BIG DATA

INFRAESTRUTURA

HADOOP
HBASE
HIVE

Humberto Sandmann
humberto.sandmann@gmail.com

Roteiro do curso

Aula 1: Introdução

Conceitos, mercado, tendências e arquitetura

Aula 2: Ferramentas

Hadoop, MongoDB, Neo4j e Solr

Aula 3: HBase, Hive e tratamento de dados

Aula 4: Machine learning

Perceptron - MLP, DeepLearning, SOM,

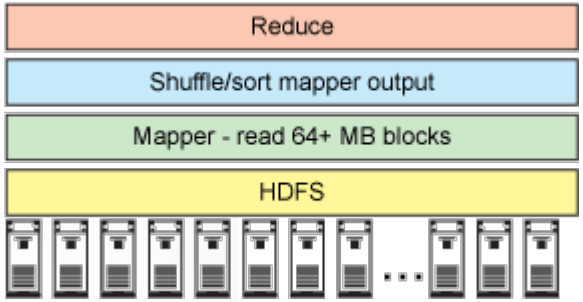
Redes probabilística (Redes Bayesianas)

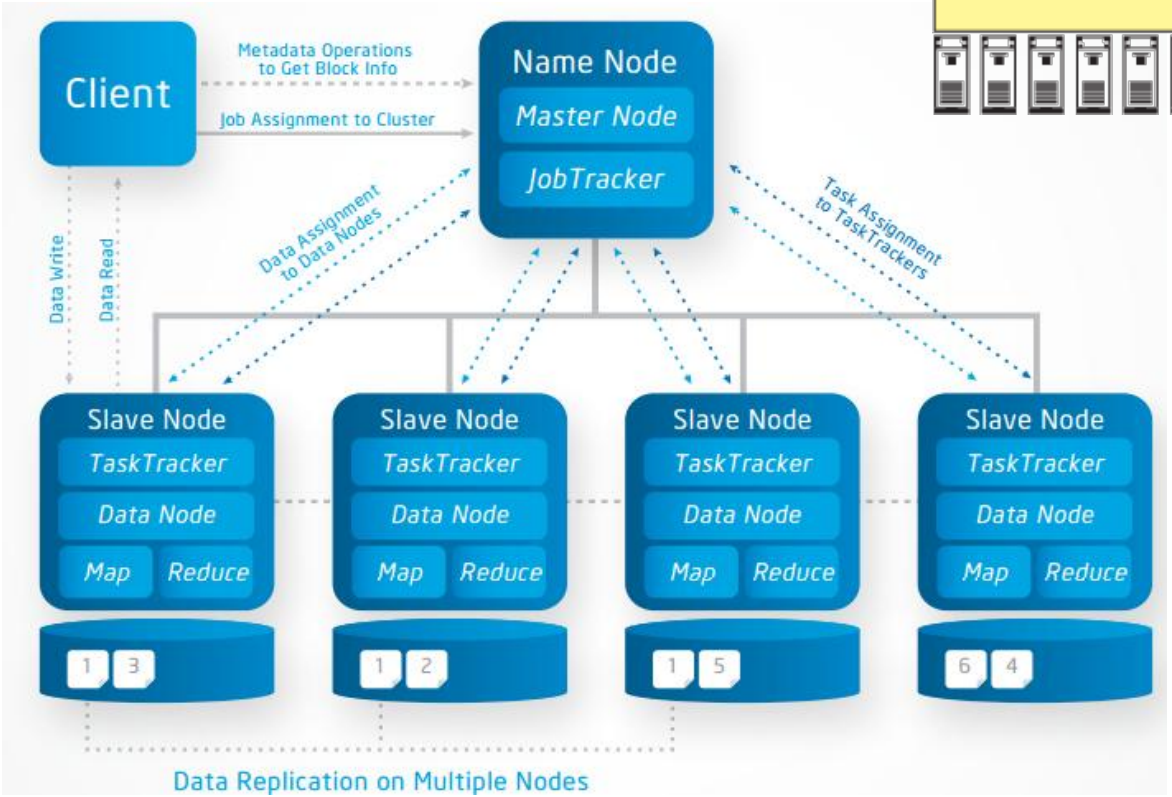
Hadoop

Gerencia e centraliza o acesso a dados espalhados por diversos nós de armazenamento (+4000 por exemplo)

- Implementa o design pattern map/reduce
  - Yarn: ferramenta para gerenciamento do cluster
  - Map/reduce
- Implementa um repositório unificado de nós
  - HDFS: Hadoop File System
- Apenas viável se a performance não for problema
- OpenSource: servidor de referência para outras implementações

Arquitetura





Ferramental Apache

Ambari™: Monitoramento Web da plataforma

Avro™: Serialização de dados

Cassandra™: Banco de dados escalável multi-master e tolerante a falhas

Chukwa™: Coletor de dados do sistema distribuído

HBase™: Banco de dados escalável para grande volume de dados estruturados

Hive™: Infraestrutura de data warehouse para sumarização e consultas

Mahout™: Maquinaria de aprendizado escalável

Pig™: A high-level data-flow language and execution framework for parallel computation

