Bootcamp: Engenheiro de Dados

Desafio Prático

Módulo 3: Soluções de Big Data e Data Lake

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Spark SQL
2. UDFs (*user-defined functions*)
3. Formato Parquet
4. Conhecimentos gerais sobre a utilização do Apache Spark

Enunciado

Oi, turma!

Bem-vindos ao Desafio do Módulo 3!

Nele, você vai utilizar o Spark SQL e UDFs para responder a algumas perguntas a partir de dados de cadastro de estabelecimentos brasileiros.

Você vai trabalhar com os seguintes arquivos, disponíveis em: http://www.dcc.ufmg.br/~pcalais/XPE/engenharia-dados/big-data-spark/desafio.

São dois conjuntos de arquivos:

- CNAEs: contém o código do CNAE (Classificação Nacional de Atividades Econômicas) e a descrição textual de cada um. Por exemplo, o CNAE 0116403 indica que a atividade de um estabelecimento com este CNAE é “Cultivo de Mamona”.
- Estabelecimentos: contém o registro de estabelecimentos e diversos metadados, como CNPJ, o código do CNAE, endereço, telefone, entre outros.

Consulte o arquivo NOVOLAYOUTDOSDADOSABERTOSDOCNPJ.pdf para checar o esquema de dados e explicação mais detalhada de cada campo.

Utilize o Apache Spark para ler os dados e responder às questões propostas. Divirta-se!

Prof. Pedro Calais.

Atividades

Os alunos deverão desempenhar as seguintes atividades:

- Utilizar Spark SQL para ler os dados fornecidos;
- Responder às questões propostas.