
Semantix
Introdução ao Big Data - Ecossistema Hadoop

Ecossistema Hadoop

Semantix

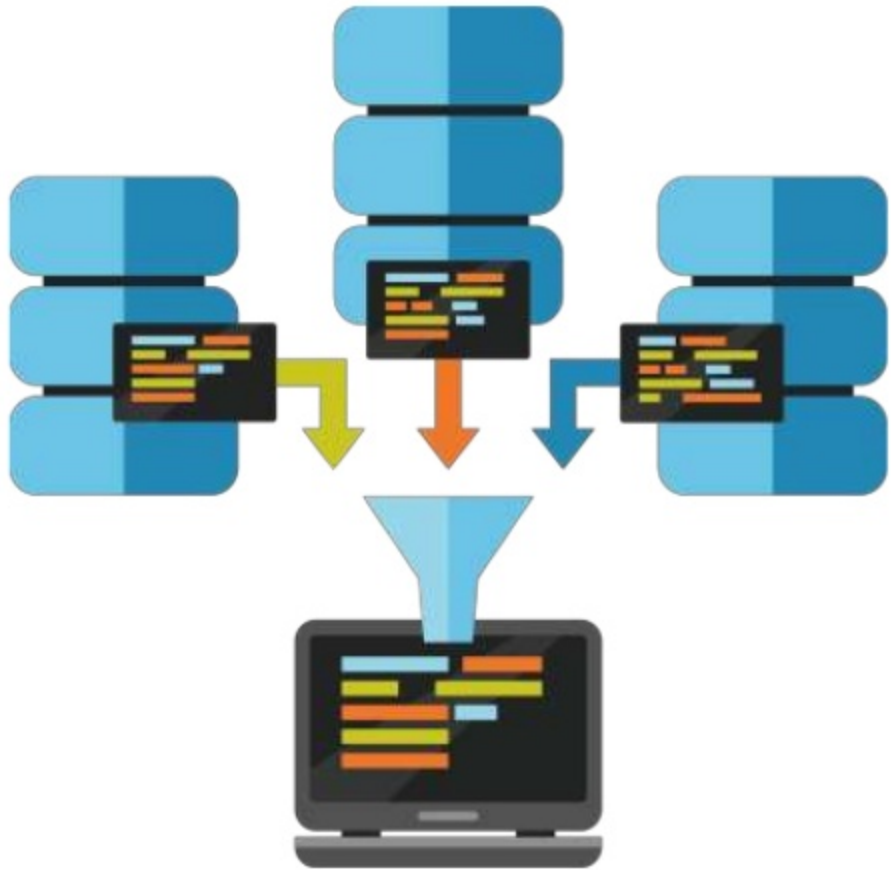
Agenda

Introdução ao Big Data.

- Hadoop HDFS.
- Hadoop Map Reduce.
- Spark.
- Hive.
- Sqoop.
- Flume.
- Kafka.
- Hbase.
- Zookeeper.

O que é Big Data ?

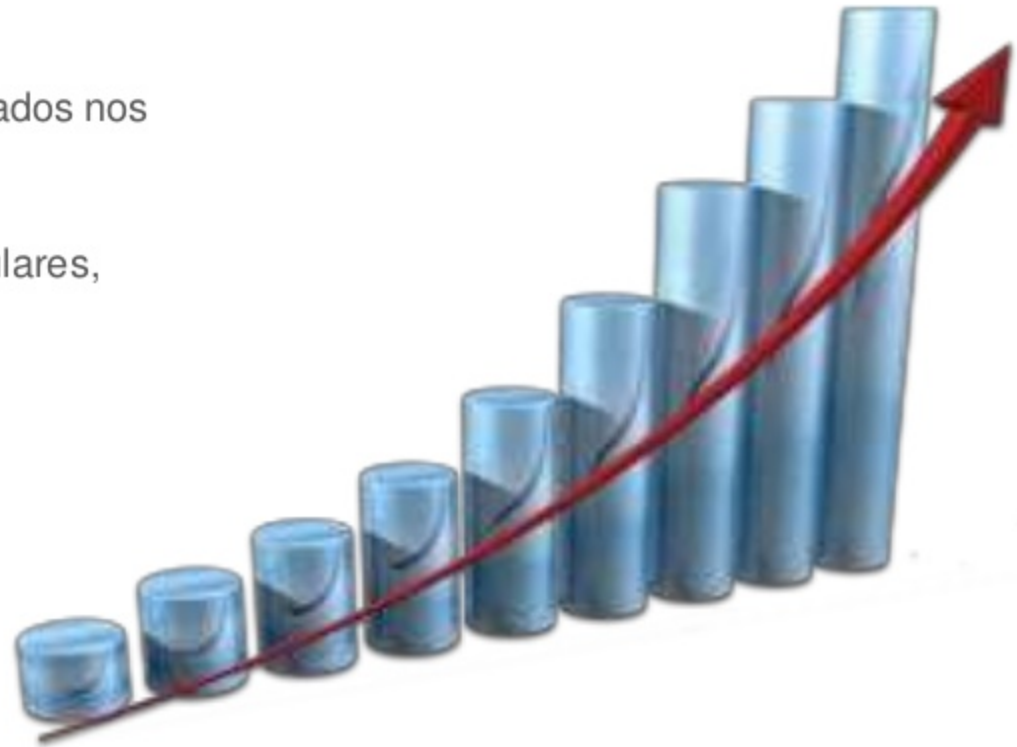
A ideia de Big Data está diretamente relacionada ao grande volume de dados gerados diariamente, não sendo possível armazenar e processar pelos métodos tradicionais.



