



Fundamentos de Big Data

Centro Universitário 7
Setembro - Uni7

Especialização em
Ciência de Dados com Big
Data, BI e Data Analytics

Prof. Manoel Ribeiro PhD

Prof. Manoel Ribeiro

- Formação
 - Doutor em Computação Big Data, Machine Learning e Sistemas Distribuídos (UFC)
 - GPS2GR:Optimized Urban Green Routes based on GPS Trajectories
 - Temas: Trajectory Pattern Mining, Green Routes, Traffic-Light Scheduler
 - Mestre em Sistemas de apoio a decisão (UECE)
 - FastClass: Classificação Automática Fuzzy, enfase em Data mining; Análise de agrupamentos; Clustering; Análise de Componentes Principais; Fuzzy.
 - Publicações relevantes
 - GPS2GR:Optimized Urban Green Routes based on GPS Trajectories, 8th ACM SIGSPATIAL Workshop on GeoStreaming, 2017
 - LB-RLT Approach for Load Balancing Heterogeneous Storage Nodes. XXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, 2016.
 - DMM: A Distributed Map-matching algorithm using the MapReduce Paradigm. Intelligent Transportation Systems Society Conference Management System, 2016.
 - Bacharel em Computação (UFC)
 - MBA em Finanças, Controladoria e Auditoria (FGV)
 - Especialista em Projetos (CETRED)

Prof. Manoel Ribeiro

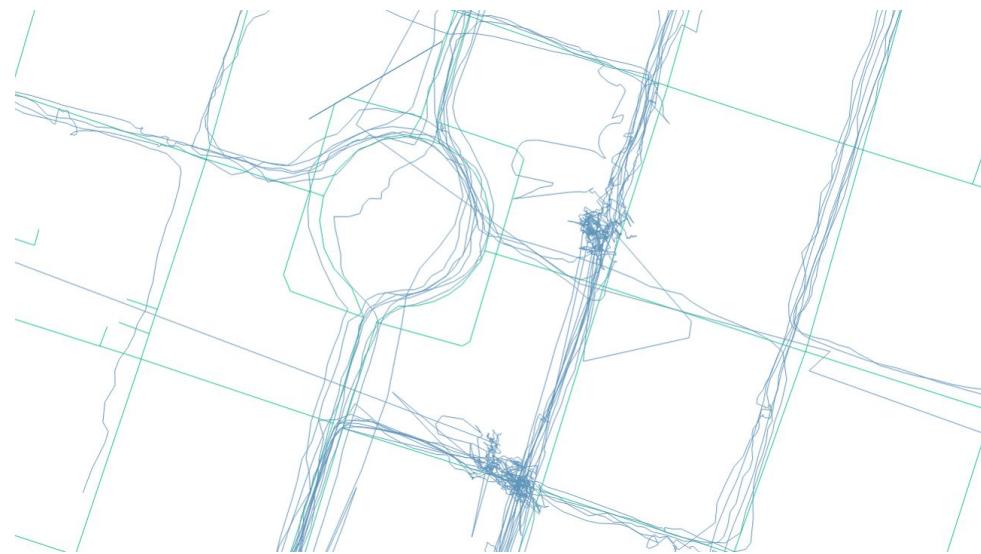
- Experiência
 - Foi executivo de TI por 25 anos no Grupo J.Macêdo e Grupo Marquise
 - Grupo J.Macêdo
 - Implantação do BI
 - Implantação do ERP SAP (SEM/BPS e BW)
 - Implantação da automação da força de venda
 - Desenvolvimento de sistema Inteligência de negócio - Navigator
 - Mudança de paradigma de formação de preço dos produtos
 - Grupo Marquise
 - Implantação ERP E-Business Suite (Oracle)
 - Implantação BI Cognos (IBM)
 - Terceirização de commodities de TIC
 - Terceirização de processos de negócios -ADP
 - Foi fundador e presidente do Grupo de Gestores de TIC do Ceará - GGTIC-CE
 - Foi sócio fundador da www.softium.com.br
 - Foi diretor de relações institucionais do I3D.org.br

Prof. Manoel Ribeiro

- Atuação
 - Professor adjunto Unilab
 - Professor de pós-graduação nas áreas de Data Science, BI e governança de TIC
 - Pesquisador associado no Instituto de Tecnologia da Informação e Comunicação (ITIC) com ênfase em IIoT, Big Data e Data Analytics
 - Possui quatro patentes em Sistemas Embarcados (INPI)
 - Consultoria em Data Science na **OPENCARE**
 - Empreendedor em IIoT com ênfase em:
 - **Data Logger** para sensores sem fio de longo alcance utilizando protocolo **LoRaWAN** (Mash) e com fio utilizando barramento **I2C** para uso industrial
 - Computação embarcada para acessibilidade

Prof. Manoel Ribeiro

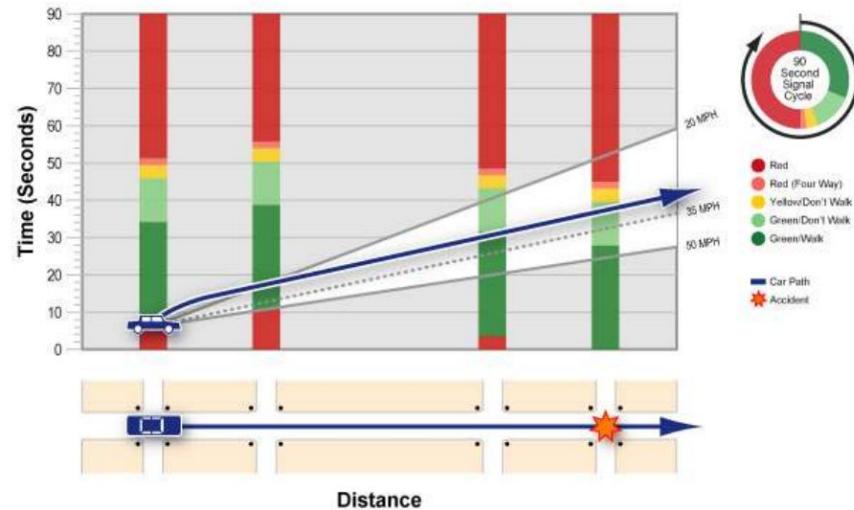
- **DMM: A Distributed Map-matching algorithm using the MapReduce Paradigm**
 - Intelligent Transportation Systems Society Conference Management System, 2016.
 - Processamento em larga escala de trajetórias de GPS para descobertas de caminhos
 - Spark/Scala num cluster com 8 nós



Prof. Manoel Ribeiro

- **GPS2GR:Optimized Urban Green Routes based on GPS Trajectories**

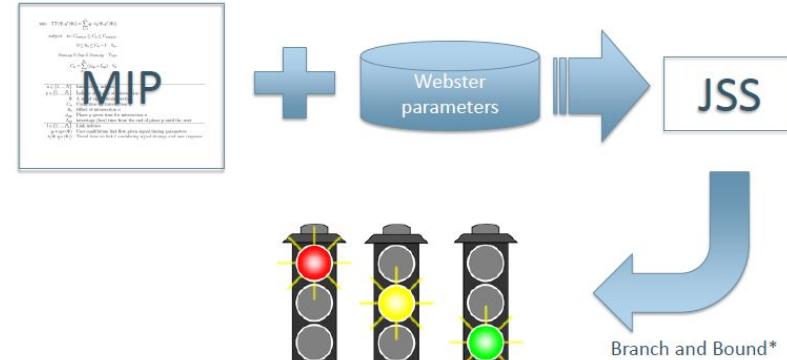
- 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2017)
- Processamento de BigData de trajetórias de veículos de uma grande cidade durante uma semana visando otimizar os semáforos para um padrão de deslocamentos diários
- Pipeline/C#



Prof. Manoel Ribeiro

- **Optimization of urban semaphore times turning into JSSP**

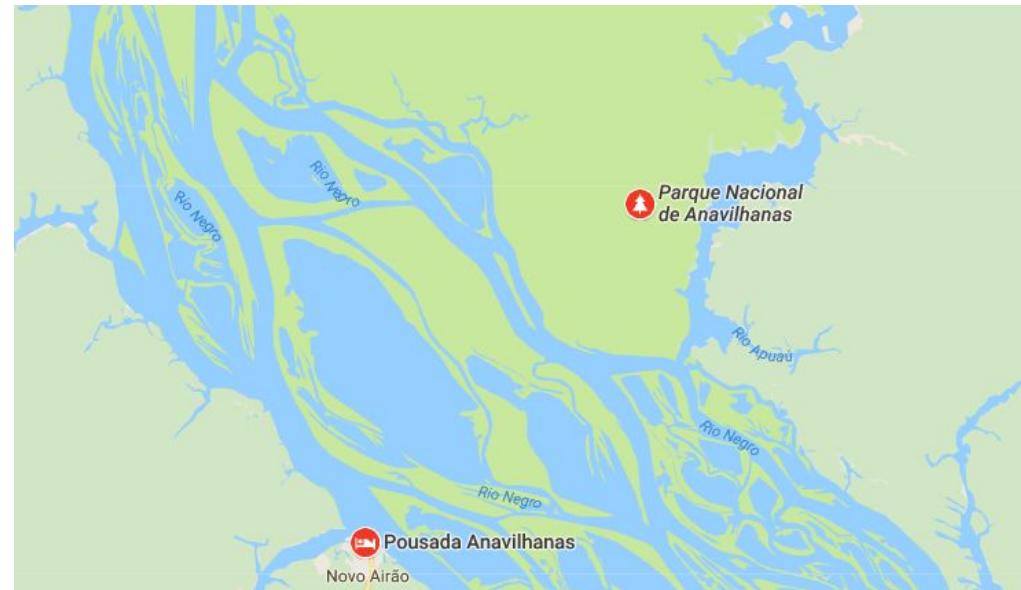
- 44th International Conference on Very Large Data Bases (VLDB 2018)
- Processamento de BigData de semáforos e rotas frequentes
- Google Optimization Tools



Prof. Manoel Ribeiro

- **Internet on the Forest - IoT**

- ITIC/RNP/MCTI/EU
- Sensores para captura de características específicas da região
- Desafios intempéries, bateria, transmissão, armazenamento e análise
- MongoDB/Sofia2



Conteúdo da disciplina (20 h/a)

- Dia 1 (4h/a)
 - Introdução e Conceitos sobre **Big Data**.
 - Como as Empresas estão Utilizando Big Data para a gestão competitiva dos negócios.
- Dia 2 (4h/a)
 - Princípios de sistemas distribuídos e os impactos destas arquiteturas no processamento de grandes massas de dados.
 - Principais paradigmas para armazenamento e processamento de dados distribuídos.

