Simulator Télos -
Nível #2 (Trilha de dados)

Seja muito bem vindo ao Simulador - Nível #2! Neste nível, você explorou o fascinante mundo da análise de dados com PySpark, aprendendo sobre as funcionalidades essenciais do PySpark e como ele pode ser utilizado para processar e analisar grandes conjuntos de dados de forma eficiente. Está pronto para aplicar esse conhecimento em um projeto prático?

Contextualização

A análise de dados é crucial para as organizações modernas, permitindo-lhes tomar decisões informadas rapidamente. Utilizar ferramentas como PySpark para processar e analisar dados não só aumenta a eficiência operacional, mas também fornece insights valiosos que podem ser transformados em ações estratégicas.

Portanto, a capacidade de aplicar técnicas de análise de dados é fundamental para qualquer profissional no campo da ciência de dados. Neste simulador, sua squad foi contratada pela DataTech Solutions para desenvolver uma série de análises em um dataset de notas de alunos, utilizando PySpark para extrair insights que ajudarão na tomada de decisões educacionais.

Descrição do Simulador

O simulador será desenvolvido utilizando PySpark, e as tarefas serão executadas em um notebook do Google colab ou similar. Os alunos interagirão com o notebook para realizar análises de dados, manipulação de dados e visualização (opcional).

Histórias de Usuário a serem Implementadas

1) Visualização e Análise Inicial dos Dados

Como analista de dados, desejo tratar os dados enviados ao sistema para padronizá-los e assim, ser possível fazer uma análise futura, entendendo a correlação entre cada variável e tipo de dado.

Critérios de Aceitação:

- Carregar o dataset e mostrar as primeiras linhas e o esquema do DataFrame.
- Realizar uma análise descritiva básica (média, mediana, desvios).

- **DicaTélos 💡:** Utilize o método `.show()` para visualizar as primeiras linhas do DataFrame e `.printSchema()` para ver o esquema dos dados.

- **Conceitos Básicos:**
  - **Média:** A média é o valor obtido pela soma de todos os dados dividida pelo número de dados.
  - **Mediana:** A mediana é o valor que separa a metade superior da metade inferior de um conjunto de dados, ou o ponto central.
  - **Desvio Padrão:** O desvio padrão é uma medida da quantidade de variação ou dispersão dos dados. Um desvio padrão baixo significa que os dados tendem a estar próximos da média; um desvio padrão alto indica que os dados estão espalhados por uma gama mais ampla de valores.

2) Limpeza e Preparação dos Dados

Como analista de dados, preciso limpar e preparar os dados, garantindo que estejam prontos para análises mais complexas.

Critérios de Aceitação:

- Identificar e tratar valores nulos.
- Normalizar e formatar as colunas conforme necessário.

- **DicaTélos 💡:** Para tratar valores nulos, utilize o método `fillna()` para substituir nulos por um valor padrão, seja ele a média para não interferir no cálculo ou o apague, usando `dropna()` para remover linhas com valores nulos.

- **Ferramentas de Normalização:**
  - Para normalizar nomes e strings, use métodos como `lower()`, `upper()`, ou `trim()` para garantir consistência.

- **Exemplo:** `df.withColumn("coluna", F.trim(F.lower(df["coluna"])))` normaliza a coluna para letras minúsculas e remove espaços extras.

3) Análise Avançada: Relação entre Renda Familiar e Notas

Como analista de dados, quero explorar como a renda familiar influencia o desempenho dos alunos.

Critérios de Aceitação:

- ⍰ Agrupar os dados por renda familiar e calcular a média das notas.
- ≣ Analisar se existe uma correlação visível, por exemplo: alunos com mais horas de estudo, que fizeram aula particular, que se exercitam, tendem a ter uma média de notas maiores? Faça uma análise que traga esses resultados.

- **DicaTélos 💡:** Utilize `groupBy()` e `agg()` para agrupar os dados por renda familiar e calcular médias. Use a função `corr()` para explorar correlações entre renda e notas. Veja se cada grupo apresenta mudanças e caso sim, quanto impacta.

- **Conceitos de Correlação:**
  - **Correlação Positiva:** Significa que à medida que uma variável aumenta, a outra também aumenta.
  - **Correlação Negativa:** Significa que à medida que uma variável aumenta, a outra diminui.
  - **Sem Correlação:** As variáveis não mostram relação direta.

4) Visualização de Dados (Opcional)

Como analista de dados, desejo criar visualizações que ilustrem os insights dos dados, facilitando a interpretação para stakeholders.

Critérios de Aceitação:

- Utilizar funcionalidades de plotagem básica ou plataformas de visualização integradas para mostrar gráficos de distribuição de notas por categorias.

- **DicaTélos 💡:** Para visualização você pode usar `toPandas()` para converter o DataFrame do Spark para um DataFrame do Pandas e utilizar bibliotecas como Matplotlib para criar gráficos.
- Link para ajudar no desenvolvimento gráfico e uso da biblioteca: [Python - Introdução a Biblioteca Matplotlib e Numpy (P1)]

Cronograma de Desenvolvimento Proposto

- **Dia 1**: Desenvolvimento das Histórias de Usuário 1 e 2.
- **Dia 2**: Continuação da História de Usuário 2 e início da 3.
- **Dia 3**: Conclusão da História de Usuário 3 e início da 4 (opcional).
- **Dia 4**: Finalização das visualizações e revisão do projeto.

Avaliação

A avaliação considerará se os alunos conseguiram aplicar suas habilidades em PySpark para:

- Manipular e preparar dados adequadamente.
- Realizar análises exploratórias e avançadas.
- (Opcional) Criar visualizações eficazes dos dados.
- Organizar o código de forma clara e eficiente.