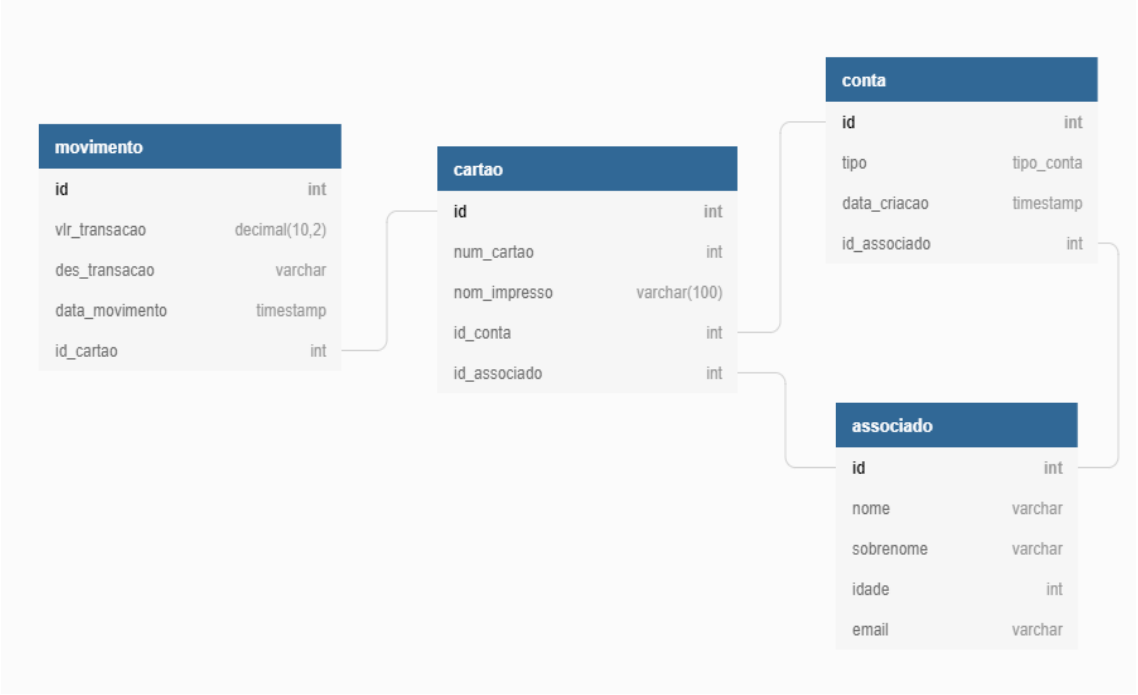
Desafio técnico engenharia de dados

Problema

Digamos que você faz parte da equipe de engenharia de dados e uma empresa, chamada SiCooperative LTDA, que tem como propósito oferecer soluções financeiras de maneira justa para as pessoas. Porém, um grande problema estava impactando a velocidade e a assertividade na tomada de decisões por parte de gestores e diretores. Esse problema pode ser explicado, principalmente, na dificuldade em agregar diferentes informações em um único ponto, muito tempo estava sendo perdido em criar relatórios individuais, e após isso, tentar correlacionar eles manualmente. Além disso, uma nova equipe de Data Science estava sendo montada com o objetivo de criar modelos preditivos que tenham a capacidade de analisar o momento atual de cada associado(cliente), e dessa forma oferecer soluções mais coerentes com a realidade de cada associado. Após uma reunião estratégica, foi decidido por iniciar uma POC para criação de um **Data Lake,** com objetivo de resolver os problemas já citados. Após isso, foi feito um mapeamento inicial para definir quais eram as prioridades de ingestão, ou seja, quais seriam as primeiras informações a serem trazidas para a nova estrutura, e como resultado os dados de movimentações dos cartões, foi a escolhida. Ela contém informações de cada **movimentação** feita por cartões dos associados, com informações de valor e data da movimentação. Além disso, também devem ser carregadas as tabelas na qual ela faz relacionamento: a tabela de **cartão** que contém informações adicionais do cartão**,** a **conta** que aquele cartão e associado estão vinculados, e informações do **associado**. Um associado pode ter vários cartões e várias contas. O diagrama dessa estrutura está organizado da seguinte forma:



Desafio

Você deve primeiramente modelar a estrutura descrita anteriormente em um banco de dados de sua preferência, após isso criar um componente capaz de ler os dados das tabelas criadas e escrever e um único arquivo flat, em um diretório local, ou seja, todas as tabelas devem ser agregadas para formar uma visão única, onde a estrutura do arquivo é representada da seguinte forma:

movimento_flat	
nome_associado	string
sobrenome_associado	string
idade_associado	string
vlr_transacao_movimento	string
des_transacao_movimento	string
data_movimento	string
numero_cartao	string
nome_impresso_cartao	string
data_criacao_cartao	string
tipo_conta	string
data_criacao_conta	string

Requisitos:

- Criar a estrutura do banco a ser ingerido, utilizando a tecnologia de sua preferência (MySQL, PostgreSQL).
- Inserir uma massa de dados fictícia nas tabelas, não precisa ser um volume tão grande.
- Utilizar a linguagem de programação de sua preferência.
- Para realizar o ETL dos dados, deve ser utilizado algum framework de processamento distribuído para Big Data, Ex: Hadoop, Spark, Flink e Storm.
- Escrever um arquivo CSV, em um diretório parametrizado pelo usuário.
- Disponibilizar esse projeto em um repositório privado no seu GitHub.

Bônus:

- Provisionar o ambiente e realizar uma execução inicial do processo de maneira automatizada utilizando Docker.

- Criar testes unitários para cobrir as principais partes do fluxo.

Observações

Ao criar o seu projeto você pode documentar tudo no **README**, porquê optou por determinado design, o que faria se tivesse mais tempo para concluir o desafio, dificuldades que encontrou no desenvolvimento. Importante salientar, que todos os scripts utilizados para criação das estruturas devem ser salvos dentro do seu projeto.