Teste Prático - Engenheiro(a) de Dados Pleno

Objetivo:

Avaliar habilidades em orquestração de pipelines, integração de dados, transformação e sustentação de processos de dados.

Cenário:

Você foi contratado para criar um pipeline de dados que extrai informações públicas, processa e prepara para análise.

Tarefas:

1. Ingestão de Dados:

- o Extraia dados de uma API pública, como:
 - JSONPlaceholder (dados de usuários, posts, etc.)
 - Dados Abertos do IBGE
 - OpenWeatherMap (dados climáticos)
- Salve os dados brutos em arquivos no formato JSON ou CSV.

2. Armazenamento:

- Armazene esses arquivos localmente ou em um bucket S3 na sua conta pessoal (caso você tenha acesso à AWS Free Trier).
- o Caso não queira usar a AWS, o armazenamento pode ser local.

3. Transformação de Dados:

- O Use Python com Pandas ou PySpark para:
 - Limpar e transformar os dados (ex: filtrar colunas, normalizar campos, criar colunas derivadas).
 - Converter para um formato otimizado, como Parquet ou CSV particionado.

4. Orquestração:

- Crie uma DAG no Apache Airflow (pode ser executado localmente com Docker ou ambiente próprio) que automatize as seguintes etapas:
 - Extração dos dados
 - Transformação e limpeza
 - Armazenamento do dataset final

5. Preparação para Data Visualization:

- O dataset final deve estar em um schema que possa ser facilmente conectado a uma ferramenta de Data Visualization (ex: Power BI, Tableau, Looker, etc.).
- Foque na organização e modelagem dos dados para garantir facilidade de uso, performance em consultas e integridade relacional (quando aplicável).

Critérios de Avaliação:

- 1. **Qualidade do Código:** Organização, legibilidade, modularização e boas práticas.
- 2. **Orquestração:** Configuração correta da DAG no Airflow, com dependências bem definidas.
- 3. Processamento de Dados: Eficiência na transformação e modelagem dos dados.
- 4. **Resiliência:** Tratamento de erros e logs básicos no pipeline.
- 5. Documentação:
 - o README explicando o pipeline, tecnologias usadas e instruções de execução.

ntrega:

- Código hospedado em um repositório Git (público ou privado com acesso compartilhado).
- Inclua o README com:
 - Descrição do fluxo de dados
 - Tecnologias utilizadas
 - o Instruções passo a passo para rodar o pipeline

Gravação da Tela:

- Durante a execução do teste, o candidato deverá gravar a tela mostrando todo o processo de desenvolvimento (pode usar qualquer ferramenta de gravação, como OBS Studio, Loom, etc.).
- o O arquivo da gravação deverá ser enviado junto com o código.
- O vídeo deve mostrar claramente o ambiente de desenvolvimento, comandos executados e etapas do raciocínio.
- Prazo de entrega: 48 horas após o recebimento do teste.