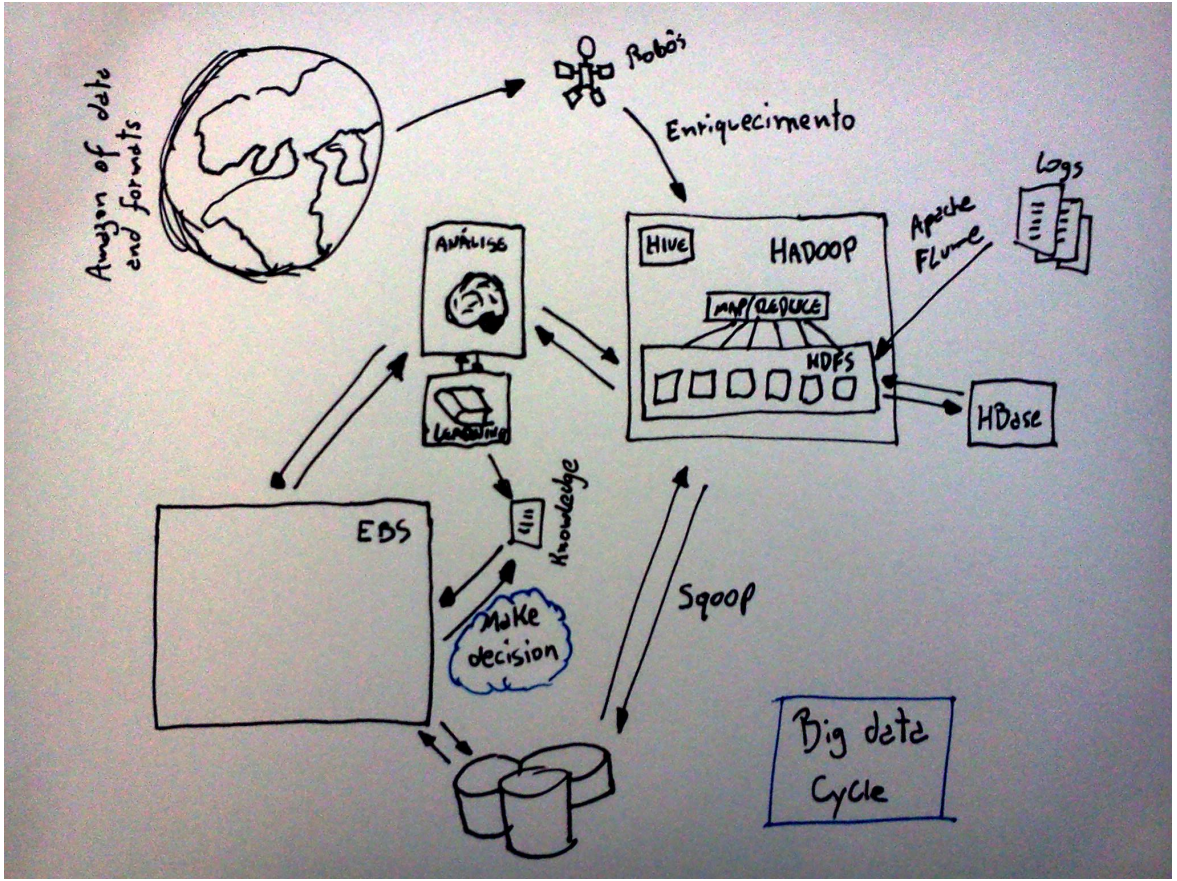
Spark™: Máquina de computação para dados em Hadoop

Ex: ETL, machine learning, stream processing e computação gráfica

ZooKeeper™: Coordenador para distribuição de processos

Infraestrutura

(Hadoop + Hive + HBase)



Enriquecimento de Dados

- Facebook
- Google
- Twitter
- Receita
- Serasa
- TSE
- Personal data (IMEI)
- Localização (GPS, Rede GPRS)
- Voz
- Imagens
- Biometria
- Outros

Dados deixam de ser dados e passam a ser informação. Para isto, é necessário correlacionar os dados, agregando-os a um contexto

Robôs de busca

- Flume Apache Project (tratamento de logs)
- wget (Linux native)

