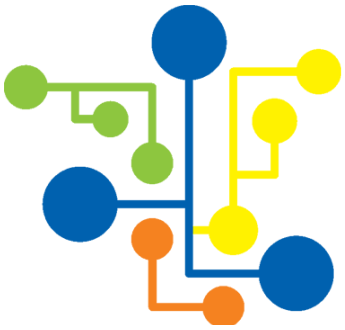


Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Componentes do Apache Hadoop

Existem três componentes principais do Apache Hadoop:

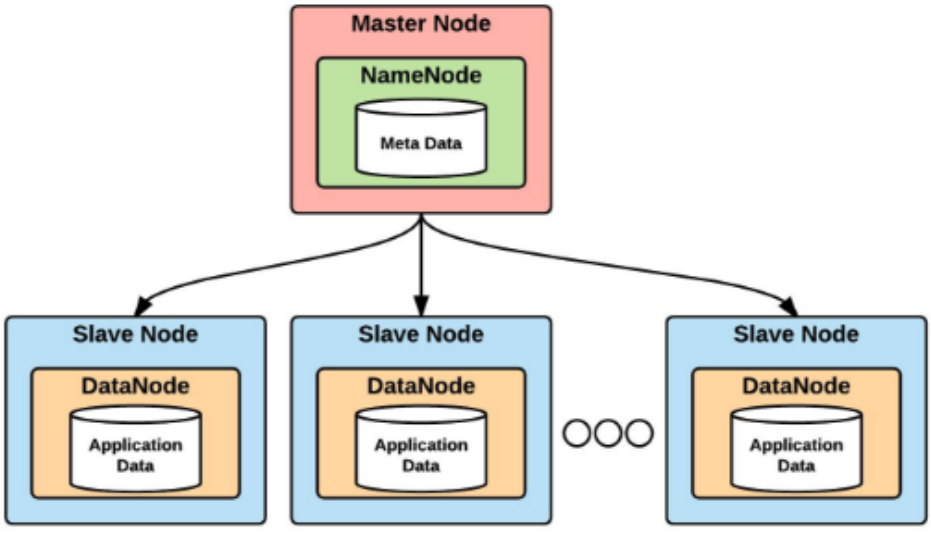
- HDFS
- MapReduce
- Yarn

Vamos discutir cada um deles.

HDFS

O Hadoop HDFS ou o Hadoop Distributed File System é um sistema de arquivos distribuído que fornece armazenamento no Hadoop de maneira distribuída.

No nó principal da Arquitetura Hadoop, um daemon chamado namenode é executado para o HDFS. Em todos os escravos, um daemon chamado datanode é executado para o HDFS. Portanto, os escravos também são chamados de datanodes. O Namenode armazena metadados e gerencia os datanodes. Por outro lado, os Datanodes armazenam os dados e executam as tarefas.

