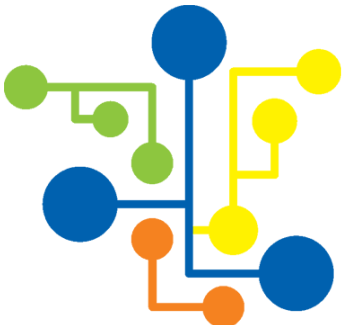

Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Mini-Projeto 1
Importando Dados do Banco de Dados Oracle para o HDFS

A Oracle é a líder mundial em banco de dados e uma das gigantes de tecnologia. Bancos de dados Oracle podem ser encontrados em Data Warehouses ou sistemas ERP como SAP, PeopleSoft e JDEdwards. Esses sistemas podem conter bancos de dados com muitos milhões de registros.

Por esta razão, trouxemos para este projeto, todo o processo passo a passo de como importar dados do Oracle para o HDFS utilizando o Sqoop, uma ferramenta ETL gratuita e um dos componentes do ecosistema Hadoop.

O Sqoop permite conectar via JDBC ao banco de dados Oracle, executar uma query, extrair dados e carregar no HDFS, para posterior processamento analítico, através de um cluster.

A Oracle tem investido bastante no Hadoop e sua solução de Big Data, chamada Big Data Appliance, é totalmente baseada no Hadoop. A Oracle fornece ainda conectores e ferramentas de análise com linguagem R, que permitem manipulação de dados no Hadoop.

Para este projeto, vamos inicialmente carregar 15 milhões de registros no banco de dados Oracle, a partir de um dataset de avaliação de filmes, disponível no site http://grouplens.org. Na sequência, vamos criar uma instrução Sqoop e importar os dados para o HDFS. O Manual de instalação e configuração do Oracle, bem como o script com todos os comandos usados para este projeto, estão em anexo.