse e Business Intelligence) ainda tem latência em D-1, ou seja, carregamos o dia anterior. Ainda acreditamos que essa solução se aplique a muitos negócios, porém, para algumas análises, quanto mais próximo do tempo real, maior pode ser o incremento de negócio.

**Variety – Variedade:** Big Data também poderia ser considerado como Any Data (qualquer dado), hoje temos capacidade de capturar e analisar dados estruturados e não estruturados, texto, sensores, navegação Web, áudio, vídeo, arquivos de logs, catracas, centrais de ar condicionado, entre outros.

## 3Vs - Detalhes



<http://beyondplm.com/2013/10/14/will-plm-data-size-reach-yottabytes/>

Detalhamento sobre os 3 V's.

## 5 Vs – Uma definição

- Volume
- Velocidade
- Variedade
- Virtude
- Valor

Alguns estudiosos acrescentaram mais V's a definição de big data

## 10 Vs – é necessário ?

- |              |                  |
|--------------|------------------|
| • Vast       | - Vasto, Amplo   |
| • Volume     | - Alto Volume    |
| • Vigorosity | - Vigor          |
| • Verified   | - Verificados    |
| • Vexingly   | - “Atormentador” |
| • Variable   | - Variaveis      |
| •Verbose     | - “Eloquente”    |
| • Valuable   | - Valiosos       |
| • Visualized | - Visualizados   |
| • Velocity   | - Velocidade     |

Existe até essa definição de 10 V's do Big Data.

Mas seria ela necessária ?

## Big Data – Definição Simples

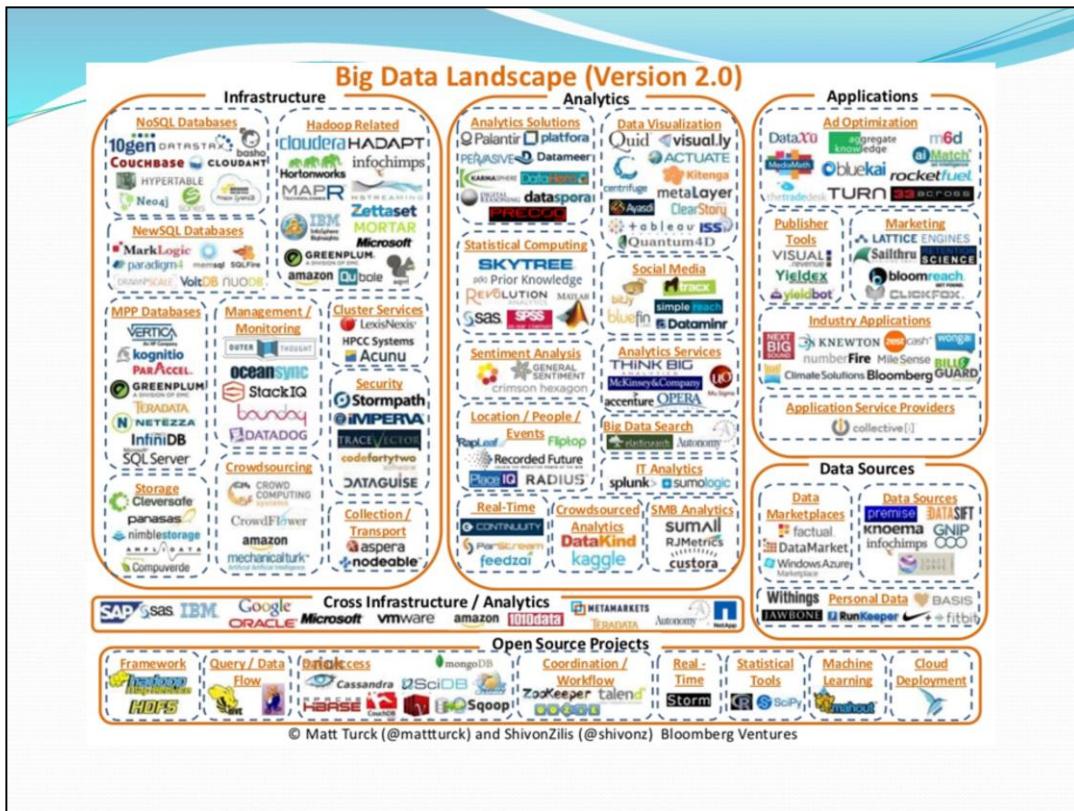
- Big Data representa um conjunto de dados que não pode mais ser facilmente gerenciado ou analisado com as ferramentas atuais de dados, métodos ou arquitetura disponível até então.

