Avaliação Confitec

PySpark

Ler o arquivo parquet e manipular os dados de acordo com os próximos passos:

1. Transformar os campos "Premiere" e "dt_inclusao" de string para datetime.
2. Ordenar os dados por ativos e gênero de forma decrescente, 0 = inativo e 1 = ativo, todos com número 1 devem aparecer primeiro.
3. Remover linhas duplicadas e trocar o resultado das linhas que tiverem a coluna "Seasons" de "TBA" para "a ser anunciado".
4. Criar uma coluna nova chamada "Data de Alteração" e dentro dela um timestamp.
5. Trocar os nomes das colunas de inglês para português, exemplo: "Title" para "Título" (com acentuação).
6. Testar e verificar se existe algum erro de processamento do spark e identificar onde pode ter ocorrido o erro.
7. Criar apenas 1 .csv com as seguintes colunas que foram nomeadas anteriormente "Title, Genre, Seasons, Premiere, Language, Active, Status, dt_inclusao, Data de Alteração" as colunas devem estar em português com header e separadas por ";".
8. Inserir esse .csv dentro de um bucket do AWS s3
9. Subir o código no github com o nome TESTEPYSPARK-Confitec

Algoritmo

Crie um algoritmo de multiplicação de matriz quadrada. O resultado do programa deverá apresentar os valores da Matriz A, Matriz B e o Produto.
Exemplo:

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

$$product = \begin{bmatrix} 34 & 44 & 54 & 64 \\ 82 & 108 & 134 & 160 \\ 34 & 44 & 54 & 64 \\ 82 & 108 & 134 & 160 \end{bmatrix}$$

1. Utilize a linguagem que possui maior familiaridade e facilidade;
2. O programa deverá instanciar uma Matriz A e Matriz B com números aleatórios;
3. O Output do programa deverá conter os valores da Matriz A, Matriz B e Produto
4. Subir o código no github com o nome TESTEPYSPARK-Confitec

Um bônus será considerado nos casos em que o algoritmo utilizar maior recurso de CPU e menor recurso de memória em sua execução. Porém não é obrigatório.

Obs: Separar os desafios por pasta no repositório.