DESAFIO BIG DATA/MODELAGEM

📌 ESCOPO DO DESAFIO

Neste desafio serão feitas as ingestões dos dados que estão na pasta /raw onde vamos ter alguns arquivos .csv de um banco relacional de vendas.

- VENDAS.CSV

- CLIENTES.CSV

- ENDERECO.CSV

- REGIAO.CSV

- DIVISAO.CSV

Seu trabalho como engenheiro de dados/arquiteto de BI é prover dados em uma pasta desafio\_curso/gold em .csv para ser consumido por um relatório em PowerBI que deverá ser construído dentro da pasta 'app' (já tem o template).

**OBS. Quem quiser entregar a solução em outro banco (Hive, Hbase, Kudu), terá que montar a documentação para tal. Com o passo a passo e scripts de automação.**

📑 ETAPAS

Etapa 1 - Enviar os arquivos para o HDFS

- Nesta etapa lembre de criar um shell script para fazer o trabalho repetitivo (não é obrigatório)

Etapa 2 - Criar o banco DEASFIO\_CURSO e dentro tabelas no Hive usando o HQL e executando um script shell dentro do hive server na pasta scripts/pre\_process.

- DESAFIO\_CURSO (nome do banco)

- TBL\_VENDAS

- TBL\_CLIENTES

- TBL\_ENDERECO

- TBL\_REGIAO

- TBL\_DIVISAO

Etapa 3 - Processar os dados no Spark Efetuando suas devidas transformações criando os arquivos com a modelagem de BI.

OBS. o desenvolvimento pode ser feito no jupyter porem no final o codigo deve estar no arquivo desafio\_curso/scripts/process/process.py

Etapa 4 - Gravar as informações em tabelas dimensionais em formato cvs delimitado por ';'.

- FT\_VENDAS

- DIM\_CLIENTES

- DIM\_TEMPO

- DIM\_LOCALIDADE

Etapa 5 - Exportar os dados para a pasta desafio\_curso/gold

Etapa 6 - Criar e editar o PowerBI com os dados que você trabalhou.

No PowerBI criar gráficos de vendas.

* Indicadores de quantidade de vendas
* Vendas por cliente
* Vendas por ano
* Vendas por estado
* Valor total de vendas

Etapa 7 - Criar uma documentação com os testes e etapas do projeto.

**REGRAS**

Campos de strings vazios deverão ser preenchidos com 'Não informado'.

Campos decimais ou inteiros nulos ou vazios, deverão ser preenchidos por 0.

Atentem-se a modelagem de dados da tabela FATO e Dimensão.

Na tabela FATO, pelo menos a métrica <b>valor de venda</b> é um requisito obrigatório.

Nas dimensões deverá conter valores únicos, não deverá conter valores repetidos.

Para a dimensão tempo considerar o campo da TBL\_VENDAS <b>Invoice Date</b>

INSTRUÇÕES

vocês deveram me entregar o projeto no github e por email (zip)

meu email: cgomesf@minsait.com

nome do email: DESAFIO MINSAIT BI/BIGDATA (Aluno)

dentro deste email o zip e o link para o github onde estará o projeto.

prazo limite: até <b>24/06/2023<b>

Apos esta data não poderei considerar para a nota.