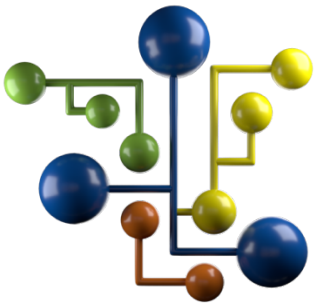

Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Apache Spark SQL

O Spark SQL, é parte integrante do framework Apache Spark, utilizado para processamento de dados estruturados, que permite executar consultas SQL no conjunto de dados do Spark. É possível realizar tarefas de ETL sobre os dados em diferentes formatos, como por exemplo JSON, arquivos csv, bancos relacionais e não relacionais. O Spark SQL permite que facilmente usemos SQL em nossas aplicações de análise de dados escritas em Python para Spark, por exemplo. Vejamos as principais funcionalidades do Spark SQL.

As versões mais recentes do Spark fornecem uma abstração de programação chamada de DataFrames, que pode agir como um motor de consultas SQL. DataFrames podem fornecer funções de alto nível, permitindo que o Spark compreenda melhor a estrutura dos dados, assim como o cálculo a ser executado. Esta informação adicional permite que o otimizador e o mecanismo de execução do Spark acelerem automaticamente as análises de Big Data.

Com a inclusão de uma API para fontes de dados, a biblioteca Spark SQL permite que a computação de informações sobre dados armazenados de forma estruturada seja facilitada e mais abrangente.

Com o servidor JDBC interno do Spark, a conexão com banco de dados relacionais, para consultar dados estruturados em tabelas e realizar análises de Big Data, pode ser feita com ferramentas de BI tradicionais.

O Spark SQL se integra com o Spark MLLib, o módulo de Machine Learning do Spark, servindo como fonte de dados para a construção de modelos preditivos. Ou seja, você pode usar o Spark SQL para extrair dados de Data Warehouses por exemplo, realizar alguns tratamentos e entregá-los para processamento de Machine Learning.