Conteúdo da disciplina (20 h/a)

- Dia 3 (4h/a)
 - Introdução ao **OpenStack**
 - Visão da Arquitetura e ao Ecossistema **Hadoop**.
 - Imersão no Apache **Spark**.
- Dia 4 (4h/a)
 - Visão dos Bancos de Dados **NoSQL**
 - Imersão no **MongoDB**
- Dia 5 (4h/a)
 - Carreira em Big Data
 - Avaliação

Repositório

<https://github.com/antoniomralmeida/BigData>

Entidades



[HTTP://WWW.DATASCIENCEINSTITUTE.ORG/](http://www.datascienceinstitute.org/)

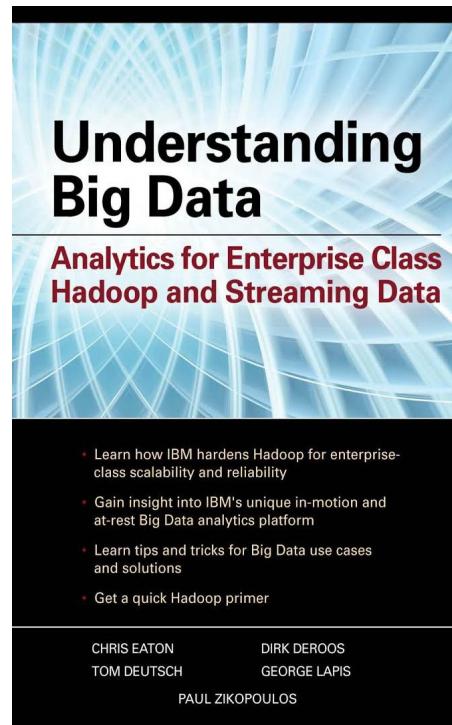
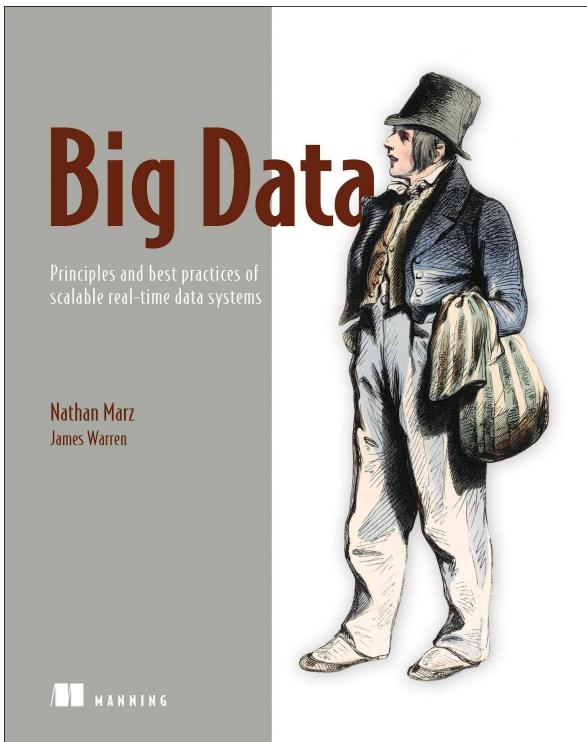


[HTTP://WWW.DATASCIENCEINSTITUTE.COM.BR/](http://www.datascienceinstitute.com.br/)

Pré-requisitos da disciplina

- Pré-requisitos da disciplina
 - Fundamentos de Rede
 - Fundamento de Sistemas Distribuídos
 - Bancos de Dados Relacional
 - Linguagem de Programação Java / Phyton

Bibliografia Básica



Paulo Polzonoff Junior

BIG DATA

**COMO EXTRAIR VOLUME, VARIEDADE,
VELOCIDADE E VALOR DA AVALANCHA
DE INFORMAÇÃO COTIDIANA**

VIKTOR MAYER-SCHÖNBERGER
KENNETH CUKIER



Contextualização



Contextualização



7 BILHÕES De Pessoas

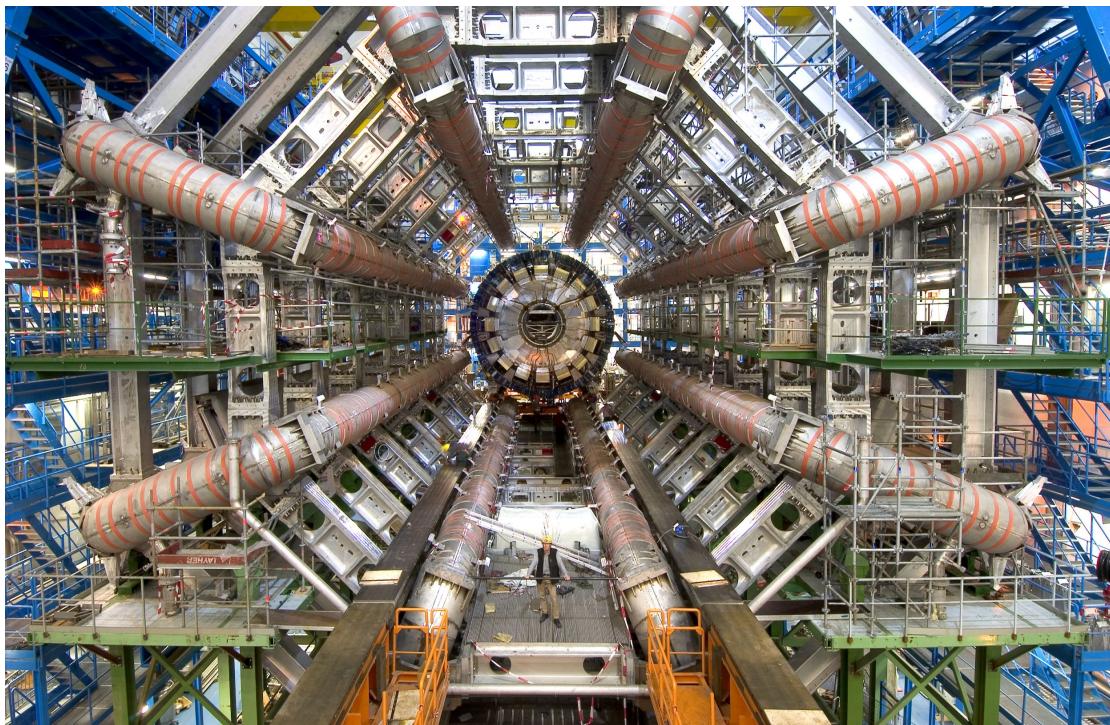


2 BILHÕES De USUÁRIOS (17%)
4 BILHÕES DE LIKES
300 MILHÕES DE FOTOS
83 MILHÕES FAKEs



2,3 BILHÕES

Larger Hadron Collider



1 Petabyte por segundo

1% dos dados úteis
gerando 25 Petabytes por
ano

Fonte: <http://home.cern/about>

Motor de buscas

- ★ 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.