Definição simples e direta, algo que não pode mais ser feito com as ferramentas atuais !

## E então ?

- E então ?
- Quais softwares serão utilizados ?
- Quais devo aprender ?

O que temos que fazer ? Quais softwares serão usados ?



## Muitos softwares ?

Por favor, se acalme, vamos falar disso um pouco mais para frente.

# Data is the new Oil !

- “Dados são o novo Petróleo”



**Perry Rotella**  
Contributor

**FOLLOW**

[full bio →](#)

Opinions expressed by Forbes Contributors are their own.

**TECH** 4/02/2012 @ 11:09AM | 10,791 views

## Is Data The New Oil?

[+ Comment Now](#) [+ Follow Comments](#)

Recently, on a CNBC Squawk Box segment, “[The Pulse of Silicon Valley](#),” host Joe Kernen posed the question, “What is the next really big thing?” to [Ann Winblad](#), the legendary investor and senior partner at Hummer-Winblad. Her response: “Data is the new oil.”

- Como petróleo, precisam ser refinados !

Os dados podem ser o novo petróleo, a nova corrida que as empresas vão enfrentar para multiplicar seus lucros!

A correta coleta, processamento e análise dos dados podem ser um diferencial competitivo a todos os negócios.

Claro, como petróleo, os dados também precisam ser refinados para um melhor resultado.

