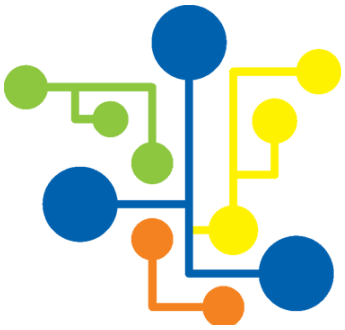

Data Science Academy

www.datascienceacademy.com.br

Data Lake – Design, Projeto e Integração

Um Pouco Mais Sobre o Hadoop

O Apache Hadoop é um framework open source, escalável e tolerante a falhas escrito em Java. Ele processa eficientemente grandes volumes de dados em um cluster de hardware commodity (baixo custo). O Hadoop não é apenas um sistema de armazenamento, mas também uma plataforma para armazenamento de dados e processamento de dados.

O Hadoop é uma ferramenta de código aberto da ASF - Apache Software Foundation. Um projeto de código aberto significa que ele está disponível gratuitamente e podemos até mesmo alterar seu código-fonte de acordo com os requisitos. Se determinada funcionalidade não atender a sua necessidade, você poderá alterá-la de acordo com sua necessidade. A maior parte do código do Hadoop foi escrita pelo Yahoo, IBM, Facebook e Cloudera.

Ele fornece uma estrutura eficiente para executar tarefas em vários nós de clusters. Cluster significa um grupo de máquinas conectadas via LAN. O Apache Hadoop fornece processamento paralelo de dados enquanto trabalha em várias máquinas simultaneamente.

A linguagem de programação básica do Hadoop é Java, mas isso não significa que você pode codificar apenas em Java. Você pode codificar em C, C ++, Perl, Python, Ruby, etc. Você pode codificar a estrutura do Hadoop em qualquer linguagem, mas será melhor codificar em java, já que você terá um controle de nível mais baixo do código.

O Hadoop pode ser configurado em uma única máquina (modo pseudo-distribuído), mas mostra seu poder real com um cluster de máquinas. Podemos escalá-lo para milhares de nós em tempo real, sem nenhum tempo de inatividade. O Hadoop consiste em três partes principais:

- Hadoop Distributed File System (HDFS) - É a camada de armazenamento do Hadoop.
- Map-Reduce - É a camada de processamento de dados do Hadoop.
- YARN - É a camada de gerenciamento de recursos do Hadoop.

O Hadoop funciona com a arquitetura mestre-escravo (master-slave). Existe um nó mestre e existem n nós escravos onde n pode ser mais de 1000 por exemplo.

O mestre gerencia, mantém e monitora os escravos, enquanto os escravos são os nós reais que fazem o trabalho. Na arquitetura do Hadoop, o mestre deve ser implementado em um bom hardware, pois ele é a peça central do cluster Hadoop.

O mestre armazena os metadados (dados sobre dados) enquanto os escravos são os nós que armazenam os dados. O cliente (aplicação cliente) se conecta ao nó mestre para executar qualquer tarefa.

No próximo item de aprendizagem, vamos discutir um pouco mais sobre os componentes do Hadoop. Até lá!