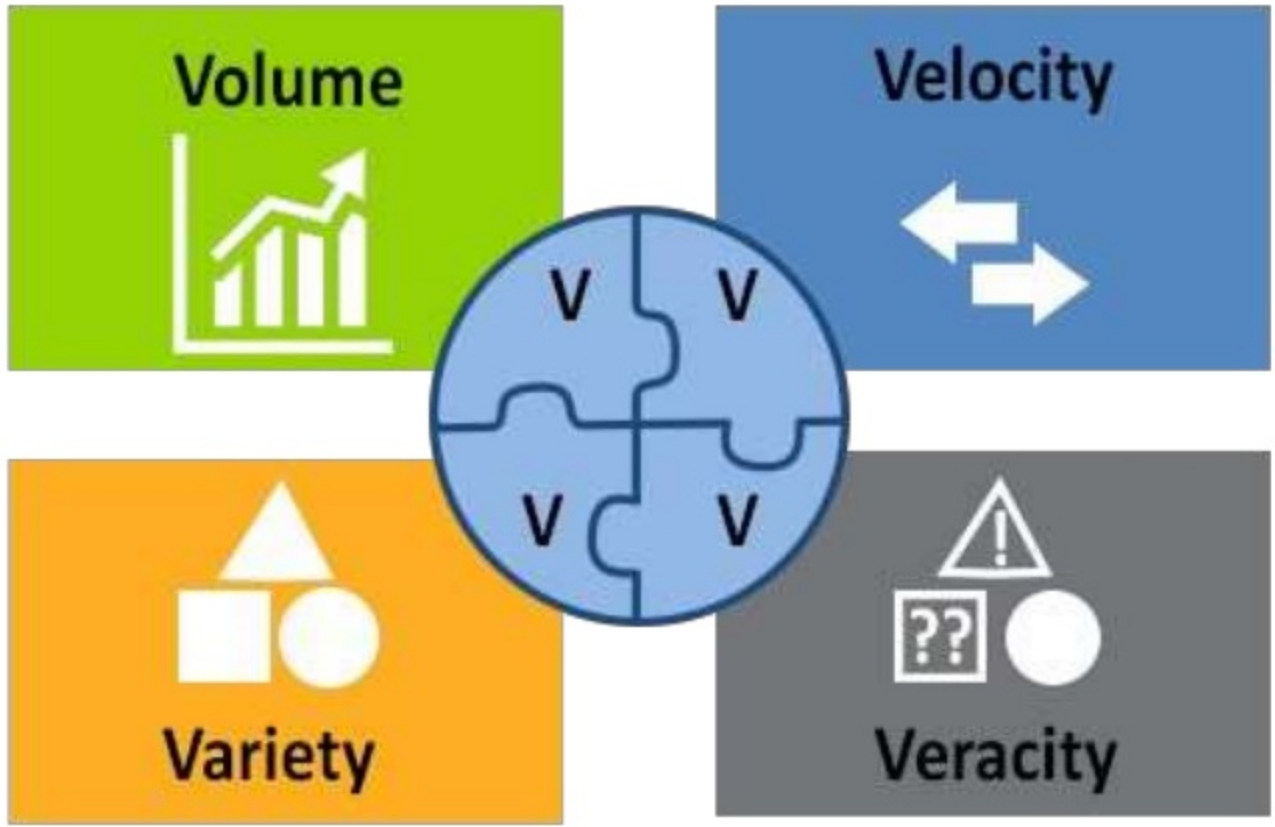
Crescimento dos Dados

Estima-se que 90% dos dados criados foram gerados nos últimos 2 anos.

Esses dados são gerados por: redes sociais, celulares, sensores, e-commerce e etc.



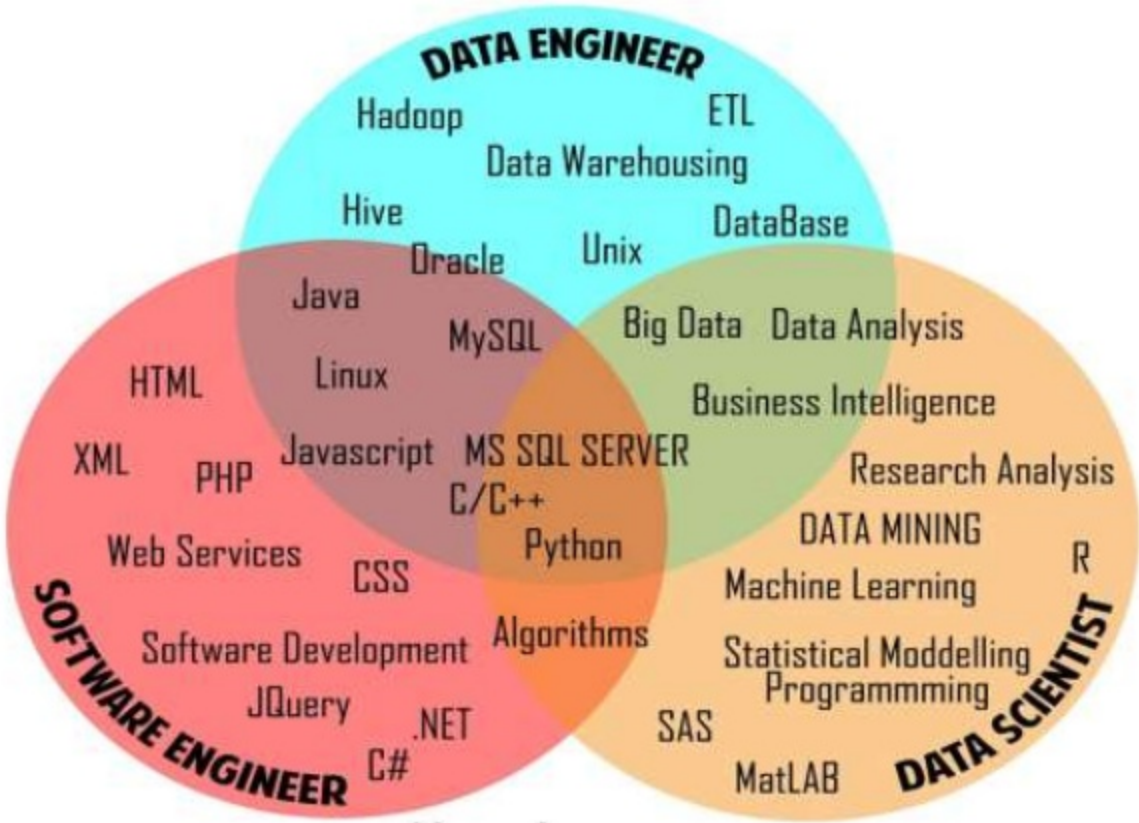
Quando devo utilizar Big Data ?