Popularização dos Gadgets

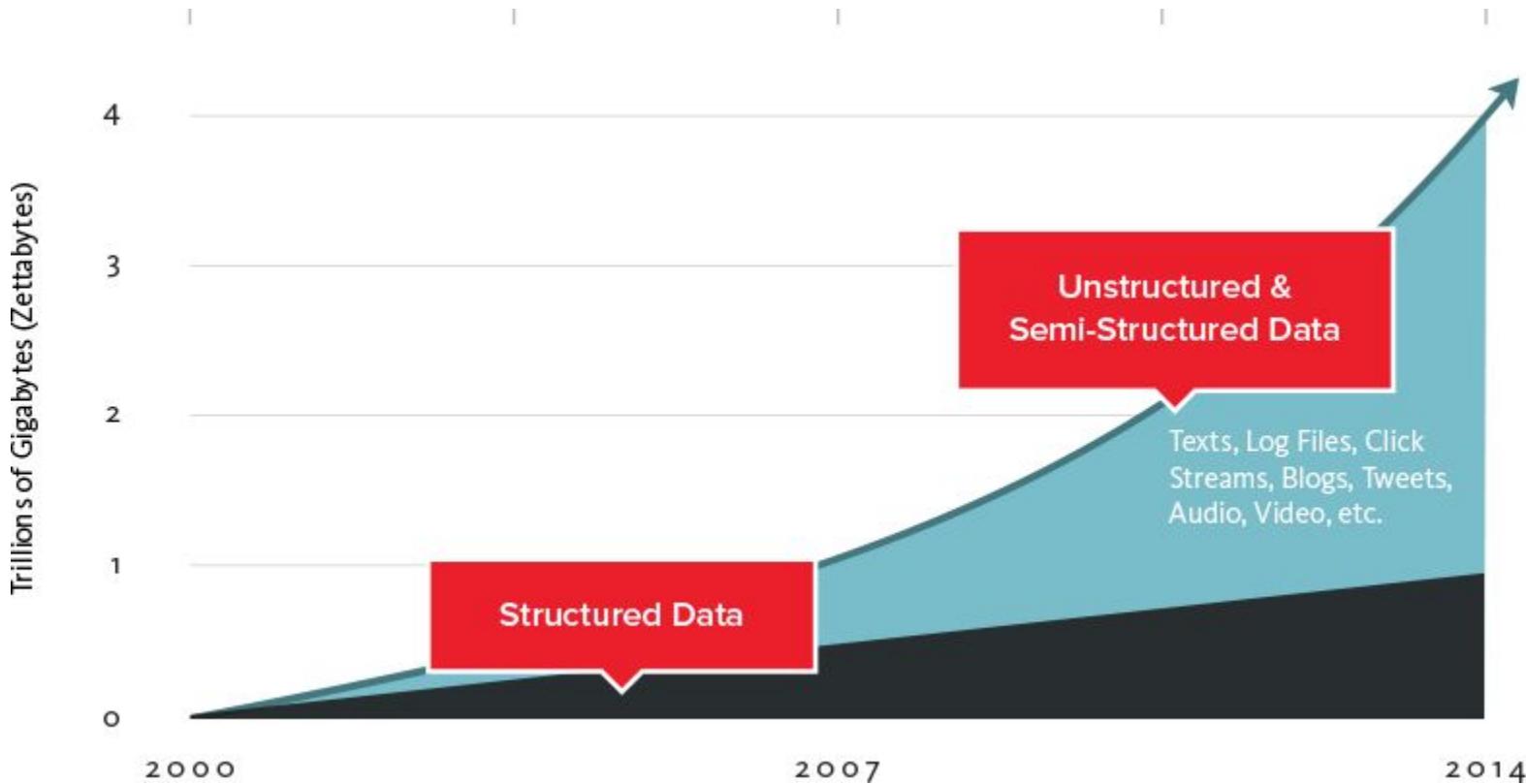


2005

2013



Natureza dos Dados



BIG DATA

A central 3D graphic features the words "BIG" and "DATA" in large, bold, white letters. Surrounding this central text are numerous other data-related terms, each rotated diagonally and placed within a circular path:

- VOLUME**: Above "BIG" and "DATA".
- STRUCTURED**: Above "DATA".
- UNSTRUCTURED**: Above "VOLUME".
- SEMI-STRUCTURED**: Above "STRUCTURED".
- METADATA**: Above "UNSTRUCTURED".
- USEFUL**: Above "SEMISTRUCTURED".
- ZETTABYTE**: To the left of "BIG".
- DATA ANALYSIS**: Below "BIG".
- PEOPLE DRIVEN**: Between "DATA ANALYSIS" and "ZETTABYTE".
- DECISION MAKING**: Between "ZETTABYTE" and "DATA ANALYSIS".
- DATA FRAMEWORK**: Between "DATA ANALYSIS" and "SEMANTIC METADATA".
- UNSTRUCTURED**: Between "DATA FRAMEWORK" and "TEXT ANALYTICS".
- SMART CONTENT**: Between "DATA FRAMEWORK" and "SEMANTIC METADATA".
- TEXT ANALYTICS**: Between "UNSTRUCTURED" and "SEMANTIC METADATA".
- SEMANTIC METADATA**: Between "UNSTRUCTURED" and "STRUCTURED".
- STRUCTURED**: Below "DATA".

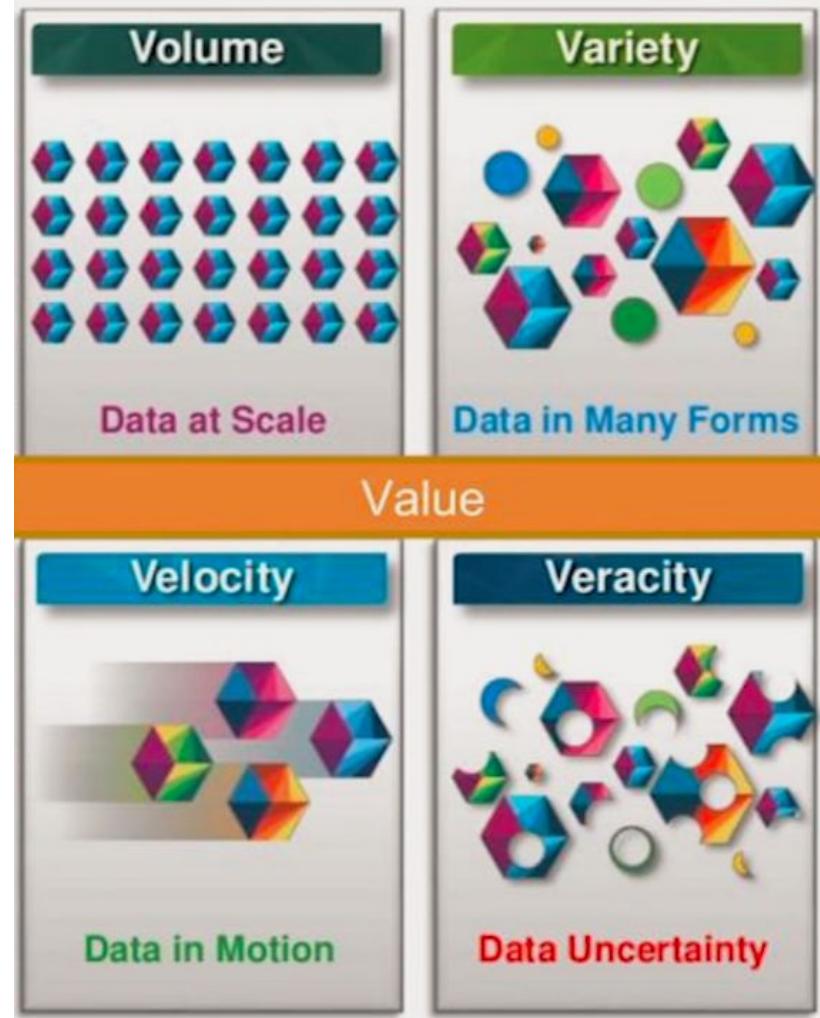
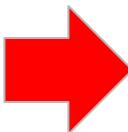
Tá na crista da onda?



Definição

Big Data é um termo amplamente utilizado na atualidade para nomear conjuntos de dados muito grandes ou complexos, onde as ferramentas e o paradigma tradicional de processamento de dados como banco de dados relacionais, planilhas, etc. são incapazes gerenciar.

Definição (5V's)



Big Data - Volume

- O Boeing 737 gerará 240 terabytes de dados de vôo durante um único vôo em todo os EUA.
- Os telefones inteligentes, os dados que eles criam e consomem;
- Os sensores incorporados em objetos do cotidiano logo resultarão em bilhões de novos feeds de dados atualizados constantemente, contendo informações ambientais, de localização e outras, incluindo o vídeo.

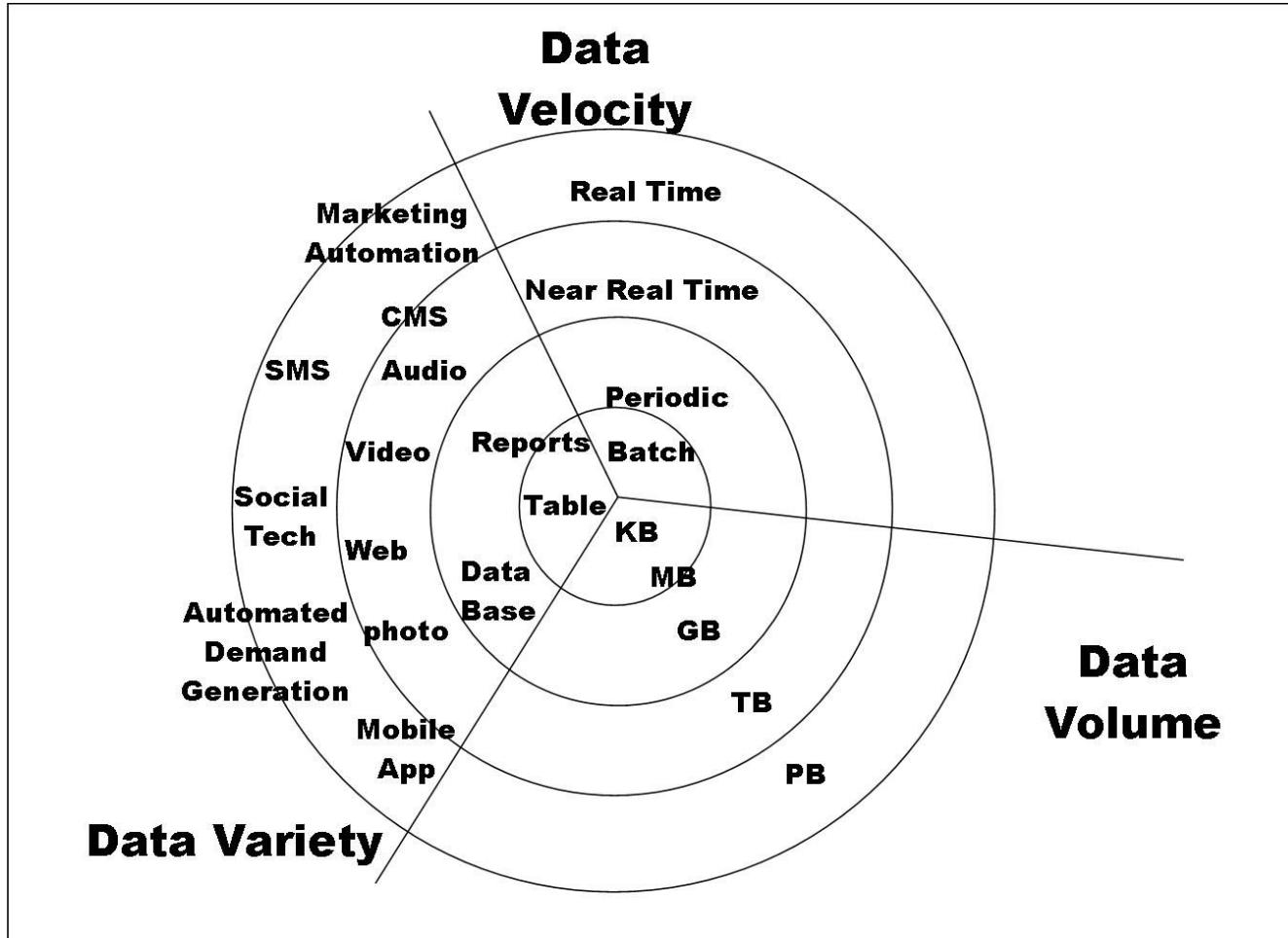
Big Data - Velocidade

- Click's e buscas em anúncios capturam o comportamento do usuário em milhões de eventos por segundo
- Os algoritmos de negociação de ações de alta freqüência refletem as mudanças no mercado em microssegundos
- Processos de máquina para máquina trocam dados entre bilhões de dispositivos
- infra-estrutura e sensores geram dados de registro maciços em tempo real
- Os sistemas de jogos on-line suportam milhões de usuários simultâneos, cada um produzindo várias entradas por segundo.

