Ambiente Livre
Tecnologia

Hadoop

Treinamento Hadoop – Big Data Open Source - Fundamental.

Instrutor: Marcio Junior Vieira.
marcio@ambientelivre.com.br

O que é HDFS

- **Hadoop Filesystem**
- Um sistema de arquivos distribuído que funciona em grandes aglomerados de máquinas de commodities.

Características do HDFS

- Inspirado em GFS
- Projetado para trabalhar com arquivos muito grandes
- Executado em hardware commodity
- Streaming de acesso a dados
- Replicação e localidade

Hadoop Filesystem

- **Sistema de Arquivos Distribuído para grande Volumes**

  10K nodes, 100 milhões de arquivos, 10 PB
- **Hardware Comum (comodite)**

  – Os arquivos são replicados esperando falha de hardware

  – Detecção de falhas e recuperação
- **Otimizado para Batch Processing**

  – Os dados ficam expostos, a computação pode ser movida onde os dados estiverem

HDFS

- Projetado para escalar a petabytes de armazenamento, e correr em cima dos sistemas de arquivos do sistema operacional subjacente.
-

“NameNode” - Master

- Gerencia o sistema de arquivos **namespace** (metadados dos arquivos)

  - FSImage e EditLog

- Conhece todos os blocos de localização
- Encaminha os blocos aos nós escravos
- Mantém as informações em memória.
- Controla a replicação, exclusão, criação

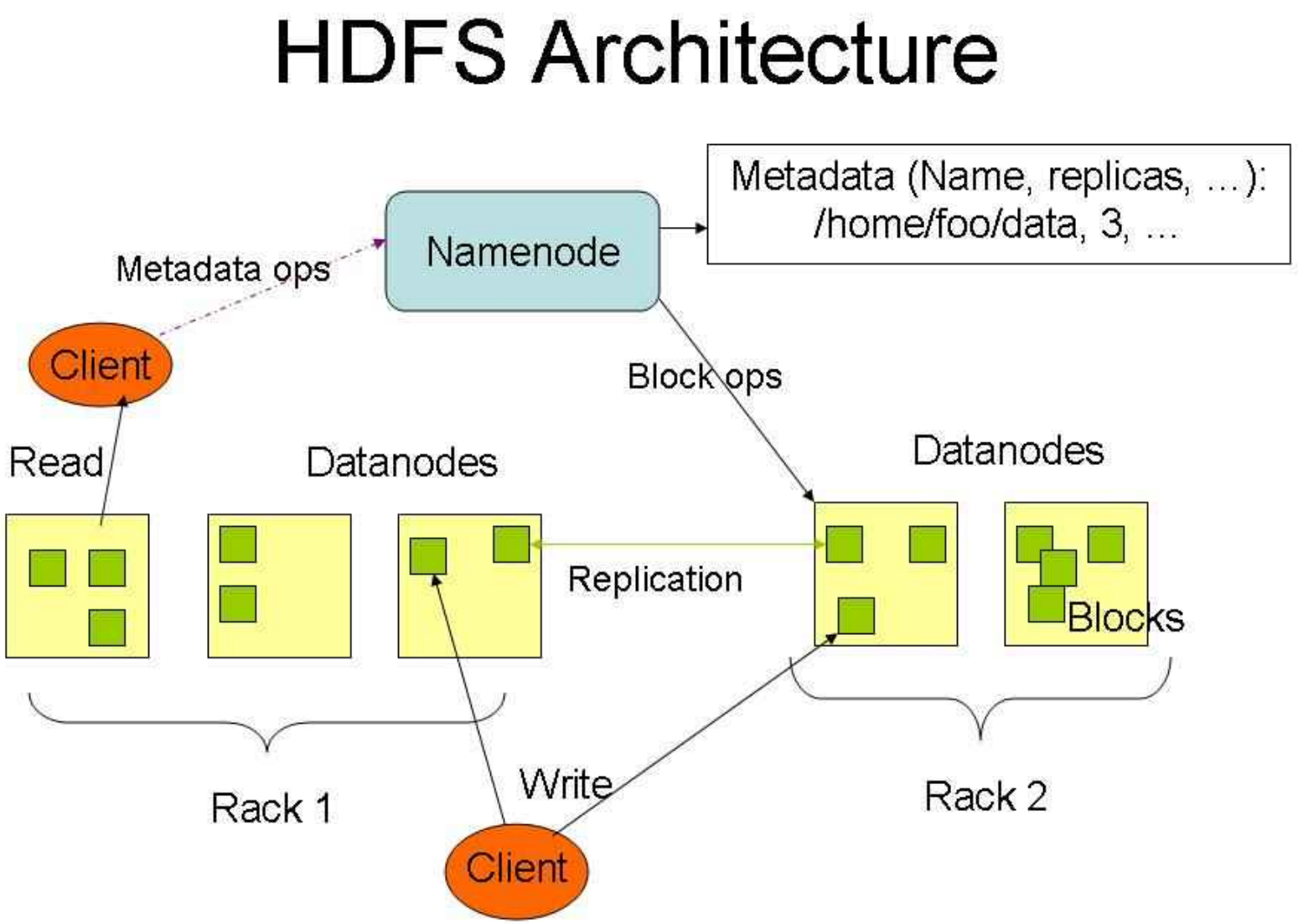
“DataNode” - Slave (workers)