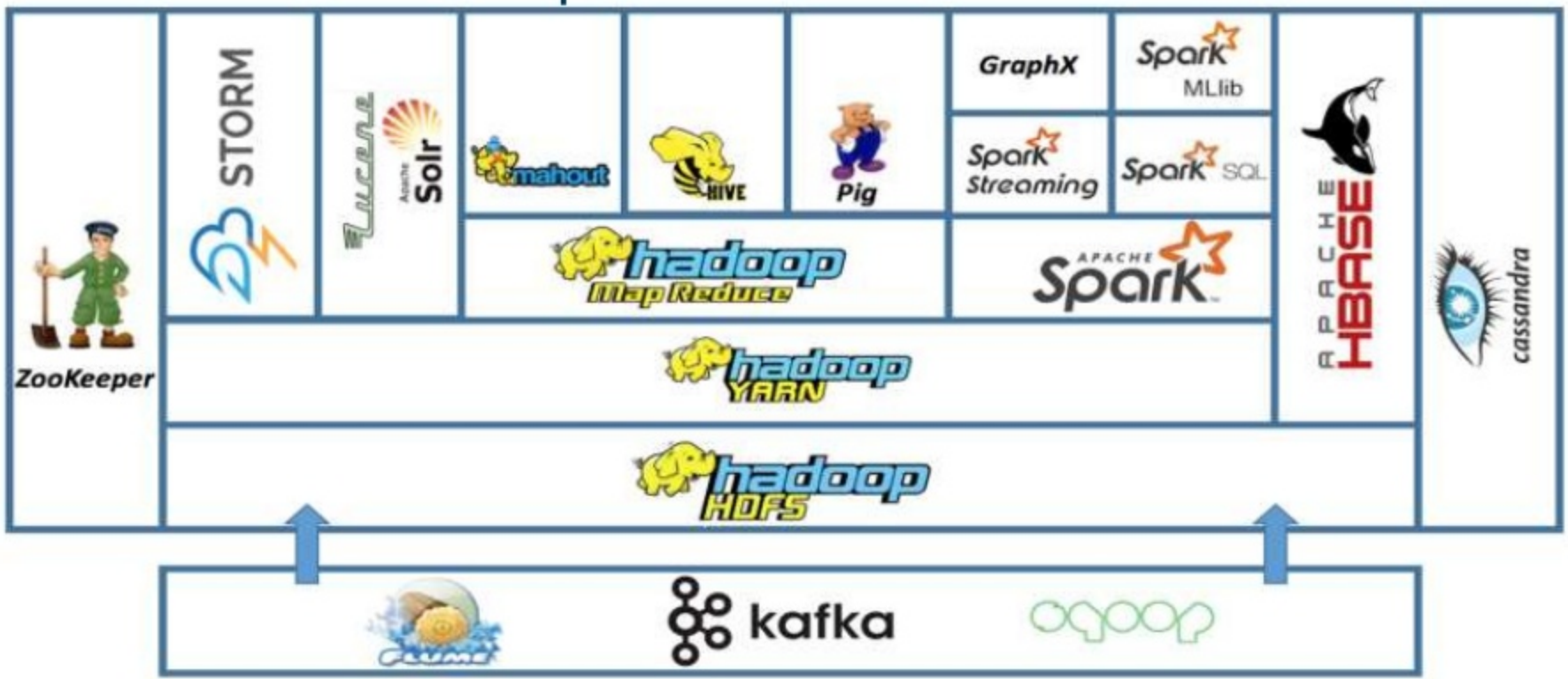
Profissões

- Arquiteto de Soluções Big Data (Arquitetura)
- Engenheiro de Big Data (Infra)
- Engenheiro de Dados (Data Lake / ETL)
- Cientista de Dados (Machine Learning)
- Estatístico / Matemático (Modelos)

Skills – Engenheiro - Cientista



Ecossistema Hadoop



Hadoop HDFS (Distributed File System)

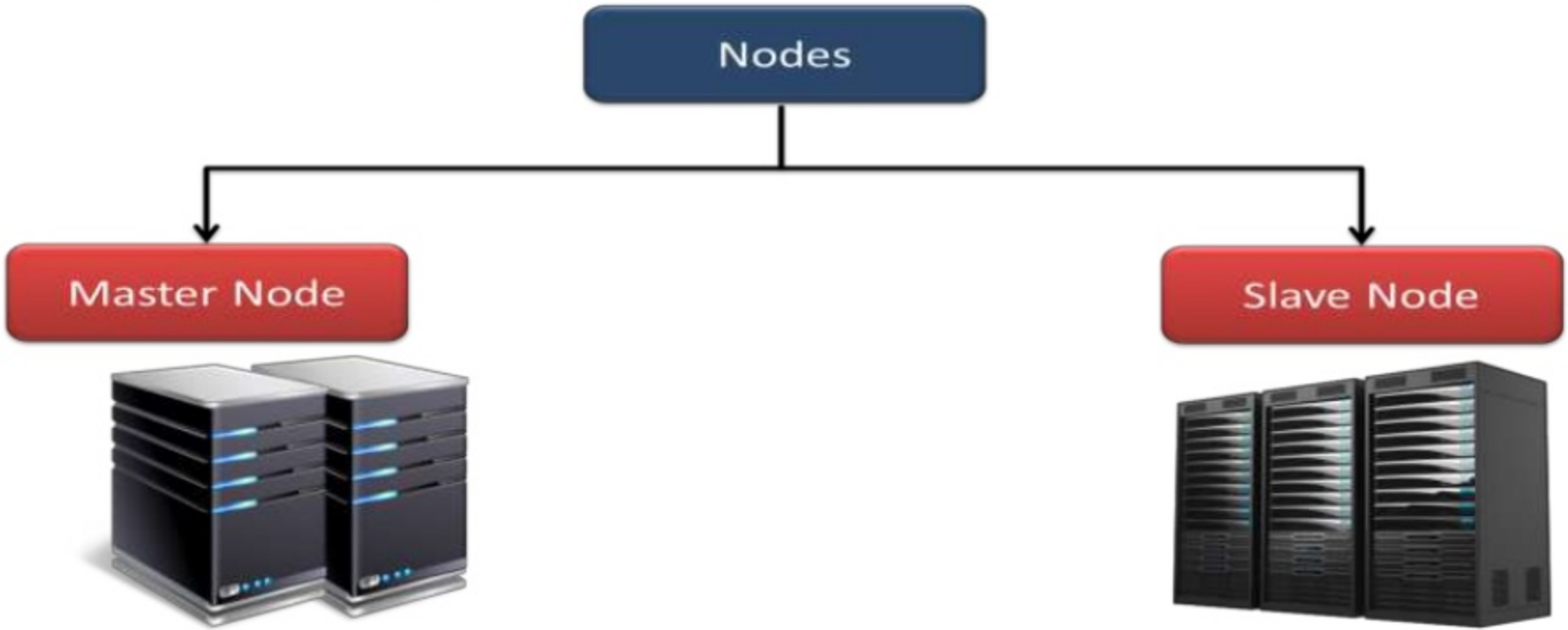
O que é o Hadoop HDFS?

Nodes, NameNodes, DataNodes?

