Data Engineer - Desafio

Objetivo do Desafio

A receita federal disponibiliza alguns dados abertos sobre as empresas existentes hoje no Brasil. Esses dados são de domínio público e livre acesso, porém eles podem sofrer um desatualização de até 3 meses. O objetivo é avaliar como você trabalha com dados cadastrais e entende os levantamentos de requisitos feitos por outros times técnicos ou de negócios. Você deverá realizar a ingestão de duas tabelas, a partir de um end-point e então irá realizar um processamento de dados para chegar em um output esperado. Para isso siga as seguintes instruções:

End-point: https://arquivos.receitafederal.gov.br/dados/cnpj/dados_abertos_cnpj/?C=N;O=D

Fazer a chamada de UMA das pastas Empresas.ZIP

Processar os dados desses arquivos e disponibilizar em uma tabela seguindo o seguinte schema:

Nome da coluna	Tipo de dado	Descrição
cnpj	string	CONTEM O NÚMERO DE INSCRIÇÃO NO CNPJ (CADASTRO NACIONAL DA PESSOA JURÍDICA).
razão_social	string	CORRESPONDE AO NOME EMPRESARIAL DA PESSOA JURÍDICA
natureza_juridica	int	CÓDIGO DA NATUREZA JURÍDICA
qualificacao_responsavel	int	QUALIFICAÇÃO DA PESSOA FÍSICA RESPONSÁVEL PELA EMPRESA
capital_social	float	CAPITAL SOCIAL DA EMPRESA
cod_porte	string	CÓDIGO DO PORTE DA EMPRESA

Fazer a chamada de UMA das pastas Socios.ZIP

Processar os dados desses arquivos e disponibilizar em uma tabela seguindo o seguinte schema:

Nome da coluna	Tipo de dado	Descrição
cnpj	string	CONTEM O NÚMERO DE INSCRIÇÃO NO CNPJ (CADASTRO NACIONAL DA PESSOA JURÍDICA).
tipo_socio	int	IDENTIFICADOR DE SOCIO
nome_socio	string	CORRESPONDE AO NOME SOCIO PESSOA FISICA, RAZÃO SOCIAL E/OU NOME EMPRESARIAL DA PESSOA JURÍDICA E NOME DO SÓCIO/RAZAO SOCIAL DO SOCIO ESTRANGEIRO
documento_socio	string	É PREENCHIDO COM CPF OU CNPJ DO SOCIO, NO CASO DE SÓCIOESTRANGEIRO É PREENCHIDO COM 'NOVES' O ALINHAMENTO PARA CPF É FORMATADO COM ZEROS À ESQUERDA.
codigo_qualificacao_socio	string	CODIGO QUALIFICACAO SOCIO

Com a ingestão realizada você deve tratar seus dados e disponibilizar uma tabela com o seguinte formato:

Nome da coluna	Tipo de dado	Descrição
cnpj	string	CONTEM O NÚMERO DE INSCRIÇÃO NO CNPJ (CADASTRO NACIONAL DA PESSOA JURÍDICA).

qtde_socios	int	NUMERO DE SOCIOS PARTICIPANTES NO CNPJ
flag_socio_estrangeiro	boolean	True: Contem pelo menos 1 sócio estrangeiro False: Não contém sócios estrangeiros
doc_alvo	boolean	True: Quando porte da empresa = 03 & qtde_socios > 1 False: Outros

Requisitos Técnicos

- Utilizar Python em sua última versão para o desenvolvimento;
- O arquivo utilizado como raw pode estar no projeto do github ou em qualquer fonte que o código possa baixar para processar; Faça um bom uso das técnicas de engenharia de dados e software;
- Utilize o modelo medalhão (camadas bronze, silver e gold) para o processamento e descreva o que é cada camada nesse projeto; Todo ambiente deve ser containerizado (Docker) e preferencialmente executado com todos serviços dependentes em um único comando;
- Para o armazenamento do output, escolha o banco que achar mais adequado pensando que o foco seria um banco que permita aplicações transacionais plugarem, seja SQL ou NoSQL;
- Use frameworks com sabedoria, sem poluir seu código;
- Ter um repositório organizado, documentação e fácil de utilizar, um ou poucos comandos para levantar todo ambiente e suas dependências;
- Deve estar funcional e cumprir os requisitos que foram apresentados no desafio;
- O repositório deve ser públlico;

Entrega do desafio

A entrega deverá ser feita através de um repositório público criado no seu GitHub contendo todo o seu código final na branch **main**. O link pode ser enviado para a pessoa que está tocando o seu processo seletivo. O time técnico irá fazer a avaliação e em seguida, iremos entrar em contato para dar um feedback em cima do seu desafio. Use e abuse da sua criatividade e tudo que quiser.