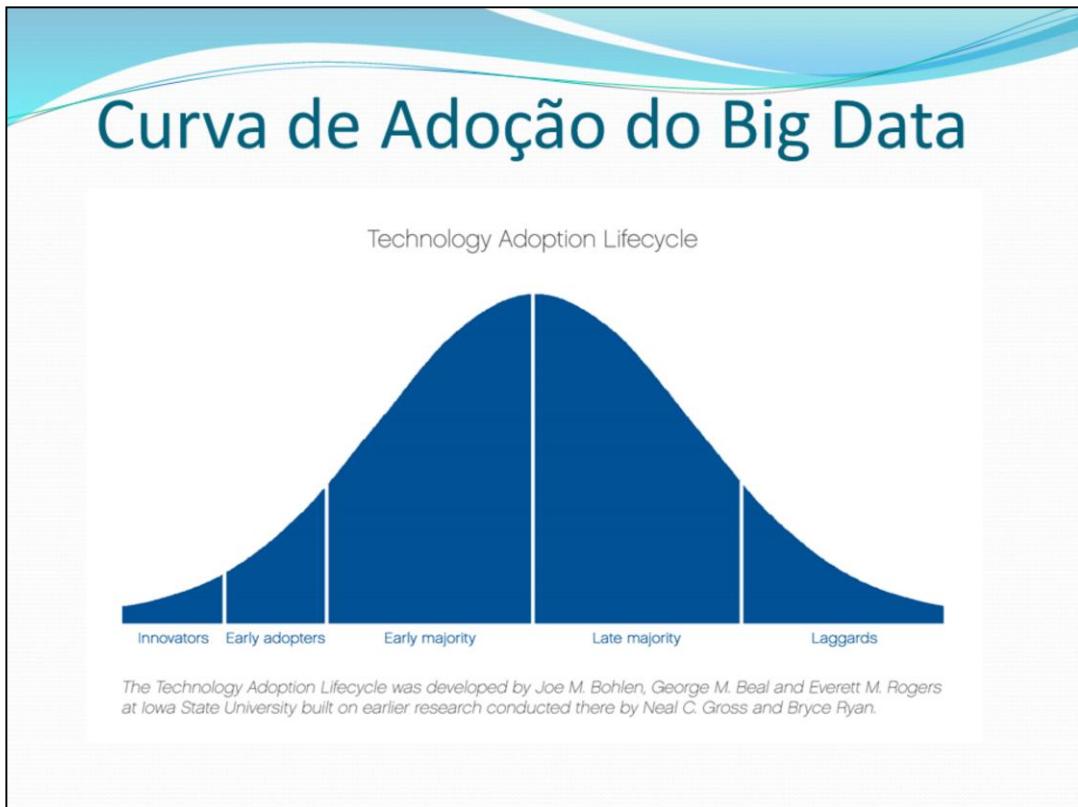
## Fontes para o Big Data

- Web log
- Click stream
- Sensor data
- Email
- Call center voice logs
- Images/video
- Dados RFID
- Dados de Localização e Geográficos
- Dados adquiridos no mercado

Essa lista é um exemplo de possíveis fontes, mas deveremos ter muito mais fontes.

As novas ferramentas permitem conexão e captura de dados em diversas categorias de softwares ou mesmo equipamentos eletrônicos que permita captura de dados.

Claro que além dos dados tradicionais que hoje buscamos em outros sistemas, bancos de dados e arquivos de texto.



Curva de Adoção de Big Data :

**Innovators / Inovadores** – nessa categoria temos as empresas de Internet, Tecnologia e Varejistas

**Early Adopters / Adiantados** – Mercado Financeiro

**Early Majority / Adiantados a Maioria** – Utilities, Infraestrutura, Serviços Públicos

**Late Majority / Atrasados a Maioria** – Manufatura, Saúde

**Laggards / Preguiçosos**

## Exemplos por Setor

- Serviços de Informação – Imagens de Satélites
- Varejo – Otimização de Preço, Inteligência de Vídeo
- Utilities/Utilidades – Consumo em tempo real
- Propaganda – Content Targeting OnLine
- Seguros – Detecção de Fraudes
- Saúde – Diagnósticos e Detecção.
- Manufatura – Controle de Qualidade, Controle de Máquinas

Alguns Exemplos de Big Data por Setor

## Aplicações Possíveis - Big Data

- Machine Learning
- Sentiments
- Text Processing
- Image Processing
- Video Analytics
- Log Parsing
- Collaborative Filtering
- Context Search
- Email & Content

Algumas aplicações possíveis ( atualmente ) em Big Data

## Aplicações Possíveis - Big Data

- Machine Learning :
  - Aprendizado de Maquina
  - Sistemas que podem aprender com os dados e tomar decisões de acordo com modelos previamente estudados.
  - Existem diversos algoritmos que podem ser aplicados em Machine Learning.

Aplicações Possíveis – Machine Learning

Algumas referências

[http://en.wikipedia.org/wiki/Machine\\_learning](http://en.wikipedia.org/wiki/Machine_learning)

<http://research.google.com/pubs/ArtificialIntelligenceandMachineLearning.html>

## Aplicações Possíveis - Big Data

- Sentiments :

- Sentimentos ou Análise de Sentimentos
- Também conhecida como Mineração de Opinião ou Opinion Mining
- São algoritmos utilizados para identificar o “sentimento” através de uma informação textual.

Aplicações Possíveis – Sentiments Analytics

Algumas referências

[http://en.wikipedia.org/wiki/Sentiment\\_analysis](http://en.wikipedia.org/wiki/Sentiment_analysis)

[https://developers.google.com/prediction/docs/sentiment\\_analysis](https://developers.google.com/prediction/docs/sentiment_analysis)

## Aplicações Possíveis - Big Data

- Text Processing :
  - Processamento de Textos
  - Pode ser utilizado para identificação, reconhecimento ou análise de dados através de textos processados.
  - Pode ser usado na identificação de comportamentos ou sentimentos

### Aplicações Possíveis – Text Processing

Além das citadas acima, pode ser um meio de identificar keywords, semantica e outros detalhes sobre textos.

## Aplicações Possíveis - Big Data

- Image Processing / Video Analytics :
  - Imagens e Videos podem ser analisados em ferramentas de Big Data.
  - As imagens podem ser decompostas e através de processos de reconhecimento, identificar pessoas, padrões suspeitos ou mesmo padrões que possam ser analíticos.
  - Os videos podem ser decompostos em uma série de imagens e essas imagens processadas.

