Big Data

Prof. Álisson R. Arantes

PUC Minas Virtual

Big Data

- Atual crescimento na geração de dados:
  - redes sociais
  - sensores de IoT
  - dados de aplicações específicas
  - dados de satélites
  - dados geográficos
  - plataformas de streaming

PUC Minas Virtual

Big Data

- Volumes massivos de dados que excedem a capacidade de tecnologias típicas como SGBDs relacionais

- Volume de dados na casa dos terabytes, petabytes, ...

Big Data

- 3 Vs caracterizam o Big Data:
- Volume: refere-se ao enorme volume dos dados gerados
- Velocidade: refere-se à velocidade na criação dos dados, velocidade na qual são gerados e, consequentemente, velocidade na qual devem ser processados

Big Data

- 3 Vs que caracterizam o Big Data:

- Variedade: refere-se à variedade dos tipos de fontes de dados (dados estruturados, semiestruturados e não estruturados)

Big Data

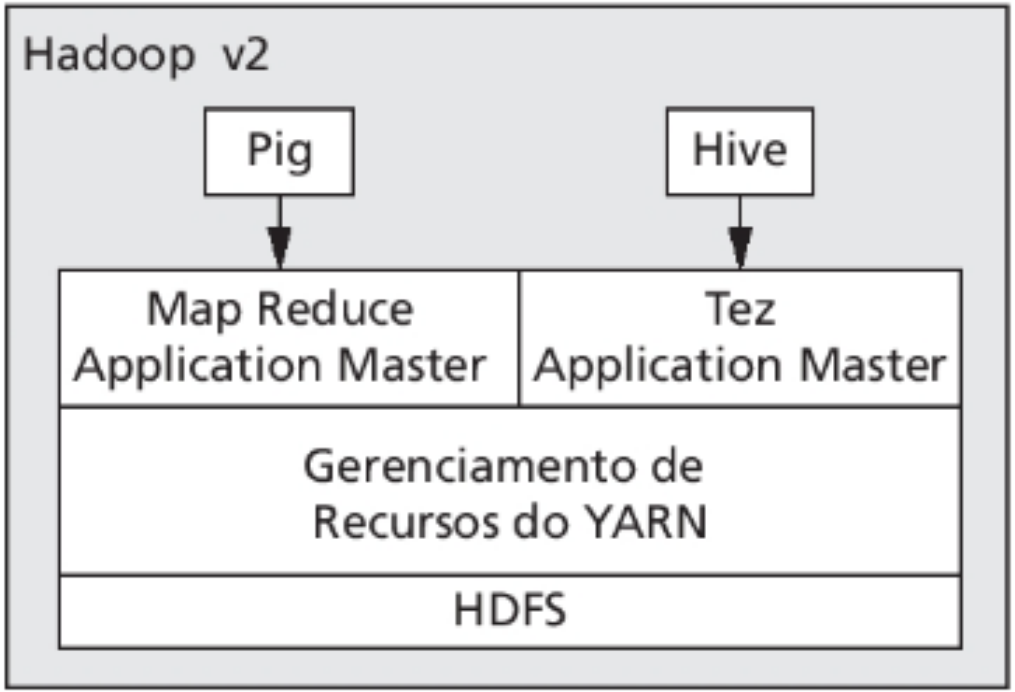
- Mais 2 Vs:
- Veracidade: refere-se à credibilidade das fontes de dados
- Valor: refere-se ao valor agregado pelos dados para a organização

PUC Minas Virtual

Hadoop

- Framework composto pelo HDFS (Hadoop Distributed File System) e pelo modelo de programação MapReduce

- Ecossistema de aplicações para Big Data



Hadoop (Yarn)

Fonte: Elmasri e Navathe (2019)

PUC Minas Virtual

Hadoop Distributed File System (HDFS)

