Teste para Engenheiro de Dados

Objetivo:
Criar um pipeline de dados que processe informações brutas, armazene os dados em um Data Warehouse (real ou simulado). Em seguida calcular as métricas indicadas e desenvolver - um dashboard para visualização de dados.

Cenário do Teste

Você trabalha em uma empresa que recebe um arquivo CSV diário com informações de transações financeiras referentes a um ativo. O arquivo possui as seguintes colunas:

- `transaction_id` (STRING): Identificador único da transação.
- `customer_id` (STRING): Identificador único do cliente.
- `transaction_date` (DATE): Data da transação.
- `transaction_amount` (FLOAT): Valor da transação.
- `transaction_status` (STRING): Status da transação (`approved`, `rejected`, `pending`).
- `transaction_type` (STRING): Tipo de transação (`buy,sell`), representando compra ou venda.
- `qtty` (FLOAT): Quantidade de unidades transacionadas.
- `price` (FLOAT): Preço unitário..

Sua tarefa é criar um pipeline de dados que:

1. Armazene os dados brutos em um sistema de armazenamento (local ou em cloud).
2. Processe e limpe os dados, eliminando inconsistências, como valores nulos ou incorretos.
3. Carregue os dados em um Data Warehouse.
4. Desenvolva consultas para responder às perguntas do negócio:
    - Qual o total de transações aprovadas por mês?
    - Qual cliente teve o maior volume de transações aprovadas nos últimos 3 meses?
    - Qual a média de transações rejeitadas por mês no último ano?
    - Qual o preço médio do estoque do ativo em questão, desconsiderando a abertura de clientes, e como ele se comporta conforme as transações são realizadas.
5. Crie um dashboard com visualizações dos resultados.

Instruções

Você pode escolher entre duas opções para realizar o teste:

1. **Utilizando o GCP (Nível Gratuito)**
    - Crie uma conta gratuita no Google Cloud Platform. O GCP oferece créditos iniciais de $300 e recursos gratuitos, como:
        - **BigQuery:** 1 TB de consultas e 10 GB de armazenamento por mês.
        - **Cloud Storage:** 5 GB de armazenamento.
    - Realize o teste utilizando os serviços abaixo:
        - **Cloud Storage** para armazenar o CSV.
        - **Dataflow** ou **Cloud Functions** para processar os dados.
        - **BigQuery** como Data Warehouse.
        - **Looker Studio** para criar o dashboard.
2. **Realizando Localmente (Simulação)**
    - Caso não tenha acesso ou prefira não usar o GCP, você pode realizar o teste localmente, utilizando ferramentas equivalentes:
        - Utilize uma pasta local para simular o **Cloud Storage**.
        - Use **Apache Beam** ou scripts Python para simular o pipeline de processamento.
        - Armazene os dados em um banco de dados local, como **PostgreSQL** ou **SQLite**, simulando o BigQuery.
        - Crie o dashboard em ferramentas como **Tableau Public** ou **Power BI**, simulando o Looker Studio.

Entregáveis

- **Pipeline de Dados:**
    - Código do pipeline (em Python).
- **Estrutura de Dados:**
    - Script SQL para criar e popular a tabela no Data Warehouse (BigQuery ou banco local).
- **Consultas SQL:**
    - Consultas que respondem às perguntas do cenário.
- **Dashboard:**
    - Link do dashboard no Looker Studio (ou arquivo equivalente em Tableau/Power BI).
- **Documentação:**
    - Explique a arquitetura do pipeline, as ferramentas utilizadas e as decisões técnicas tomadas.
    - Inclua instruções claras para replicar o teste.

Critérios de Avaliação

1. **Estrutura do Pipeline:**
    - Uso de boas práticas no pipeline (processamento e carregamento).
    - Clareza e eficiência do código.
2. **Modelagem de Dados:**
    - Organização da tabela e qualidade das consultas SQL.
3. **Visualização de Dados:**
    - Clareza, organização e usabilidade do dashboard.
4. **Documentação:**
    - Qualidade da explicação técnica e facilidade para replicação.

Prazo de entrega:

- 72 horas.