Big Data - Variedade

- Big Data não é apenas números, datas e strings. Big Data também é dados geoespaciais, dados 3D, áudio e vídeo e texto não estruturado, incluindo arquivos de log e redes sociais.
- Os sistemas de banco de dados tradicionais foram projetados para abordar volumes menores de dados estruturados, menos atualizações ou uma estrutura de dados previsível e consistente.
- A grande análise de dados inclui diferentes tipos de dados

3V's básicos



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

Dados estruturados x não-estruturados

- Big Data não é apenas números, datas e strings. Big Data também é dados geoespaciais, dados 3D, áudio e vídeo e texto não estruturado, incluindo arquivos de log e redes sociais.
- Os sistemas de banco de dados tradicionais foram projetados para abordar volumes menores de dados estruturados, menos atualizações ou uma estrutura de dados previsível e consistente.
- A grande análise de dados inclui diferentes tipos de dados

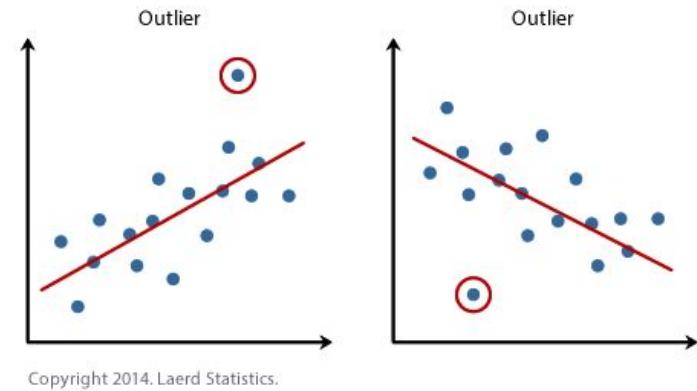
Dados estruturados x não-estruturados

- | | |
|--|---|
| <ol style="list-style-type: none">1. Esquema fixo2. Formato bem definido3. Conhecimento prévio da estrutura4. Simples de relacionar5. Dificuldade para alterar a estrutura | <ol style="list-style-type: none">1. Sem tipo predefinido2. Não possui estrutura regular3. Pouco ou nenhum controle sobre a forma4. Manipulação mais simplificada5. Facilidade de alteração |
|--|---|

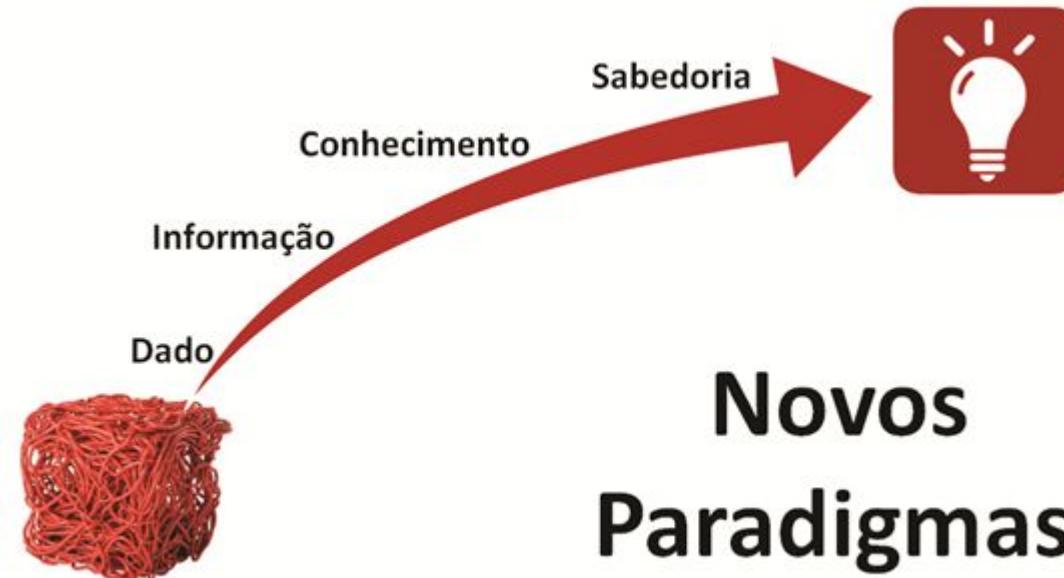


Big Data - Veracidade (incerteza)

- Possibilidade de dados não estar mais disponível
- Baixa qualidade dos dados
 - Dados faltantes (null)
 - Dados sujos, atípicos, inconsistentes (outlier)
 - Tipos inadequados
 - Falta de chaves para ligação
- Possibilidade de dados não confiáveis
 - Fontes não confiáveis
 - análise textual

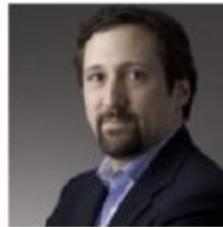


Big Data - Valor



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 Kernan 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 !

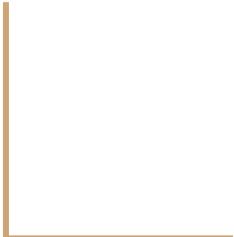
A nova economia do compartilhamento dos dados

Jeremy Rifkin: The sharing economy is the future of the society

The economist explained to the Forum PA audience why the future of human race is in danger and how the only choice available to public and private organizations is a change towards the sharing economy

Barbara Bosco

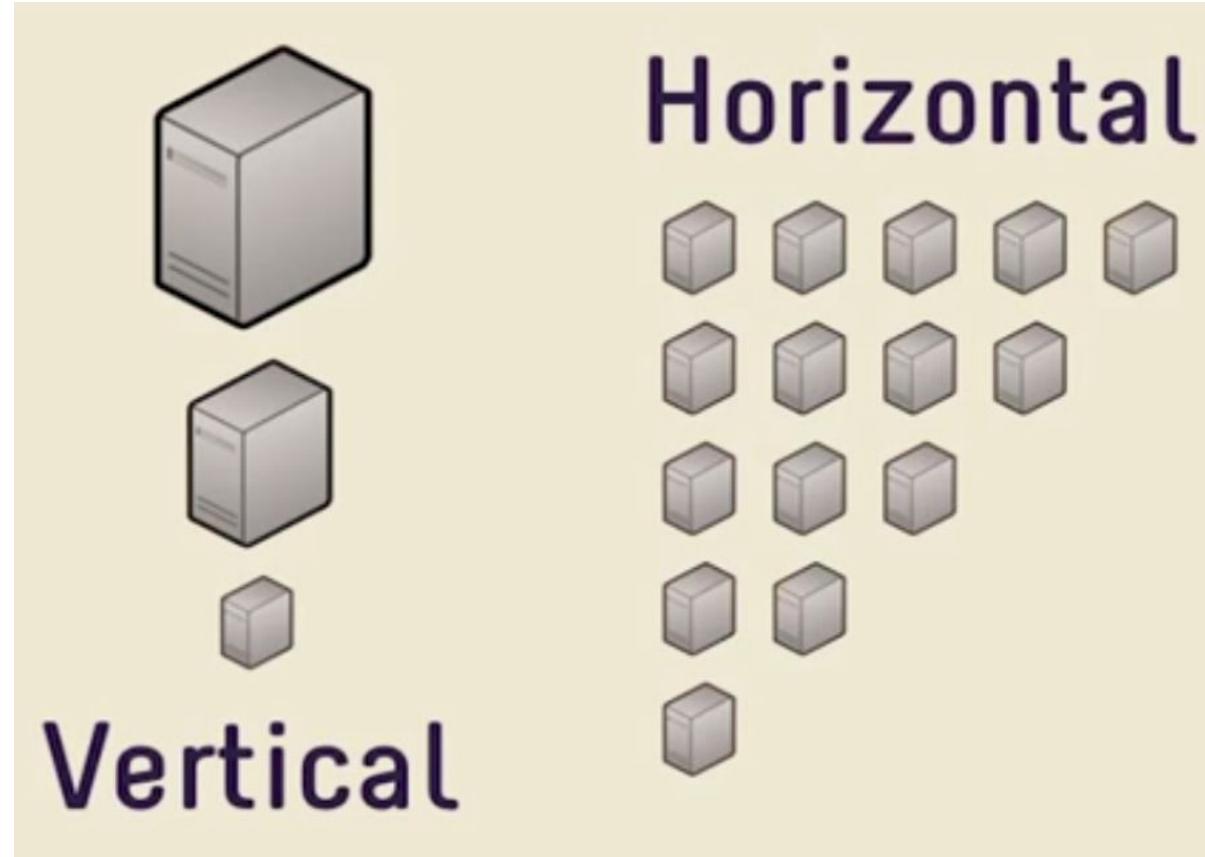




Como se processa algo tão desafiador?

