Avaliação N2 – Questões 2 e 3

Nome: José Victor de Farias
RA: 1328701
Disciplina: Análise Preditiva
Professor: Luiz Carlos Camargo

Questão 2 – Implementação do Ambiente de Dados Estruturados

O ambiente definido na Avaliação N1 foi um Data Warehouse utilizando o PostgreSQL.
A implementação foi feita criando uma estrutura relacional para armazenar os dados do dataset sintético "Heart Failure Prediction" (Kaggle).

A tabela principal foi definida como `heart_data`, com os seguintes campos:
- age (idade)
- gender (gênero biológico)
- chest_pain_type (tipo de dor no peito)
- resting_bp (pressão arterial em repouso)
- cholesterol (nível de colesterol)
- fasting_blood_sugar (glicemia de jejum)
- resting_ecg (resultado do ECG em repouso)
- max_heart_rate (frequência cardíaca máxima)
- exercise_induced_angina (angina induzida por exercício)
- oldpeak (depressão do segmento ST)
- slope (inclinação do segmento ST)
- num_major_vessels (número de vasos maiores)
- thalassemia (presença de talassemia)
- diabetes (diagnóstico de diabetes)
- smoking_history (histórico de tabagismo)
- alcohol_consumption (nível de consumo de álcool)
- physical_activity_level (nível de atividade física)
- family_history (histórico familiar de doenças cardíacas)
- bmi (índice de massa corporal)
- heart_failure (presença de insuficiência cardíaca)

Essa estrutura foi criada no PostgreSQL com suporte total para manipulação de dados históricos e integração com ferramentas de análise. O nome do

banco de dados é `analise-preditiva` e todas as tabelas foram criadas e populadas utilizando scripts SQL executados manualmente via terminal.

Questão 3 – Inserção de Amostras e Processo ETL

a) Origem dos Dados

O dataset utilizado é o "Heart Failure Prediction Dataset", disponível publicamente no Kaggle.
É um conjunto de dados sintético com 10.000 registros de pacientes, contendo informações clínicas e variáveis preditivas de insuficiência cardíaca.
Link para o dataset:
https://www.kaggle.com/datasets/miadul/heart-failure-prediction-synthetic-dataset

b) Processo ETL

O processo de ETL (Extract, Transform, Load) foi estruturado da seguinte forma:

EXTRAÇÃO:
Leitura do arquivo CSV `heart_failure_prediction.csv` contendo os dados dos pacientes, localizado na pasta `questao-2-e-3/`.

TRANSFORMAÇÃO:
- Adaptação da estrutura da tabela `heart_data` para incluir todas as colunas presentes no dataset original (20 colunas).
- Conversão de colunas booleanas (como `Fasting_Blood_Sugar`, `Exercise_Induced_Angina`, `Diabetes`, `Family_History`, `Heart_Failure`) de inteiros para tipos BOOLEAN.
- Verificação e mapeamento de tipos de dados como `VARCHAR`, `INT` e `NUMERIC`.

CARGA:
- Inserção dos dados transformados diretamente na tabela `heart_data` do PostgreSQL utilizando o comando SQL `COPY`.
- O comando foi executado com sucesso através do script `load_heart_data.sql` rodando dentro do container Docker.

Esse processo permitiu a ingestão de mais de 10.000 amostras reais do dataset, organizadas de forma estruturada no PostgreSQL.
O uso do comando `COPY` proporcionou alta performance na carga e garantiu integridade com o modelo do Data Warehouse.