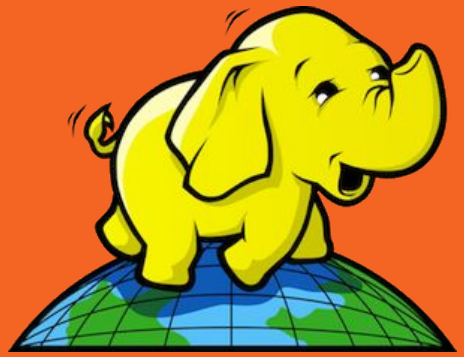
HADOOP - HDFS



EQUIPE:
Gabriel Wagner
Kauly Rosa Bohm

INSTITUTO FEDERAL
SANTA CATARINA

SISTEMA DISTRIBUÍDO

- Acessível
- Escalável
- Transparente

BIG DATA

Big Data é uma maneira de solucionar os os problemas não resolvidos relacionados ao gerenciamento e manipulação de grandes volumes de dados, com alta velocidade e variedade extensível de dados.

HADOOP

- Software Livre da Apache (Java)
- Computação Distribuída
- Tolerância a Falhas
- Alta Escalabilidade
- Grande Volume de Dados
- Cluster de Community Hardware
- Computação Paralela

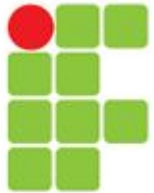
INSTITUTO FEDERAL
SANTA CATARINA

HADOOP - HISTÓRIA

- Fev/03 - Primeira biblioteca Map/Reduce na Google
- Out/03 - Artigo sobre GFS (Google File System)
- Dez/05 - Doug Cutting implementa MR e DFS no Nutch
- Fev/06 - Hadoop se torna um projeto oficial da Apache,
- Abr/06 - Hadoop classifica 1,8 TB em 188 nós em 47,9 horas
- Abr/07 - Yahoo! roda Hadoop em um cluster de 1000 nós

HADOOP - HISTÓRIA

- Jan/08 - Hadoop se transforma em um projeto principal da Apache
- Out/08 - Yahoo chega a marca de 10 TB/dia em seus clusters
- Jan/11 - Facebook, LinkedIn, eBay e IBM contribuem com 200,000 linhas de código.
- Mar/11 - Apache Hadoop ganha o prêmio de Media Guardian Innovation
- Nov/11 - Apache disponibiliza versão 1.0.0
- Jun/14 - Apache disponibiliza versão 2.4