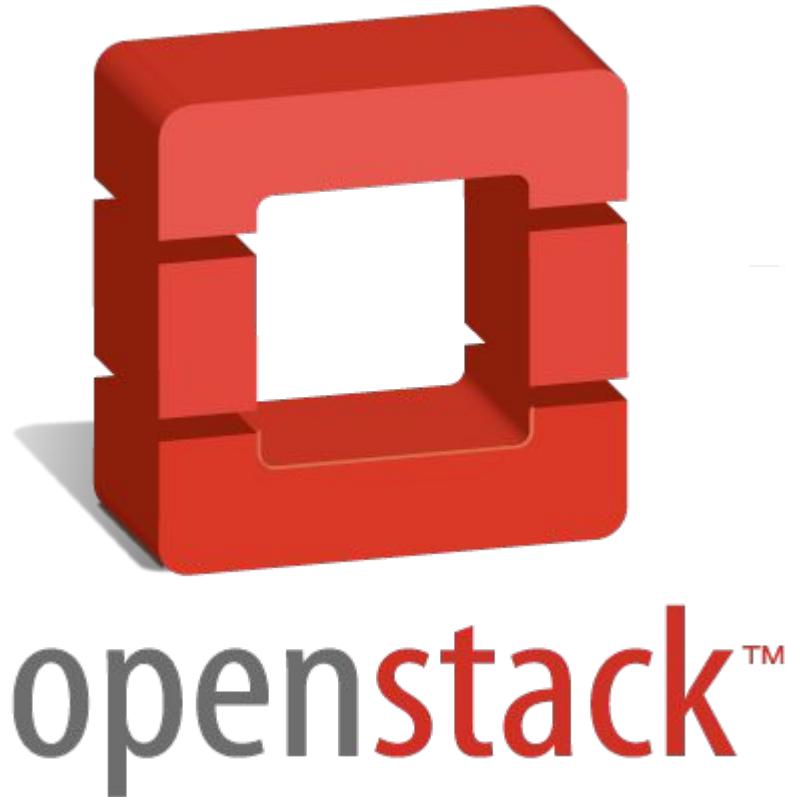
Escalabilidade Horizontal

Novo paradigma
da computação
moderna



Virtualização (Cloud)

Novo paradigma
da computação
distribuída





Novo paradigma
para resolução de
problemas
complexos

“Dividir e Conquistar” é uma técnica de projeto de algoritmos que consiste em resolver um problema a partir da solução de “sub-problemas menores” do mesmo tipo.

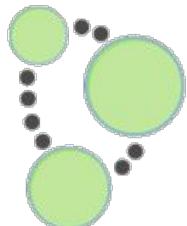
Dividir para conquistar (“Divide et impera” ou “Divide et Vinces”) é um clássico nas estratégias de guerra para enfraquecer e subjugar os povos. O termo, embora já era conhecida na Antiguidade, foi cunhado por Júlio César em seu livro ” De Bello Gallico ” (Guerra das Gálias), que explicou como a vitória romana na guerra gaulesa era essencialmente uma política de “dividir” seus inimigos, aliar com tribos individuais durante suas disputas com adversários locais.

Not Only SQL

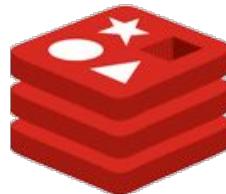
Novo paradigma para
de banco de dados de
estrutura mais simples
e altamente escalar



HYPERTABLE^{INC}



Neo4j

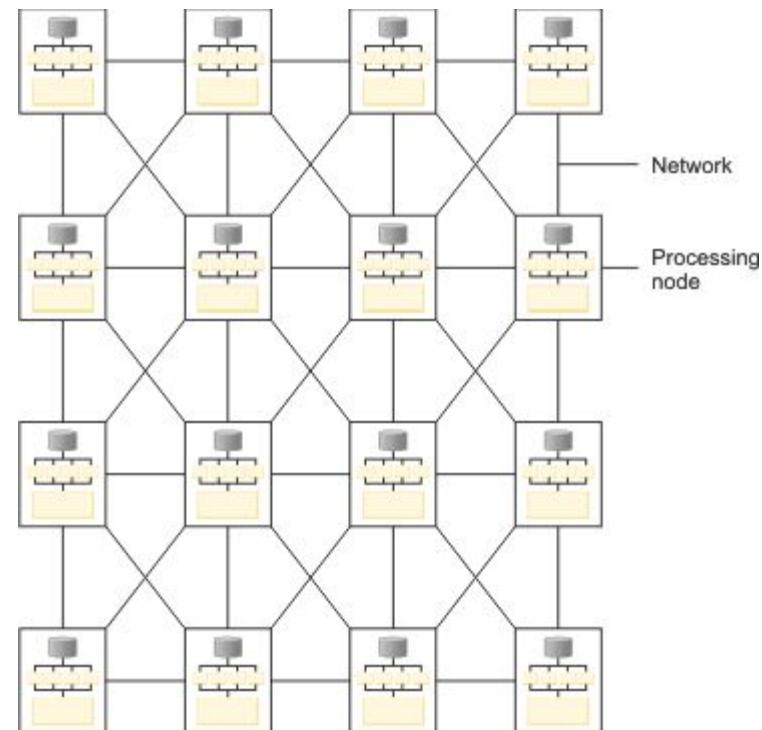


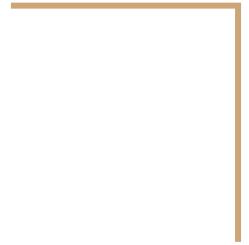
redis

MASSIVE PARALLEL PROCESS - MPP

Remodelagem do conceito de Sistemas Distribuídos com gestão e recuperação de falhas: Cluster e Grid

HDFS - Modelo eventualmente consistente





Como isso tudo aconteceu?

2003

GFS

<http://research.google.com/archive/gfs.html>

2004

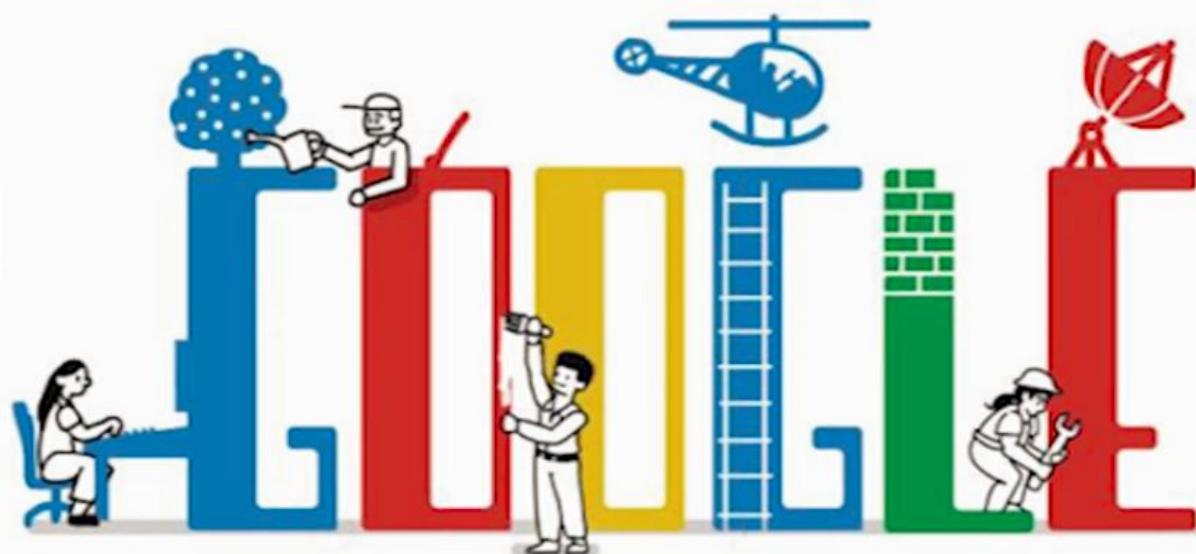
MapReduce

<http://research.google.com/archive/mapreduce.html>

2006

Big Table

<http://research.google.com/archive/bigtable.html>





- Surgiu em 2006 no Yahoo
- Escrito em Java
- Contém:
 - **HDFS** - Sistema de Arquivos Distribuídos
 - **YARN** - gerenciador de recursos
 - **MapReduce** - Biblioteca de suporte a aplicações distribuídas



<http://hadoop.apache.org/releases.html>

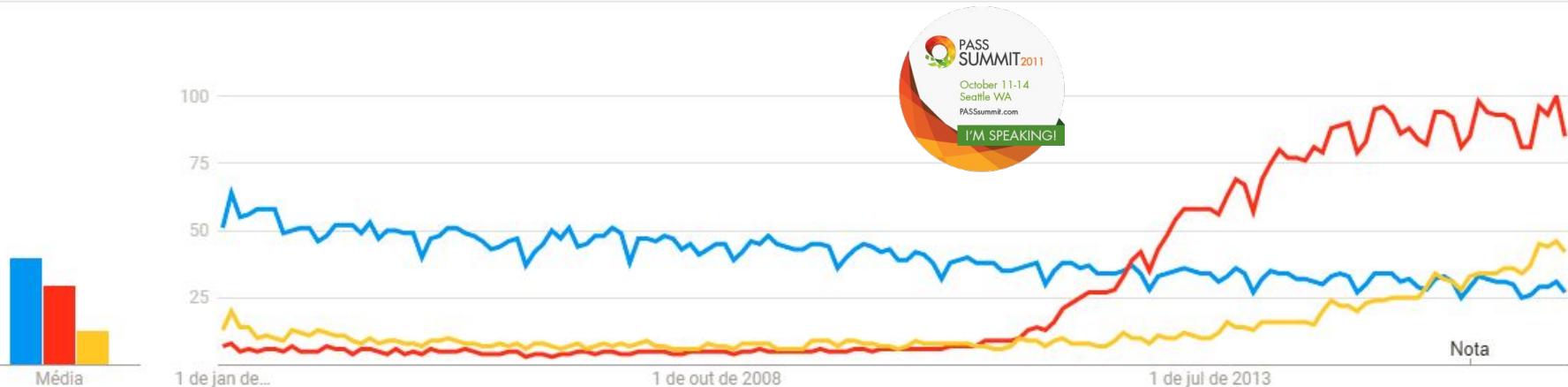
● Business Intelligen... : Termo de pesquisa

● Big Data : Termo de pesquisa

● data science : Termo de pesquisa

+ Adicionar comparação

Interesse ao longo do tempo ?



