ENGENHARIA DE DADOS COM HADOOP E SPARK

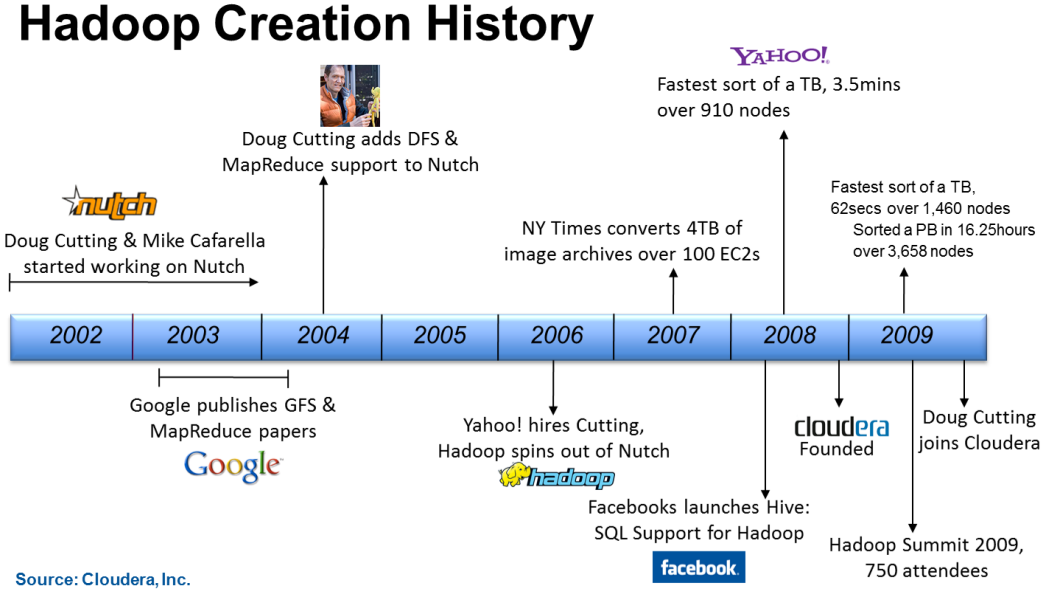
CONTEÚDO

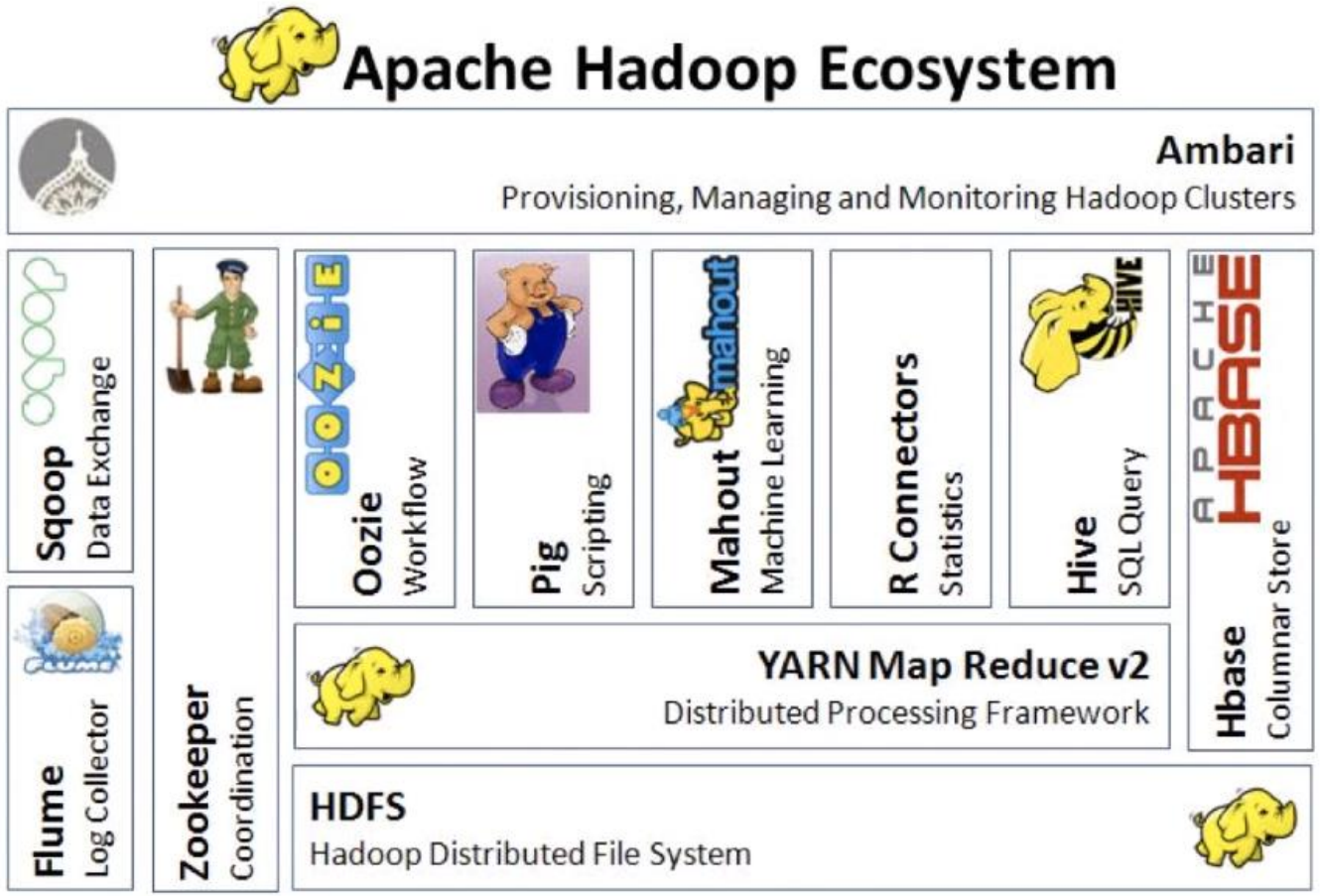
- Cluster Hadoop
- Armazenamento de Dados
- Machine Learning
- Hadoop e Spark

