

# Big Data e Machine Learning com Hadoop e Spark



# Sobre a Konfidentia

## Sobre nós

---

A Konfidentia IT Solutions é uma empresa com 10 anos de mercado, atuante em serviços de Consultoria e Treinamento em Governança e Gestão de TI, BigData e Machine Learning, DevOps e Desenvolvimento para os seguintes temas:

- E-Commerce
- Inteligência de Negócios com BigData
- Modelos de Aprendizagem de Máquina
- Times de TI de Alta Performance
- Apoio a Pesquisa e Desenvolvimento
- Gestão da Inovação
- Gestão de Continuidade de Negócios / BCP
- Gestão de Riscos
- Gestão de Segurança da Informação
- Auditoria de TI
- Implantação de Escritórios de Projetos e Escritórios de Processos
- Implantação de Modelos de Gestão e Governança



Contato:

Paulo César Rodrigues

<https://www.linkedin.com/in/rodrigpc/>

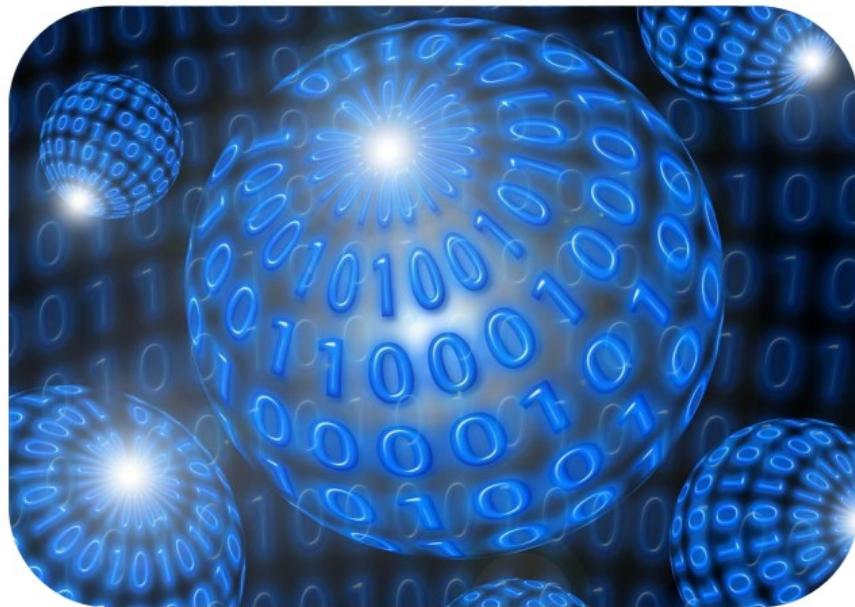
paulo.rodrigues@konfidentia.com

11-3280-6194



# Conceituação

BIG DATA é um conjunto de metodologias utilizadas para capturar, armazenar e processar um volume imenso de informações de várias fontes (dados estruturados e não estruturados) com o objetivo de acelerar a tomada de decisão e trazer vantagem competitiva.



# Os 7 V's do Big Data

VOLUME

VARIEDADE

VELOCIDADE

VISUALIZAÇÃO

VERACIDADE

VALOR

VULNERABILIDADE



# Tipos de Dados – Dados Estruturados



Inicialmente, os modelos eram construídos com base em informações armazenadas em bancos de dados com **dados estruturados**.

# Tipos de Dados – Dados Estruturados

	<b>produto character varying(70)</b>	<b>valor numeric</b>	<b>segmento character varying(70)</b>	<b>data date</b>
<b>1</b>	DVD	2.00	Papelaria e informática	2013-09-12
<b>2</b>	HD 500 GB	300.00	Papelaria e informática	2013-10-20
<b>3</b>	Tonner	250.00	Papelaria e informática	2013-11-01
<b>4</b>	Cadeira	50.00	Marcenaria	2013-09-19
<b>5</b>	Mesa	600.00	Marcenaria	2013-10-21
<b>6</b>	Armário	900.00	Marcenaria	2013-11-02
<b>7</b>	Corrimão	400.00	Serralheria	2013-09-12
<b>8</b>	Portão	1500.00	Serralheria	2013-10-22
<b>9</b>	Grade de proteção para janela	800.00	Serralheria	2013-11-03
<b>10</b>	Detergente	5.00	Limpeza e higiêne	2013-09-20
<b>11</b>	Desinfetante	40.00	Limpeza e higiêne	2013-11-23
<b>12</b>	Papel toalha	60.00	Limpeza e higiêne	2013-11-04

# Tipos de Dados – Dados Semiestruturados

Um exemplo de arquivo com dados semiestruturados é o arquivo:  
**XML (eXtensible Markup Language)**

```
<estoque>
    <item>
        <nome>Livro</nome>
        <preco>12</preco>
    </item>
    <item>
        <nome>Ventilador</nome>
        <preco>23</preco>
    </item>
    <item>
        <nome>Bolsa</nome>
        <preco>123</preco>
    </item>
</estoque>
```



# Tipos de Dados – Dados Semiestruturados

Neste caso, os dados são irregulares com uma estrutura embutida. A estrutura dos dados é heterogênea.

Sua principal característica é a **facilidade de compartilhamento** de informações pela internet.



# Tipos de Dados – Dados não estruturados

- Um dado não estruturado é um dado **sem uma estrutura pré-definida.**
  - **Textos** são exemplos de dados não estruturados. Podem ser oriundos de várias fontes como:



# Tipos de Dados – Dados não estruturados



# Definição – Machine Learning

**Machine Learning** pode ser traduzido simplesmente como **Aprendizado** (ou Aprendizagem) **de Máquina** (ou Computacional). O termo se refere a um enorme conjunto de técnicas que visam construir sistemas computacionais cujo comportamento seja definido com base em dados existentes.

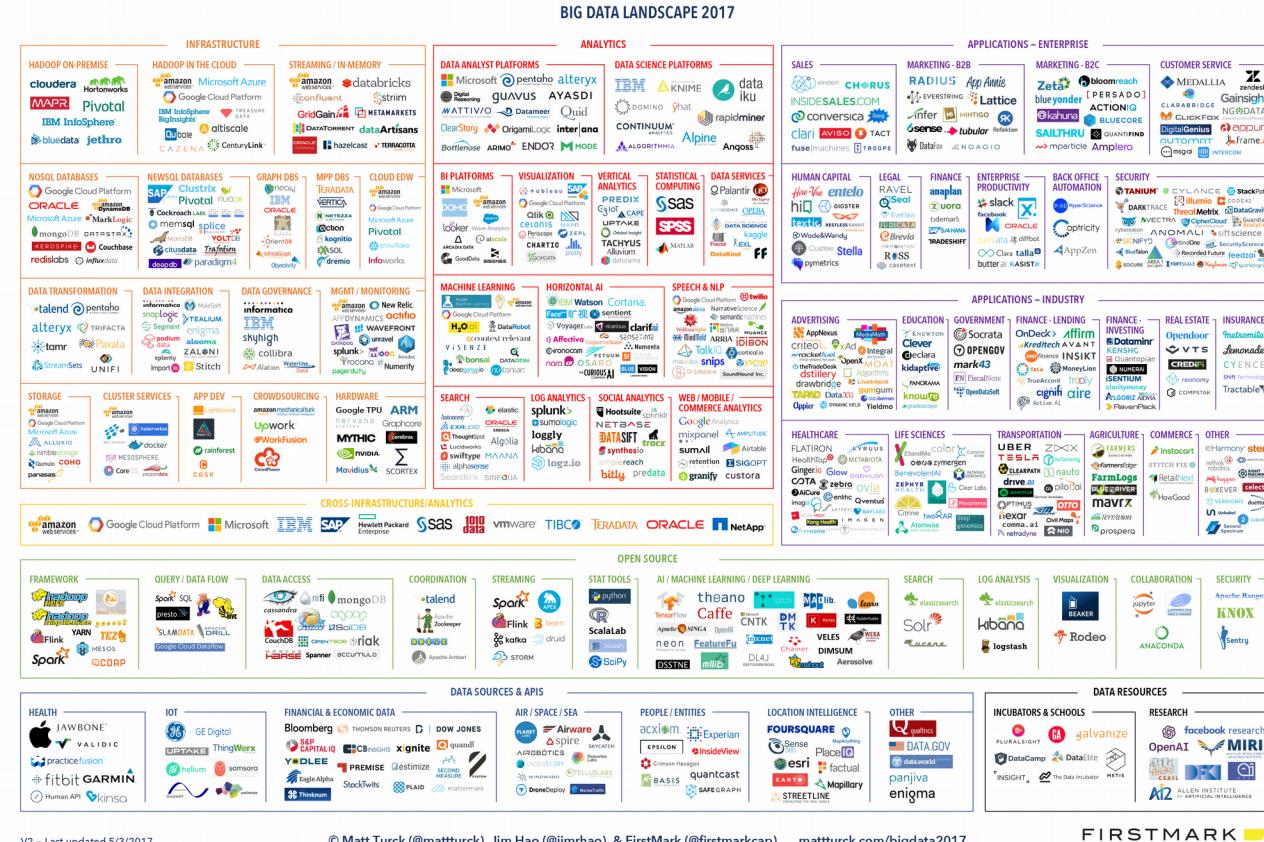
Como o comportamento do sistema não seria diretamente programado, mas sim adaptado de algum "conhecimento" previamente adquirido, essa abordagem teria similaridade com a forma como animais (entre eles, nós humanos) aprendem com a experiência.



# Business Cases



# Big Data & Machine Learning Landscape



paulo.rodrigues@konfidentia.com

# Big Data e Machine Learning com Hadoop e Spark



# O Profissional – Cientista de Dados

Um profissional de Big Data toma decisão com base em todas as informações disponíveis sobre o tema em estudo.

Este profissional considera as informações internas da empresa, as informações disponíveis na internet, nas redes sociais e nos dados gerados por sensores e imagem.



# O Profissional – Cientista de Dados



O cientista de dados é o profissional adequado para trabalhar com Big Data.

O cientista de dados é um profissional que precisa capturar informação, armazenar informação e elaborar modelos para a adequada tomada de decisão.

Para realizar estas tarefas o profissional precisa ter conhecimento de programação, banco de dados, segurança de informação, tecnologias de Big Data e modelagem.

# O Mercado de Big Data

CARGO (JOB TITLE)	2017		2018		%
<b>APLICAÇÃO &amp; INTEGRAÇÃO DE SISTEMAS (APPLICATION &amp; SYSTEM INTEGRATION) (B)</b>					
Gerente de Sistemas - Systems Manager	R\$ 13.000	- R\$ 25.000	R\$ 15.000	- R\$ 25.000	5,26%
Coordenador de Sistemas - Systems Coordinator	R\$ 9.500	- R\$ 13.000	R\$ 10.000	- R\$ 13.000	2,22%
Analista de sistemas Senior - Senior systems analyst	R\$ 5.200	- R\$ 8.000	R\$ 5.500	- R\$ 8.200	3,79%
Analista de sistemas Pleno - Systems analyst	R\$ 3.300	- R\$ 5.500	R\$ 3.500	- R\$ 5.700	4,55%
Analista de sistemas Junior - Junior systems analyst	R\$ 2.000	- R\$ 3.500	R\$ 2.200	- R\$ 3.500	3,64%
Arquiteto de aplicações - Applications architect	R\$ 8.500	- R\$ 12.000	R\$ 9.000	- R\$ 12.500	4,88%
Analista de Devops - Devops analyst	R\$ 4.000	- R\$ 11.000	R\$ 4.500	- R\$ 11.500	6,67%
Administrador de sistemas - Systems admin	R\$ 4.000	- R\$ 11.000	R\$ 4.500	- R\$ 11.200	4,67%
DBA	R\$ 3.000	- R\$ 9.500	R\$ 3.200	- R\$ 9.700	3,20%
<b>BIG DATA (BIG DATA) (C)</b>					
Especialista de Big Data / Cientista de dados - Big Data / Data Scientist specialist	R\$ 11.500	- R\$ 18.000	R\$ 12.000	- R\$ 22.000	15,25%
Analista de Big Data/ Cientista de dados - Big Data / Data Scientist analyst	R\$ 5.300	- R\$ 10.500	R\$ 5.500	- R\$ 12.500	13,92%
Especialista de BI - Business Intelligence specialist	R\$ 8.000	- R\$ 16.000	R\$ 10.000	- R\$ 16.500	10,42%
Analista de BI - Business Intelligence analyst	R\$ 4.000	- R\$ 10.000	R\$ 4.500	- R\$ 11.000	10,71%

# O Mercado de Big Data

**LOVE MONDAYS**

## Salários de Cientista de Dados na Nubank

[Ver todos os salários](#)

Ver salários publicados:

nos últimos 12 meses     nos últimos 24 meses     todos    [ATUALIZAR](#)

Todas as cidades [▼](#) [FILTRAR](#)

**Salário médio bruto**  
2 Postados

R\$ 24.406/mensal

min.     máx.

**Gráfico de variação salarial**



Não existem dados suficientes para exibir o gráfico.

# O Mercado de Big Data

LOVE MONDAYS

## Salários de Cientista de Dados

Último salário postado - 19/04/2018

Ver salários publicados:

nos últimos 12 meses

nos últimos 24 meses

todos

[ATUALIZAR](#)

### Salário médio bruto

33 Postados

R\$ 9.974/mensal

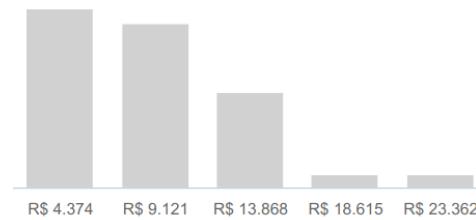
min.

R\$ 2.000

máx.

R\$ 25.735

### Gráfico de variação salarial



# O Mercado de Big Data



The screenshot shows a web browser window with the URL <https://canaltech.com.br/carreira/cientista-de-dados-e-a-profissao-com-as-melhores-oportunidades-de-carreira-101818/>. The page title is "Cientista de dados é a profissão com as melhores oportunidades de carreira". The text discusses the data scientist profession being listed as one of the most relevant for the world economy by the World Economic Forum until 2020. It quotes Renato Souza from FGV EMAP, mentioning applications in various sectors like finance, education, health, agriculture, geology, and industry. A sidebar on the left has icons for search, video, audio, fix, shopping cart, and more.

Seguro | https://canaltech.com.br/carreira/cientista-de-dados-e-a-profissao-com-as-melhores-oportunidades-de-carreira-101818/

Home > Carreira

## Cientista de dados é a profissão com as melhores oportunidades de carreira

Por Redação | 17 de Outubro de 2017 às 13h48

A carreira de cientista de dados foi listada pelo Fórum Econômico Mundial como uma das mais relevantes para o mercado até 2020. Para o professor da Escola de Matemática Aplicada da Fundação Getúlio Vargas (FGV EMAP), Renato Souza, as possibilidades de atuação estão nos mais diversos setores, como finanças, educação, saúde, agricultura, geologia e indústria.

"Qualquer empresa que gere dados pode contratar um profissional para analisá-los e tomar decisões com base em informação, não na intuição. Vamos ver as aplicações disso no dia a dia de governos, sociedade, hospitais e indústrias. O Brasil está entre os grandes produtores e consumidores de informação e, de maneira geral, tem iniciativas nessa área pipocando no mundo todo", afirma Renato Souza.

# Conteúdo

## CONTEÚDO PROGRAMÁTICO

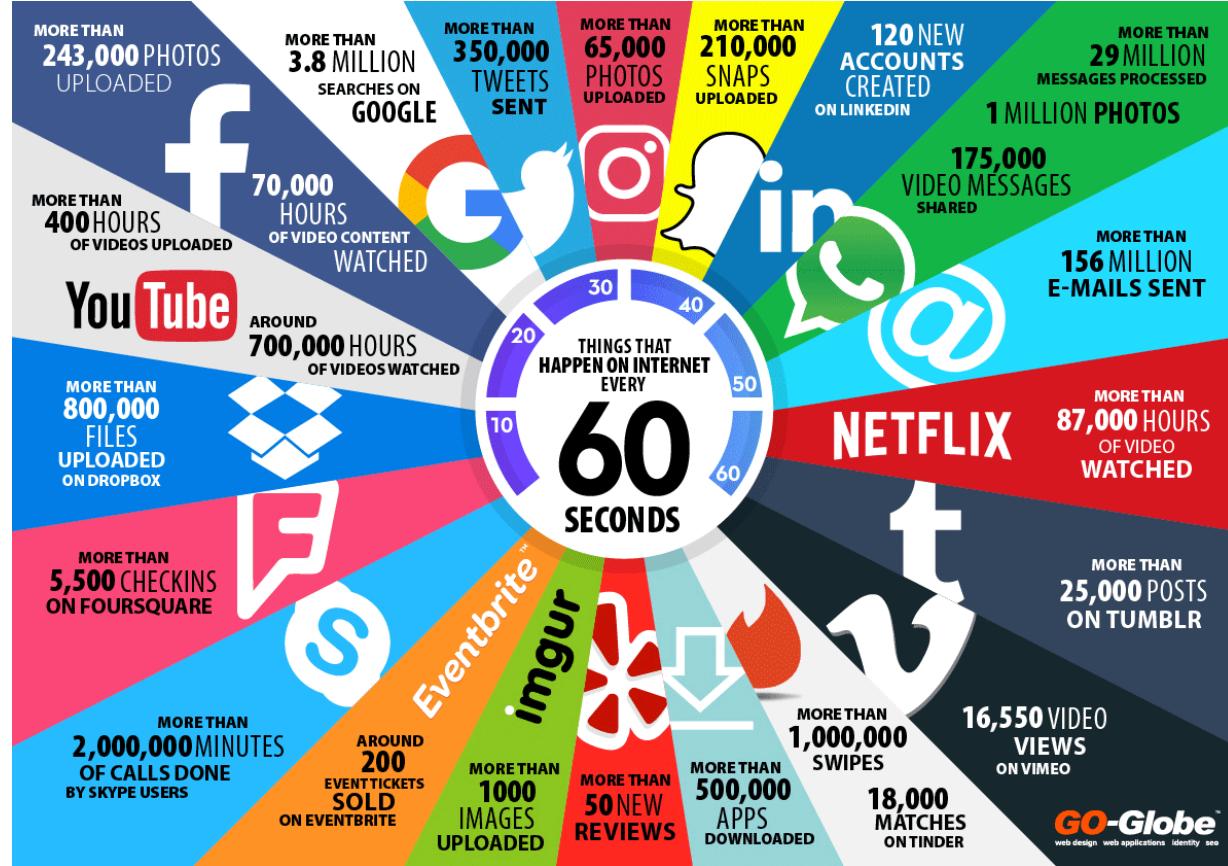
- Visão geral da ciência de dados e aprendizado de máquina em escala
- Visão geral do ecossistema do Hadoop
- Instalação de um Cluster Hadoop
- Trabalhando com dados do HDFS e tabelas do Hive usando o Hue
- Visão geral do Python
- Visão geral do R
- Visão geral do Apache Spark 2
- Leitura e gravação de dados
- Inspeção da qualidade dos dados
- Limpeza e transformação de dados
- Resumindo e agrupando dados
- Combinando, dividindo e remodelando dados
- Explorando dados
- Configuração, monitoramento e solução de problemas de aplicativos Spark
- Visão geral do aprendizado de máquina no Spark MLlib
- Extraíndo, transformando e selecionando recursos
- Construindo e avaliando modelos de regressão
- Construindo e avaliando modelos de classificação
- Construindo e avaliando modelos de cluster
- Validação cruzada de modelos e hiperparâmetros de ajuste
- Construção de pipelines de aprendizado de máquina
- Implantando modelos de aprendizado de máquina

## MATERIAL DIDÁTICO

- Slides do treinamento em PDF
- GitHub com exercícios e códigos exemplo
- Máquinas virtuais para exercícios simulados
- Gravação das aulas disponível durante 3 meses



# Por que?



# Skills

## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.



