ROITBANK

Prova técnica – Engenheiro de Dados

Pipeline de Processamento para Dados Públicos de CNPJ

Esse teste tem como intuito avaliar conhecimentos básicos, sendo que o principal

objetivo é analisar o raciocínio lógico e a capacidade de solucionar problemas,

mesmo que ainda desconhecidos pelo desenvolvedor.

O problema proposto é o desenvolvimento de um pipeline de dados que realiza o

processamento dos dados de empresas e estabelecimento da base pública de

CNPJs da Receita Federal e disponibiliza em um banco de dados estruturados para

consumo.

O dataset possui 3 arquivos principais:

• empresa-part-00.csv e estab-part-00.csv são os dados a serem trabalhados.

schema-tabelas.pdf contém o descritivo de cada tabela.

Download: dataset-test-dados.zip

1. Realizar processamento nas tabelas de estabelecimentos e empresas,

ambas as tabelas possuem uma série de inconsistências, utilize seus

conhecimentos em engenharia de dados para disponibilização de um dado

qualificado para consumo.

A principal parte do teste é a limpeza e enriquecimento dos dados,

os processamentos realizados devem ser legíveis e

explicáveis (baseado em números do dataset).

Na tabela de estabelecimentos podemos gerar a coluna de CNPJ

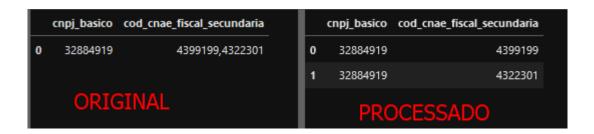
com as seguintes colunas: cnpj_basico, cnpj_ordem, cnpj_dv.

Nem todos os CNPJS presentes na base de dados são válidos, fica

a encargo do desenvolvedor encontrar meios de validação.



 Na tabela de estabelecimentos a coluna cod_cnae_fiscal_secundaria pode possuir uma ou mais informações. Remover a coluna da tabela principal e transformar em uma tabela auxiliar.



- 2. Disponibilizar a base enriquecida em um banco de dados SQL.
- 3. A partir do banco de dados criado montar as seguintes views com SQL:
 - Quantidade de estabelecimentos que não possuem
 cod_cnae_fiscal_secundaria e estão presentes no estado do Paraná;
 - Quantidade de estabelecimentos que possuem cod_cnae_fiscal_secundaria igual a 4530703;
 - Quantidade de estabelecimentos que possuem
 cod_cnae_fiscal_secundaria igual a 4530703 e estão na tabela de empresa
 - Top 10 estabelecimentos que possuem a maior quantidade de cod_cnae_fiscal_secundaria e estão presentes no estado de Santa Catarina e São Paulo. Trazer as colunas de cnpj, estado, qtde_cnae_secundario.
- 4. Hospedar o código em um **repositório público** do github utilizando conceitos de Gitflow e Commit Semântico, criar um REAME.md sobre o código.

Observações:

 O software pode ser desenvolvida em qualquer linguagem (Recomendamos Python);



- O desenvolvedor pode se sentir à vontade para tirar eventuais dúvidas com os avaliadores;
- O desenvolvedor deve ser capaz de explicar o código desenvolvido.
- Após a finalização encaminhar o link do repositório para o email:

employee.experience@roit.com.br

"O simples é melhor que o complicado" - Jorge Paulo Lemann