grão direto

Engenheiro de Dados – Exercício Técnico

Exercício

Ingestão, Transformação e Consumo de Dados em Arquitetura Medallion usando Databricks

INSTRUÇÕES

1. SUAS RESPOSTAS AO EXERCÍCIO SERÃO AVALIADAS APÓS A ENTREGA POR E-MAIL.

2. CASO ACHE NECESSÁRIO, MONTE UMA APRESENTAÇÃO (POWERPOINT, PDF, GOOGLE SLIDES OU COM PRÓPRIO BI...) PARA FACILITAR SUAS EXPLICAÇÕES.

3. CASO NECESSISTE, INFORMAÇÕES NÃO CONTIDAS NO DESAFIO PODERÃO SER SUPOSTAS.

4. CASO NÃO CONSIGA CONCLUIR TOTALMENTE O DESAFIO, NÃO SE PREOCUPE, QUEREMOS APENAS VER ATÉ ONDE VOCÊ CONSEGUE CHEGAR!

grãodireto

Suponha que você foi contratado por uma empresa fictícia chamada Grain Logistic para **estruturar o pipeline de dados** que alimenta as análises sobre entregas e envios de pacotes. A empresa tem múltiplas fontes de dados e enfrenta problemas de qualidade, duplicação e falta de padronização. Seu papel é organizar todo o fluxo de dados seguindo o conceito de "**Medallion Architecture**" (Bronze, Silver, Gold) em uma plataforma **Databricks**.

ESCOPO DO DESAFIO

➢ **Ingestão (Bronze):**

- Criar uma estrutura Bronze em um Data Lake (Databricks File System ou similar) para receber os dados brutos sem grandes transformações;
- Garantir que todos os dados sejam armazenados;
- Registrar metadados como data/hora da ingestão

➢ **Limpeza e Padronização (Silver):**

- Criar a camada Silver, padronizando esquemas, removendo duplicatas, lidando com campos ausentes, corrigindo tipos de dados, etc.

➢ **Enriquecimento e Agregações (Gold):**

- Criar a Gold para consumo analítico e dashboards;
- Nesse estágio, é esperado realizar agregações, cálculos de métricas, junções entre tabelas e criação de dimensões e fatos se apropriado;
- Demonstrar queries analíticas que possam ser facilmente usadas por times de BI ou Data Science.

REQUISITOS OBRIGATÓRIOS

Além dos três níveis de transformação, o candidato deve implementar:

- **Data Quality:**

Aplicar checagens de qualidade nos dados durante os processos de ingestão e transformação. Alguns exemplos:

- Verificação de schema: garantir que os dados seguem o formato esperado;
- Detecção de valores nulos/anômalos;
- Monitoramento de duplicatas;
- Aplicação de regras de negócio para validar consistência dos dados.

- **Testes:**

Criar testes automatizados para validar a pipeline, incluindo:

- Testes de validação de schema;
- Testes para garantir tratamento adequado de valores nulos e duplicatas;
- Testes de integração para validar a movimentação dos dados entre camadas.

BÔNUS (DIFERENCIAIS)

Dois itens bônus para enriquecer o desafio, que não são obrigatórios, mas serão grandes diferenciais:

- **Consumo e Visualização:**

Integração e criação de um Power BI gerando um relatório simples.

- **Uso do Unity Catalog:**

Demonstrar como o Unity Catalog pode ser utilizado para governança e controle de acesso.

grãodireto

THANK YOU AND GOOD LUCK!