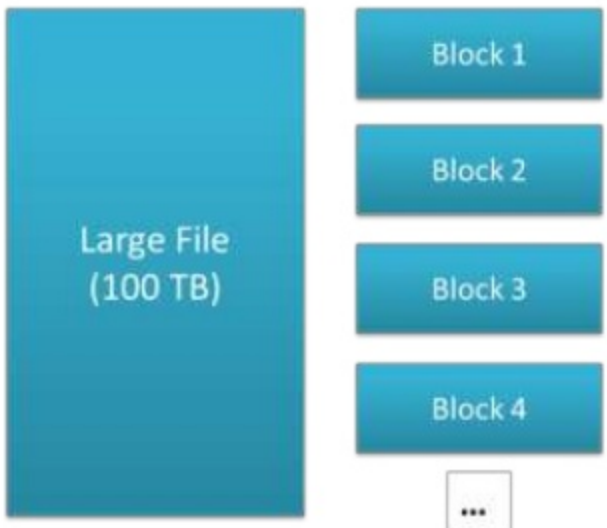
Data Storage?

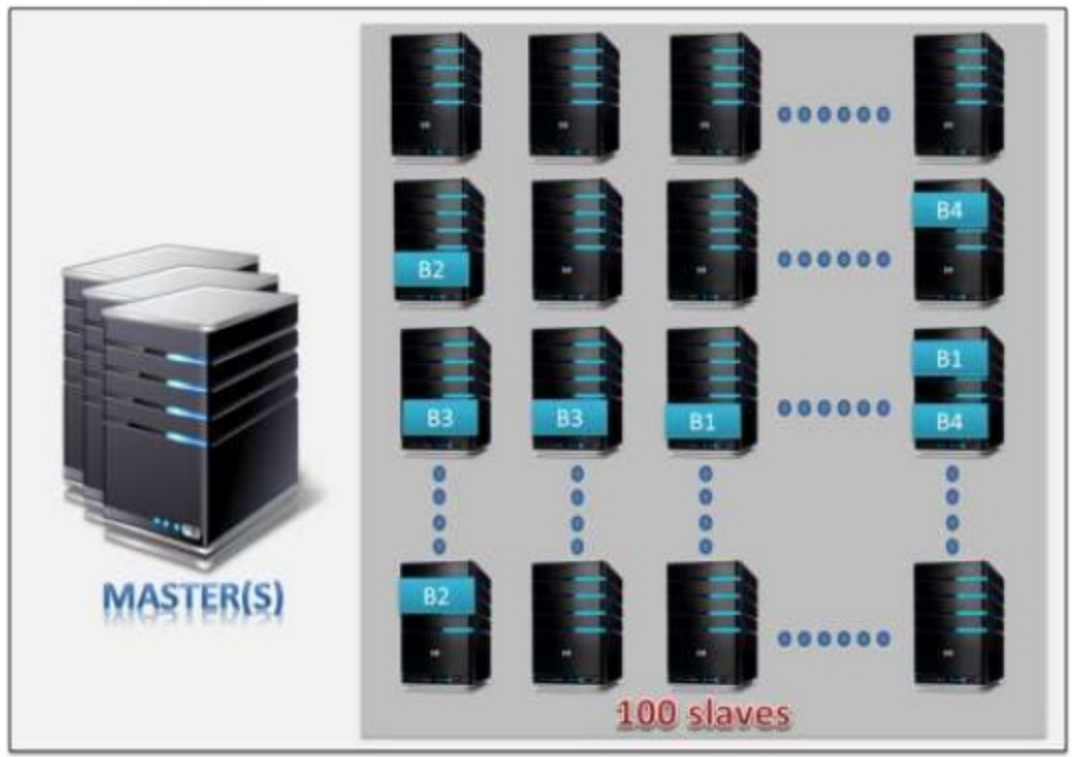
Arquitetura, Bloco, Replicação?

Hadoop HDFS (Nodes)

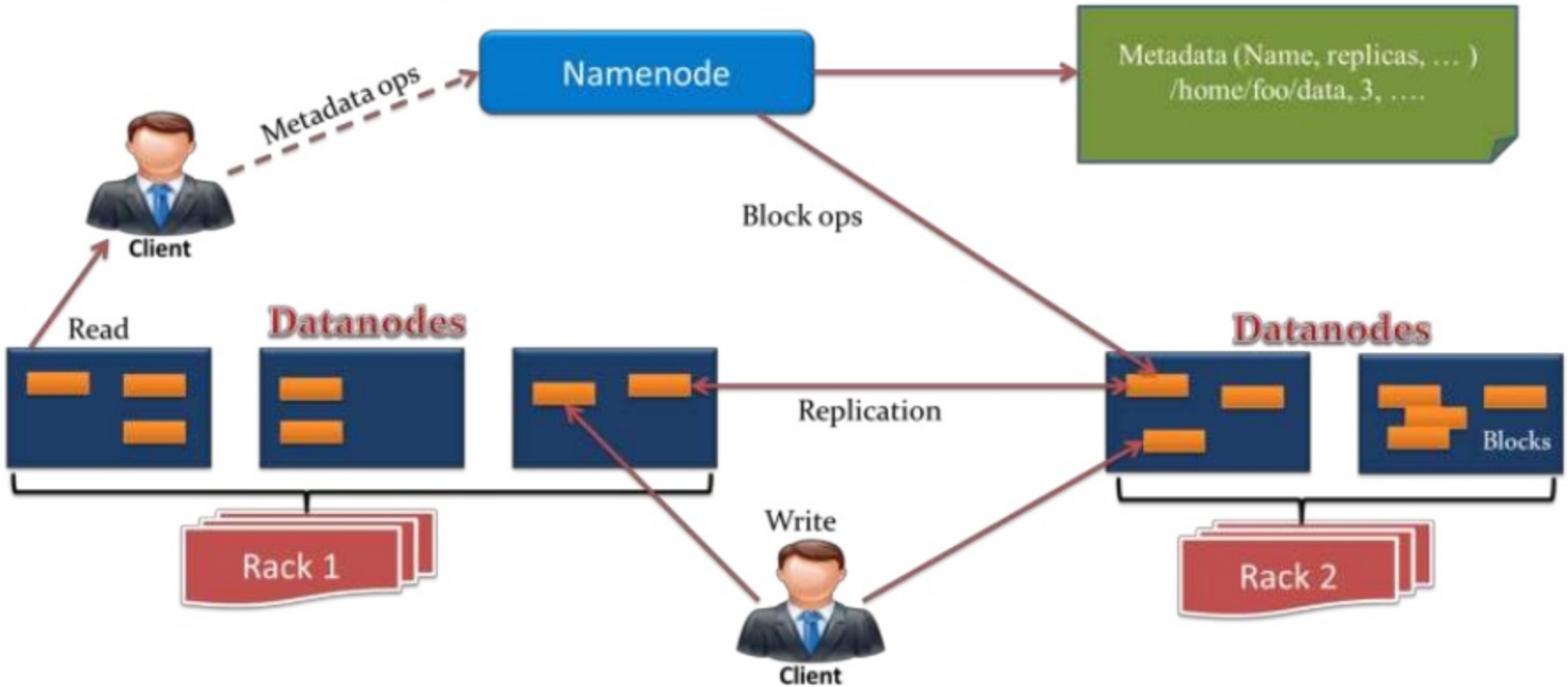


Hadoop HDFS (Data Storage)