<https://trends.google.com.br/trends/?hl=pt-PT>



- HBase é um banco de dados distribuído open-source orientado a coluna, modelado a partir do Google BigTable e escrito em Java.
- O Hbase tem fácil integração com o Hadoop, sendo assim, pode utilizar o MapReduce para distribuir o processamento dos dados, podendo processar facilmente vários terabytes de dados.
- Foi criado pela empresa PowerSet em 2006

<https://hbase.apache.org/>



- Apache Spark é um framework de código fonte aberto para computação distribuída.
- Foi desenvolvido no AMPLab da Universidade da Califórnia e posteriormente repassado para a Apache Software Foundation que o mantém desde então.
- Spark provê uma interface para programação de clusters com paralelismo e tolerância a falhas.
- Spark não está preso ao paradigma MapReduce
- Spark possui estrutura de dados em memória (RDD)

<https://spark.apache.org/>

O Spark e o aprendizado de máquina dão gás ao Big Data

O [Apache Spark](#), anteriormente um componente do ecossistema do Hadoop, está se tornando a plataforma preferida de Big Data das empresas. Em uma [pesquisa](#) com arquitetos de dados, gerentes de TI e analistas de BI, aproximadamente 70% dos entrevistados preferiam o Spark ao tradicional MapReduce, que é baseado em lote e não pode ser usado com aplicativos interativos ou no processamento de fluxo em tempo real.

Esses recursos de processamento de Big Data provocaram uma evolução nas plataformas, que agora oferecem aprendizado de máquina intensivo, IA e algoritmos de gráfico. O aprendizado de máquina do Microsoft Azure, em particular, emplacou graças à sua interface simples de usar e facilidade de integração com plataformas Microsoft existentes. Disponibilizar o aprendizado de máquina para as massas resultará na criação de mais modelos e aplicativos que, por sua vez, gerarão petabytes de dados.

Qual o benefício disso tudo?

Sequência do Genôma humano

1998 → 2001
2,8 Bilhões USD

2016 1.000 USD

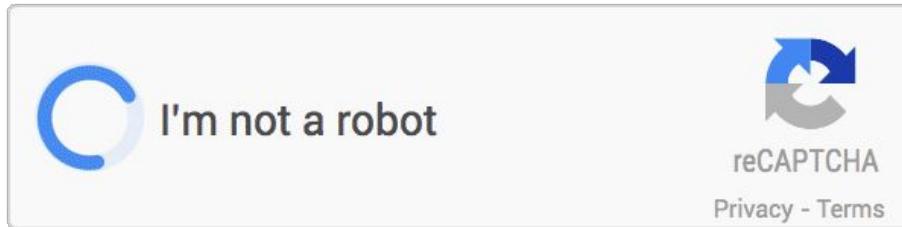


POWERED BY

<https://databricks.com/blog/2016/05/24/genome-sequencing-in-a-nutshell.html>

Bots Cybercrime e prova de humanidade

- reCaptcha - ferramenta de prova de humanidade que utiliza de reconhecimento de imagem ou ontologia para reconhecimento humano
- Pergunta:
 - Quanto custaria hoje para combater manualmente as tentativas de acesso indevido via Bots?



- Estimativas falam em bilhões de dólares de economia para muitas empresas

Sistema de recomendação

- O objetivo dos sistemas de recomendação (SR) é gerar recomendações válidas para um conjunto de usuários, de itens que possam interessá-los



NETFLIX

amazon.com® [Help](#) | [Close window](#)

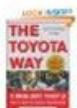
Recommended for You

 [LOOK INSIDE!](#)
Inside Apple
How Steve Jobs, the Genius and the Genius He Created
by Tony Hsieh

Inside Apple: How America's Most Admired--and Secretive--Company Really Works
Our Price: \$9.99
Used & new from \$9.99
[See all buying options](#)

Rate this item
 
 I own it
 Not interested

Because you purchased...

 [LOOK INSIDE!](#)
THE TOYOTA WAY
14 Management Principles from the World's Greatest Manufacturer
(Kindle Edition)

The Toyota Way : 14 Management Principles from the World's Greatest Manufacturer
(Kindle Edition)
[See all buying options](#)

Rate this item
 
 This was a gift
 Don't use for recommendations

Dados abertos governamentais



- Dados produzidos pelo governo e colocados à disposição das pessoas de forma a:
 - Cumprir metas de transparência
 - tornar possível não apenas sua leitura e acompanhamento,
 - mas também sua reutilização em novos análise
 - seu cruzamento com outros dados de diferentes fontes;
 - sua disposição em visualizações interessantes e esclarecedoras.

Como avaliar o que não tem preço?

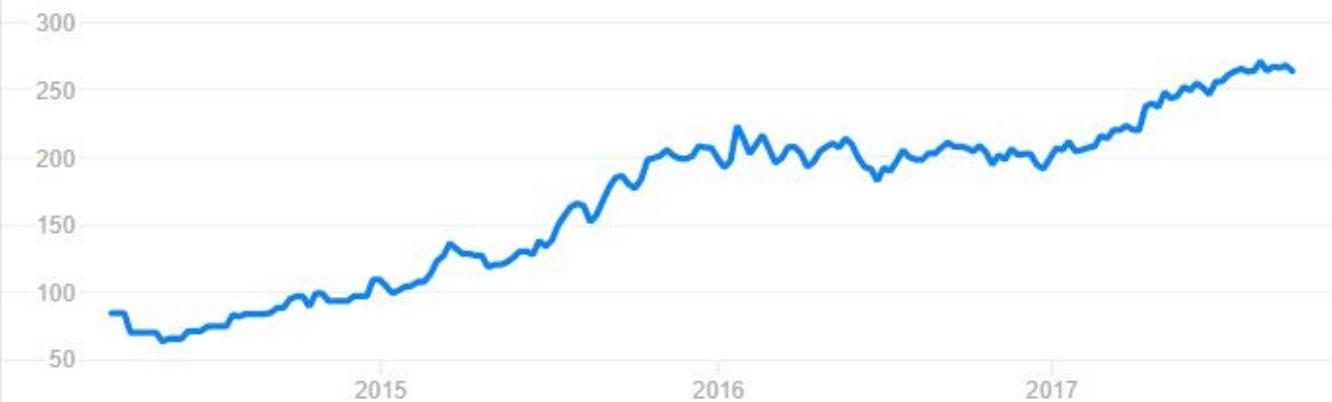
- Em 18 de maio de 2012 Facebook lançou ações na NASDAQ
- O mercado fixou o preço da ação em US\$38, o que significa US\$104 BI
- Neste mesmo ano o balanço da empresa registrava bens no valor de US\$6,3 bilhões
- Facebook não registrou nada de bens intangíveis usando a alegação de que seria impossível valorizar a sua base de dados (Big Data)!
- Estima-se porém que cada usuário gere para Facebook US\$ 100 por suas informações, isso corresponde a um valor do seu banco de dados de mais de US\$200 BI!!!

Facebook Unsp BDR

BVMF: FBOK34 - 4 de out 16:00 BRT

263,99 BRL 0,00 (0,00%)

Um dia Cinco dias Um mês Três meses Um ano **Cinco anos** máx



Abertura

-

Cap. merc.

-

Alta

-

Pr./lucro

-

Baixa

-

Rend. div.

