Atividade de Aprofundamento

O objetivo dessa atividade de aprofundamento é trabalhar com o PySpark. Ela é composta por duas questões, na primeira você deve manipular um arquivo de texto e na segunda você trabalhará com um arquivo csv. Caso você não consiga executar a máquina Cloudera, não se preocupe, pois você conseguirá responder as questões com base no tutorial "Conhecendo o PySpark" apresentado nesta trilha.

1. Trabalhando com arquivo de texto

Você utilizará o arquivo "python_conceito.txt". Para baixar esse arquivo acesse:

https://drive.google.com/open?id=1rfnz97W6-BOm7iAQbmxL5KKIfY5BTOGC

Salve o arquivo em uma pasta local, em seguida resposta as seguintes questões utilizando o PySpark:

a) Quantas linhas tem o arquivo?
b) Exiba todas as linhas que contém a palavra Python. Importante: você não deve mostrar o número de linhas e sim o conteúdo.
c) Aplique o processo de contagem de palavras e exiba as 10 primeiras ocorrências da contagem de palavras (não ordene os dados).

2. Trabalhando com um arquivo csv

Baixe o arquivo vgsales_mod2.csv por meio do seguinte link:

https://drive.google.com/open?id=1BIhybwnVz1KQdINMJH7X0V9WTj045TMk

Salve o arquivo em uma pasta local, em seguida resposta as seguintes questões utilizando o PySpark:

a) Carregue a base em um DataFrame do PySpark e informe a quantidade de linhas que a base possui. Os campos da base são: Id, Rank, Name, Platform, Year, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales.
b) Exibe os dois primeiros registros da base.
c) Apresente a quantidade de registros que contém o gênero (Genre) do tipo Sports.
d) Selecione apenas os jogos com venda global (Global_Sales) superior a 20. Em seguida apresente as 3 primeiras linhas.

e) Selecione lançados após 2010 (use o campo Year). Em seguida apresente os 10 primeiros registros.