**MATH & STATISTICS**

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants

**PROGRAMMING & DATABASE**

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing packages, e.g. R
- ★ Databases: SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

**DOMAIN KNOWLEDGE & SOFT SKILLS**

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

**COMMUNICATION & VISUALIZATION**

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization, customer tracking and on-site analytics, predictive analytics and econometrics, data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing  
DISTILLERY  
(c) Krzysztof Latawski

paulo.rodrigues@konfidentia.com

Big Data e Machine Learning com Hadoop  
e Spark



# Times

Para trabalhar com Big Data é necessário muito mais do que o cientista de dados, as empresas hoje já possuem uma visão evoluída acerca deste tema e vem buscando montar equipes de Big Data.

Estas equipes tendem a trazer mais resultados quando multidisciplinares, normalmente são compostas pelo cientista de dados, um profissional de infraestrutura\DBA, um engenheiro de dados, um gerente de projetos e um especialista funcional.



# Times



## The Big Data Dream Team

**Dr. Raúl Arrabales**

Academic / Innovation Director

@MbitSchool



paulo.rodrigues@konfidentia.com

Big Data e Machine Learning com Hadoop  
e Spark



# Times

## Demystifying the Big Data Superheroes Mindset

**Gifted Individuals**  
**Vs.**  
**Talented Teams**

## How do Big Data Projects Succeed?

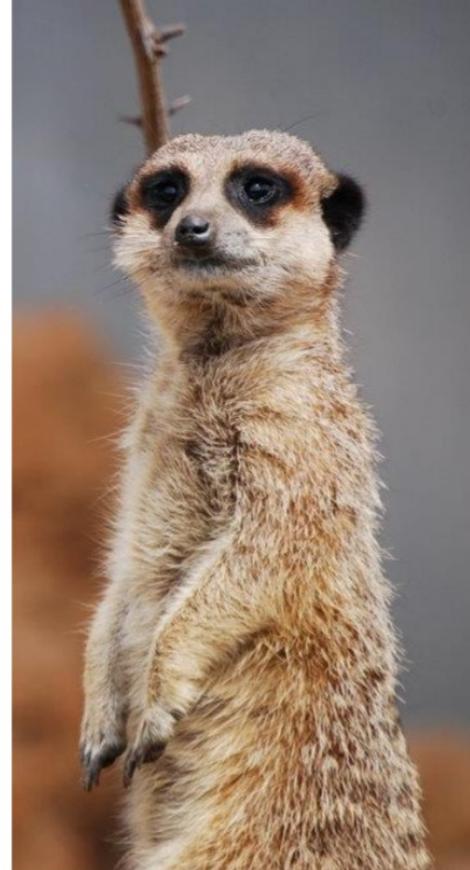
- Team
- Technology
- Project Mngmt
- Business Objective!
- Alignment!!



## Business Analyst

- Consultancy background
- Degree in Business
- Strong IT literacy
- Desirably MBA

**BA understands data life-cycle and context, identifies opportunities, boosts profit, improves competitiveness, reduces costs.**



# Times



## Project Manager

- PM background
- Degree in IT/Engineering
- Strong IT literacy
- Agile/Lean Methodologies
- Desirably PMP/PRINCE2

**Synergy/comm/flow of the team. Facilitates innovation and team talent growth. Manages customer expectations & requirements and ensures quality in solutions delivery.**

## Big Data Architect

- Storage/DB/Cloud background
- Degree in CS/IT/Engineering
- Datacenter/DWH Architecture
- OS/Network Administration.
- NoSQL, MapReduce, Shell Script

**Designs and optimize reliable, robust, fail-tolerant, safe, on demand, elastic computer systems infrastructure that enables big data applications to run and be delivered smoothly.**



# Times



## Big Data Developer

- Software Eng. background
- Degree in Computer Science
- Hardcore coder / Dif. Lang.
- NoSQL, Hadoop API, MR APIs
- Backend developer, scripting
- Real-time / in-memory app.

**Develops and integrates  
software applications to run  
efficiently and process petabytes  
of data across big data clusters.**

## Data Scientist

- Data mining background
- Degree in CS/Statistics/Math
- Hardcore algorithm designer/coder
- Analytical problem solver
- Expert in Machine Learning
- Ideally Ph.D. in CS/AI/Analytics

**Able to develop/design/apply new efficient algorithms to extract value from big data. Understands the opportunities hidden in massive datasets.**



# Times



## Cybersecurity Engineer

- Information Security Backgr.
- Degree in Computer Science.
- Ideally M.Sc./Ph.D. Security.
- IDS, Datacenter, SysAdm.
- Backend and frontend security.

**Able to assure the security of information across all big data platform and infrastructures, detecting and foreseeing cyber attacks end to end.**

# Times



## Folks from other lands

- Members of the board.
- Legal affairs. Digital ecosystem.
- BI/Datawarehouse Mngmt.
- IT infrastructures Mngmt.
- Business Partners.
- Data Owners & Agencies.
- Users & Customers!!

**Seamless deployment and integration coordinated with operations, delivery, etc. And external stakeholders.**

# Times

## Teamwork & Specialization



# Arquitetura e Componentes

『Data Lake é um termo recente, Lago de Dados em tradução livre, trata-se de uma área de armazenamento de dados estruturados ou não.

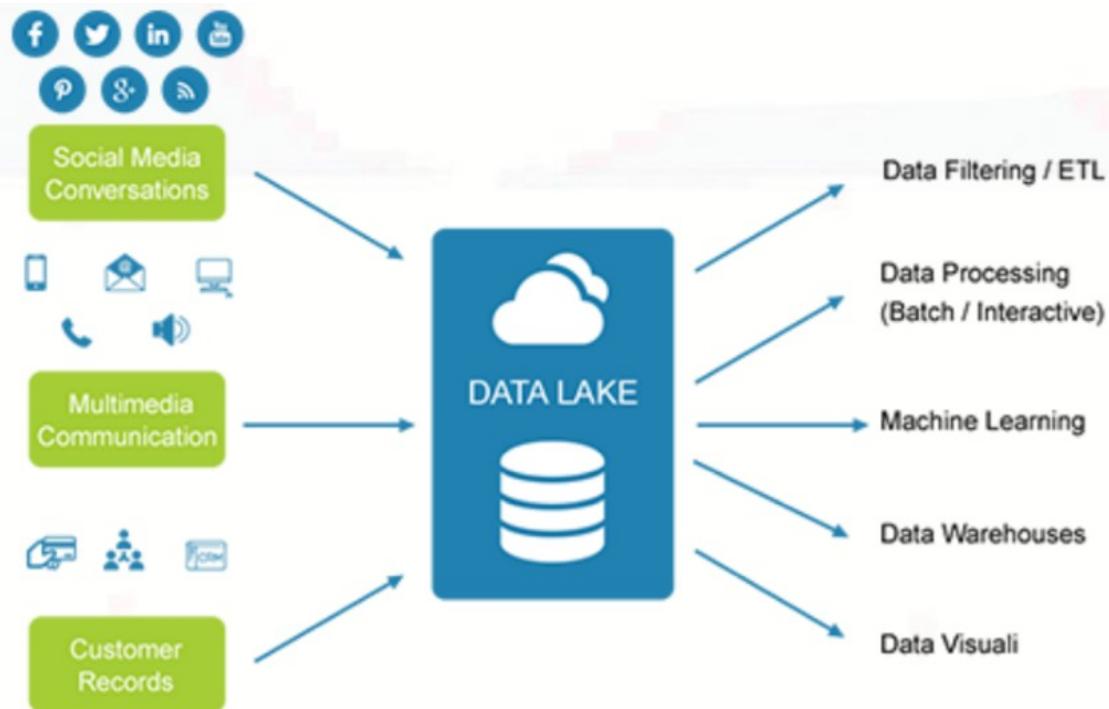
O Data Lake é um ponto essencial para a arquitetura de um ambiente Big Data, ele é o ponto de centralização dos dados da empresa e fonte de dados para os engenheiros e cientistas de dados.