Hadoop HDFS (Arquitetura)



Hadoop MapReduce

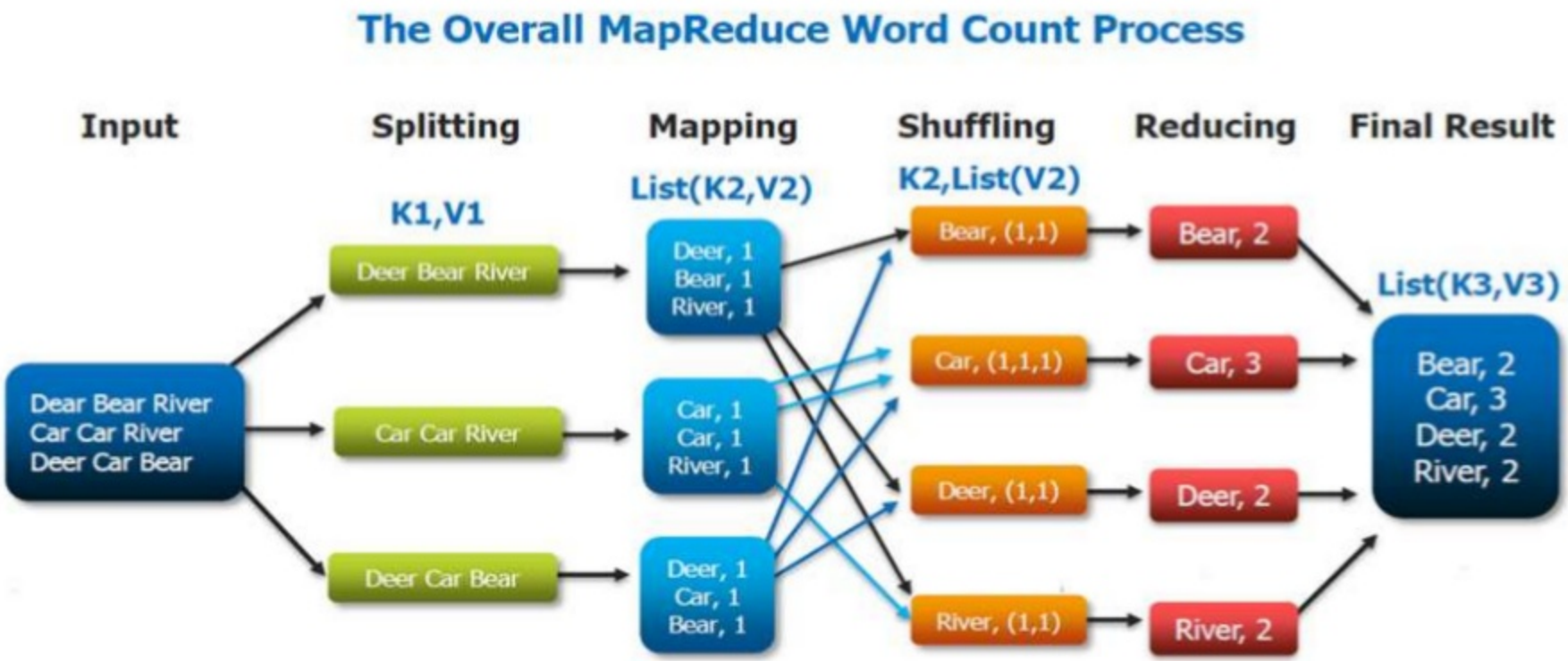
O que é o MapReduce?

Como ele trabalha?

Estrutura interna?

Etapas de um Job MapReduce?

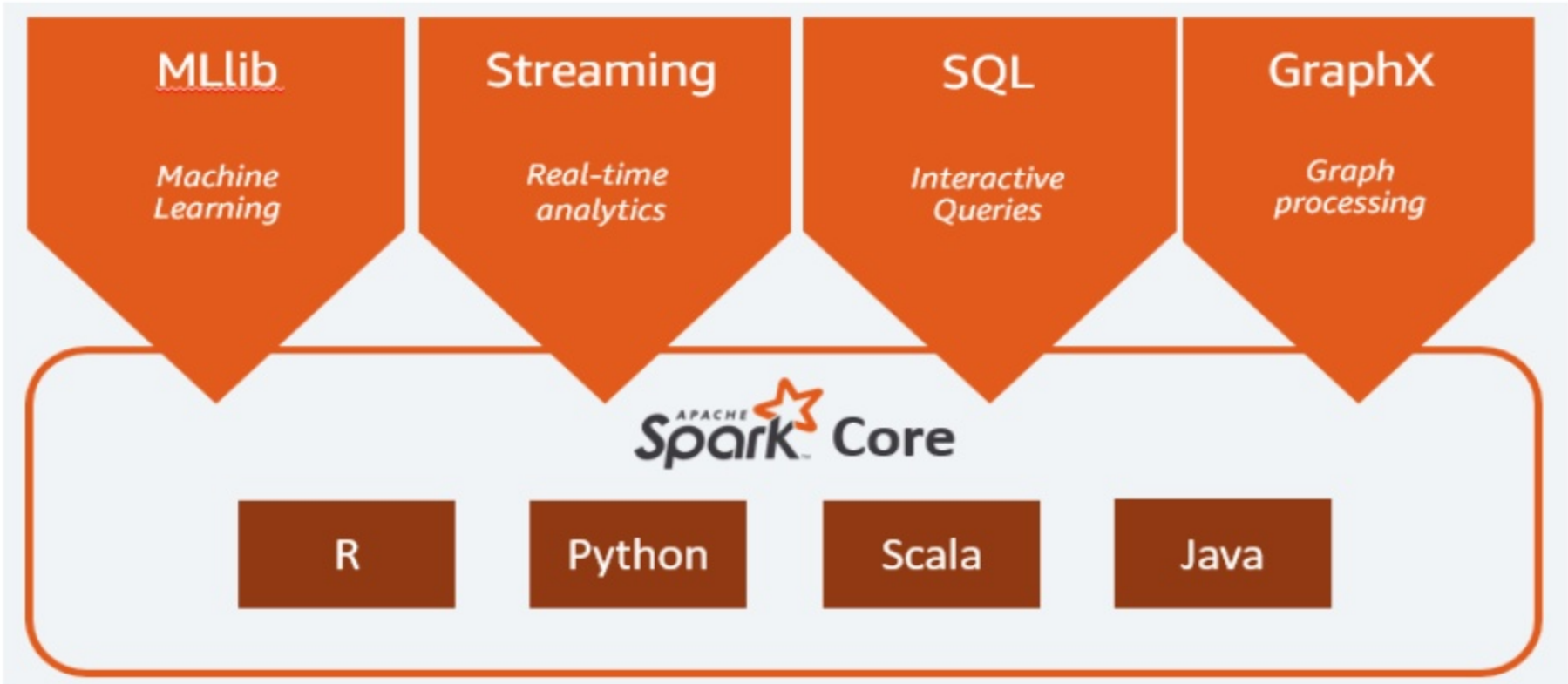
Hadoop MapReduce - WorkFlow



Spark

- Processamento de dados alta escala.
- Execução rápida.
- Pode ser usado com Java, Scala, Python, R.
- Pode ser usado para processamento Batch ou Streaming.