- Mantem os blocos de dados
- Reporta de volta para namenode suas listas de blocos periodicamente
- lida com a recuperação de dados

SecondaryNameNode

- Nó auxiliar do HDFS
- Realiza pontos de checagem em intervalos configuráveis
- Permite manter nivel de desempenho do NameNode

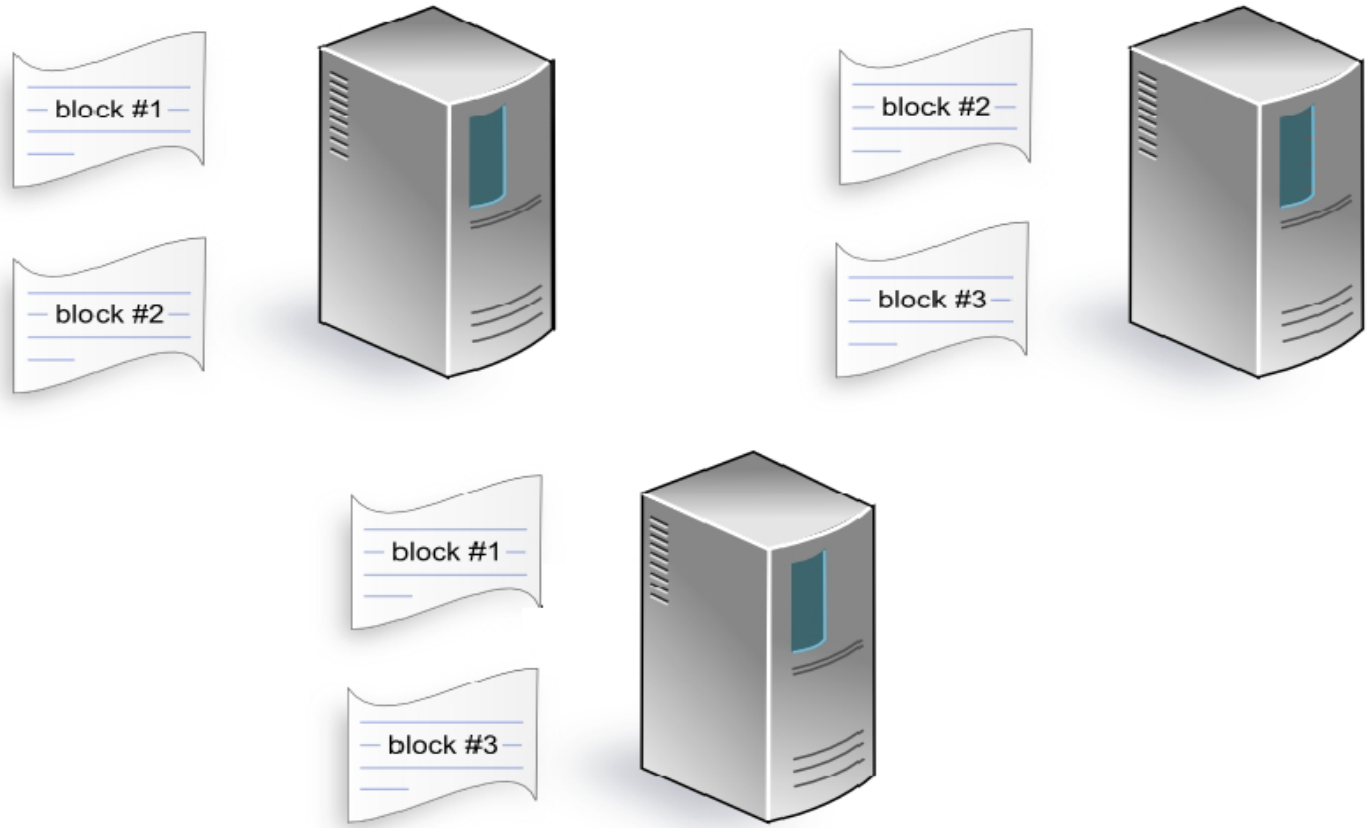
Arquitetura



Ambiente Livre
Tecnologia

HDFS - Replicação