# Arquitetura e Componentes

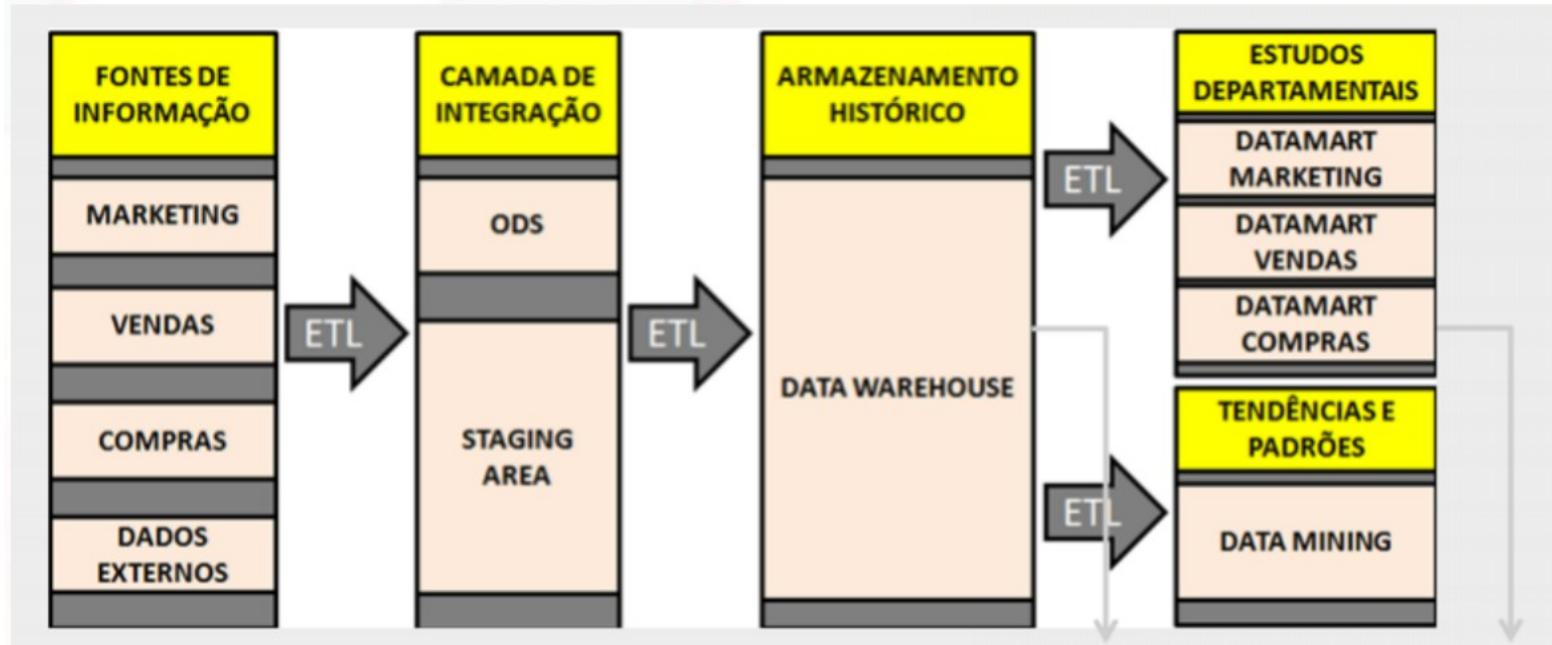
«DataLake tem como objetivo armazenar dados de diversos tipos e origens diferentes para ser utilizado como fonte principal de dados.

Ele também recebe os dados em seu formato original com o objetivo de otimizar o desempenho da ingestão dos dados.

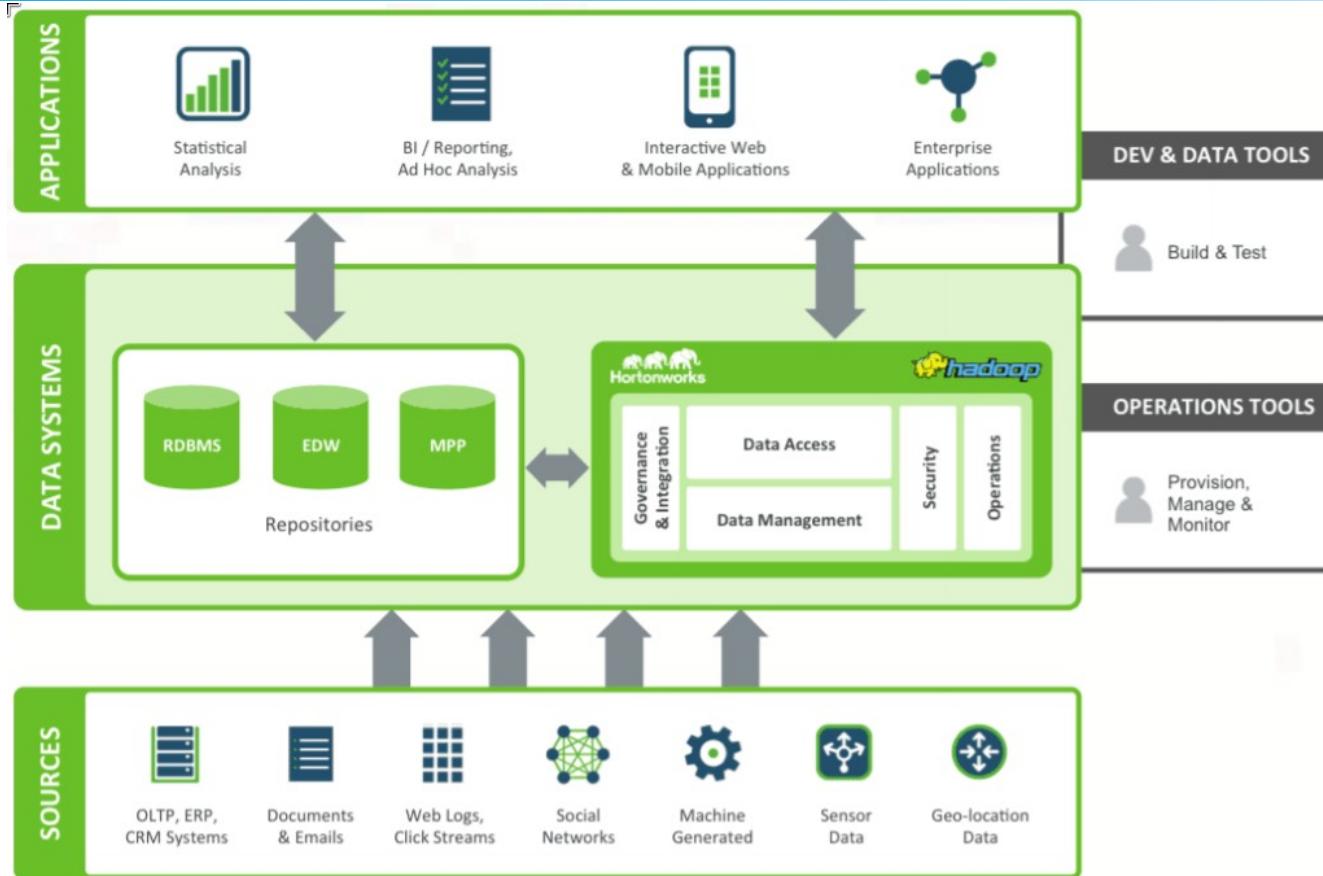


# Arquitetura e Componentes

## Arquitetura DataWare House tradicional

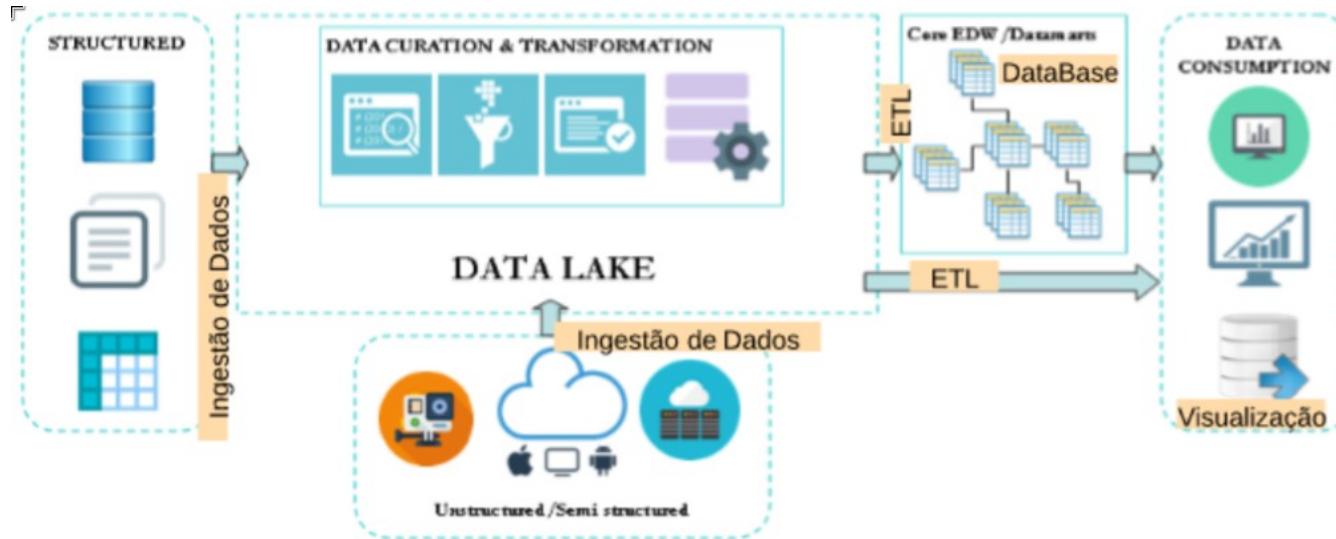


# Arquitetura e Componentes



# Arquitetura e Componentes

Agora, é hora de entender suas necessidades, objetivos e pesquisar as ferramentas mais adequadas, a seguir temos uma relação com as principais ferramentas de acordo com as etapas da arquitetura básica de Big Data aqui apresentada. Sendo estas Ingestão de Dados, DataLake, ETL, DB's\Datamart's e Visualização.