Spark Componentes



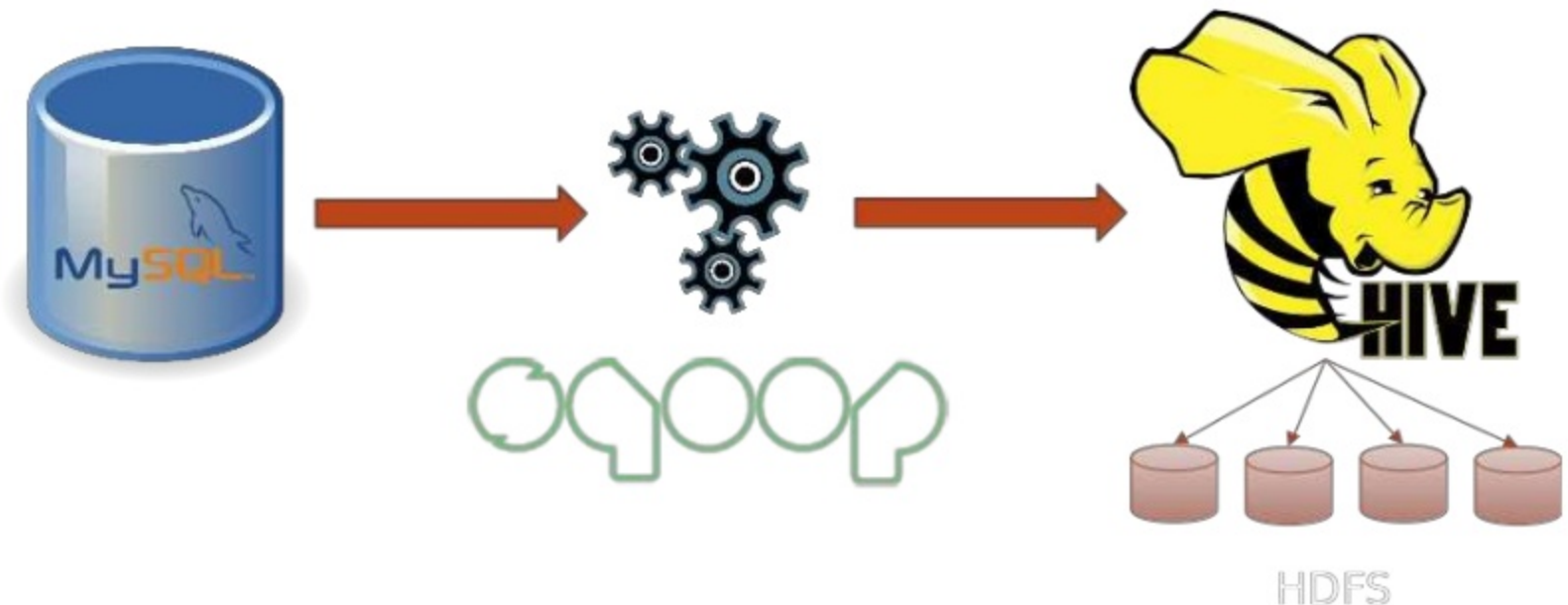
Hive

- Surgiu nos laboratórios do Facebook.
- Permite conexões ODBC/JDBC.
- Operações Batch.
- Utiliza a linguagem HQL.
- Converte SQL para MapReduce.
- Suporta vários tipos de arquivos: Avro, Parquet, ORC, TXT.
- Suporta conversões: snapy e gzip.

Apache Sqoop - WorkFlow

O Apache Sqoop é uma ferramenta para transferir dados de um RDBMS para Hadoop.

SQL-to-Hadoop.



