Informações Gerais Teste

O objetivo do teste é avaliar habilidades técnicas e comportamentais na criação de pipelines de dados para atender a necessidade de negócio de disponibilizar dados úteis para consumo.

O teste tem duração de 1 semana. As instruções, dúvidas e entregas serão feitas via email.

Os requisitos de entrega são:

- código fonte em linguagem de programação (preferência PySpark);
- desenho da arquitetura utilizando dados armazenados na AWS e sendo consumidos no Snowflake;
- documento (estilo READ.me) descrevendo principais perguntas levantadas, decisões tomadas e justificativas.

Cenário Proposto

Você atua em uma empresa que precisa consolidar informações de vendas, produtos e categorias econômicas para análise estratégica no desenvolvimento de produtos. Sua missão é construir uma solução de pipeline de dados que extraia dados de diferentes fontes, transforme-os conforme necessidade e armazene os dados prontos para criação de análises refinadas a partir de um Data Lake.

Você recebeu os seguintes requisitos iniciais:

1. Extrair dados de duas fontes:
    a. Arquivos CSV de tabelas relacionadas a pedidos, produtos, pagamentos e clientes do Kaggle
    Disponível aqui: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data

    b. Uma API REST com dados descrição das categorias econômicas dos códigos CNAE
    Disponível aqui:
    https://servicodados.ibge.gov.br/api/docs/CNAE?versao=2#api-Classes-classesGet

2. Transformar dados para:
    a. Consolidar vendas por estado e categoria de produto;
    b. Enriquecer os dados dos vendedores com as descrições dos códigos CNAE;
    c. Categorizar produtos em faixas de valor médio (baixo, médio, alto);

d. Calcular o valor total de vendas por estado.

3. Carregar os dados processados em uma ou mais tabelas no formato Delta a serem integradas em uma plataforma de consumo.

Tarefas

- **Modelagem do Pipeline:** Desenhe ou descreve a arquitetura mais adequada para atender as necessidades do projeto e negócio;
- **Desenvolvimento:** Construa o pipeline utilizando técnicas de ETL/ELT como Python, Airflow, etc. Busque incluir boas práticas de tratamentos e monitoramento de dados;
- **Entrega:** Script que retorna tabela delta;
- **Perguntas e Reflexões:** Para incluir no READ.me:
  Quais perguntas você faria para o time de analistas de negócio ou para os stakeholders sobre os requisitos?
  Que outras melhorias ou extensões poderiam ser feitas nesse pipeline?
  Quais diferentes técnicas ou ferramentas você considera que poderiam ser utilizadas para resolver o mesmo problema?