# Ferramentas de Ingestão de Dados



# Data Lake



ETL



talend\*



alteryx

# Database / Datamart



# Visualização



# Streaming Processing



Apache Flink



beam



# SQL Data Flows / Pipelines



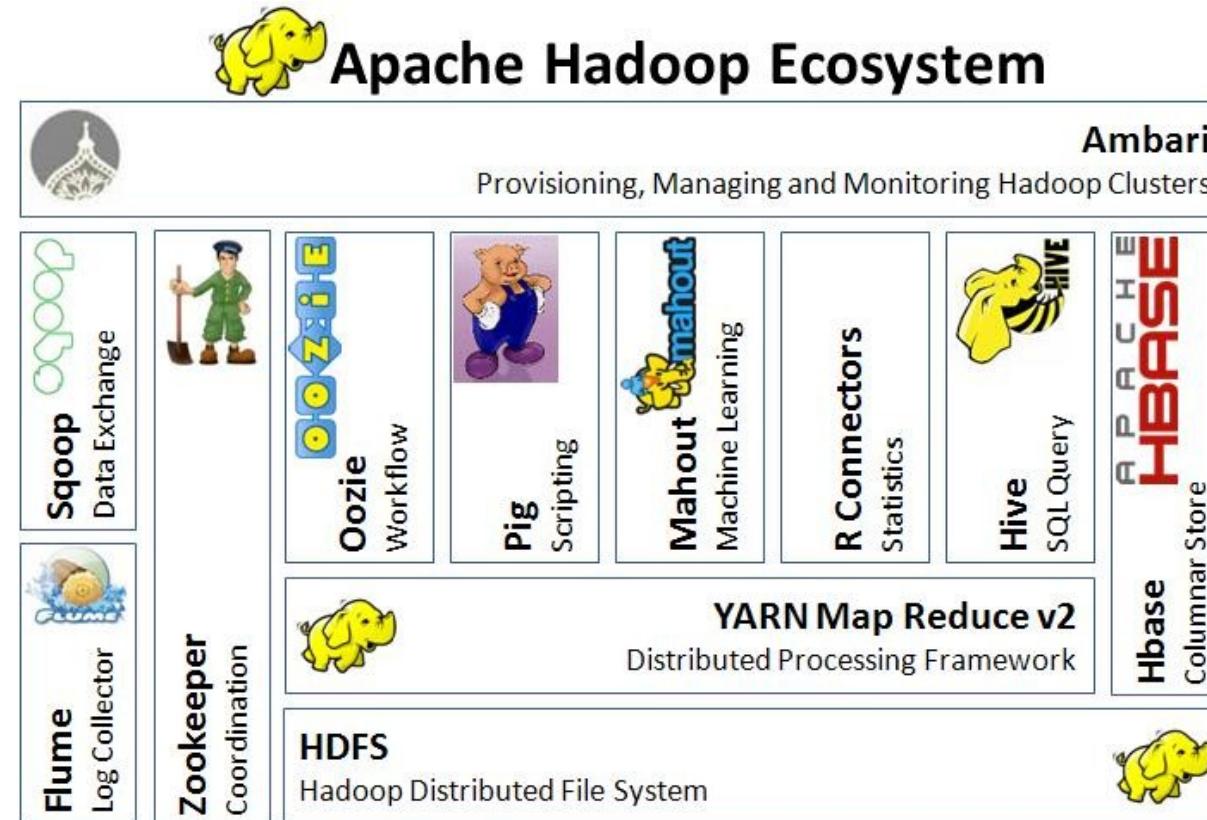
# Apache Software Foundation

A Apache Software Foundation (ASF) é uma organização sem fins lucrativos criada para prover suporte a projetos de código aberto, principalmente os Softwares Apache (Apache HTTP Server).

Os softwares criados pela fundação Apache são distribuídos como Software Livre e com a licença Apache.

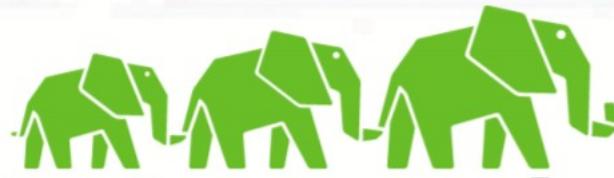


# Apache Software Foundation



# Distribuições

cloudera



databricks

DATASTAX

# Visão Geral



Apache Hadoop é um FrameWork de Software aberto distribuído através da Apache Software Foundation (ASF), tem como objetivo o armazenamento e processamento de dados de forma distribuída. Ele permite o armazenamento e processamento de grandes volumes de dados utilizando hardware commodity através de seu poder de distribuição.

Dentre os serviços e funcionalidades do Hadoop é possível além de armazenar e processar, realizar gestão, governança, segurança e operações de dados.

O Hadoop foi desenvolvido pela Yahoo e posteriormente entregue para a Apache.

# História

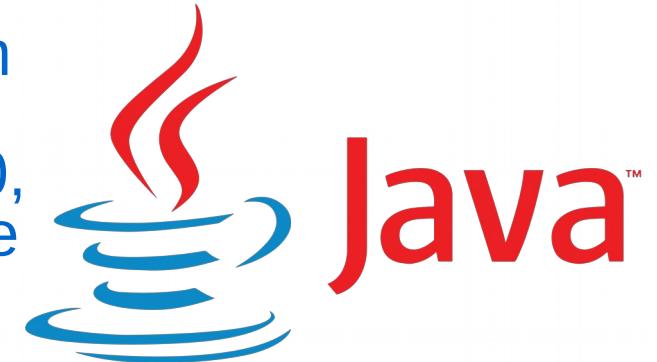


- 2003 Google publica artigo do GFS (SOSP'03)
- 2004 Google publica artigo do MapReduce (OSDI'04)
- 2005 Doug Cutting cria uma versão do MapReduce para o projeto Nutch
- 2006 Hadoop se torna um subprojeto do Apache Lucene
- 2007 Yahoo! Inc. se torna o maior contribuidor e utilizador do projeto (aglomerado com mais de 1.000 nós)
- 2008 Hadoop deixa a tutela do projeto Lucene e se transforma em um projeto top-level da Apache
- 2010 Facebook anuncia o maior aglomerado Hadoop do mundo (mais de 2.900 nós e 30 petabytes de dados)
- 2011 Apache disponibiliza a versão 1.0.0

# História

O Ecossistema Hadoop é desenvolvido em Java, linguagem orientada a objetos desenvolvida em meados da década de 90, um dos diferenciais dessa linguagem é que ela utiliza uma Maquina Virtual que interpreta seus códigos compilados em bytecode, diferente de outras linguagens que são compiladas diretamente para a linguagem nativa do Sistema Operacional.

A Maquina Virtual que interpreta os códigos Java é conhecida como JVM e é responsável também pela execução do Hadoop.



# HDFS – Hadoop Distributed File System

Trata-se do gerenciador de arquivos do Hadoop, é um dos subprojetos dentro do projeto Hadoop e tem por objetivo permitir o armazenamento de grandes quantidades de dados de forma distribuída.

É altamente recomendado para quantidades de armazenamento em proporções de Tera e PetaBytes. O HDFS armazena os dados de maneira contínua e o acesso aos dados é realizado por modo de fluxo, ou seja os dados são resgatados através de comandos MapReduce.



# HDFS– Hadoop Distributed File System

O HDFS possui diversas características específicas que o diferenciam de outros sistemas de arquivos distribuídos. A característica que mais chama atenção é o fato dele trabalhar sob o modelo WORM (Write-Once-Ready-Many), o uso deste modelo permite que ele não dependa de forte controle de simultaneidade, simplifica a persistência dos dados e habilita o acesso de alto rendimento.



# HDFS– Hadoop Distributed File System

Outra característica positiva do HDFS é que ele orienta o processamento a ser realizado o mais próximo dado possível, ao invés de movimentar o dado para um nó específico e para realizar o processamento, ou seja se você precisa processar o Dado 1 e o Dado 5 e estes estão localizados nos Nós 1 e 5 consecutivamente, não há necessidade de movimentar o Dado 1 e o Dado 5 para o Master realizar o processamento. O próprio nó 1 deve processar o Dado 1 enquanto o Nó 5 irá processar o Dado 5 e paralelamente ambos retornaram o resultado desejado, reduzindo assim o tempo de processamento de acordo com o quanto espalhados estão os dados nos Cluster.



# HDFS– Objetivos

- Tolerância a falhas pela detecção de falhas e aplicação de recuperação rápida, automática;
- Acesso a dados por meio do fluxo MapReduce;
- Modelo de simultaneidade simples e robusto;
- Lógica de processamento próxima aos dados, ao invés dos dados estarem próximos à lógica de processamento;
- Portabilidade entre sistemas operacionais e hardware padrão heterogêneos;



# HDFS– Objetivos

- Escalabilidade para armazenar e processar de modo confiável grandes quantidades de dados;
- Economia pela distribuição de dados e pelo processamento entre clusters de computadores pessoais padrão;
- Eficiência pela distribuição de dados e pela lógica para processá-los em paralelo nos nós em que os dados estão localizados;
- Confiabilidade pela manutenção automática de várias cópias dos dados e pela reimplementação automática da lógica de processamento no caso de falhas;



# Hadoop 2.x vs 3.x

Attributes	Hadoop 2.x	Hadoop 3.x
Handling Fault-tolerance	Through replication	Through erasure coding
Storage	Consumes 200% in HDFS	Consumes just 50%
Scalability	Limited	Improved
File System	DFS, FTP and Amazon S3	All features plus Microsoft Azure Data Lake File System
Manual Intervention	Not needed	Not needed
Scalability	Up to 10,000 nodes in a cluster	Over 10,000 nodes in a cluster
Cluster Resource Management	Handled by YARN	Handled by YARN
Data Balancing	Uses HDFS balancer for this purpose	Uses Intra-data node balancer



# Componentes mais usados

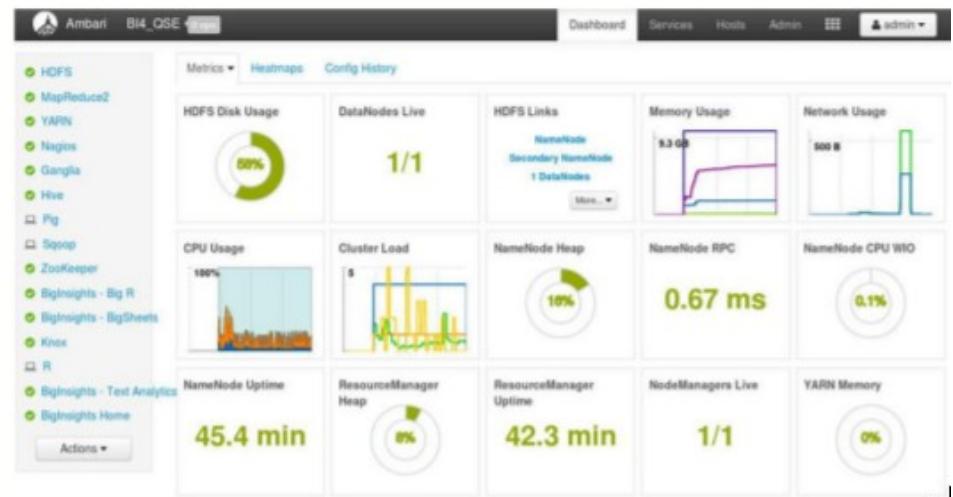
## Ambari

Tem o objetivo de simplificar a monitoração e gerenciamento dos Hadoop Clusters.