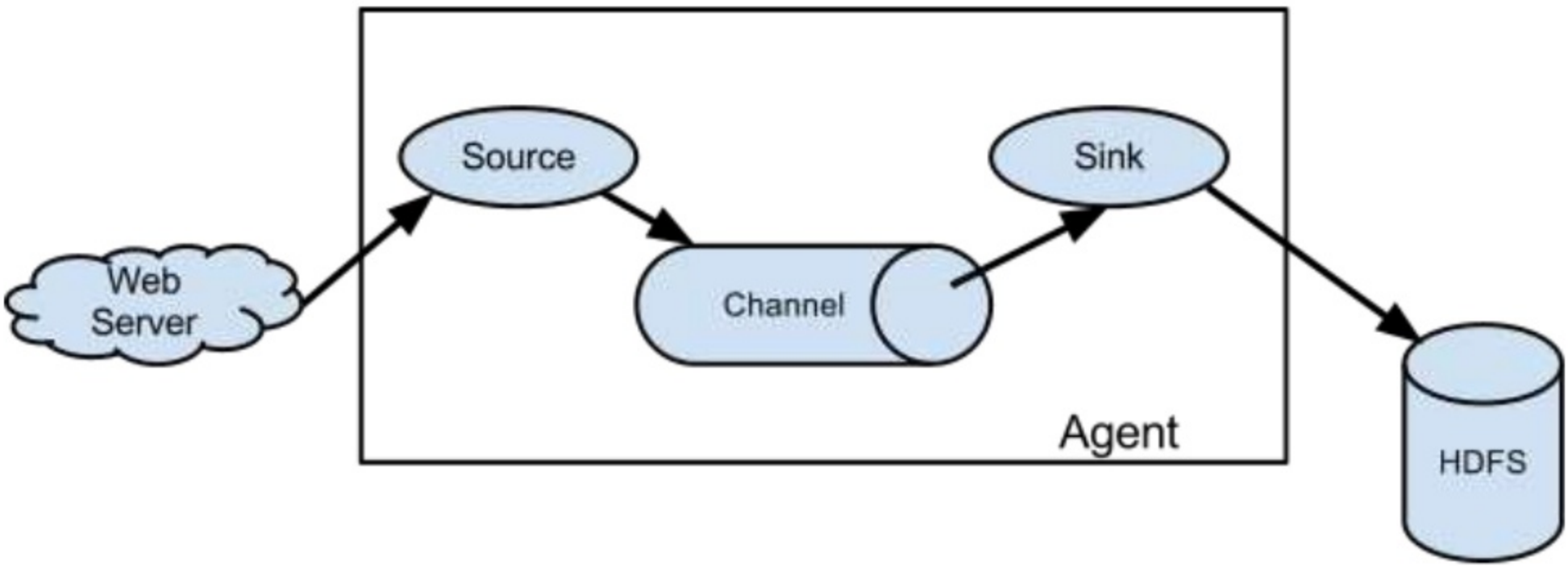
Apache Flume

O Apache Flume é um mecanismo de ingestão de dados para coletar, agregar e transportar um grande volume de dados e armazenar em um storage centralizado.

O objetivo principal do Flume é ingerir dados de eventos no HDFS de forma simples e automatizada. Porém, seu uso não se limita apenas ao HDFS, é possível enviar também dados para um arquivo ou banco de dados, entre outros.

Sink com suporte nativo ao HDFS e Hbase.

Apache Flume - Componentes



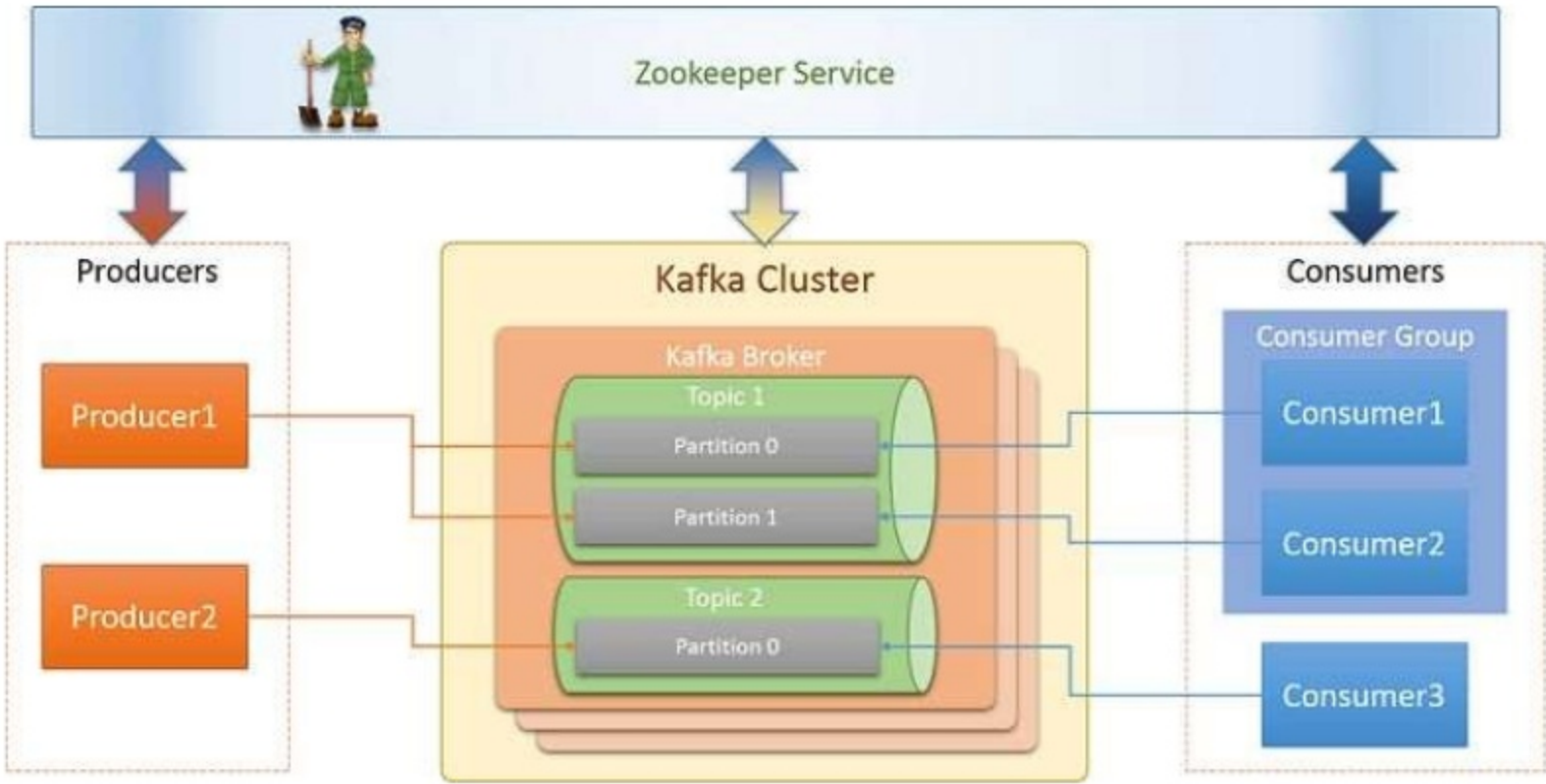
Apache Kafka

Sistema de mensagens de uso geral baseado em uma arquitetura de
Publish e Subscribe.

Inicialmente desenvolvimento pelo LinkedIn, quando surgiu a necessidade de
Processamento massivo de dados.

Em 2012 foi absorvido pela Apache e se tornou open-source.