GUIA DE ESTUDO

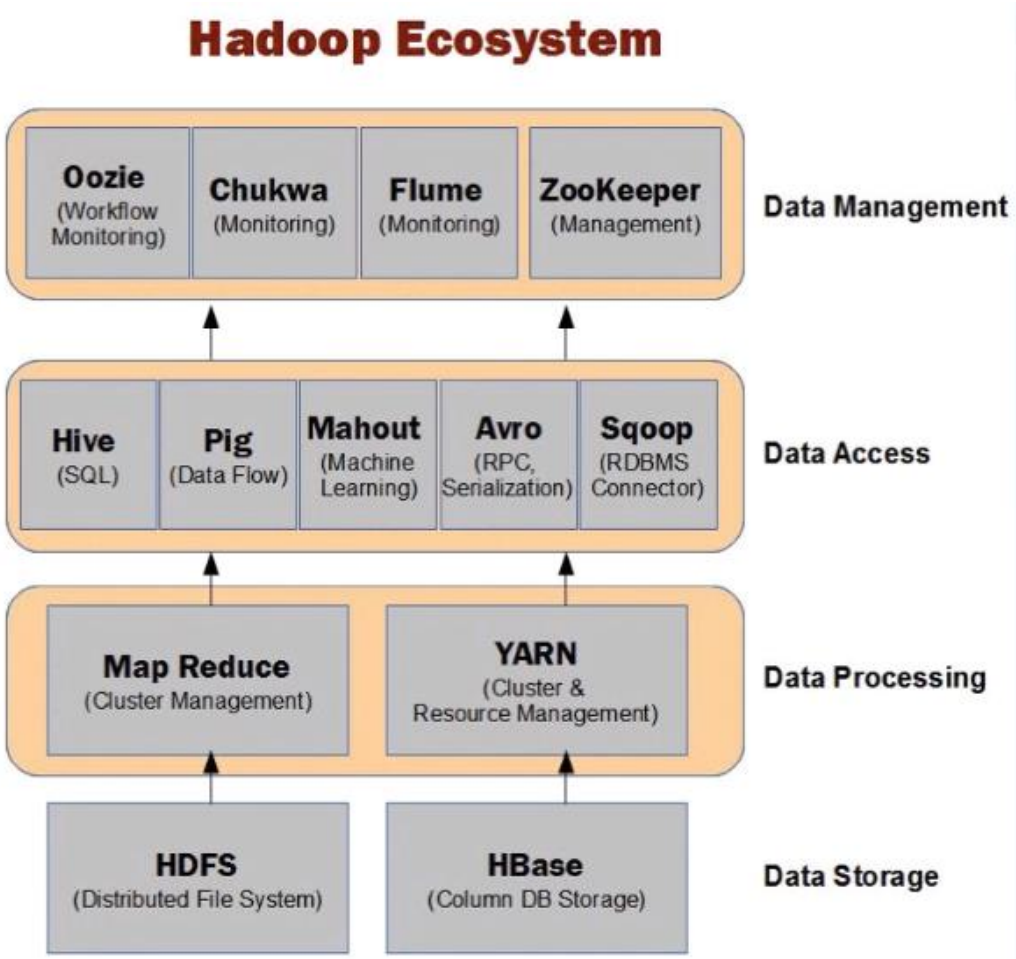
1. Conceitos e definições de Big Data, Hadoop, Ecossistema Hadoop e Spark.
2. Como planejar, instalar e configurar um cluster Hadoop.
3. Como planejar, instalar e configurar o Ecossistema Hadoop (Hive, Hbase, Zookeeper, Flume, Oozie, Ambari, Sqoop, Spark e Storm).
4. Configurações e utilização do HDFS e configurações avançadas do cluster Hadoop.
5. Administração e Manutenção do Hadoop e Spark.
6. Machine Learning com Apache Mahout.
7. Importação/exportação de dados e ETL com Sqoop.
8. Principais distribuições Hadoop do mercado: Cloudera e Hortonworks.
9. Infraestrutura de Big Data.
10. Análise de Big Data.

ECOSSISTEMA HADOOP





ECOSSISTEMA HADOOP



Principais projetos do Ecossistema Hadoop:

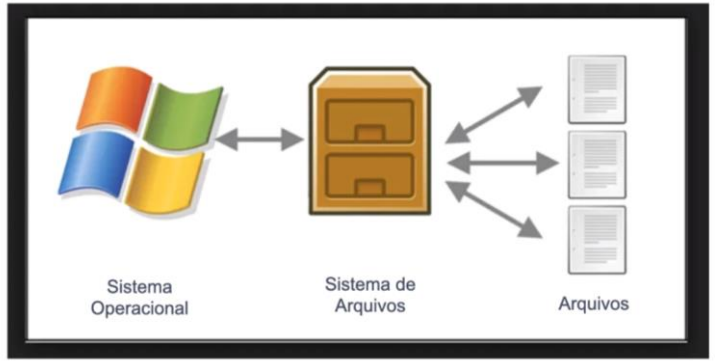
- Hadoop MapReduce
- Hadoop YARN
- Hadoop HDFS (Hadoop Distributed File System)

Outros projetos importantes:

- Zookeeper (Coordenação)
- Hive (SQL Query)
- Hbase (Banco de dados NoSQL)
- Pig (Linguagem orientada a fluxo de dados)
- Sqoop (Transferência de dados)
- Mahout (Machine Learning)
- Flume (Coletor de logs de diversas fontes)
- Oozie (Fluxo de trabalho)

APACHE HDFS

- HDFS (Hadoop Distributed File System) é o Sistema de Arquivos Distribuído para Big Data.

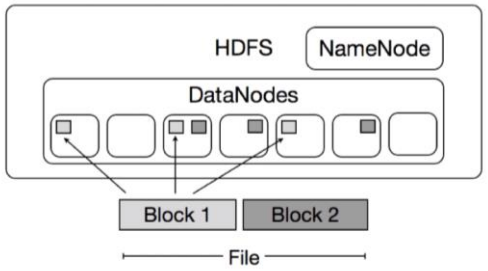


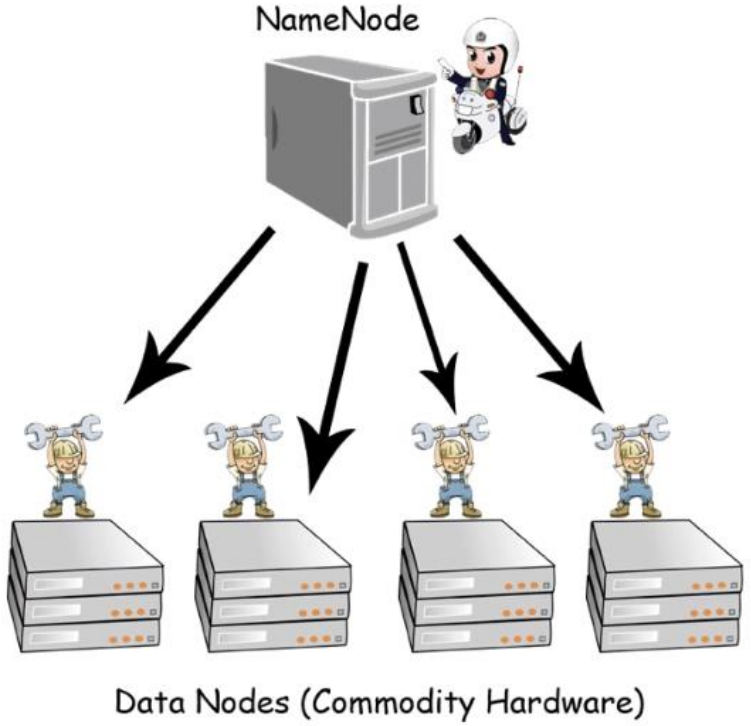
Os tipos de Sistemas de Arquivos são:

Tipo	Descrição
ext2	Sistema de arquivos padrão do Linux
ext3	Sistema de arquivos ext2 melhorado
reiserfs	Sistema de arquivos do tipo Journaling
msdos	Sistema de aruqivos FAT da Microsoft DOS
vfat	Sistema de arquivos FAT-32 do Microsoft Windows
iso9660	Sistema de arquivos do CD-ROM
nfs	Network File System. Usado par a montar dispositivos em computadores remotos.
swap	Sistema de arquivos de troca utilizando para memória virtual.
proc	Uma janela especial dentro do Kernel do Linux. Utilizada pelos usuários, programas e utilitários para escrever ou ler parâmetros do Kernel. Geralmente montado no diretório /proc.

Estruturas de controle de arquivos distribuídos:

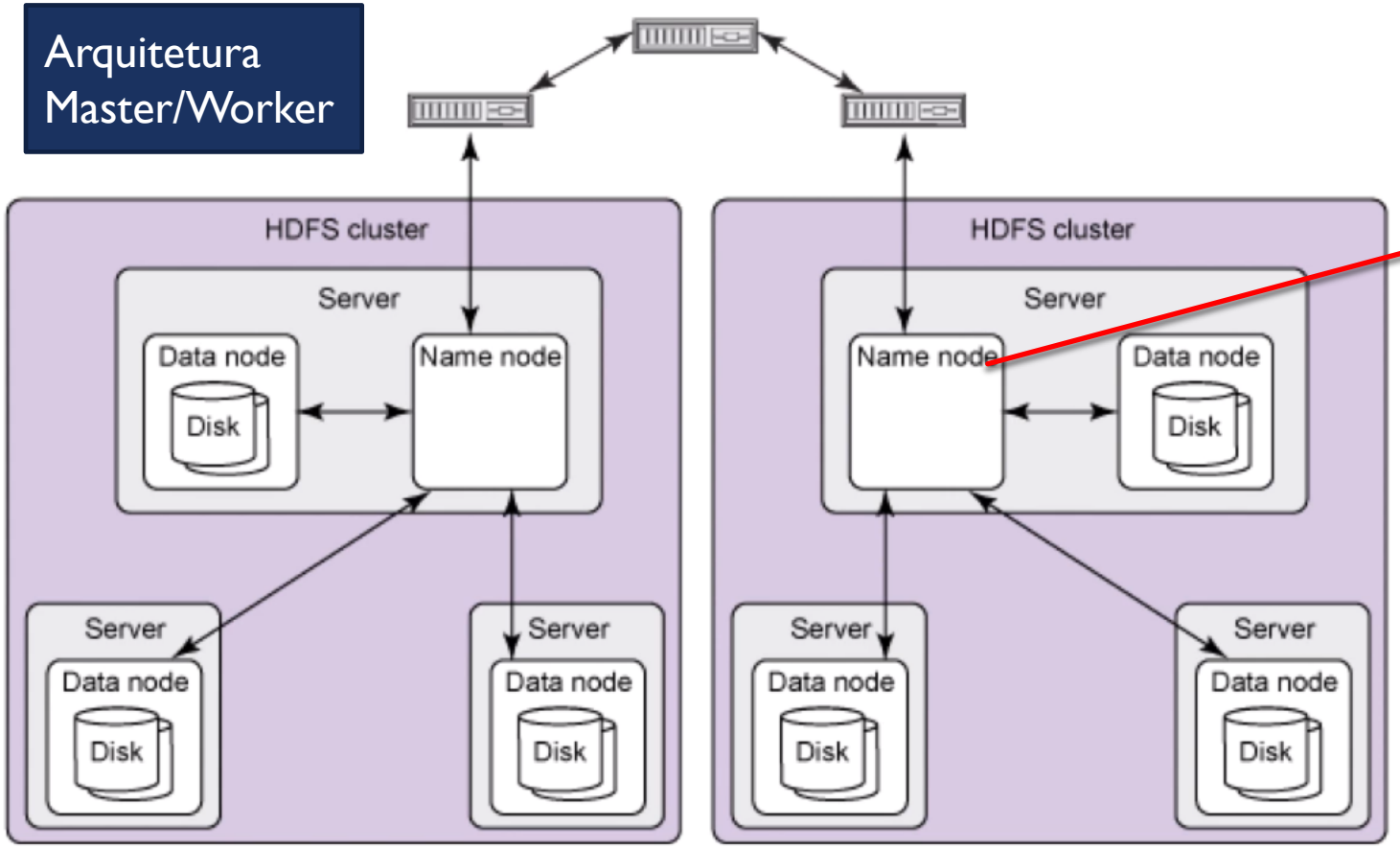
- Tolerância a falhas
- Integridade
- Segurança
- Desempenho
- Consistência





