

Formação Cientista de Dados


Introdução a Big Data

Hadoop

- Processamento em Batch
- Baseado no conceito de MapReduce
- Desenvolvido em Java
- Open source
- Distribuído
- Hardware commodity
- Capaz de distribuir o processamento em dezenas ou milhares de nós em um cluster
- Suporte a dados estruturados ou não estruturados
- Petabytes ou mais de dados



Hadoop



- Opera no conceito Master/Slave
- Master: (namenode, jobtracker, tasktracker)
 - Gestão: mantém metadados, logs, adiciona, encontra, exclui e copia arquivos, distribui as tarefas de mapeamento e redução entre os nós, agendamento, balanceamento etc.
- Slaves: (datanodes)
 - Mantém dados, replica blocos

Hadoop

- NameNode pode ser replicado (Hadoop 2)
- Datanodes são configurados em modo ativo e standby
- Heartbeat: enviado do DataNode ao NameNode regularmente, como sinal de “saúde”



Yarn

- Alocação de Recursos de forma global e unificada no cluster
- Agendamentos
- Priorização
- Tolerância a Falhas
- Componentes:
 - ResourceManager: um por cluster
 - ApplicationManager: gerencia atividades, otimização, distribuição de recursos etc.
 - Scheduler
 - NodeManager: um por nó
 - Responsável pela execução dos Jobs
 - Application Master:
 - Distribui tarefas aos containers
 - Container: mantem as tarefas

Cases

- 2008: Yahoo Coloca em Produção Cluster Hadoop com 10 mil cores
 - 5 Petabytes de Dados
 - <https://yahooohadoop.tumblr.com/post/98098649696/yahoo-launches-worlds-largest-hadoop-production>