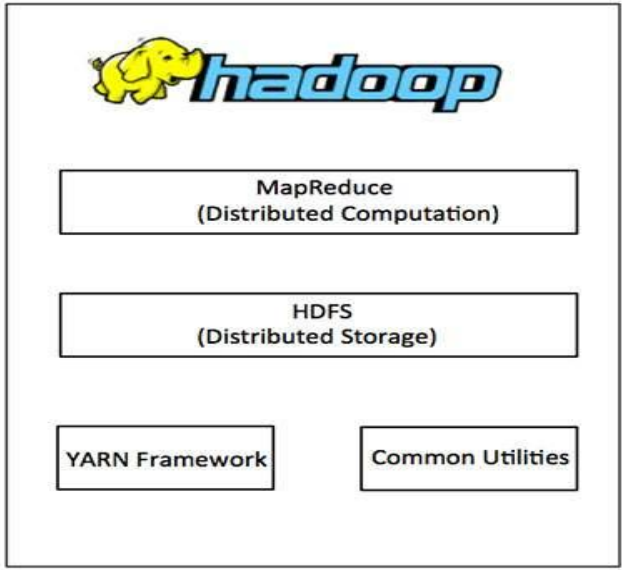


INSTITUTO FEDERAL
SANTA CATARINA

Quem utiliza

HADOOP - ARQUITETURA

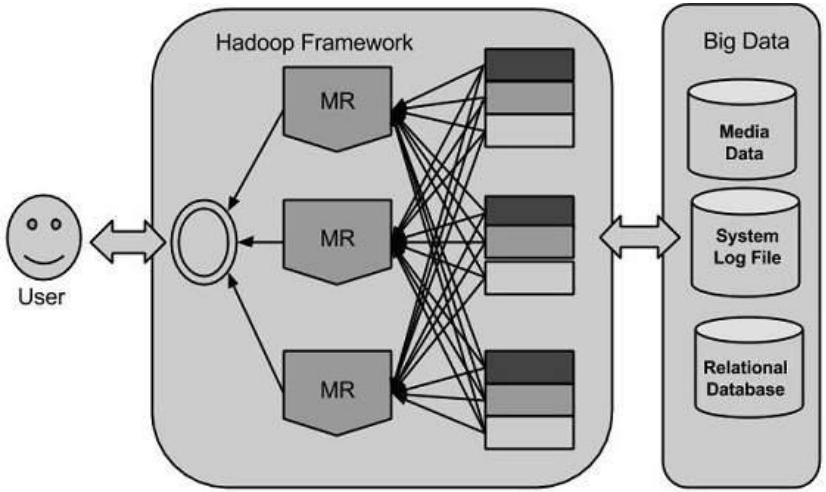
- MapReduce
- **HDFS**
- YARN
- Common



INSTITUTO FEDERAL
SANTA CATARINA

HADOOP - ARQUITETURA

- Visão Geral



INSTITUTO FEDERAL
SANTA CATARINA

HADOOP - HDFS

- Tolerância a Falhas.
- Acesso Paralelo a Dados.
- Replicação de Dados.
- Escalabilidade.
- Master/Slave.
- Write-Once-Read-Many
- CLI.

“ Hadoop Distributed File System - O mais confiável sistema de armazenamento do mundo.”

