Disciplina:
Big Data

Exercício:
Utilização do Apache Hadoop e Spark na Série de Preços de Combustíveis do Brasil

Objetivo:
Parte 1
Escrever um script que realiza um web scraping e capture todos os preços dos últimos 6 anos (Janeiro de 2020 até 2025 Janeiro) do Portal de Dados Abertos. Você deve coletar todos os arquivos CSV que possua dados referentes ao preço do combustível. Fique atento que a nomenclatura dos arquivos muda conforme o tempo.

Link do Portal:
https://dados.gov.br/dados/conjuntos-dados/serie-historica-de-precos-de-combustiveis-e-de-glp

Parte 2
Repasse todos os arquivos CSV capturados pelo web scraping para o HDFS do Hadoop. Em seguida, faça um Map Reduce (com a linguagem de sua preferência) para calcular o preço médio do diesel do estado de SP.

Parte 3
Faça um script que analise todos os dados coletados utilizando o Apache Spark. O script deve ser capaz de responder às seguintes perguntas:

- Média, mediana e desvio padrão dos preços de venda da gasolina, etanol, diesel e diesel S10.
- Quais são os 3 principais postos de São Paulo que têm a maior média de venda da gasolina, etanol e Diesel
- Qual o estado que possui a maior média de venda para diesel e diesel S10
- Quais foram os valores de venda mais alto atrelados a cada bandeira do estado de São Paulo

- Qual o município apresentou o maior e o menor preço médio do diesel?
- Informe os 3 bairros de Recife que apresentaram a maior média de preço para diesel e diesel S10, e seus respectivos preços

Observação:
Sobre a parte 1 do trabalho, o aluno deverá obrigatoriamente utilizar o selenium para baixar os arquivos CSV da página.

Sobre a parte 2 do trabalho, o aluno poderá executar o hadoop em um container Docker. Como apresentado em sala de aula, o aluno poderá utilizar esse projeto como referência.

Imagem Docker:
https://github.com/mjstealey/hadoop

Referente a parte 3, deve existir duas implementações, uma utilizando o Apache Spark Core e outra utilizando o Apache Spark SQL.

A parte 2 e 3 deste trabalho só poderá ser implementada caso o aluno realize a extração dos CSV na parte 1. É importante ressaltar que a parte 2 e 3 são independentes, ou seja, você não precisa executar o Apache Spark conectado ao Hadoop para computador os dados.

Formação de Equipes:
O trabalho deverá ser feito com as mesmas equipes

Observação:
O aluno poderá utilizar estender o script apresentado em sala de aula:

Data de Entrega:
01/04/2025