Aplicações Possíveis – Image Processing e Video Analytics

Imagens e Videos também pode ser fontes de análises de big data.

## Aplicações Possíveis - Big Data

- Log Parsing
- Collaborative Filtering
- Context Search
- Email & Content

Algumas aplicações possíveis ( atualmente ) em Big Data

## Requerimentos do Big Data

- Load ( Carga )
- Structure ( Estrutura )
- Response ( Resposta )
- Complex Workload ( Processamentos Complexos )
- Economics ( Retorno ao Investimento )

Requerimentos dos Processos de Big Data

# Dados por Validade para Big Data

Source: VoltDB, Inc.

Data Age				
Interactive	Real time Analytics	Record Lookup	Historical Analytics	Exploratory Analytics
<i>Milliseconds</i>	<i>Hundredths of seconds</i>	<i>Second(s)</i>	<i>Minutes</i>	<i>Hours</i>
. Place trade . Serve ad . Enrich stream . Examine packet . Approve trans.	. Calculate risk . Leaderboard . Aggregate . Count	. Retrieve click stream . Show orders	. Backtest algo . BI . Daily reports	. Algo discovery . Log analysis . Fraud pattern match

Referencia - <http://voltdb.com/blog/big-data/big-data-value-continuum/>

## Como atendemos hoje ?

- Bancos de Dados Relacionais
- Ferramentas de ETL
- Ferramentas de BI
- Mas esse é o melhor ferramental ?
- Com certeza não !

## Qual o impacto de tudo isso no DW Atual ?

- Novos Tipos de Dados
- Novos Volumes
- Novas Análises
- Novas Cargas de Processamento
- Novos Metados

Tudo isso = Menos Performance !

As tecnologias correntes não estão prontas para todo o impacto que o Big Data está causando nos negócios.

É necessário combinar novas soluções com aquilo que estamos fazendo hoje.

## RDBMS – SGBDR (ACID)

- Nossos bancos de dados são baseados em ACID
  - Atomic – Todo trabalho será salvo ao commit
  - Consistent – o banco de dados trata consistencia de dados.
  - Isolated – o resultado das transações não são visíveis até a transação ser completada (commit).
  - Durable - o resultado “pós” commit é durável, sobrevive a falhas

Atualmente utilizamos bancos de dados relacionais para armazenamento de Data Warehouse e Sistemas de Inteligência.

Porém o foco dessa tecnologia é controlar transações e não grandes volumes de dados.

# Inovações são necessárias

Categorias	Novas Tecnologias
Infra Estrutura	Big Data e Data Warehouse Appliances Tecnologias In-Memory SSD Storage Fast Networks Cloud Computing Tecnologias Móveis
Softwares	In-Memory Databases Hadoop, Cassandra e NoSQL Databases Colunar DBMS ETLs com integração Hadoop – Informatica, Talend, etc.
Algoritimos	Mahout
Arquiteturas Pré-Configuradas	IBM, Teradata, Kognitio, EMC, Cloudera, HortonWorks, Cirro, Intel, Cisco UCS, Pivotal, Oracle, MapR