-

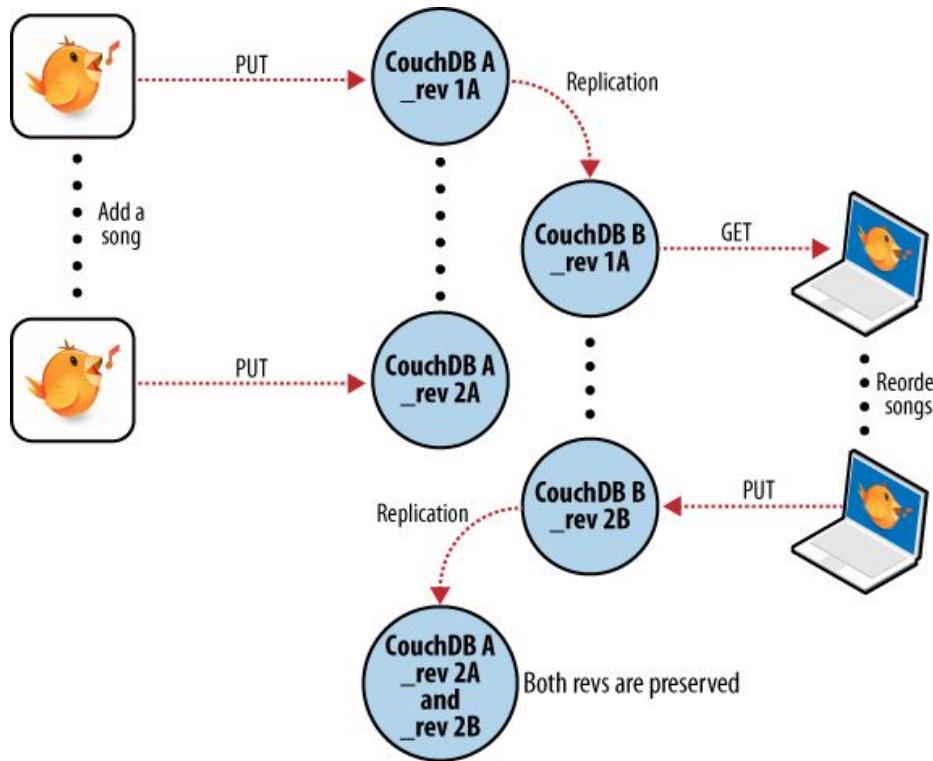
Paradigmas de Sistemas Distribuídos

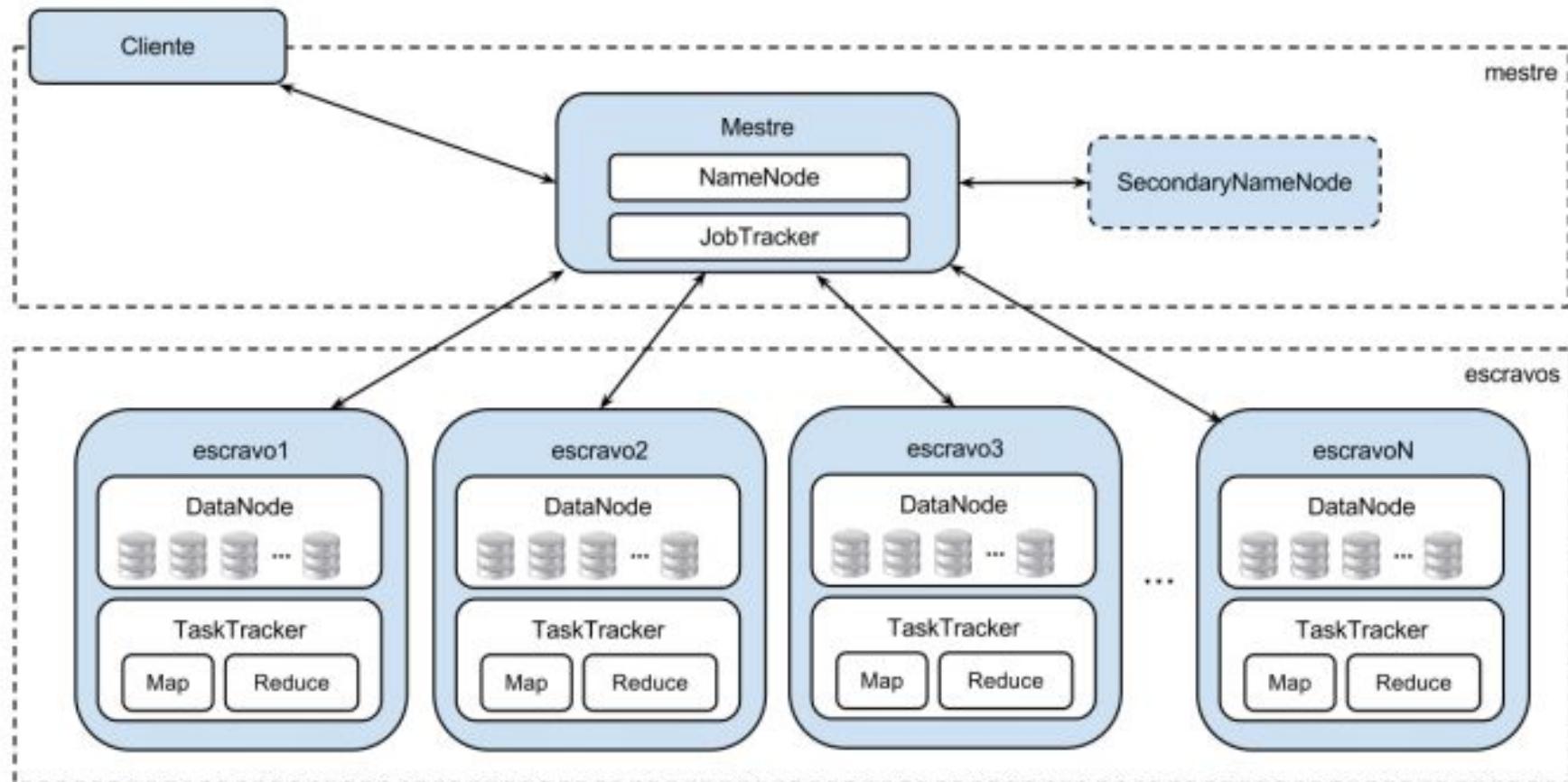


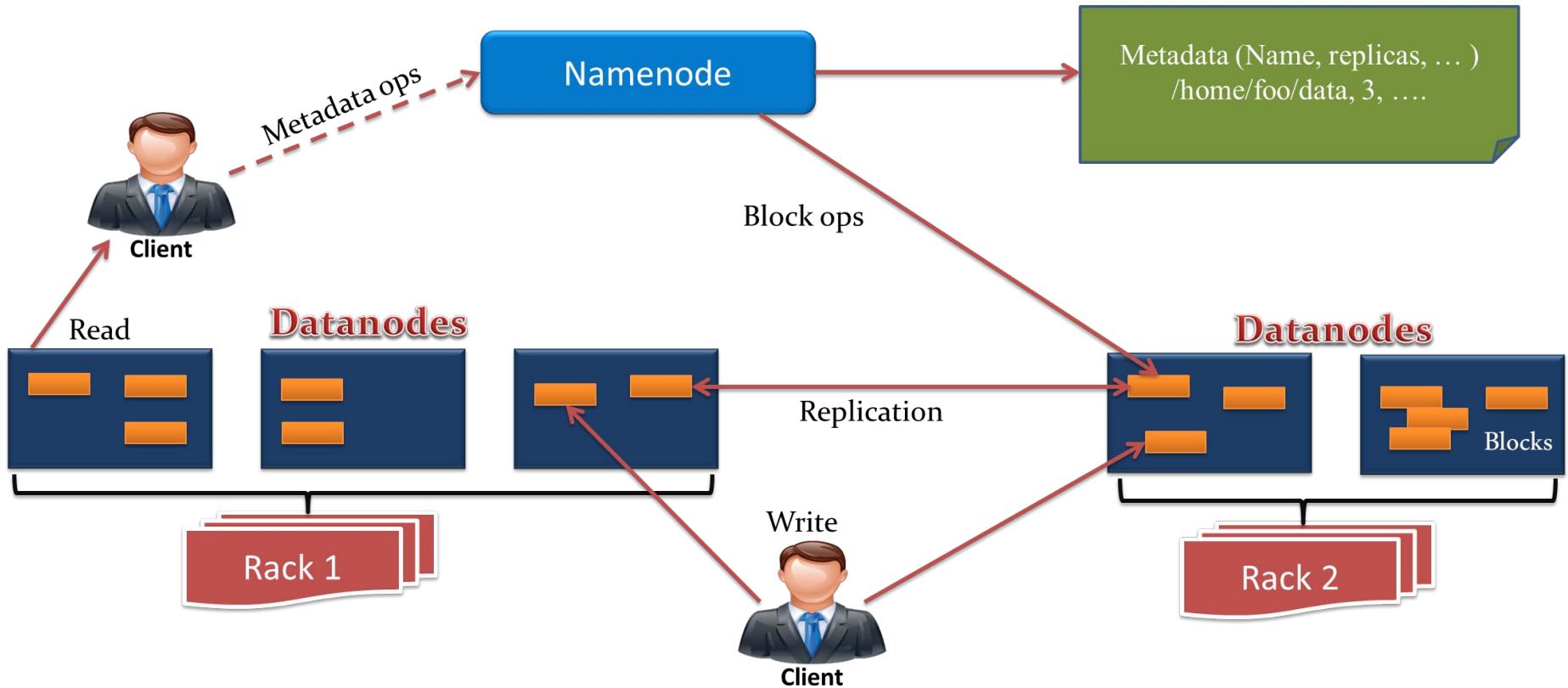
Sistema de arquivos distribuídos

- **WORM (write-once-read-many)** do HDFS que afrouxa as exigências do controle de simultaneidade, simplifica a persistência de dados e habilita acesso de alto rendimento
- **Eventually consistent** - A consistência eventual é um modelo de consistência usado na computação distribuída para alcançar alta disponibilidade, também chamada de replicação otimista é amplamente implantada em sistemas distribuídos e tem origens em projetos iniciais de computação móvel. Um sistema que alcançou consistência eventual é freqüentemente dito ter convergido , ou conseguido convergência de réplica. (OpenStack, EC2)

