INTRODUÇÃO

agro-e

Parabéns Candidato!

Por ter avançado para esta etapa! Este teste tem como objetivo avaliar suas habilidades práticas em Analista de Dados, especialmente em Webscraping, SQL e Python, além de sua capacidade de lidar com problemas reais e entregar soluções eficientes.

REGRAS E ORIENTAÇÕES

1. Todas as respostas devem ser organizadas em um repositório Git (público ou privado), contendo:
   - Scripts ou projetos.
   - Documentação explicando o raciocínio adotado.
   - Testes unitários para códigos relevantes.
2. A entrega será avaliada pela originalidade, clareza e aderência às melhores práticas.
3. Respostas prontas ou genéricas que não demonstrem pensamento crítico serão penalizadas.

INSTRUÇÕES

- Utilize Python (bibliotecas como **requests**, **BeautifulSoup**, **Scrapy**) ou **N8N** para o webscraping.
- Utilize PostgreSQL para criação de tabelas e análises em SQL.
- Utilize Python com Streamlit para visualização interativa dos resultados.
- Insira e trate os dados manualmente, simulando um cenário real, com diversidade de categorias, regiões e preços.
- Estruture o processo em camadas (raw, processed, curated) como em um Data Lake.
- Documente a avaliação no GitHub, incluindo:
  - Prints das consultas em SQL.
  - Prints da aplicação em Streamlit.
  - Justificativas de modelagem, índices e otimizações propostas.

QUESTÕES WEBSCRAPING E ETL ?

agro-e

1 Coleta de Dados Externos (30 pontos)

Tarefas:

- Capture dados de uma fonte pública do agronegócio relacionada a commodities agrícolas (ex.: soja, milho, trigo, algodão, café, arroz, cana-de-açúcar, entre outros).
- Utilize Python (requests, BeautifulSoup, Scrapy) ou N8N.
- A fonte escolhida deve envolver desafios reais de scraping, como paginação, tabelas fragmentadas (caption), captchas ou inconsistências nos dados.
- Salve os dados brutos em CSV ou JSON.

---

2 Estruturação da Camada Raw (15 pontos)

Tarefas:

- Organize os arquivos brutos de forma livre (ex.: diretórios locais, diferentes formatos de arquivo).
- Explique as vantagens de formatos como CSV, JSON e Parquet e justifique sua escolha.
- Explique também como faria essa mesma organização em um bucket AWS S3.

---

3 Criação de Tabelas no PostgreSQL (20 pontos)

Tarefas:

- Crie tabelas normalizadas a partir dos dados capturados (commodities, regiões, preços, datas).
- Insira os dados coletados.
- Justifique as chaves primárias e estrangeiras utilizadas.

---

4 Tratamento e ETL (25 pontos)

Tarefas:

- Desenvolva um script em Python ou fluxo em N8N que:
- Corrija tipos de dados (datas, números).
- Trate valores ausentes.
- Padronize categorias (ex.: “soja”, “SOJA”, “Soja”).
- Carregue o resultado tratado em tabelas PostgreSQL (camada processed).

QUESTÕES SQL ?

agro-e

Estruturação do Data Lake (10 pontos)

5

Tarefas:

- Explique como estruturaria um Data Lake simples em três camadas:
- raw (dados brutos)
- processed (dados tratados)
- curated (dados prontos para análise)
- Mostre um exemplo de diretório.

Análises SQL – Tendências e Indicadores (25 pontos)

6

Tarefas:

a) Calcule o preço médio mensal por commodity, mostrando a variação percentual em relação ao mês anterior (função LAG).
b) Liste os 5 produtos mais negociados no último ano.
c) Identifique registros anômalos (ex.: preços negativos ou fora de faixa).

Otimização e Indexação (15 pontos)

7

Tarefas:

- Avalie a performance das consultas da questão 6.
- Sugira índices ou ajustes de modelagem para otimização.
- Justifique suas escolhas.

Análise Exploratória em Pandas (20 pontos)

8

Tarefas:

- Calcule estatísticas descritivas (média, mediana, desvio padrão).
- Detecte outliers.
- Exiba gráficos com Pandas/Matplotlib (boxplot, histogramas, scatter).

QUESTÕES PYTHON (STREAMLIT) ?

agro-e

Visualização em Streamlit (30 pontos)

Tarefas:

9

- Crie uma aplicação em Streamlit que:
  - Exiba dashboards com preços e tendências.
  - Permita filtros por produto, região e período.
  - Mostre gráficos interativos (linhas, barras, boxplot).

Insights e Documentação (10 pontos

Tarefas:

10

- Documente os principais padrões identificados nos dados.
- Sugira aplicações práticas para o agronegócio.
- Comente limitações da fonte utilizada.

Critérios de Avaliação

Critério	Pontos
1. Coleta de Dados Externos (Webscraping)	30
2. Estruturação da Camada Raw	15
3. Criação de Tabelas no PostgreSQL	20
4. Tratamento e ETL	25
5. Estruturação do Data Lake	10
6. Análises SQL – Tendências e Indicadores	25
7. Otimização e Indexação	15
8. Análise Exploratória em Pandas	20
9. Visualização no Streamlit	30
10. Insights e Documentação	10
Total	**200**

Instruções para Aplicação

- Use um ambiente controlado.
- Tempo limite: 96 horas.
- **A avaliação deve ser entregue e documentada no GitHub, com prints das consultas e do Streamlit.**