Desafio de Engenharia de Dados

Introdução

Este desafio busca simular um projeto de engenharia de dados de um banco fictício chamado Banco Vitória que tem o objetivo de amadurecer na sua cultura e no uso de dados dentro da organização. Você será avaliado de acordo com sua capacidade de raciocínio lógico, programação e orquestração de pipelines de dados.

Contexto

BanVic - A Jornada dos Dados Financeiros. O Banco Vitória S.A., também conhecido como BanVic, foi fundado em São Paulo em 2010, com uma visão inovadora de oferecer serviços bancários eficientes tanto em agências físicas quanto no ambiente digital. Com uma equipe de 100 colaboradores dedicados, o BanVic cresceu para se tornar uma instituição financeira nacional de destaque.

O banco sempre foi focado em proporcionar aos clientes experiências bancárias transparentes e convenientes. No entanto, à medida que a instituição expandiu suas operações e serviços, surgiu a necessidade de aprimorar a compreensão de seus dados para impulsionar ainda mais a excelência em seus serviços.

Nossa história começa quando a CEO do BanVic, Sofia Oliveira, percebe que utilizar dados para a tomada de decisão é a chave para elevar o banco a novos patamares. Ela acredita que entender profundamente as operações e comportamentos dos clientes pode levar a melhorias significativas nos serviços oferecidos.

Sofia convoca uma reunião com a equipe de liderança, incluindo o Diretor de Tecnologia, André Tech, a Diretora Comercial, Camila Diniz, e o recém-contratado Analista de Dados, Lucas Johnson. Cada um deles traz perspectivas únicas para a mesa.

André Tech, o especialista em tecnologia, está animado com a ideia de implementar técnicas avançadas de análise de dados para otimizar as operações internas do banco.

Há tempos André e sua equipe fazem análises manuais para o banco e ele não gostaria de seguir dessa forma por já conhecer os riscos deste formato.

Camila Diniz, por outro lado, não está convencida que este é o caminho. Ela acredita que o BanVic pode investir mais em marketing e melhorar a segmentação dos clientes nas cidades que o banco já está estabelecido, sendo esse um caminho mais rápido e já conhecido pelo BanVic. Sua postura pode colocar em risco o projeto, pois sua equipe hoje detém parte dos dados comerciais importantes para a estruturação digital da empresa e isso pode acarretar em burocracias e atrasos em relação a acessos e permissões.

Por fim, Lucas Johnson, apaixonado por dados, propõe um projeto piloto para compreender os dados de crédito do BanVic, a fim de iniciar a jornada de dados gerando valor para uma área crítica do banco e convencer a Diretora que essa iniciativa pode ser muito benéfica para a companhia.

A equipe concorda que a implementação de um projeto de análise de dados bem-sucedido pode proporcionar insights valiosos, melhorando a eficiência operacional e a experiência do cliente. No entanto, todos são conscientes dos desafios técnicos e da importância de escolher as ferramentas certas.

Em um servidor da nuvem estão os dados do ERP, CRM e marketing. Atualmente as análises do BanVic são realizadas em planilhas e apresentações, sendo que não possuem nada em ferramentas de BI, mas estão abertos a utilizar ferramentas como Metabase, Looker, PowerBI, entre outros.

Para que o projeto de análise de dados seja possível será necessário a centralização dos dados das diferentes fontes em um Data Warehouse diariamente.

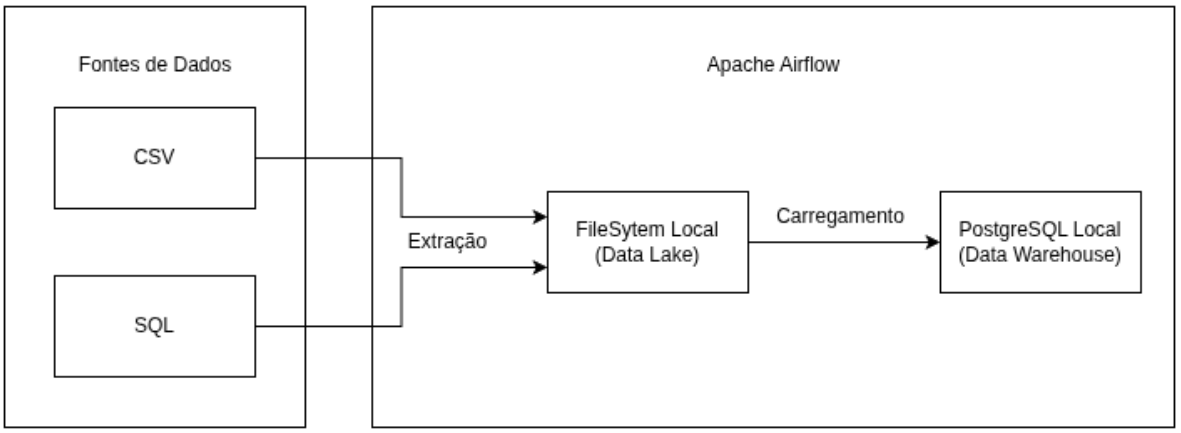
Arquivos

- 1 CSV com dados de transações;
- 1 arquivo SQL com as tabelas mais importantes para a análise de dados;
- 1 docker-compose.yml com instruções para subir o banco de dados **fonte**.

Desafio

Neste desafio você deverá realizar a extração e carregamento dos dados de ambas as fontes fornecidas (CSV e SQL) para um Data Warehouse **local** (PostgreSQL). Essa tarefa deverá respeitar os requisitos e diagrama abaixo:

Diagrama



Requisitos

- Utilize o Apache Airflow (2 ou 3) como orquestrador de tarefas;
- As extrações devem ser idempotentes;
- Devem ser extraídos todos os dados fornecidos;
- As extrações devem escrever os dados no formato CSV para seu FIleSystem Local seguindo o padrão de nomenclatura:
    - <ano>-<mês>-<dia>/<fonte-de-dados>/<nome-da-tabela-ou-csv>.csv
- As etapas de extração de dados devem ocorrer uma em paralelo à outra;
- A etapa de carregamento no Data Warehouse deve ocorrer somente se ambas extrações tenham sucesso;
- O pipeline deve ser executado todos os dias às 04:35 da manhã;
- O projeto deve ser reproduzível em outros ambientes.

Entregas

1. Todos os códigos e configurações utilizados para a execução do projeto devem ser compactados num arquivo **.zip** e anexados na entrega.
2. Um documento descritivo com instruções para execução do projeto em outros ambientes.
3. Um **vídeo** curto explicando a lógica utilizada em sua DAG do Airflow bem como uma execução manual da mesma e os resultados obtidos a partir da execução

do pipeline (arquivos escritos no formato desejado e dados adicionados ao Data Warehouse).

Prazos

- Você tem até 7 dias corridos para a entrega, contados a partir do recebimento deste desafio. O não cumprimento deste prazo implica na desclassificação do processo seletivo.
- A Indicium possui ferramentas avançadas de detecção de plágio e inteligência artificial. A utilização de IA implica na desclassificação do processo seletivo.
- Envie o seus entregáveis dentro da sua data limite para o email: selecao.lighthouse@indicium.tech

O arquivo de entrega deve ser nomeado como: LH_DE_SEUNOME

Bom trabalho!