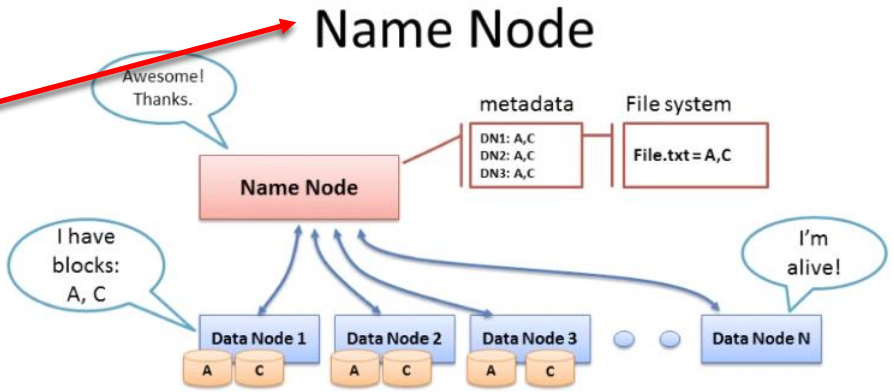
HDFS - ARQUITETURA

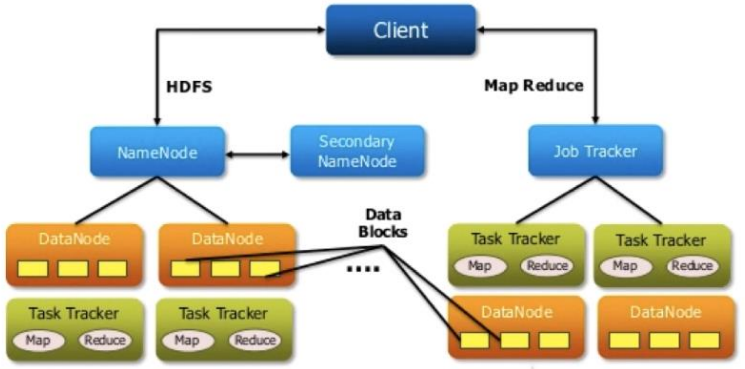
Arquitetura Master/Worker



Estruturas de dados do NameNode:

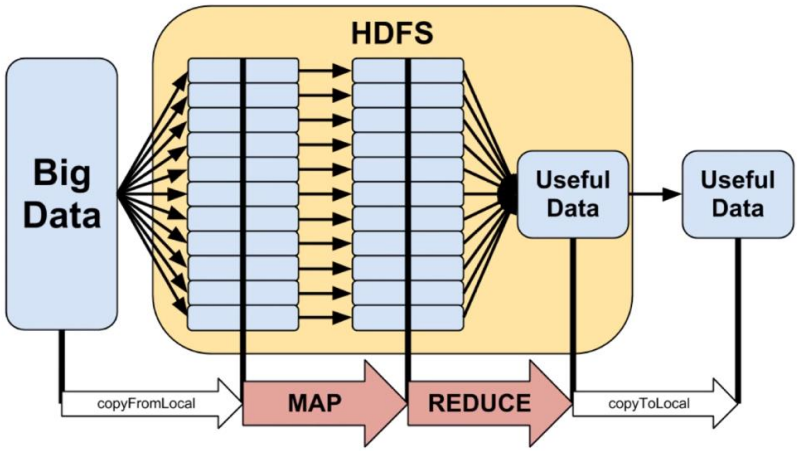
- FsImage
- EditLog



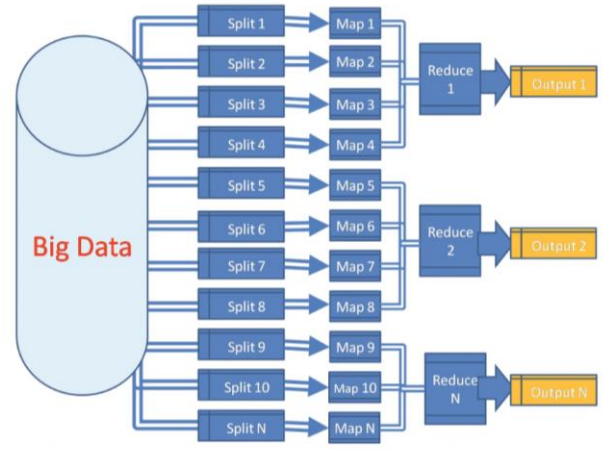


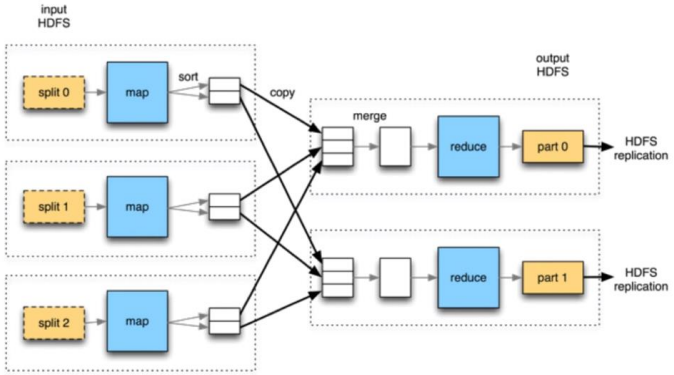
MAPREDUCE

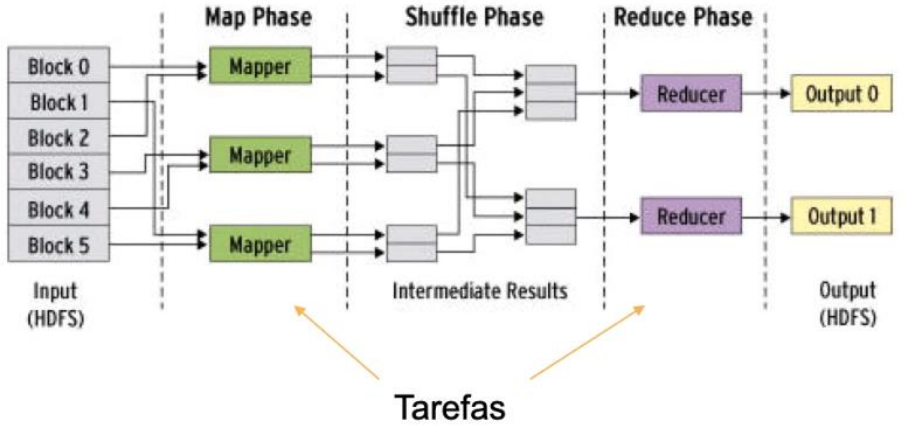
MapReduce



Processamento Paralelo e Distribuído:



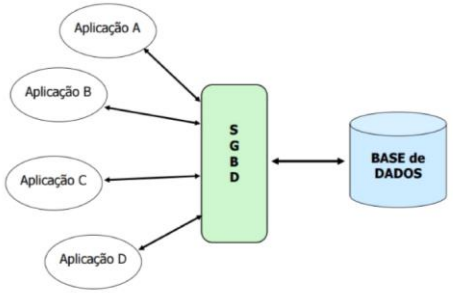


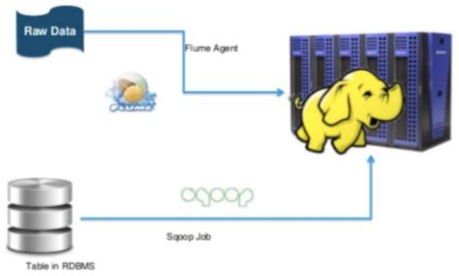


HADOOP X BANCO DE DADOS RELACIONAL

- O Hadoop não é um Banco de Dados, é um framework composto por uma camada de sistema de arquivos distribuído (HDFS) e uma camada de programação em paralelo (MapReduce, além do gerenciador de recursos YARN).

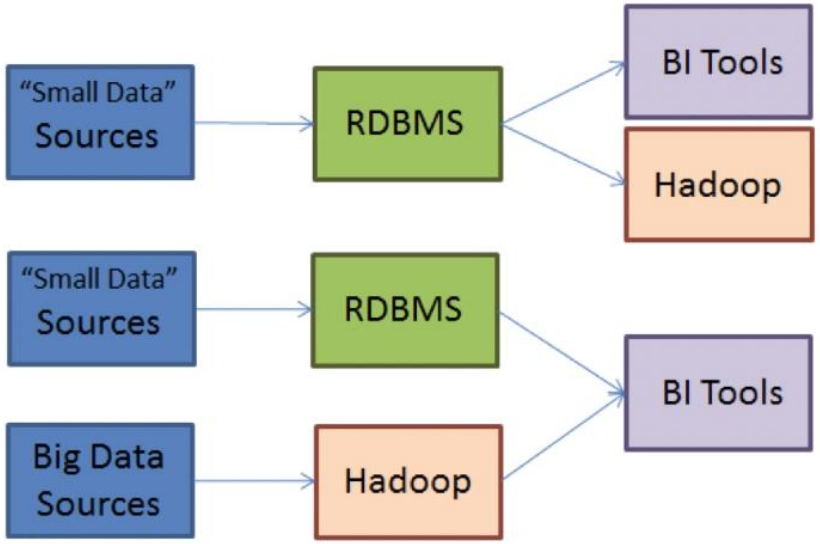
SGBD's: gerenciam um ou mais bancos de dados





RDBMS	Dados Estruturados (Transacionais)
NoSQL	Dados Não Estruturados
Hadoop	Dados Estruturados e Não Estruturados em GRANDE VOLUME

Hadoop processa dados em batch. Não deve ser usado para processor dados transacionais. Mas o Hadoop pode resolver muitos outros tipos de problemas relacionados ao Big Data



INSTALANDO O ECOSSISTEMA HADOOP

- Infraestrutura de TI para Big Data

Parte I

- Criar uma máquina virtual
- Instalar o Sistema operacional Linux
- Instalar utilitários (Java, ssh, ferramentas)
- Instalar o MySQL

Parte II

- Instalar o Hadoop
- Configurar o HDFS e o MapReduce
- Executar um job MapReduce no HDFS
- Instalar e configurar o Zookeeper
- Instalar e configurar Hbase
- Instalar e configurar Hive
- Instalar e configurar o Pig
- Instalar e configurar o Sqoop
- Instalar e configurar o Spark
- Instalar e configurar o Flume

