Bootcamp: Cientista de Dados

Desafio Prático

Módulo 3: Técnicas para o processamento do Big Data

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Pré-processamento dos dados.
2. Detecção de anomalias.
3. Processamento dos dados.
4. Correlações.
5. Spark MlLib.

Enunciado

A análise de dados é a ciência que estuda e interpreta os dados, a fim de possibilitar a tomada de decisão mais assertiva. Por meio da análise de dados, é possível identificar oportunidades, prever o impacto de decisões, escolher qual investimento é mais lucrativo e conhecer melhor o cliente.

Nesse sentido, coletar e analisar dados sobre os consumidores e o público-alvo de um negócio é essencial para garantir que as necessidades dos clientes sejam atendidas. Logo, empregar técnicas como a segmentação de clientes auxilia na tarefa de aproximar e alinhar expectativas. Assim, é possível aumentar a competitividade do negócio e a satisfação dos consumidores.

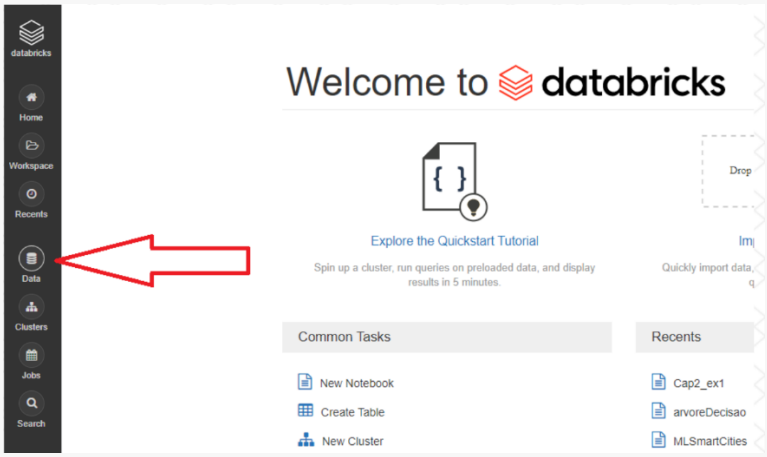
Neste desafio, vamos realizar uma análise sobre um banco de dados composto por clientes de um shopping. Nessa análise, vamos tentar compreender melhor quais são as características dos nossos consumidores. Para isso, vamos aplicar diferentes técnicas estatísticas, como correlação e regressão para identificar possíveis tendências e relacionamentos nos dados, investigar anomalias e aplicar o algoritmo K-means para

dividir o conjunto de clientes em grupos com características similares. Portanto, será possível identificar as particularidades de cada um desses segmentos de clientes e poder oferecer produtos e serviços que atendam às necessidades de cada grupo.

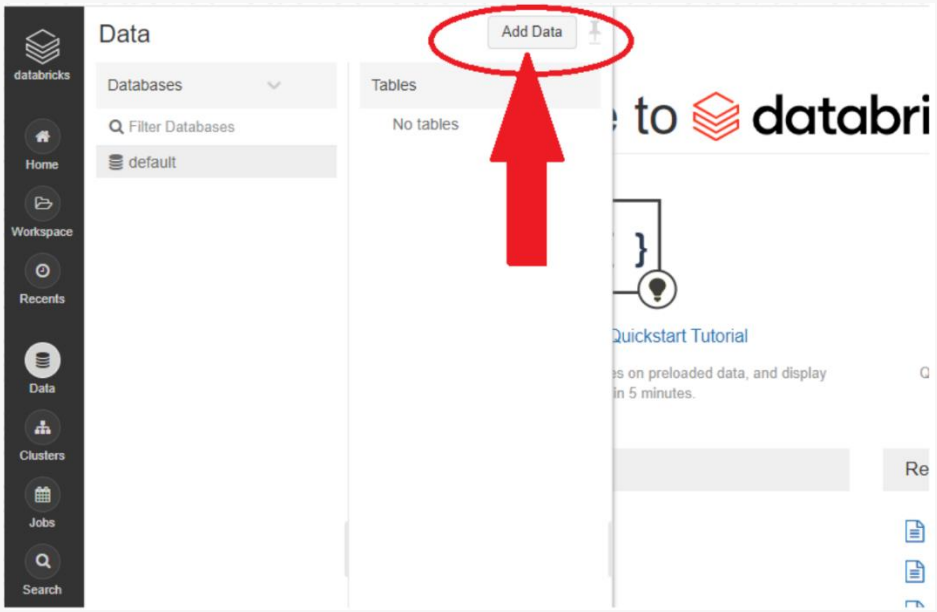
Atividades

Os alunos deverão desempenhar as seguintes atividades:

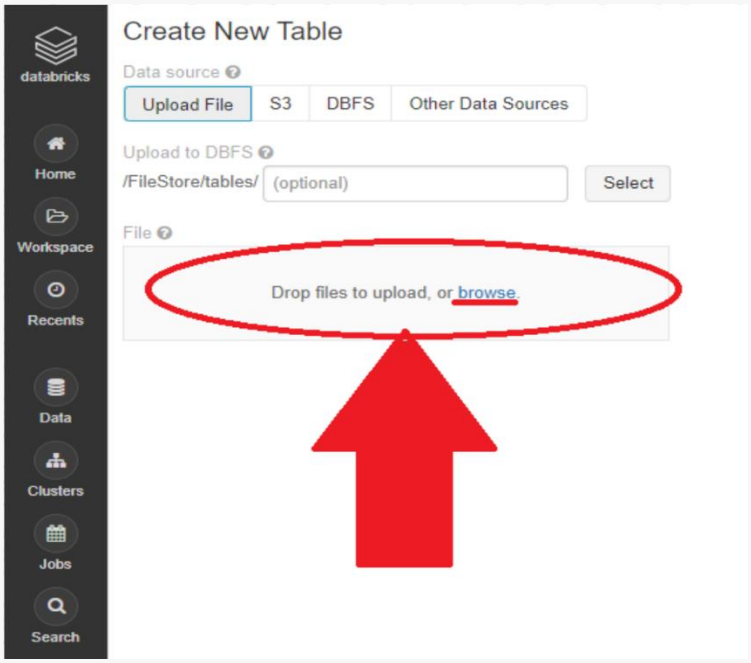
1. Acessar o site https://community.cloud.databricks.com/ e criar a conta **gratuita**. O tutorial de criação da conta **gratuita** está presente na plataforma Canvas, logo abaixo da apostila, no item "**Arquivos complementares**", com o nome "**tutorial_databricks_TPD.pdf.**
2. Acessar e baixar os arquivos "**Mall_Customers.csv**" e "**desafio_bootcamp_TPD.ipynb**", presentes na pasta: https://drive.google.com/drive/folders/1uXq_QGjGPEtvDStFx2lGy9NRP0j3cROP?usp=sharing.
3. Criar um "**Cluster**" para a atividade, seguindo os passos presentes no tutorial de criação da conta no databricks.com (arquivo **tutorial_databricks_TPD.pdf**).
4. Realizar o upload do "**Mall_Customers.csv**" para a plataforma databricks.
5. Acessar a conta criada. No menu lateral esquerdo, acesse o ícone "**Data**". A Figura 1 mostra esse ícone:



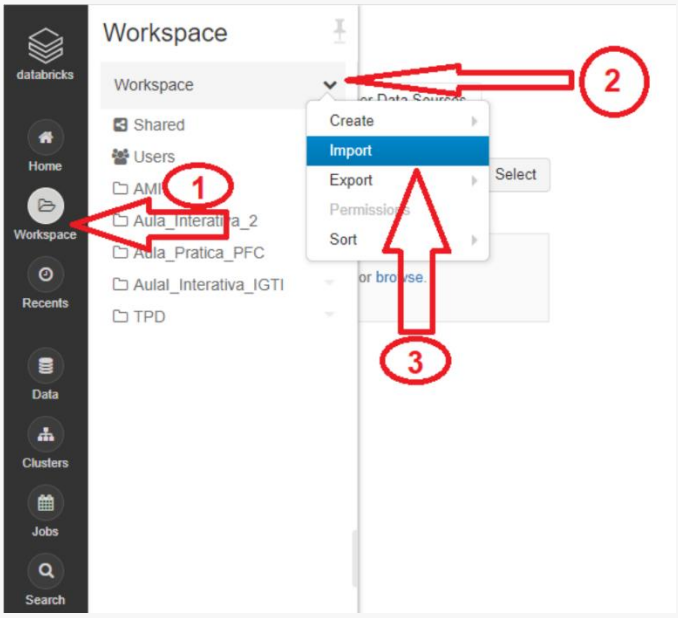
6. Na nova janela que irá abrir, clique em "**Add Data**". A Figura 2 mostra o procedimento:



7. Na tela seguinte, arraste e solte o arquivo "**Mall_Customers.csv**" dentro do campo "**Drop files to upload, or browser**". A Figura 3 mostra o procedimento:



8. Após arrastar e soltar o arquivo e o upload tiver terminado, NÃO é necessário escolher nenhuma das opções mostradas, apenas prossiga para o passo 9.
9. Após finalizado o upload do arquivo, acesse o "**Workspace**", clique na "**setinha**" no canto superior direito da aba e, depois, clique em "**Import**". A Figura 4 mostra as etapas a serem seguidas:



10. Arraste e solte o arquivo “**desafio_bootcamp_TPD.ipynb**” ou clique em “**browser**” para realizar o upload do arquivo.
11. Execute as células em sequência.
12. Responda às questões do desafio.