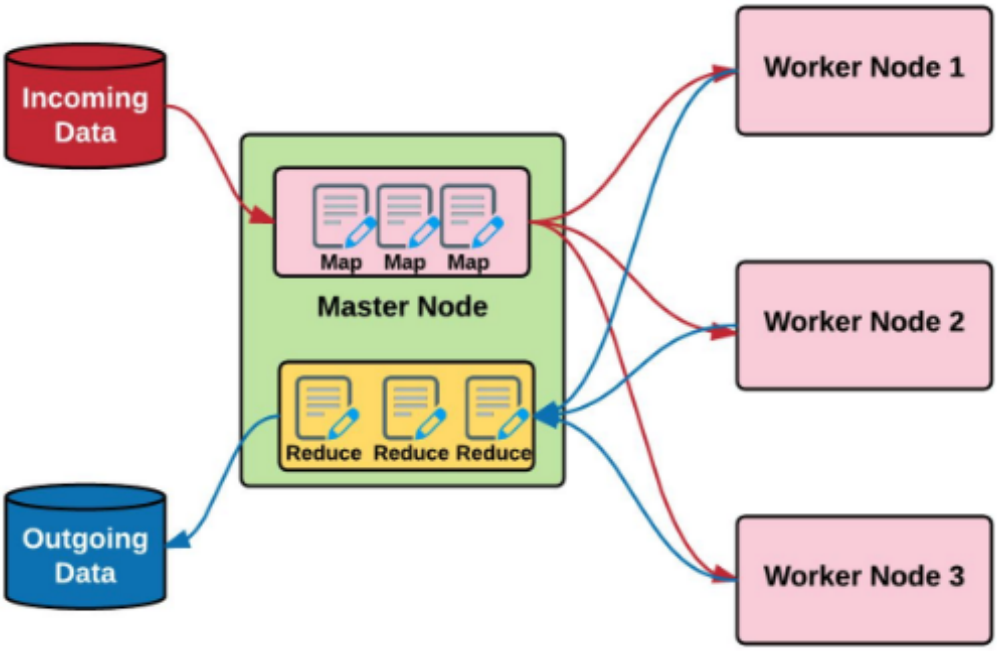


O HDFS é um sistema de arquivos altamente tolerante a falhas, distribuído, confiável e escalável para armazenamento de dados.

O HDFS foi desenvolvido para lidar com grandes volumes de dados e pode ser usado com arquivos de tamanho no intervalo de GBs a TBs. Um arquivo é dividido em blocos (padrão 128 MB) e armazenado de forma distribuída em várias máquinas. Esses blocos se replicam de acordo com o fator de replicação. Após a replicação, os blocos são armazenados em diferentes nós. Logo, se uma das máquinas do cluster onde está uma das 3 cópias do bloco falhar, ainda teremos duas cópias daquele bloco. Considerando um arquivo de 640 MB, ele será dividido em 5 blocos de 128 MB cada (se usarmos o valor padrão).

MapReduce

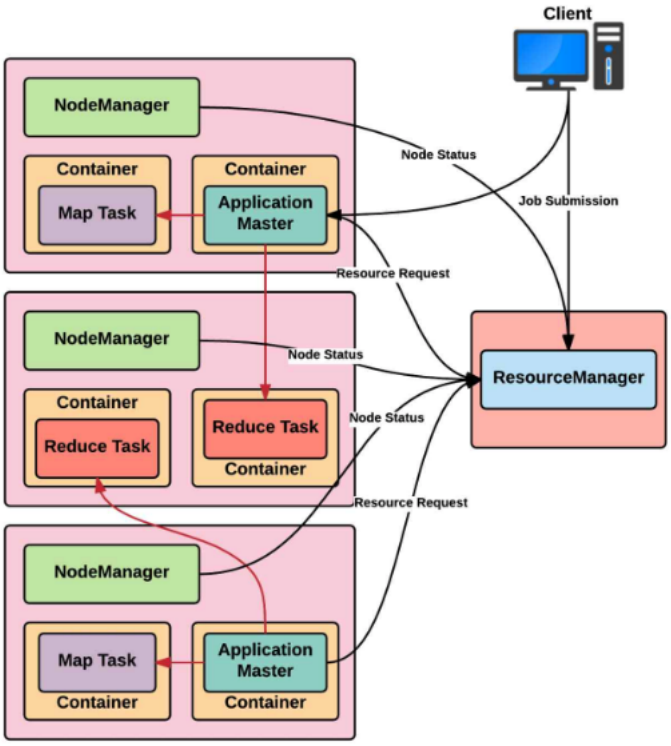
O Hadoop MapReduce é um modelo de programação. Ele foi projetado para processar grandes volumes de dados em paralelo, dividindo o trabalho em um conjunto de tarefas independentes. O MapReduce move a computação para perto dos dados (ao invés de mover os dados para perto da computação, como em arquiteturas de Data Warehouse por exemplo). Ele faz isso, pois o movimento de um enorme volume de dados será muito caro e demorado. Ele permite escalabilidade massiva em centenas ou milhares de servidores em um cluster Hadoop.



Portanto, o Hadoop MapReduce é uma estrutura para processamento distribuído de grandes volumes de dados configurados em um cluster de computadores. Como os dados são armazenados de maneira distribuída no HDFS, o MapReduce permite realizar o Map–Reduce para realizar o processamento paralelo.

YARN

YARN - Yet Another Resource Negotiator - é a camada de gerenciamento de recursos do Hadoop. No cluster de vários nós, torna-se muito complexo gerenciar/alocar/liberar os recursos (CPU, memória, disco). O Hadoop Yarn gerencia os recursos com bastante eficiência.



No nó principal, o daemon do ResourceManager é executado para o YARN e, em seguida, para todos os nós escravos, o daemon do NodeManager é executado.

Uma alternativa ao YARN é o Apache Mesos.