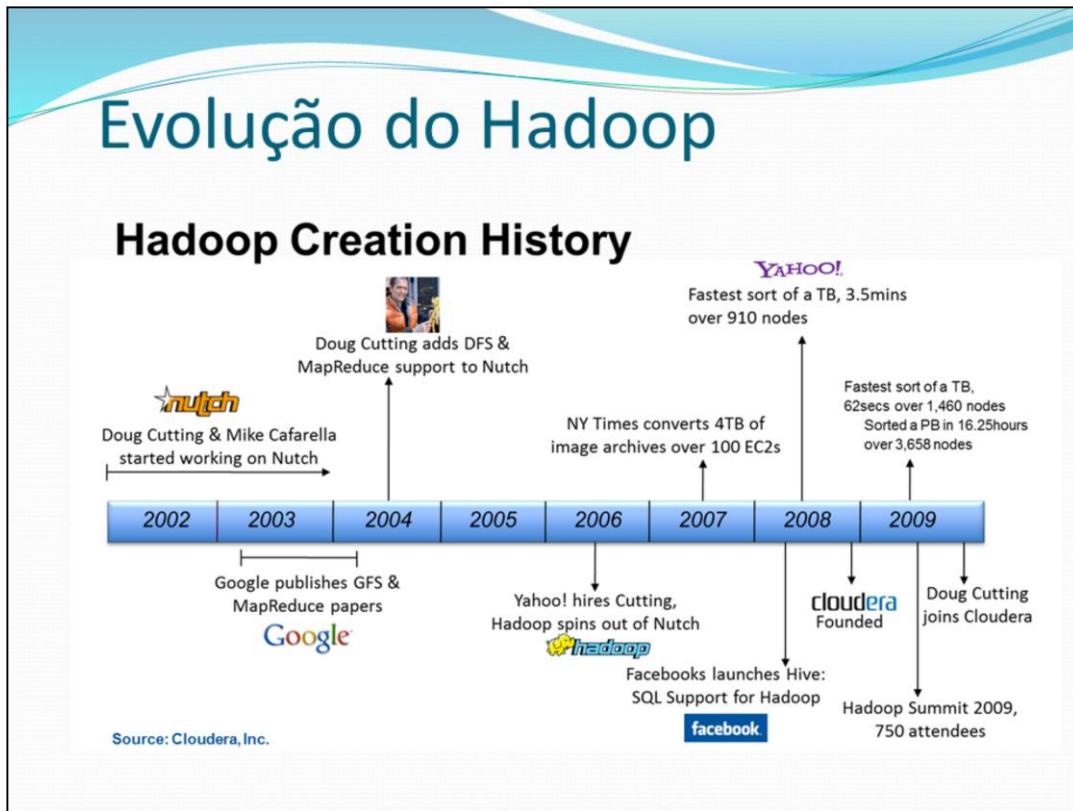
As novas tecnologias disponíveis tem auxiliado a evoluir com Big Data.

Todas as tecnologias tem permitido captura, processar e analisar grandes volumes de dados.

