Processo Seletivo

Data Engineer: desafio técnico

Desafio pipeline de dados: instruções e contexto

Prazo: 7 dias a partir do envio do email com o desafio
Para enviar a resposta, basta responder este email. **Não há restrição de formato.**

Objetivo: estruturar um pipeline de dados definindo: modelagem dos dados, tecnologias utilizadas, etapas de tratamento.
Contexto: o objetivo do pipeline é o tratamento e reestruturação dos dados que permita a realização facilitada de análises e consultas

A implementação do pipeline é um requisito para a realização do desafio

Desafio pipeline de dados: etapas e perguntas

Etapa 1: Configurar um banco de sua preferência:

- Escolha um banco de dados (ex.: PostgreSQL)
- Se preferir, pode subir em ambientes cloud (ex.: S3 + Athena)

Etapa 2: Subir os dados no banco:

- Baixar os microdados mais recentes do Censo Escolar do INEP - http://inep.gov.br/microdados
- Descompactar o zip baixado e subir todos os arquivos que comecem com “MATRICULA_”

Etapa 3: Com base nos dados brutos, crie tabelas consolidadas ou DW que ajudem a respondam essas perguntas:

- Sabendo que um dos 3 requisitos que temos para estudar na Trybe é ter o ensino médio completo, quais são os 10 municípios (pode ser apenas a sigla) que têm o maior número de pessoas no “Ensino Fundamental de 9 anos - 9º Ano”?
- Um dos nossos principais valores é a diversidade. É muito importante sabermos a distribuição de pessoas de todas as cores/raças para diminuir igualdades. Qual a distribuição de cores/raças (Branca, Pretas, Pardas, Amarelas e Indígenas) entre os estados (pode ser apenas a sigla)?

Desafio pipeline de dados: etapas e perguntas

Etapa 4: Com base no que definir nas etapas anteriores, responda às seguintes perguntas:

- Qual arquitetura você definiu?
- Qual o schema da(s) tabela(s) consolidada(s)?
- Com base na(s) tabela(s) consolidada(s), quais queries você utilizou para responder às perguntas acima?
- Quais dados você incluiria na(s) tabela(s) consolidada(s) e que seriam interessantes para as pessoas analistas usarem? e qual a razão delas?
- Caso os dados fossem 10.000 vezes maior, você manteria a mesma arquitetura? porque?

betrybe.com