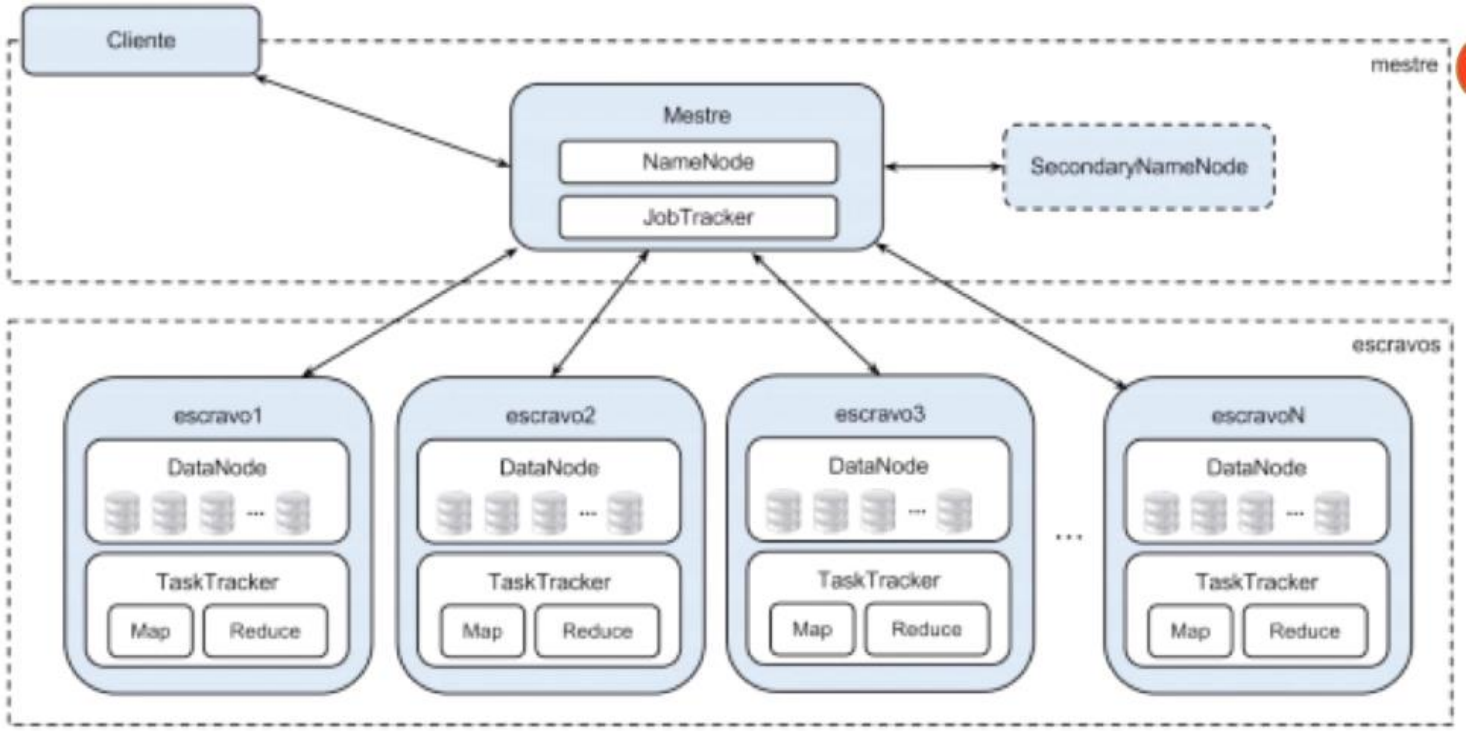
Extra

- Instalar e configurar a máquina virtual Cloudera
- Instalar e configurar a máquina virtual Hortonworks
- Instalar e configurar o Hadoop com Docker

HADOOP COMO SOLUÇÃO PARA ARMAZENAMENTO, GERENCIAMENTO E MANIPULAÇÃO DE BIG DATA

1. Hadoop é open source.
2. Hadoop oferece o framework mais completo para armazenamento e processamento de Big Data.
3. A líder mundial em banco de dados relacionais, a Oracle, oferece soluções de Big Data Analytics com Hadoop.
4. A líder mundial em sistemas operacionais, Microsoft, oferece soluções corporativas em nuvem, com Hadoop.
5. O Hadoop é mantido pela Apache Foundation, mas recebe contribuições de empresas como Google, Yahoo e Facebook.
6. Um Cientista de Dados deve conhecer bem o paradigma de processamento MapReduce.
7. Hadoop é usado por algumas das maiores empresas do mundo.

MODOS DE EXECUÇÃO DO HADOOP



Component	Property	Standalone	Pseudo-distributed	Fully distributed
Core	fs.default.name	file:/// (default)	hdfs://localhost/	hdfs://*namenode*/
HDFS	dfs.replication	N/A	1	3 (default)
MapReduce	mapred.job.tracker	local (default)	localhost:8021	*jobtracker*:8021

Modo Local (Standalone)

Modo Pseudo-Distribuído (Pseudo-Distributed)

Modo Totalmente Distribuído (Fully Distributed)

- Core-site.xml
- Hdfs-site.xml
- Mapred-site.xml

Editar arq's para alternar entre as configs acima.

END