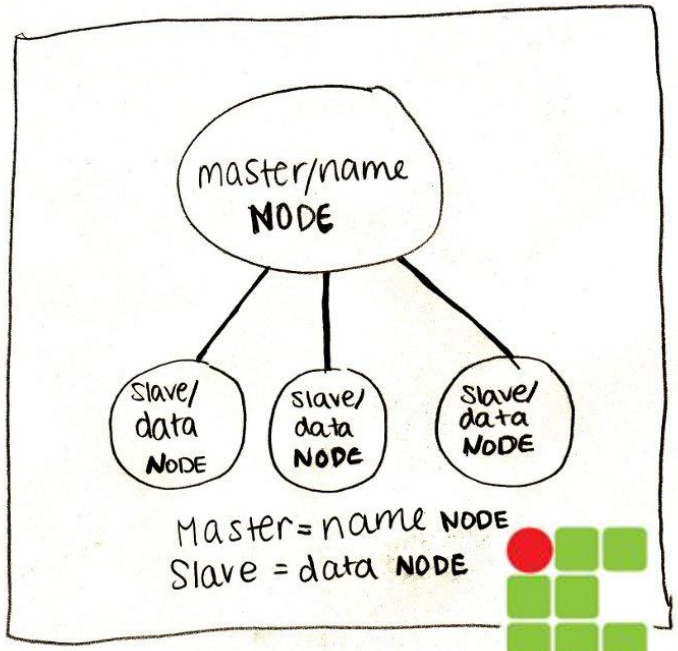
INSTITUTO FEDERAL
SANTA CATARINA

HDFS - ARQUITETURA

- Master/Slave
- NameNodes(Master)
- DataNodes(Slave)
- Blocks

HDFS - MASTER/SLAVE

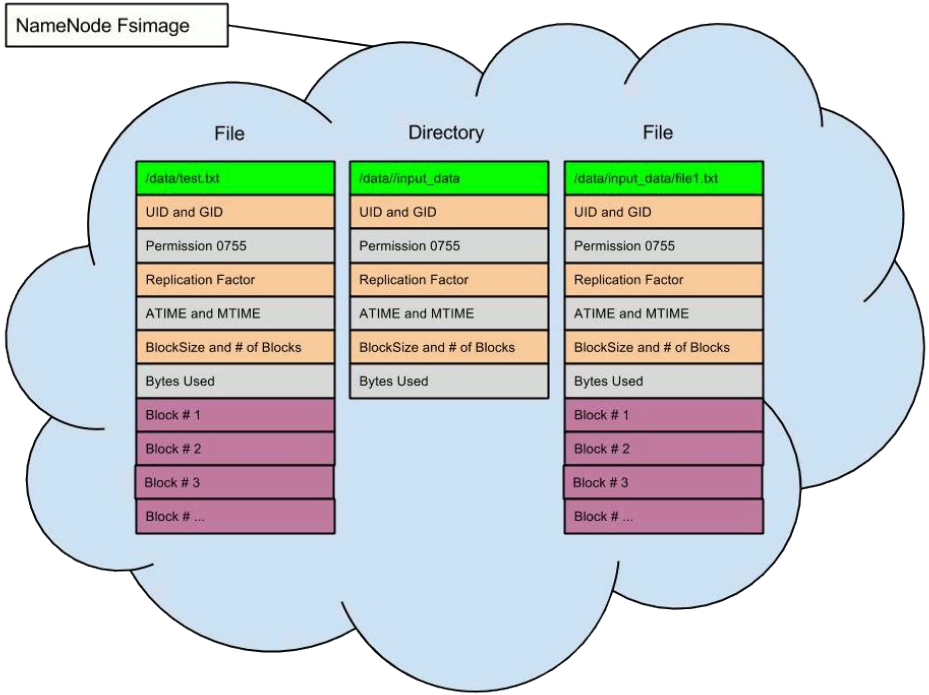
- Master/NameNode.
- Slave/DataNode.
- Comunicação Constante



HDFS - MASTER (NameNode)

- Peça central do HDFS.
- Controla o acesso aos dados.
- Controla as operações sobre os dados.
- Guarda a metadata.
- Localizado em nó de alta confiança.
- “Calcanhar de Aquiles” do HDFS.

MASTER - METADATA



INSTITUTO FEDERAL
SANTA CATARINA

MASTER - RECOMENDAÇÕES

- Nó exclusivo para o Master.
- Bastante memória RAM.
- Criar mais de um diretório para o master.
- Monitoramento constante do espaço livre em disco.

INSTITUTO FEDERAL
SANTA CATARINA

HDFS - SLAVE (DataNode)

- Armazenamento.
- Leitura
- Escrita.
- Criar
- Excluir
- Replicação.

INSTITUTO FEDERAL
SANTA CATARINA

HDFS - BLOCKS

- Unidade Mínima de Espaço.
- Fator de Replicação - Padrão 3.
- Tamanho - Padrão 128MB.
- Quanto maior, melhor.

INSTITUTO FEDERAL
SANTA CATARINA

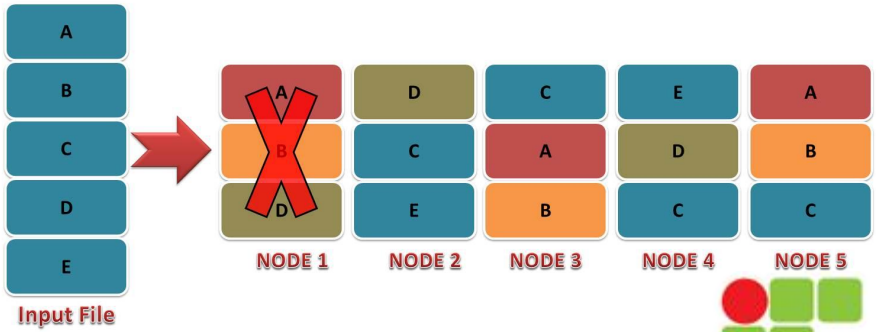
HDFS - FUNCIONAMENTO

- DADO INSERIDO NO HDFS
- DADO DIVIDIDO EM BLOCOS DE 128MB*
- OS BLOCOS SÃO REPLICADOS - Padrão 3
- METADATA ARMAZENADA NO MASTER
- OS BLOCOS SÃO ENVIADOS PARA OS SLAVES
- Ex: 1GB = 1024MB
- 1024/128 = 8 Blocos de 128MB

INSTITUTO FEDERAL
SANTA CATARINA

HDFS - REPLICAÇÃO

- Tolerância a Falhas.
- Replicação.
- Transparencia.