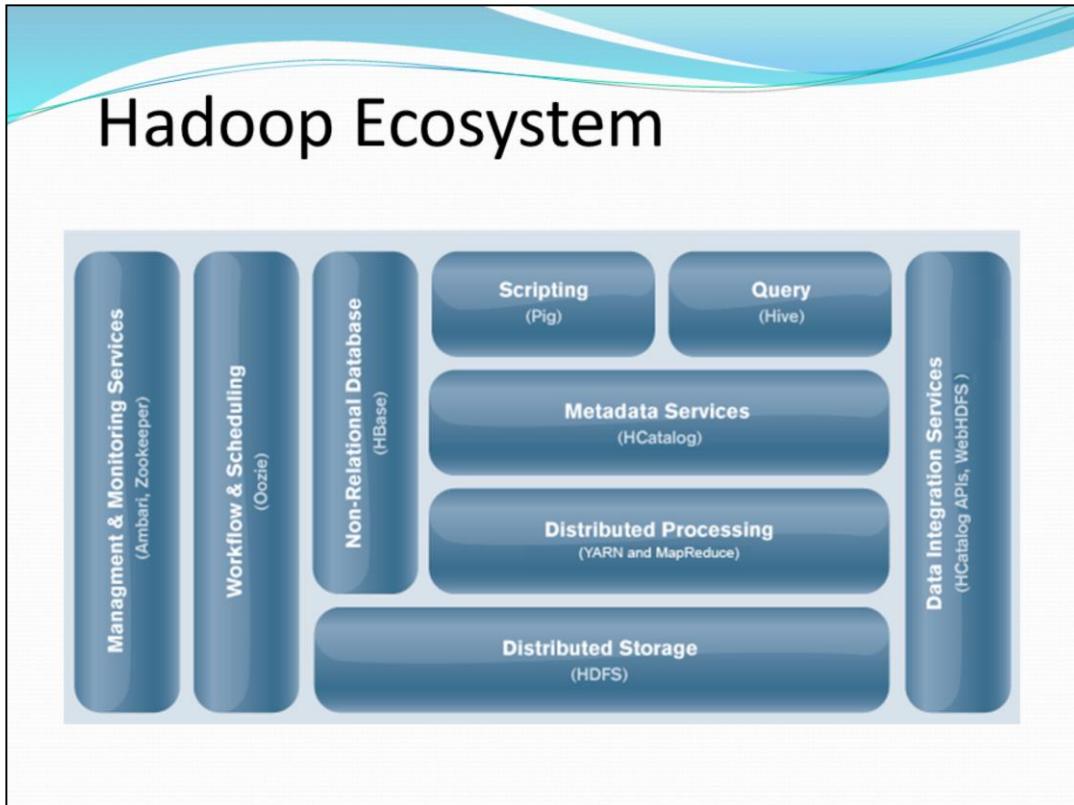
## Hadoop surge como alternativa

- Criado dentro do Yahoo em 2005.
- O nome Hadoop não é um acrônimo, é o nome do Elefante de brinquedo do filho de Doug Cutting.
- O Hadoop tem como base um file system para armazenamento rápido e barato de dados !
- Tem como objetivo o processamento de grandes datasets de dados com modelos simples de programação

O Hadoop surge para auxiliar nos “problemas” gerados pelos grandes volumes de dados.



Fonte - Cloudera



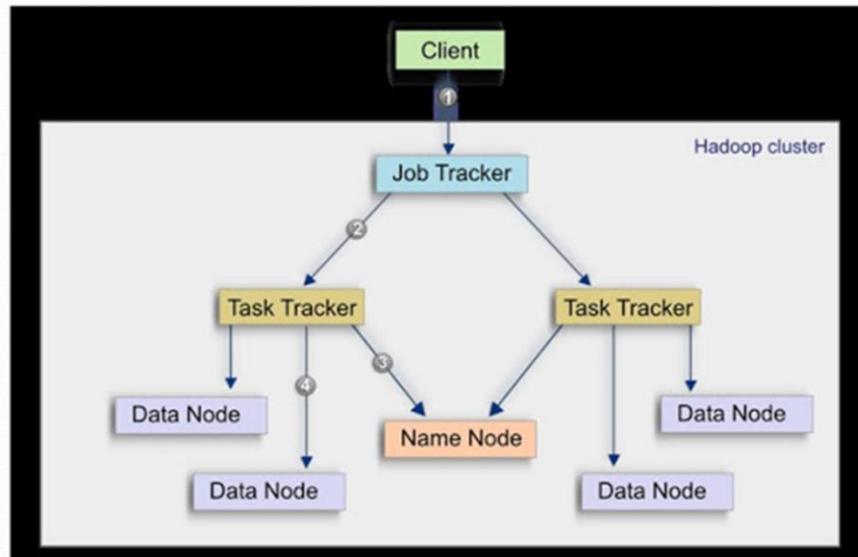
## Ecossistema do Hadoop

## Porque Hadoop

- Feito para Hardware “Commodity”
  - Servidores de baixo custo
  - Expansão modular
- Desenho Orientado a Metadados
  - NameNode mantem metadados
  - DataNodes gerenciam o local e o armazenamento
- Processamento feito junto com os dados
  - Servidores tem 2 objetivos : armazenar e processar
- Arquitetura de “File-System”
  - Foco no acesso sequencial.
  - Leitura rápida de dados

O Hadoop é uma plataforma desenhada para armazenamento e acesso a grandes volumes de dados.

# Arquitetura Hadoop



## Name Node :

- Somente um name node ativo por cluster
- Gerencia o namespace do filesystem e metadados
- Local onde se deve investir mais recursos no hardware

