Bootcamp: Arquiteto(a) de Big Data

Desafio

Módulo 4	Processamento de dados utilizando o ecossistema Hadoop

Objetivos

Exercitar os seguintes conceitos trabalhados no Módulo:

- ✓ Aprender a mexer no Databricks e no seu notebook;
- ✓ Fazer análises de dados empregando o PySpark

Enunciado

Vamos utilizar três arquivos de dados, que também serão disponibilizados separadamente, relacionados a casos de COVID. Eles se encontram internamente no Databricks, no diretório dbfs:/databricks-datasets/COVID/coronavirusdataset/

- dbfs:/databricks-datasets/COVID/coronavirusdataset/Case.csv

- dbfs:/databricks-datasets/COVID/coronavirusdataset/PatientInfo.csv

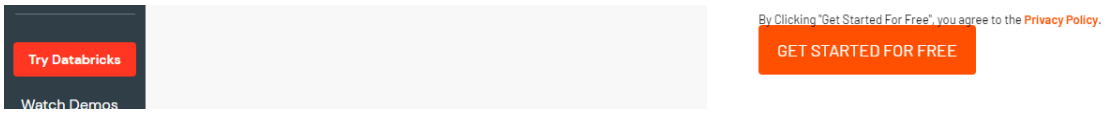
- dbfs:/databricks-datasets/COVID/coronavirusdataset/PatientRoute.csv

De posse deles, faça as análises necessárias, com o PySpark, para responder às questões do Desafio.

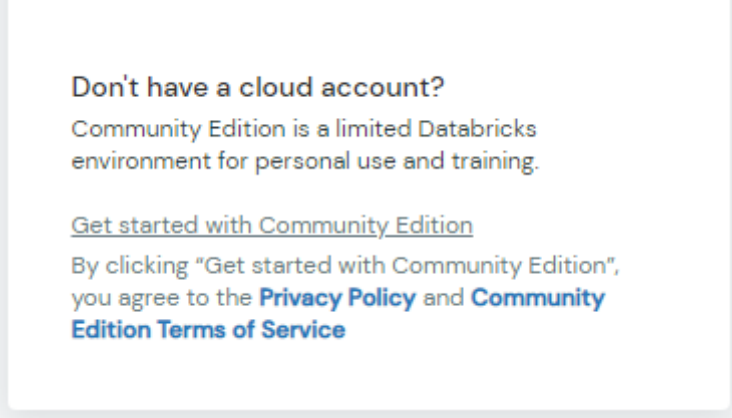
É recomendado que você crie uma conta no ambiente Databricks Community Edition (gratuito) e resolva as questões, utilizando o PySpark no Databricks Notebook. Esse ambiente não precisa de configurações e você pode começar imediatamente, aplicando para responder às questões.

Instruções para Criação de Conta no Databricks Community Edition

1. Vá para o site do Databricks e clique no botão em vermelho **Try Databricks**, preencha com seus dados e posteriormente clique em **GET STARTED FREE**.



2. No final da página há um link escrito “Get started with Community Edition”. Clique nele e será enviado um e-mail para você confirmar. É importante **NÃO** selecionar o botão azul Get started.



3. Você terá acesso à página do Databricks Community Edition, que não requer configurações e é totalmente gratuito para aprender.

Respostas Finais

Os alunos deverão desenvolver a prática e, depois, responder às seguintes questões objetivas: