Aula 2

Big Data

Prof. Luis Henrique Alves Lourenço

1

Conversa Inicial

2

Hadoop

- Ecossistema Hadoop
- HDFS
- MapReduce
- Pig
- Gerenciamento de *cluster*

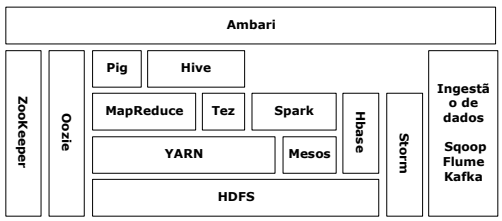
3

O ecossistema Hadoop

4

- "Hadoop é uma plataforma de *software open source* para o armazenamento e processamento distribuído de conjuntos de dados muito grandes em *clusters* de computadores" (Hortonworks)

5



6

Hadoop Distributed File System

7

HDFS

- Sistema distribuído para *cluster*
- Tolerante a falhas (partição)
- Alta disponibilidade
- Escalabilidade horizontal
- Acesso paralelo aos dados

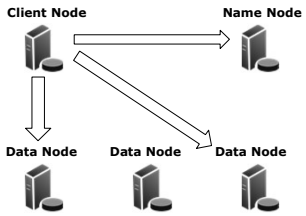
8

Arquitetura HDFS

- Os dados são distribuídos em blocos de 128 MB
- Os blocos são replicados nos nós do *cluster*
- Name Nodes
- Data Nodes

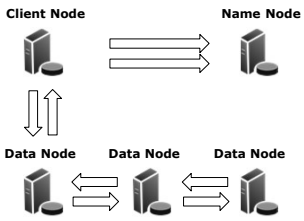
9

HDFS – leitura de arquivos



10

HDFS – escrita de arquivos



11

MapReduce

12

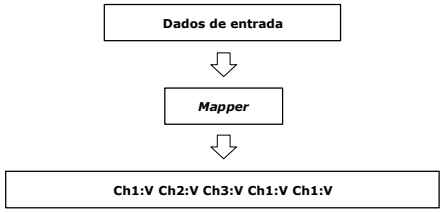
MapReduce

- Os *mappers* transformam os dados
- Os *reducers* agregam os resultados
- Tolerante a falhas

13

13

Mapper



14

14

Shuffle and sort

- Embaralhar e ordenar
- Processo automático
- Agrega dados por chave

15

15

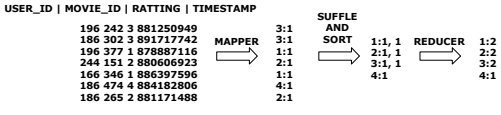
Reducer

- Extrai as informações dos conjuntos de dados

16

16

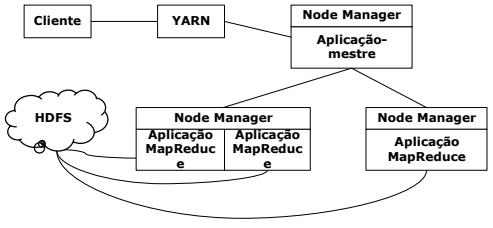
Exemplo



17

17

Fluxo de processamento MapReduce



18

18

Além do MapReduce

19

Alternativas

- Mais simples?
- Mais eficientes?
- Mais versáteis?

20

Pig

- Estende a capacidade do MapReduce
- Oferece uma linguagem semelhante à do SQL
- Executa *mappers* e *reducers* otimizados

21

Pig Latin

- Facilidade de programação
- Oportunidade de otimização
- Extensibilidade

22

- Correspondência com bancos de dados tradicionais
  - Relação → tabela
  - Tupla → registro (linha)
  - Campo → coluna

23

- Permite operações JOIN
- Pode ser executado em conjunto com substitutos do MapReduce
  - Tez, Spark
- Interfaces
  - CLI Grunt
  - Ambari
  - Hue

24

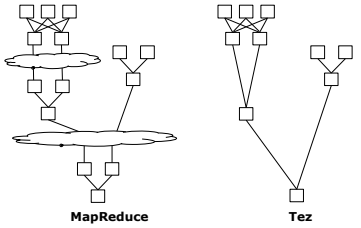
Tez

- Substitui o MapReduce
- Grafos acíclicos direcionados
- Pode otimizar a execução de outros componentes

25/32

25

Processamento distribuído baseado em grafos acíclicos direcionados



26/32

26

Gerenciamento de *cluster*

27/32

27

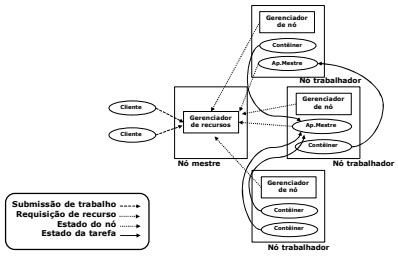
YARN

- "*Yet Another Resource Navigator*"
- Gerenciador de recursos global
- Distribui e gerencia o processamento pelo *cluster*
- Desmembramento do MapReduce original

28/32

28

Arquitetura do YARN



29/32

29

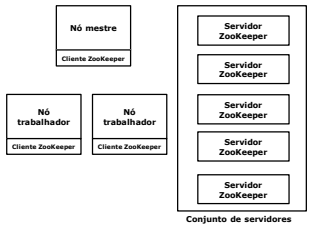
ZooKeeper

- Coordenador de aplicações distribuídas
- Alta disponibilidade
- Estrutura de árvores de diretório
- Permite manter informações dos nós e tarefas de uma aplicação

30/32

30

Arquitetura do ZooKeeper



31/32

31

Oozie

- **Orquestração de trabalhos**
- **Fluxo de dados por vários componentes Hadoop**
- **Agendamento e execução de trabalhos**
- **Grafos acíclicos direcionados**

32/32

32

33