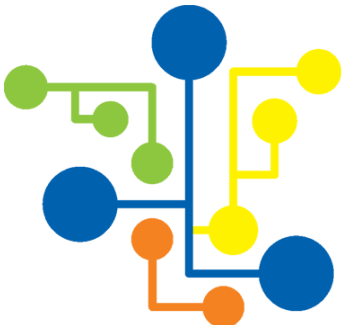

Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Como o Hadoop Funciona?

Vamos resumir como funciona o Apache Hadoop:

1) A aplicação cliente envia os dados para o Hadoop, que então divide os dados em blocos de tamanho 128 Mb (por padrão). Em seguida, os dados são enviados pelo HDFS para diferentes nós no cluster.

2) Depois que todos os blocos do arquivo são armazenados nos datanodes, um usuário pode processar os dados.

3) Quando um job de processamento é disparado com o MapReduce (ou Apache Spark), o nó mestre agenda o job (enviado pelo usuário) em nós individuais.

4) Depois que todos os nós processarem os dados, a saída será gravada de volta no HDFS.

O daemon Namenode armazena os metadados enquanto os daemons datanode armazenam os dados reais.

Os dados são divididos em pequenos pedaços chamados de blocos e esses blocos são armazenados de forma distribuída em nós diferentes no cluster. Cada bloco é replicado de acordo com o fator de replicação (por padrão, 3).

Para processar os dados, o cliente precisa enviar o algoritmo para o nó mestre. O Hadoop funciona com base no princípio da localização dos dados, ou seja, em vez de mover dados para o algoritmo, o algoritmo é movido para datanodes onde os dados estão armazenados.

Em resumo, podemos dizer que o cliente envia os dados (que serão armazenados) e o programa (que vai processar os dados). O HDFS armazena os dados enquanto o MapReduce processa os dados. Simples, não?