## Data Node:

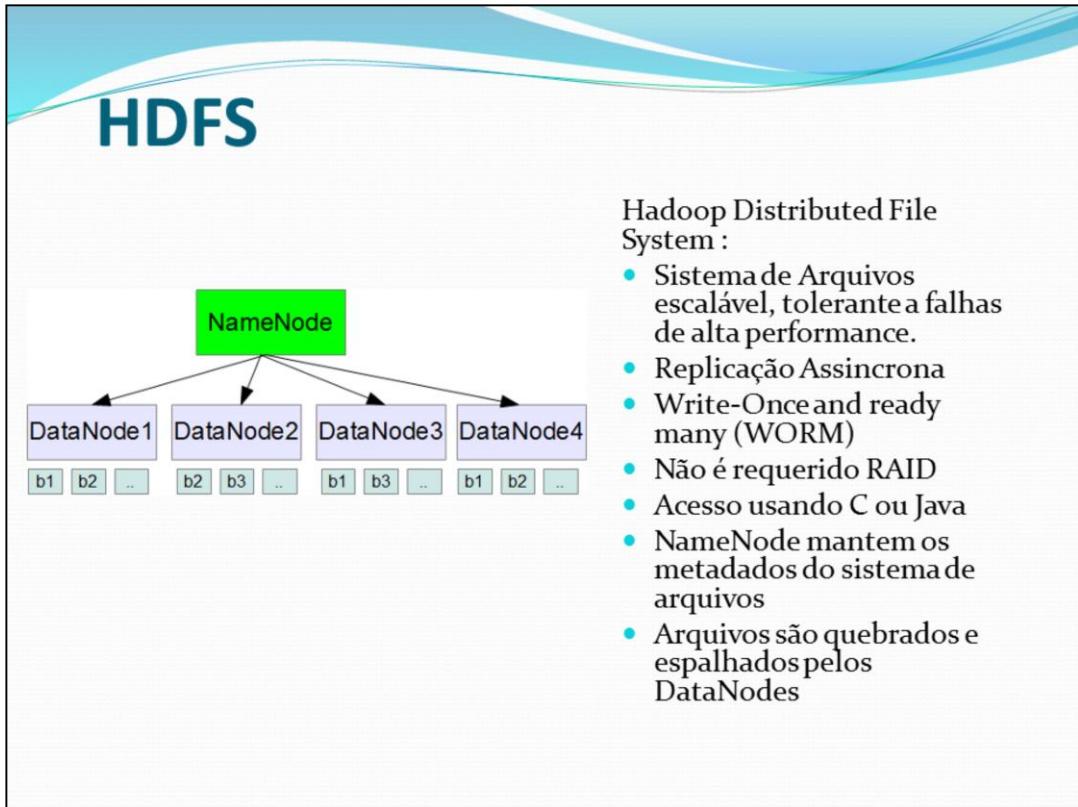
- Normalmente existem diversos data nodes.
- Ele gerencia os data blocks e gerencia a entrega de dados
- Os dados são replicados

## Job Tracker:

- Somente 1 job tracker por cluster
- Recebe as requisições enviadas ao cliente
- Agenda e monitora Map Reduce jobs

## Task Tracker:

- Normalmente existem diversos tasktrackers
- Responsável pela Execução dos Map Reduce Jobs
- Lê os blocos dos Data Nodes



Arquitetura Mestre e Detalhe :

Master “NameNode”

- Gerencia o namespace do filesystem
- Mantém a lista de blocos e o mapa de locais ( location mapping )
- Gerencia a replicação e alocação dos blocos
- Controle de acesso ao namespace

Slaves “Datanodes” – gerencia o armazenamento dos blocos

- Armazena os blocos no sistema operacional
- Clientes acessam blocos diretamente pelos datanodes
- Periodicamente envia relatório de blocos para o NameNode
- Periodicamente checa a integridade dos blocks

Comandos Hadoop são como Unix e shell scripts.

## Splits e Replication no HDFS

- Os dados são organizados em arquivos e diretórios
- Os arquivos são divididos em blocos uniformes e distribuídos através dos cluster nodes
- Blocos são replicados para suportar quaisquer falhas de hardware.
- O filesystem mantém os checksums dos dados para evitar corrupção de dados

O dado é dividido em múltiplos blocos no HDFS

Cada bloco tem 128MB

Os blocos são replicados 3 vezes para evitar falha no nodes e perda de dados

## Características do HDFS

- Name Node

- Name Node é coração do HDFS.
- Mantem o diretório de todos os arquivos do file system e acompanha a distribuição dos arquivos pelo cluster.
- Gerencia configuração do cluster
- Gerencia Transactional Log

## Características do HDFS

- Características Gerais ( Acesso )
  - Possui Java API para aplicações
  - Pode ser acessado com Python
  - Também pode utilizar C para Acessar Name Node é o coração do HDFS.
  - Pode ser acessado via browser ( HTTP ) para visualização dos arquivos