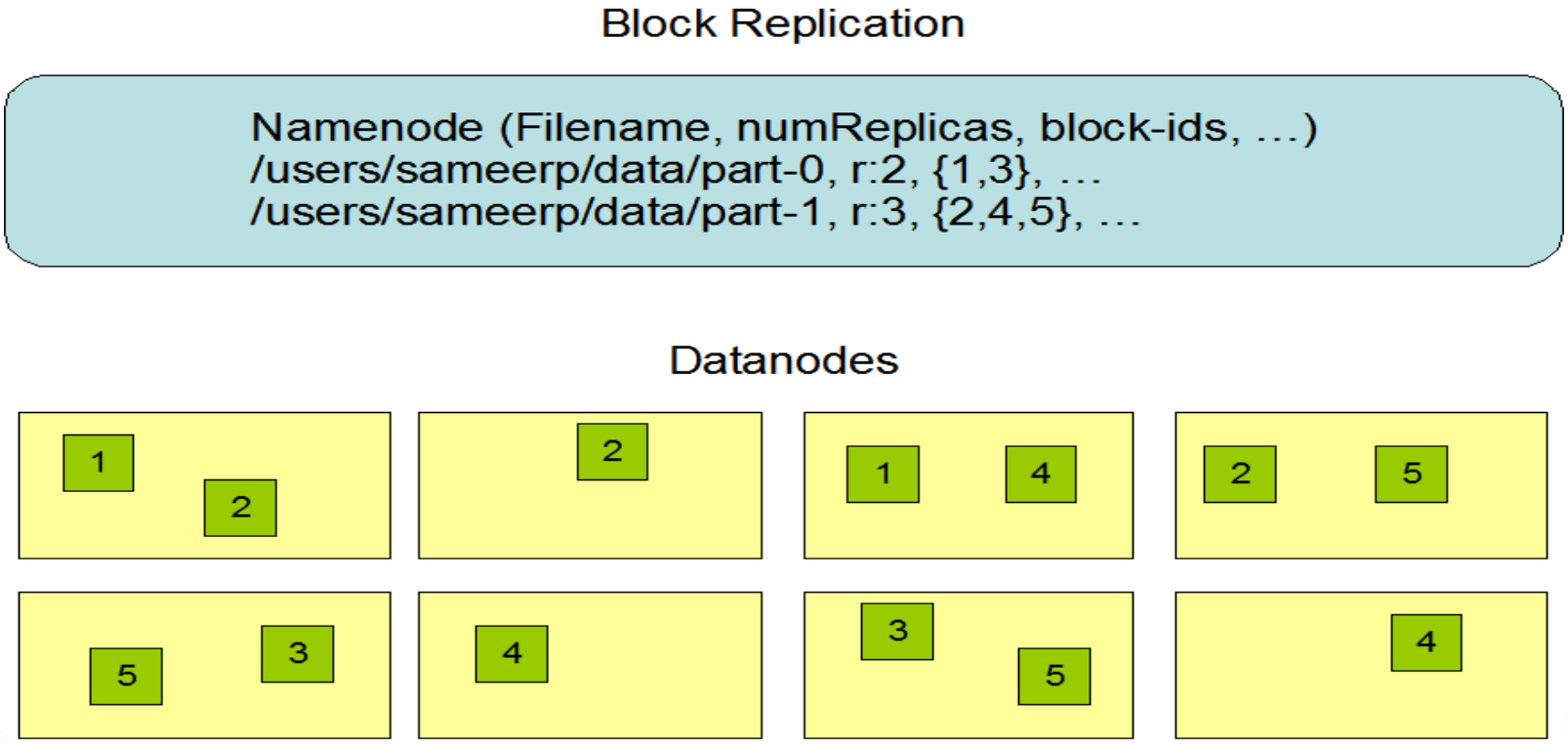
- Dados de entrada é copiado para HDFS é dividido em blocos e cada blocos de dados é replicado para várias máquinas



hadoop

HDFS - Replicação

- Dados de entrada é copiado para HDFS é dividido em blocos e cada blocos de dados é replicado para várias máquinas



Modos de Operação

- Standalone (Local)
- Pseudo-distributed
- Fully-distributed

HDFS – Fluxo do MapReduce