Por padrão a página inicial do Ambari trás o status operacional de seu cluster, além disso ainda por padrão o Ambari faz exibição das métricas do HDFS, YARN e Hbase. Também é possível incluir ou remover widgets otimizando sua monitoração do ambiente.



Apache  
Ambari

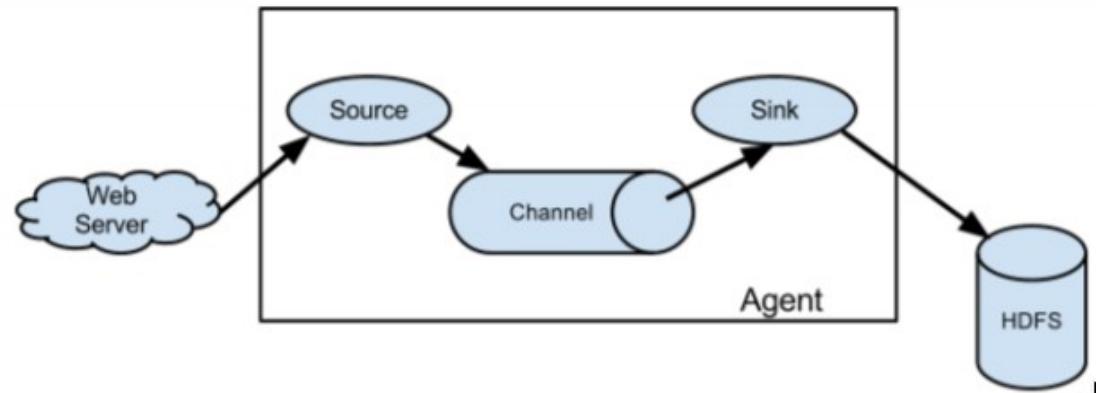


# Componentes mais usados

Serviço para coleta, agregação e transporte de grandes quantidades de dados de log para o HDFS. Foi criado especificamente para a ingestão de dados para o HDFS, permite algumas transformações no processo de ingestão como a agregação de dados



O flume é indicado não apenas na ingestão de dados de logs, mas para qualquer fluxo dados gerado em timeseries e grandes quantidades como informações de redes sociais por exemplo.



# Componentes mais usados

## HBase



Banco de dados NoSQL orientado a colunas, que roda em cima do HDFS, pode ser utilizado como fonte ou destino do para as funções de MapReduce. Permite o particionamento dos dados através do cluster e é indicado para o armazenamento e análise de grandes massas de dados.

Ao trabalhar com o Hbase, deve-se tomar muito cuidado com a modelagem dos dados, das chaves das famílias de colunas e suas partições.

	fname	lname	picture
aputrell@apache.org	"Andrew"	"Putrell"	
jdcryans@apache.org	"Jean-Daniel"	"Cryans"	downfall.jpg
stack@apache.org		"Stack"	dancing_stack.jpg
todd@apache.org	"Todd"	"Lipcon"	turbo.jpg

Row Key identifies a row across Column Families

Column Families store frequently accessed data together. Settings can be customized on a per Column Family basis

# Componentes mais usados

Sistema de Data Warehouse para o Hadoop que facilita summarização de dados e queries adhoc. Linguagem similar ao SQL. De forma simplificada o Hive é uma camada de abstração onde é possível criar comandos SQL e ele converte em operações Map Reduce que são executadas nas famílias de colunas do Hbase. Permitindo assim maior facilidade na hora de buscar dados para questões pontuais e isoladas



# Componentes mais usados

『Biblioteca escalável de machine learning e mineração de dados

Através d Mahout desenvolvedor e cientistas de dados podem realizar analise avançadas aplicando conceitos de machine learning aproveitando a escalabilidade do framework Hadoop. Está é uma ferramenta especifica para analise avançadas de grandes massas de dados.



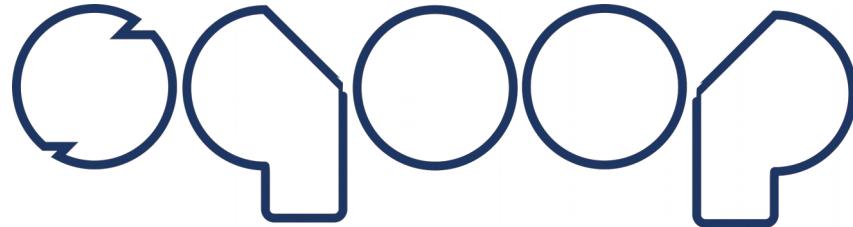
```
frank@frankthetank:~$ mahout
no HADOOP_HOME set, running locally
An example program must be given as the
first argument.
Valid program names are:
arff.vector: : Generate Vectors from an ARFF
               file or directory
canopy: : Canopy clustering
cat: : Print a file or resource as the
       logistic regression models would see
       it
...
...
```

# Componentes mais usados

Pig é uma linguagem de alto nível específica para a criação de processos Map Reduce, possui uma sintaxe simples e de fácil aprendizado. Permite não apenas realizar os processos de Map Reduce, mas também alguma transformação nos dados trabalhados. Existem alguns casos em que o Pig é utilizado como ETL para disponibilizar dados do Hbase ou HDFS para DataMarts específicos.



# Componentes mais usados



『Ferramenta para importar dados de banco de dados relacionais para o Hadoop e vice-versa.

É utilizada especificamente para a ingestão de dados para o HDFS ou Hbase, porém o sqoop foi criado especificamente para trabalhar com bancos de dados relacionais, ou seja para realizar a ingestão de dados estruturados de bases Oracle, MsSQL, PostGreSQL, entre outro o Sqoop é mais indicado que o Flume.

## Sqoop Import Examples

- `Sqoop import --connect jdbc:oracle:thin:@//dbserver:1521/masterdb  
--username hr --table emp  
--where "start_date > '01-01-2012'"`
- `Sqoop import  
jdbc:oracle:thin:@//dbserver:1521/masterdb  
--username myuser  
--table shops --split-by shop_id  
--num-mappers 16`

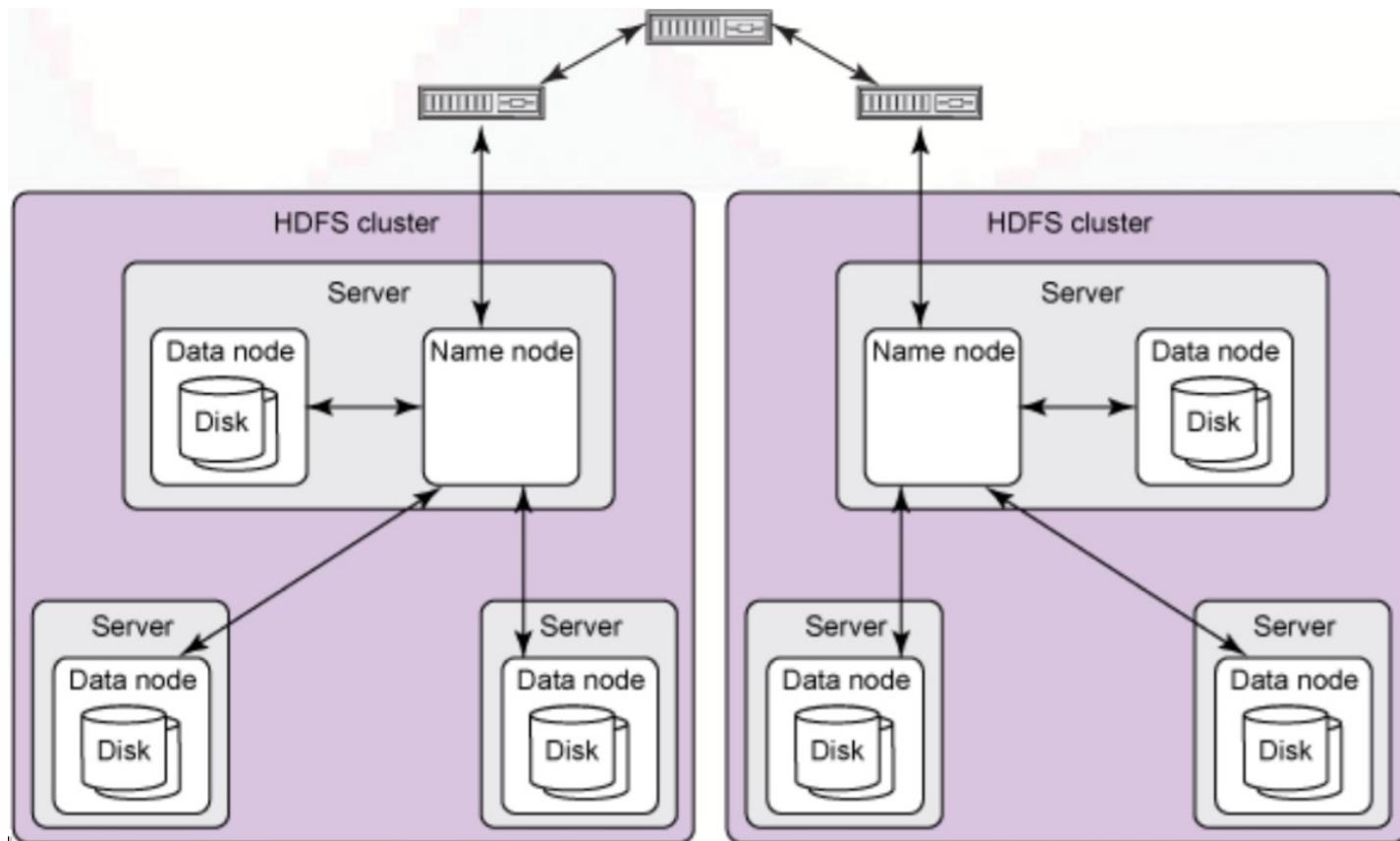
Must be indexed or partitioned to avoid 16 full table scans

