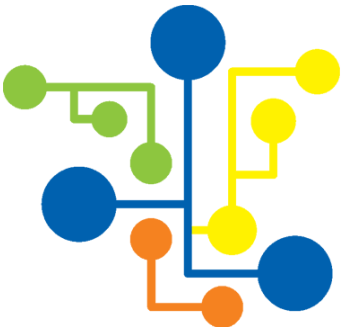

Data Science Academy

www.datascienceacademy.com.br

Formação Cientista de Dados

Projeto com Feedback 5

Implementando Slow Changing Dimensions em um DW Usando Hive e Spark

Um dos mais amplos usos do Hadoop atualmente é construir uma plataforma de Data Warehousing sobre um Data Lake, através do Apache Hive. E na construção de um Data Warehouse, as tradições deixadas por Kimball e Inmon ainda são mais válidas do que nunca (Kimball e Inmon são os “papas” do DW).

A dimensão de alteração lenta (Slowly Change Dimension) de um DW, é uma dimensão que raramente muda. No entanto, quando muda, deve haver uma abordagem sistemática para capturar essa mudança. Exemplos de SCDs são informações de clientes e produtos.

Neste projeto do Hive, você deve implementar as SCDs no Hive e no Spark.

O banco relacional completo está sendo fornecido a você em anexo a este pdf. Você deve seguir as instruções nos scripts, carregar o database no MySQL e depois construir as SCDs no Hive. Claro que para isso você precisa construir um Data Lake com Apache Hadoop e Hive. Use o Spark para o processamento dos dados, mas apenas se desejar. O mais importante é ter um DW funcionando com o Hive e as SCDs disponíveis para uso.

Você deve enviar os scripts usados para montar o Data Lake e o ambiente com Hive.

Quando concluir o projeto, envie os scripts e datasets para projeto@dsacademy.com.br. Caso os datasets usados sejam muito grandes, armazene em um diretório virtual (existem vários na internet, como Google Drive ou Dropbox) e envie o link para que nossa equipe possa baixar os datasets. Se os arquivos foram pequenos (uma amostra do dataset original), envie no anexo junto com o script. Documente seu script tanto quanto possível.

Caso prefira, disponibilize seu projeto no Github e envie o link do seu repositório para nossa equipe no e-mail projeto@dsacademy.com.br. Nesse caso, o Readme do repositório deve constar que este trata-se de um projeto da Formação Cientista de Dados da Data Science Academy.

Em até 24 horas, daremos o feedback!

Bom trabalho!