## Características do HDFS

- File Size
  - Blocos de 64 MB, podendo chegar a 128 MB
  - Os arquivos são fatiados em partes de 64MB e armazenados
- Data Node
  - Armazena os dados no HDFS
  - Pode ser encontrado diversos deles
  - Dados são replicados por diversos data nodes

## HBASE

- HBASE prove armazenamento para o Hadoop Distributed Computing Enviroment.
- Os dados são logicamente organizados em tabelas, colunas e linhas.
- Derivado do Google Big Table.
- Implementado em Java, pode ter clientes em Java, C++, Ruby, etc.
- Armazenamento orientado a colunas.
- Distribuido por diversos server.
- Tolerante a falhas de maquina (hardware).

## HBASE

APACHE  
**HBASE**

- É uma camada sobre o HDFS.
- Tem forte consistência.
- Não tem característica relacionais
- Dados esparsos ( sparse data )
- Suporta dados semi-estruturados e não-estruturados
- Extremamente Escalável – bilhões de registros x milhões de colunas

## HIVE



- Query e Sumarizacao de dados sobre o Hadoop
- Usa MapReduce para Execução & HDFS para o armazenamento
- Hive Query Language
  - SQL Basicos : Select, from, join, Group By
  - Equi-Join, Multi-Table Insert, Multi Group By
  - Query Batch
- MetaStore
  - Propriedades das tabelas e partições
  - Thrift API : Clientes em PHP, Python e Java
  - Metadados armazenados em qualquer SQL

Hive é uma ferramenta de Acesso ao Hadoop inicialmente desenvolvida no Facebook.

Permite aos usuários escrever queries SQL que são convertidas em programas MapReduce.

## PIG



- Pig é uma plataforma para analisar grandes datasets de dados, utilizando uma linguagem de alto nível para programas de análise de dados
- Pig gera e compila programas MapReduce
- Ele abstrai os detalhes específicos
  - Tem foco no processamento
  - Fluxo de dados
  - Construído para manipulação de dados
- Pig é direcionado a fluxo de dados e fácil de manter.

Pig é uma linguagem originalmente criada no Yahoo

Simples de ser usado, ele facilita a criação de programas MapReduce

## SQOOP

The logo for Sqoop, featuring the word "SQOOP" in a stylized, bold, blue font.

- Sqoop é uma ferramenta feita para ajudar na importação de grandes datasets de dados de bancos de dados relacionais para Hadoop.
- Importação automatizada de dados.
- Fácil importação de diversos databases para Hadoop.
- Gera código para ser usado em aplicações de MapReduce
- Integra com Hives

Sqoop é uma ferramenta de conectividade para mover dados de data stores não Hadoop, como bancos de dados relacionais e data warehouses para o Hadoop.

# ZOOKEEPER

```
graph TD; ZooKeeper[ZooKeeper Service] --> S1[Server]; ZooKeeper --> S2[Server]; ZooKeeper --> S3[Server]; ZooKeeper --> S4[Server]; ZooKeeper --> S5[Server]; S1 --> C1[Client]; S1 --> C2[Client]; S1 --> C3[Client]; S2 --> C4[Client]; S3 --> C5[Client]; S4 --> C6[Client]; S5 --> C7[Client]; S5 --> C8[Client]
```

- Zookeeper é um serviço centralizado para manter informações de configuração, nomeação, provendo distribuição, sincronismo e provendo grupo de serviços.
- Um servidor é escolhido como líder
- Os demais servidores “seguem” o líder

## AVRO



- Avro é um sistema de serialização que provê integração dinâmica com linguagens de script.
- Avro permite :
  - Estruturas de dados enriquecidas
  - Formatos rápidos e compactos
  - RPC(Remote Procedure Call)

# FLUME



- Flume :

- Framework escalável, configurável, extensível e gerenciável para população de dados.
- Desenvolvido Open Source
- Pode ser considerado uma solução definitiva para todos os formatos de dados.
- Permite tuning de performance
- Habilita rápida interação em estratégicas de integração de dados.

Flume pode ser utilizado para inserir dados no Hadoop.