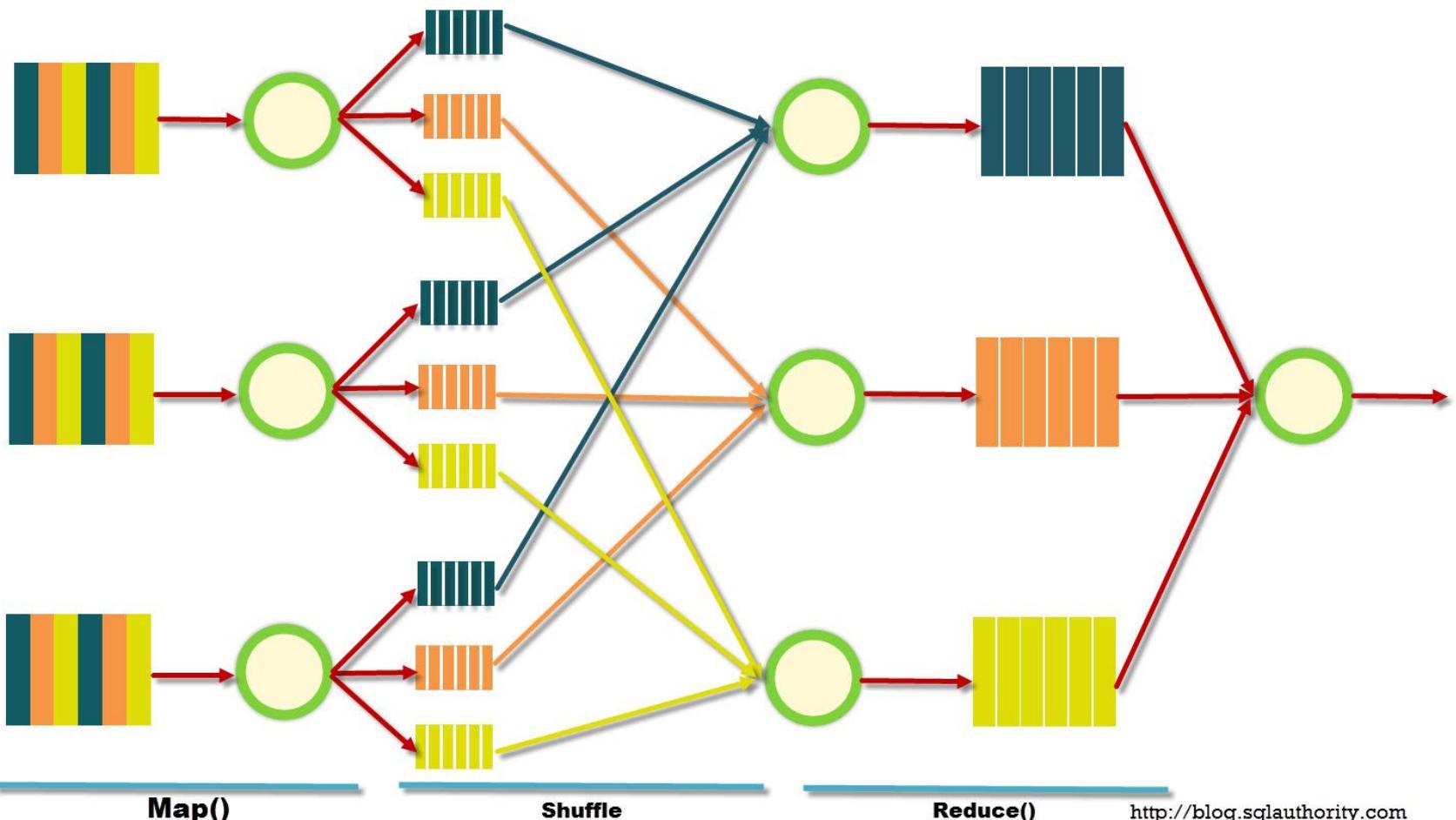
eventually consistent







How MapReduce Works?



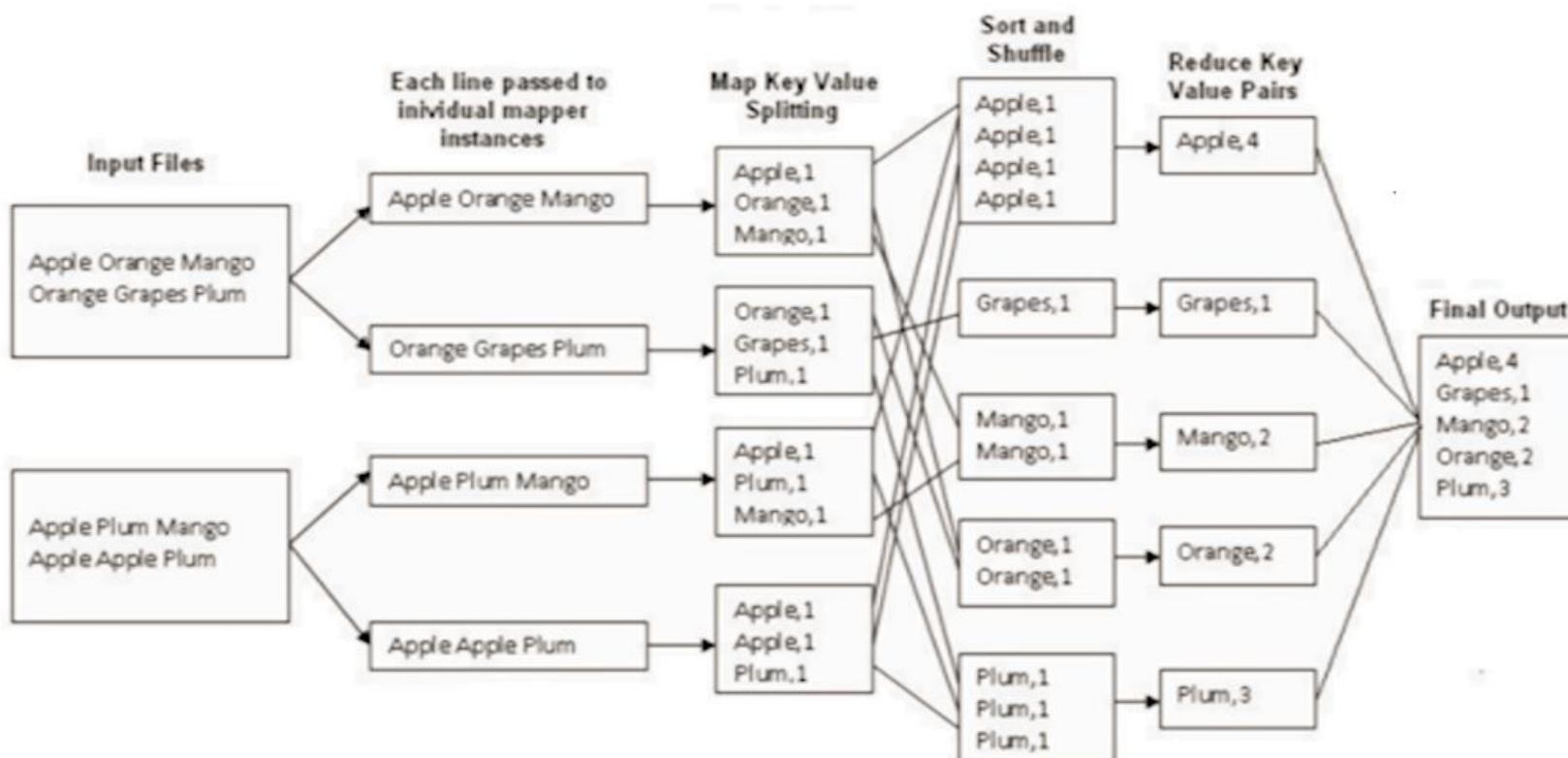
Map()

Shuffle

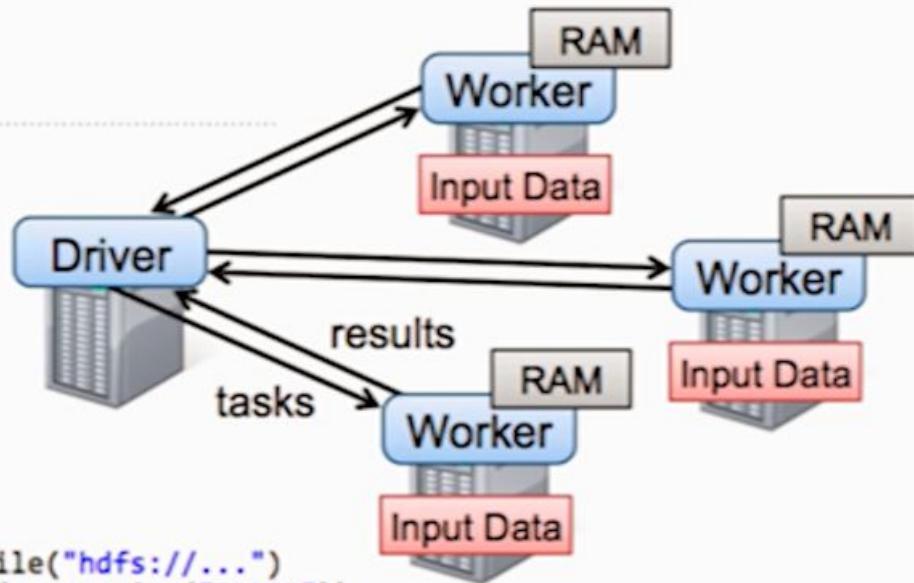
Reduce()

<http://blog.sqlauthority.com>

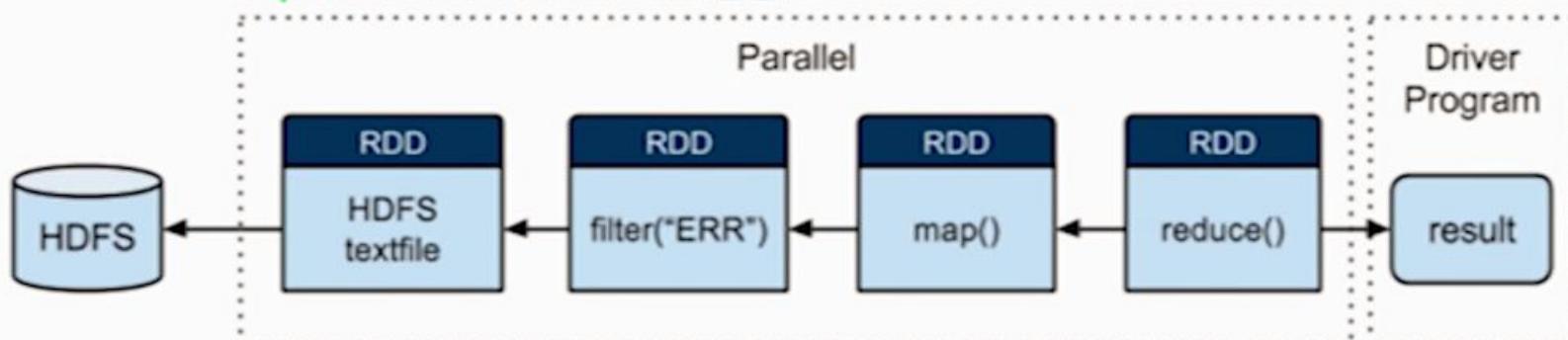
MapReduce



Spark



```
1 | val file = spark.textFile("hdfs://...")  
2 | val errs = file.filter(_.contains("ERROR"))  
3 | val ones = errs.map(_ => 1)  
4 | val count = ones.reduce(_+_)
```



Desafios para adoção do BigData



Jornada Analítica Empresarial

- **Iniciante:** dados desconectado, intervenção manual, relatórios Estáticos e sem governança
- **Executor:** Automaçã, BI em silos, traços de governança de dados
- **Líder:** Informações corporativas, metadados, funções analíticas, simulação e governança formal
- **Inovador:** Usa BigData, indicadores corporativos, Análise preditiva, DNA analítico e auditoria da governança

Baixa qualidade dos dados

- **Fraco acoplamento**
- **Dados faltantes**
- **Dados inconsistentes**
- **Falta de identificadores universais**

Data privacy

- **semi-identificadores**
- **k-anonymity**
- **Differential privacy**

Fim