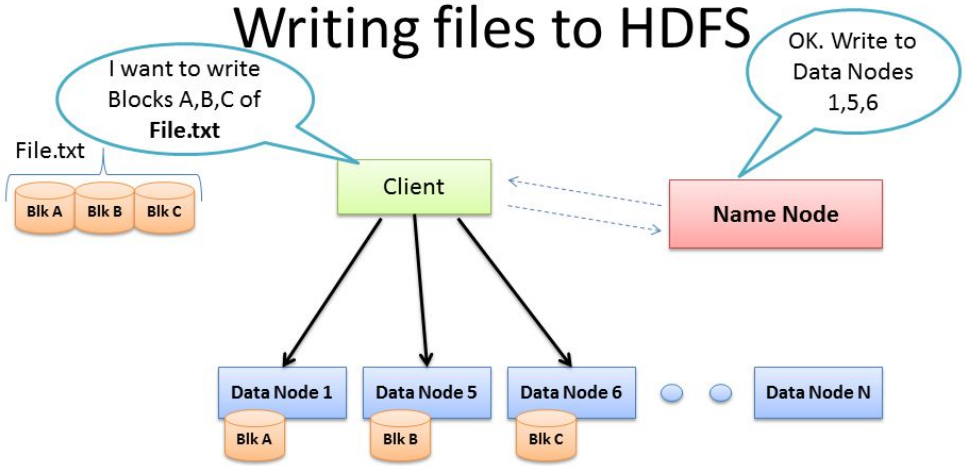
INSTITUTO FEDERAL
SANTA CATARINA

HDFS - REPLICAÇÃO

- Disponibilidade.
- Confiabilidade.
- Uso de Largura de Banda.

Mais Réplicas	Menos Réplicas
✓ Disponibilidade, Confiabilidade	✗ Disponibilidade,Confiabilidade
✗ Uso de Largura de Banda	✓ Uso de Largura de Banda
✗ Desempenho	✓ Desempenho

HDFS - FUNCIONAMENTO



- Client consults Name Node
- Client writes block directly to one Data Node
- Data Nodes replicates block
- Cycle repeats for next block

BRAD HEDLUND .com

INSTITUTO FEDERAL
SANTA CATARINA

HDFS - ESCALABILIDADE

- Expansão/Contração
- Vertical (RAM,Disk)
- Horizontal (+ nós)
- Escalabilidade horizontal feita sem down time.

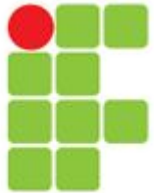
HDFS - CLI

- Execução: bin/hdfs.
- Sintaxe: hdfs [SHELL_OP] [COMM] [G_OP][COMM_OP]
- Comandos sobre FS:
- hdfs dfs -[COMM]
- Exemplo de comandos: cat, cp, ls, chmod, chown ...

https://hadoop.apache.org/docs/r2.7.1/hadoop-project-dist/hadoop-common/FileSystemShell.html

HDFS - REFERENCIAS

- http://hadoop.apache.org/
- https://www.tutorialspoint.com/hadoop/
- https://www.vivaolinux.com.br/artigo/Instalando-Apache-Hadoop
- http://data-flair.training/blogs/hadoop-hdfs-introduction-architecture-features-operations-tutorial
- https://linoxide.com/cluster/setup-hadoop-multi-node-cluster-ubuntu/#comment-2839
- http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/



INSTITUTO FEDERAL
SANTA CATARINA