  wget http://hadoop.apache.org

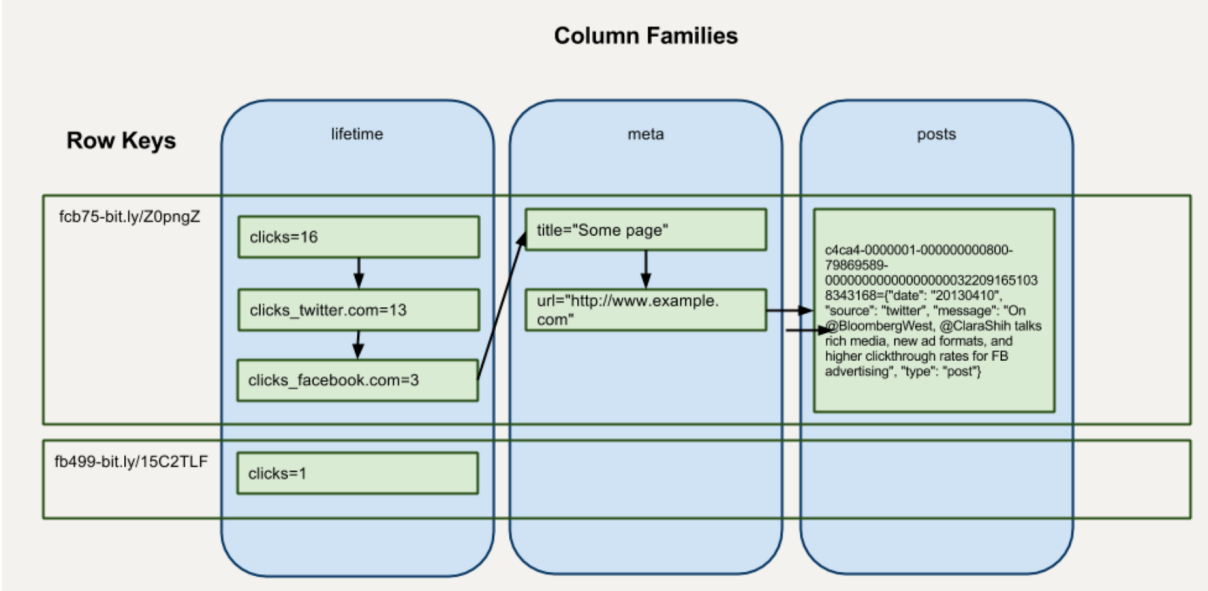
- Robots

  http://www.robotstxt.org/

Todos os dados coletados vão para o HDFS

HBase

- Originário do BigTable (projeto Google)
- Modelagem orientada a Chave-Valor
- Baseado em JSON
- Visualização temporal



Exemplo

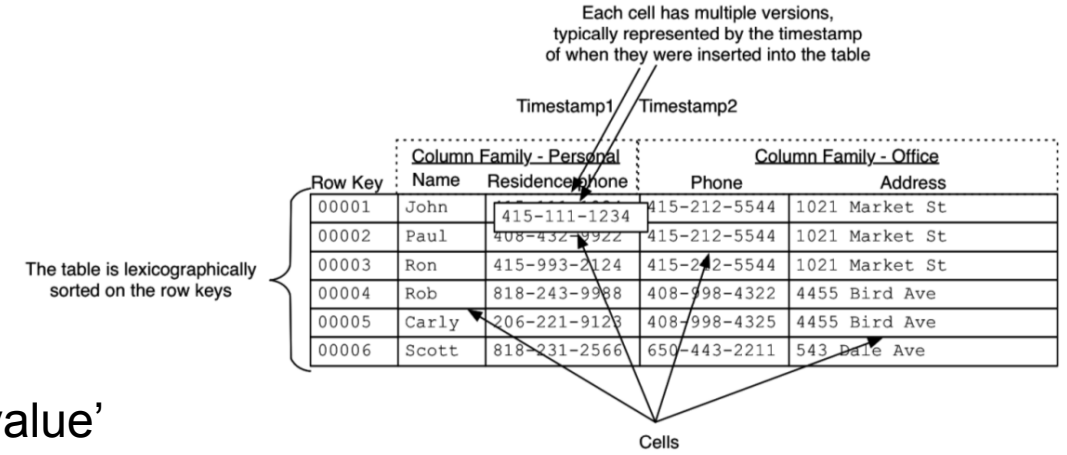
create ‘table’, ‘cf1’, ‘cf2’, ‘cf3’, …

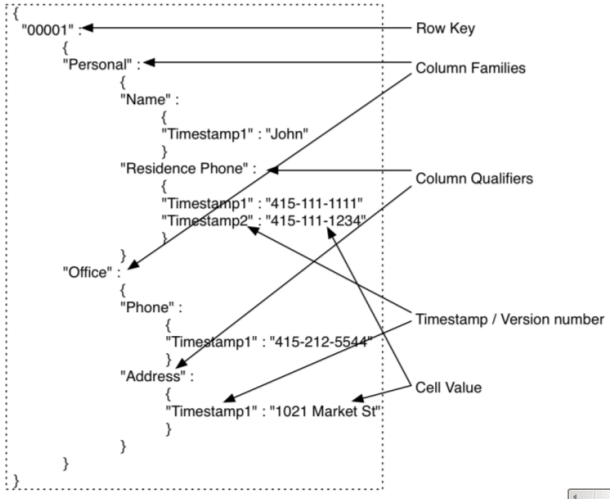
put ‘table’, ‘rowId’, ‘cf:qualifier’, ‘value’

disable ‘table’

enable ‘table’

drop ‘table’





Hive

- Executa consulta a dados formatados no HDFS através de instruções SQL-Like
- Permite joins
- Faz sumarização de dados do HDFS utilizando o Map/Reduce

Exemplo

create table state (cod int, name string, sig string, country int) row format delimited fields terminated by ‘\t’ lines terminated by ‘\n’;

describe state;

describe extended state;

select * from state;

select count(*) from state;

load data local inpath ‘/home/bigdata/state.csv’ into table state;