Apache Kafka - Componentes



Apache Hbase

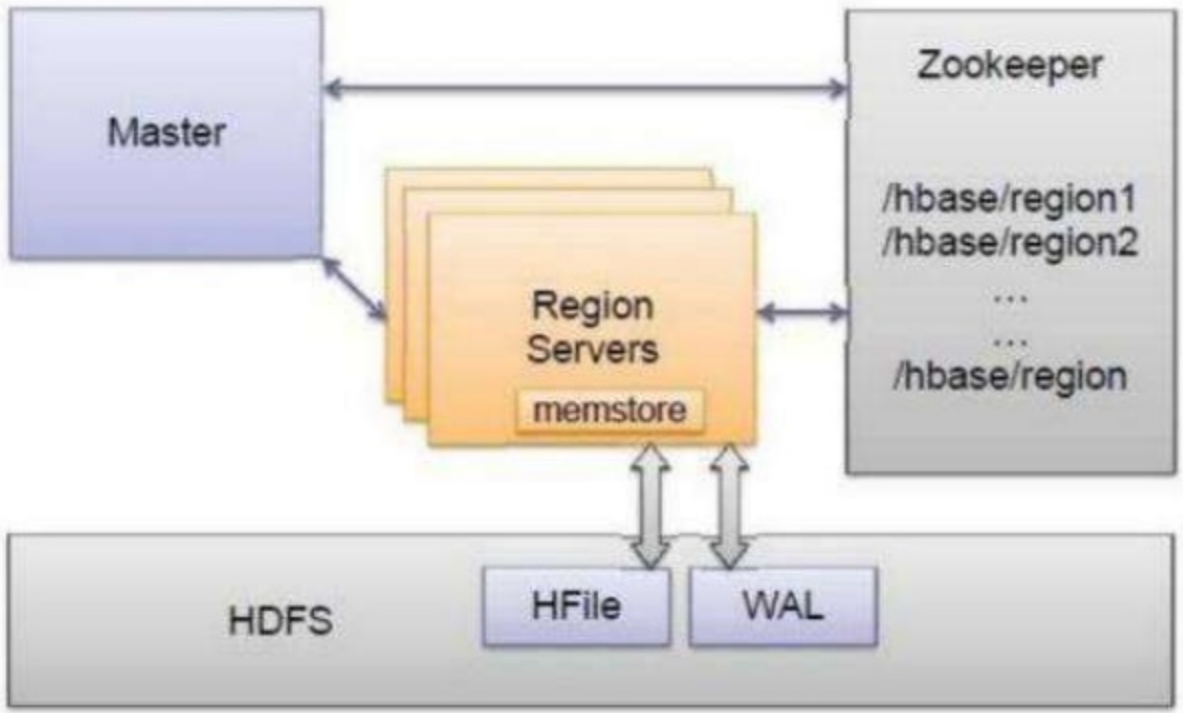
O Hbase é um banco não relacional (No-Sql), roda em cima do HDFS.
Sua escalabilidade é Horizontal.
Possui baixa latência para leitura / escrita com grande volume de dados, orientado a colunas e consegue suportar tabelas com bilhões de linhas.

Informações de topologia de cluster altamente disponíveis através de implantações de produção com múltiplas instâncias Hmaster e Zookeeper.

Não possui uma linguagem de consulta (Query language) apenas uma api própria para operações CRUD

Apache Hbase - Arquitetura

Os Dados ficam armazenados nos Region Server.
O Zookeeper gerencia todos os Region Servers.



Apache Zookeeper – Funcionalidades

Serviço centralizado para manter informações de configuração

Nomeação de servidores

Provimento de serviços de grupos

Sincronização distribuída

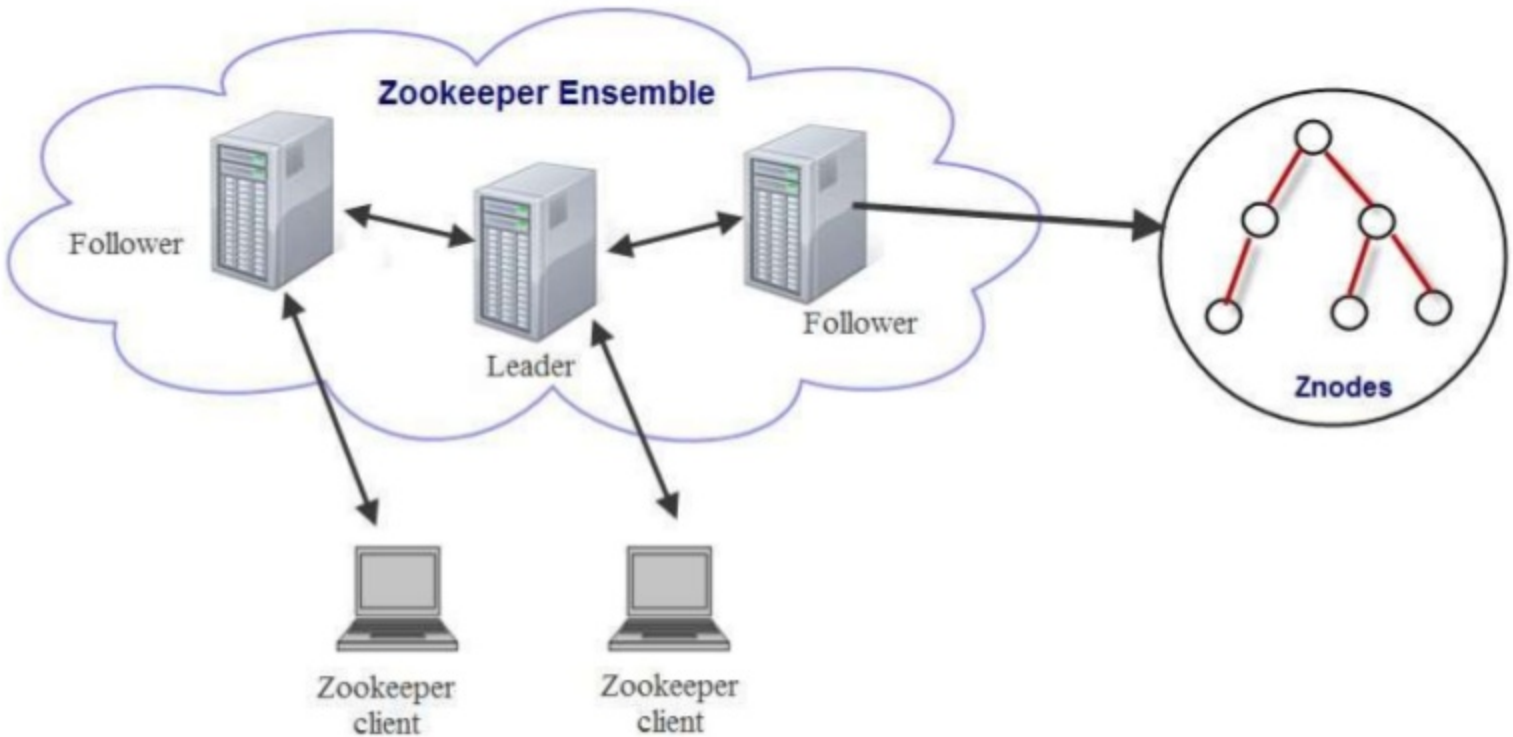
Garante o HA (High Availability) do cluster

Tolerante a falhas

APACHE ZooKeeper™

Apache Zookeeper – Sistema de Eleições

O Nó com menor número de Znode se torna líder



Obrigado

Ederson Corbari


ederson.corbari@semantix.com.br


In/ecorbari

Moisés Pereira


moises.mendes@semantix.com.br


In/moisespereira