# Componentes mais usados



- 『 Serviço de coordenação de alto desempenho para sistemas distribuídos.  
Basicamente é o responsável por manter todos os serviços do Ecossistema Hadoop em pleno funcionamento.  
ZooKeeper resolve este problema com a sua arquitectura simples e API. Permite que os desenvolvedores se concentrem na lógica do aplicativo principal sem se preocupar com a natureza distribuída da aplicação.
  - Confiabilidade - falha de um único ou alguns sistemas não faz todo o sistema falhe.
  - Escalabilidade - O desempenho pode ser aumentado como e quando necessário, adicionando mais máquinas com pequena alteração na configuração do aplicativo sem tempo de inatividade.
  - Transparência - Oculta a complexidade do sistema e mostra-se como uma única entidade / aplicação.

# HDFS - Arquitetura



# MapReduce - Conceitual

Definição: É um modelo de programação para processamento de dados com chave e valor. Não é uma linguagem ou uma plataforma. Foi inspirado no paper MapReduce do Google - 2004 .

- Consiste basicamente em três etapas:
  - » Map: Extrai algo que você espera de cada registro.
  - » Shuffle and Sort: Embaralha e ordena os registros.
  - » Reduce: Agrega, summariza, filtra ou transforma os dados.
- Objetivo: Facilitar a distribuição das tarefas e execução em paralelo entre os nodes de um cluster Hadoop.



# MapReduce - Conceitual

- Motivações:
  - Facilitar o desenvolvimento e execução de aplicações utilizando processamento paralelo.
  - Processar um volume massivo de dados utilizando uma infraestrutura de hardware commodity.



# MapReduce - Conceitual

Podemos também definir o MapReduce como um Pseudo Código das funções Map e Reduce com o objetivo de encontrar frequências de palavras ou conjuntos de caracteres.

## Map

```
map (String fileName, String document) {  
    List <String> T = tokenize(document)  
    for each token in T {  
        emitIntermediate ( token, 1)  
    }  
}
```

## Reduce

```
reduce (String token, List<Integer> values) {  
    Integer sum = 0  
    for each value in values {  
        sum = sum + value  
    }  
    emit ( token, sum)  
}
```

## Map em Java

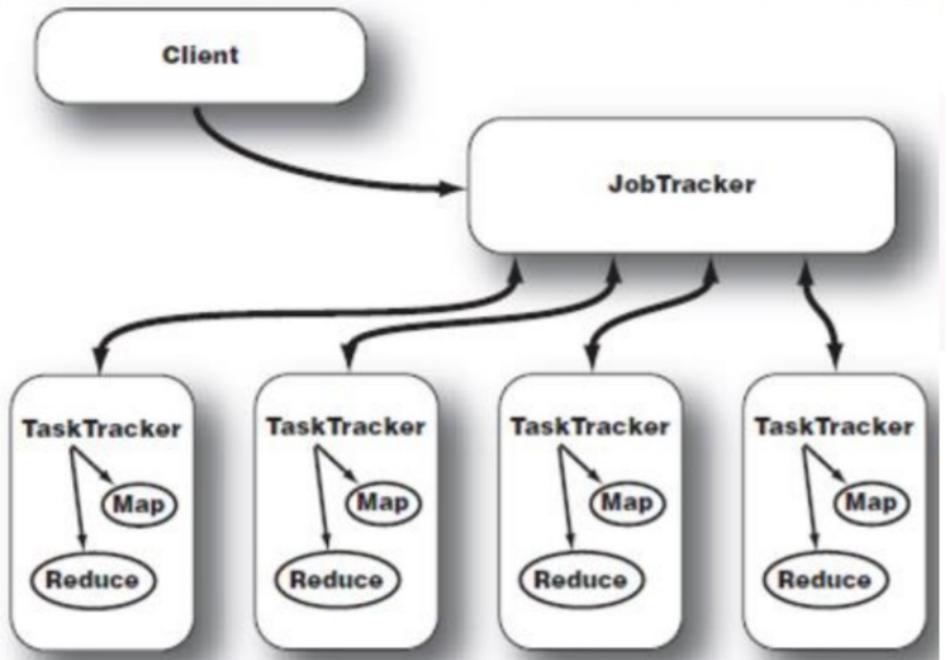


```
import java.io.BufferedReader;  
import java.util.Scanner;  
import java.util.StringTokenizer;  
public class Mapper {  
    private void map () {  
        Scanner sc = new Scanner(new BufferedReader(System.in));  
        String line = "";  
        while (sc.hasNextLine()) {  
            line = sc.nextLine();  
            StringTokenizer tokens = new StringTokenizer(line);  
            while(tokens.hasMoreTokens()){  
                emitIntermediate(tokens.nextToken(), "1");  
            }  
        }  
        sc.close();  
    }  
    private void emitIntermediate (String w, String um) {  
        System.out.println(w + "\t" + um);  
    }  
    public static void main(String[] args) {  
        new Mapper().map();  
    }  
}
```