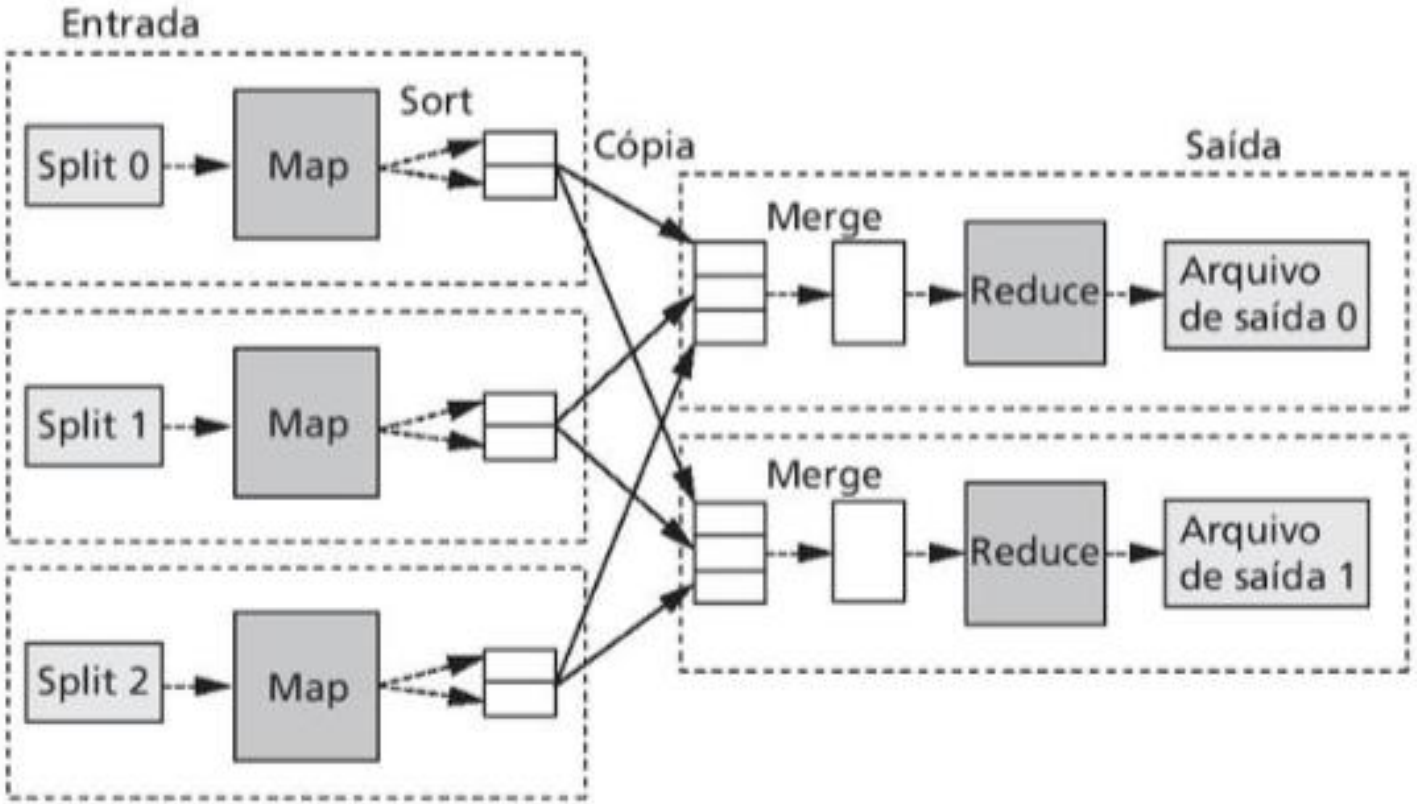
- Sistema de arquivos componente do Hadoop
- Projetado para executar em clusters de hardware barato e abundante
- Provê acesso rápido a grandes volumes de dados
- Dados replicados: confiabilidade e disponibilidade

PUC Minas Virtual

MapReduce

- Modelo de programação caracterizado por duas etapas principais:
  - Map
    - Divide o problema para ser paralelizado em diversas máquinas
  - Reduce
    - Reúne as soluções parciais

PUC Minas Virtual



Visão geral da execução do MapReduce.
Fonte: Elmasri e Navathe (2019)

PUC Minas Virtual

Ecossistema Hadoop

- Projetos relacionados com funcionalidades adicionais
- Apache Pig: interface de alto nível para interação com o Hadoop (consultas similares ao SQL), abstraindo a complexidade do MapReduce.
- Apache Hive: solução de data warehouse sobre o Hadoop (provê sumarização e consultas ad hoc)

PUC Minas Virtual

Spark

- Processa grandes volumes de dados de forma paralela e distribuída

- Componentes para diferentes tipos de processamento, disponibilizados sobre o núcleo

Spark

- Spark Streamming: processamento de dados em tempo real
- GraphX: processamento de grafos
- SparkSQL: SQL para a realização de consultas
- MLlib: biblioteca de aprendizado de máquina

PUC Minas Virtual

PUC Minas
Virtual