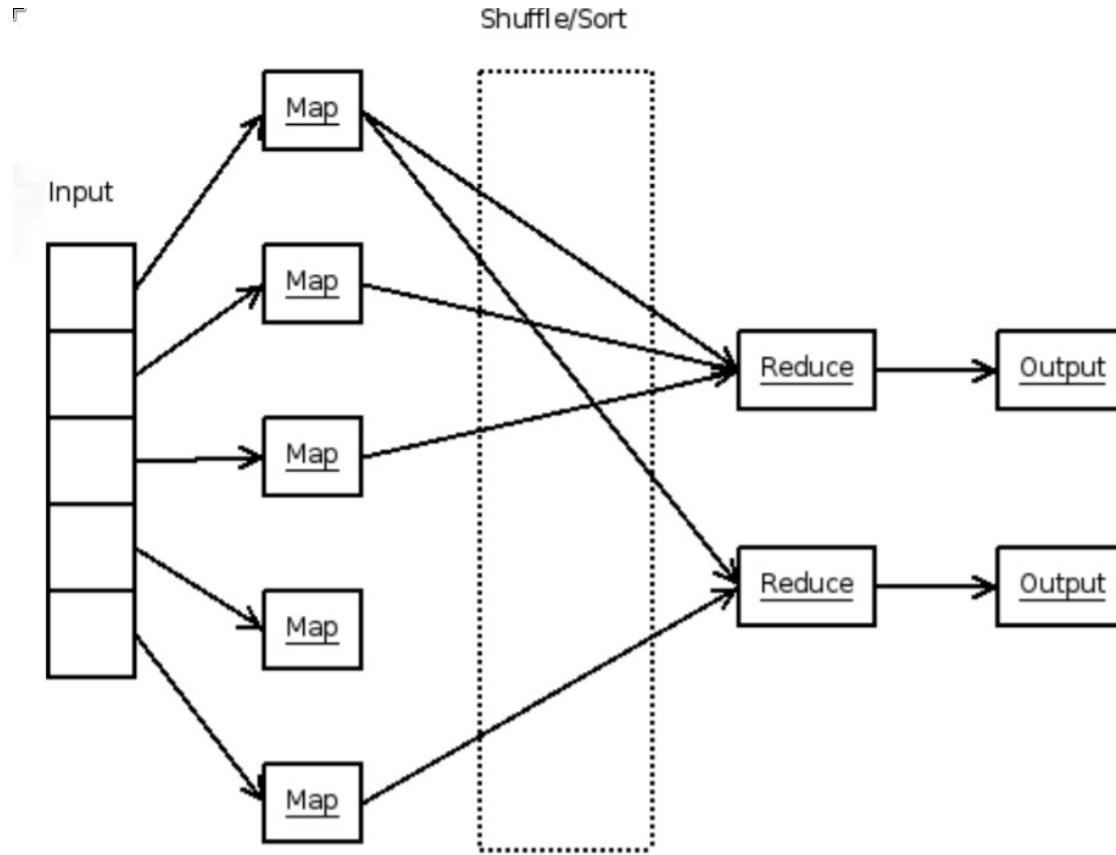
# MapReduce - Conceitual

## Componentes da arquitetura:

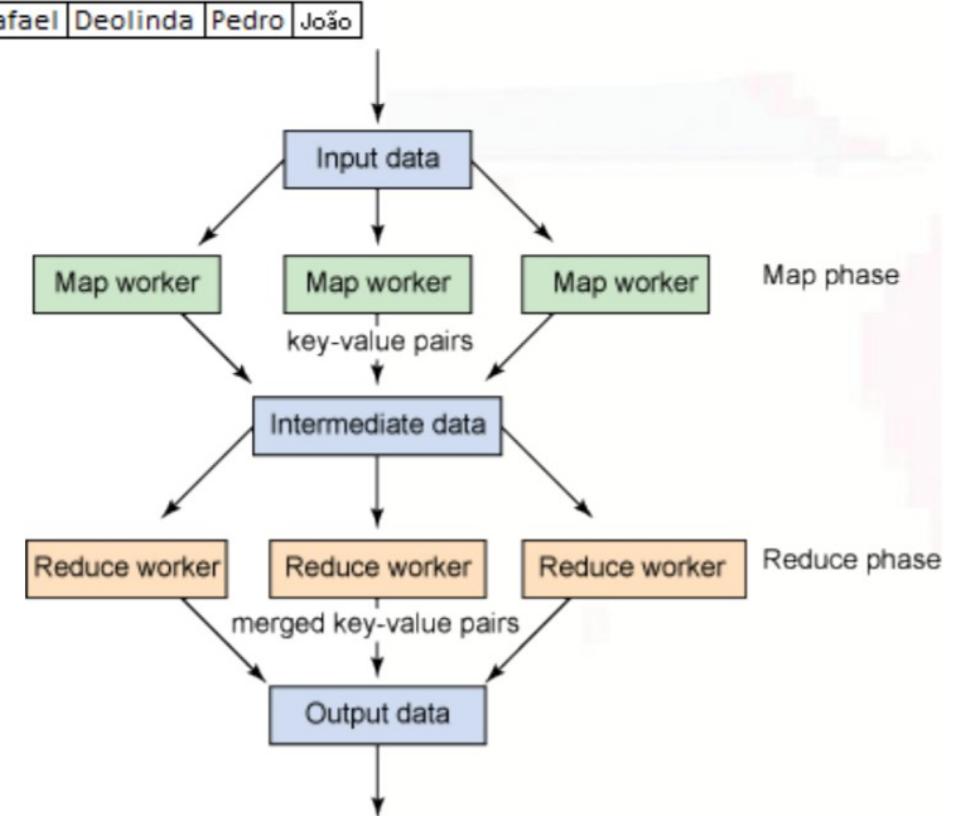
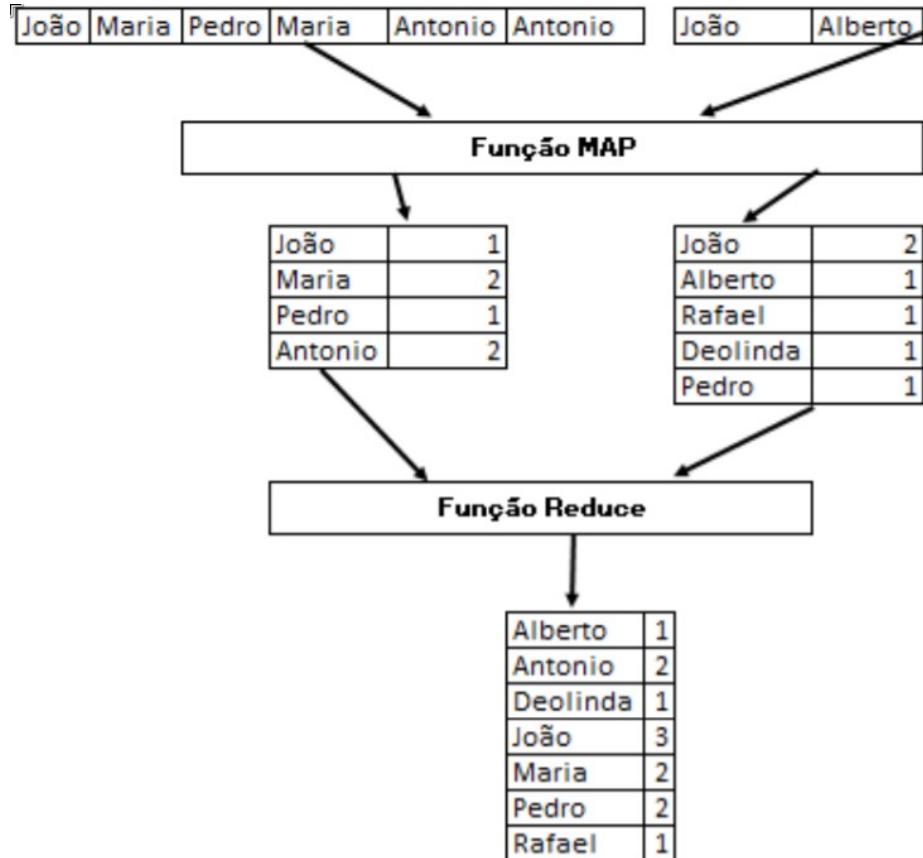
- **JobTracker:** Gerencia os Jobs e distribui as Tasks entre os TaskTrackers.
- **TaskTrackers:** Inicia e monitora a execução individual das Tasks de Map e Reduce.



# MapReduce - Conceitual



# MapReduce - Conceitual



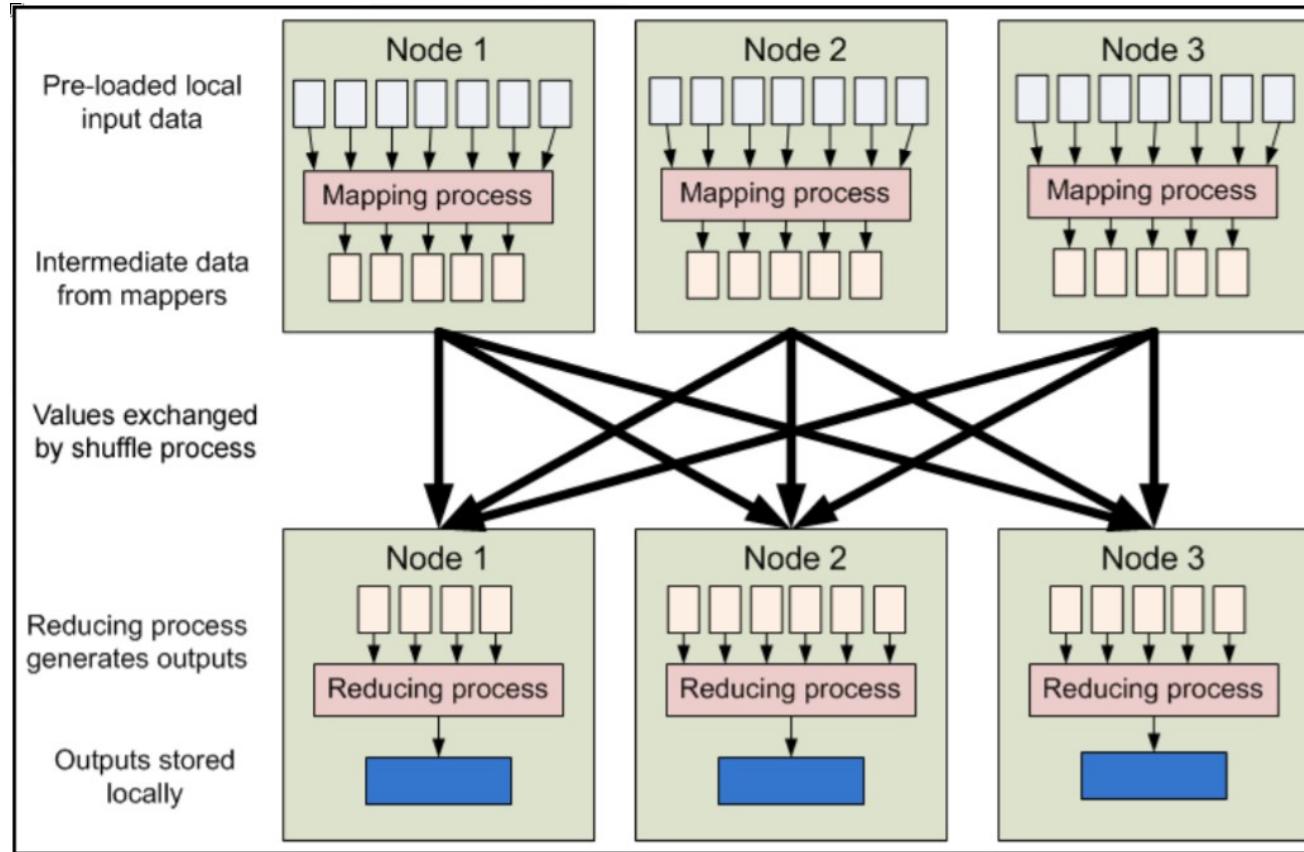
# MapReduce no Cluster

O MapReduce por ser um algoritmo orientado ao processamento paralelo e distribuído, possui uma vantagem natural ao ser executado em ambientes Clusterizados. Este é o principal motivo pelo qual o MaReduce é uma das principais funções do Hadoop.

Ele remove pontos de “gargalo” que aplicações de processamento centralizado possuem, como por exemplo o limite de CPU ou de áreas de log. Além disso ele utiliza do Cluster para garantir que você nunca irá perder seu processamento por completo. Por exemplo se um servidor de processamento centralizado perde conexão ou cai por qualquer motivo, você irá perder todos os dados já processados e terá de reiniciar seu processo, enquanto utilizando o MapReduce em ambientes clusterizados caso um node perca conexão ou fique offline, você poderá perder apenas uma pequena fração de seu processamento, porém normalmente não há perda de dados ou processamento, pois os dados estão replicados e nesta situação o MapReduce automaticamente irá encontrar outro Node que tenha aquele dado e retomar o processamento.



# MapReduce no Cluster



Obrigado!!!

Nos vemos amanhã!!!

Bom descanso!

