Centro Universitário de Brasília (CEUB)

Faculdade de Tecnologia e Ciências Sociais Aplicadas (FATECS)

Thales Rassi Porto de Matos - 22400186

Gabriel Marques da Rocha - 22451254

Gabrielle Gutierres - 22350026

Pedro Klein - 22105154

Matheus de Morais - 22352763

Henrique Lessa - 22402204

Documentação ETL – EXAME REALIZADO

Brasília

2025

Thales Rassi Porto de Matos

Gabriel Marques da Rocha

Gabrielle Gutierres

Pedro Klein

Matheus de Morais

Henrique Lessa

Documentação ETL – EXAME REALIZADO

Atividade final apresentada à Faculdade de Tecnologia e Ciências Sociais Aplicadas (FATECS)
, do Centro Universitário de Brasília (CEUB) como parte integrante do currículo da disciplina Interação Humano Computador, da graduação em Ciência da computação
Professora responsável: Kadidja Valeria Reginaldo de Oliveira

Brasília

2025

SUMÁRIO

IDENTIFICAÇÃO

Script: ETL_EXAME_REALIZADO.py
Responsável(is): Equipe Radiologia DF

OBJETIVO DO ETL

Extrair a quantidade mensal de mamografias realizadas no Distrito Federal a partir do dataset bruto, interpretar corretamente o campo "Mes/Ano", normalizar valores, mapear os identificadores de UF e tipo de exame no banco de dados e inserir os registros na tabela de fato exame_realizado.

Este ETL garante que os dados históricos de produção de mamografia do DF estejam padronizados, com chaves estrangeiras corretas e prontos para análises temporais.

FONTES DE DADOS DE ENTRADA

Arquivo principal (dataset de produção mensal de mamografias):

1. dirty_data_qtd_mamografias_df.csv
    - Caminho esperado: diretório local do projeto
    - Formato: CSV
    - Separador: vírgula (",")
    - Encoding: padrão do pandas (utf-8, salvo ajuste no arquivo)
    - Periodicidade de atualização: conforme disponibilização de novas séries históricas de produção ambulatorial (SIA/SUS ou reportes equivalentes)

Observação:

- A coluna Mes/Ano é interpretada por meio do dicionário MAPA_MESES_EXTENSO que converte meses por extenso (JANEIRO, FEVEREIRO etc.) para números de 1 a 12.
- O ETL insere diretamente novos registros, não havendo lógica de atualização (ON CONFLICT) neste script.

TABELAS DE DESTINO IMPACTADAS

1. exame_realizado
    - Descrição: Tabela onde cada linha representa a quantidade de um determinado tipo de exame realizado em um mês e ano específicos, associado a uma unidade federativa.

FLUXO RESUMIDO DO ETL

Passo 1 – Carregamento de ids necessários

- Para cada arquivo listado em DATASETS_PROF:
    - Carrega id_uf associado à sigla "DF".
    - Carrega id_tipo_exame associado ao nome "Diagnostico Por Mamografia".
    - Caso DF ou o tipo de exame não estejam cadastrados, o ETL encerra com erro.

Passo 2 – Validação e leitura do dataset

- Verifica se o arquivo definido em CAMINHO_EXAMES existe.
- Lê o CSV com pandas.
- Imprime no console as colunas encontradas para conferência.
- Valida a presença das colunas:
    - "Mes/Ano"
    - "Exames"

Passo 3 – Processamento e normalização dos dados

- Para cada linha do dataset:
    - Interpreta o campo "Mes/Ano":
    - Separa mês e ano pelo caractere /
    - Converte o mês por extenso usando MAPA_MESES_EXTENSO

    - Converte o ano para inteiro
    - Caso qualquer parte seja inválida, a linha é ignorada
- Normaliza o campo "Exames":
    - Converte strings, floats e números para inteiro
    - Valores vazios, inválidos ou não numéricos são descartados
- Gera um registro no formato: (id_tipo_exame, mes, quantidade, id_uf, ano)

Passo 4 – Inserção no banco de dados

- Se não houver registros válidos, informa no console e encerra.
- Caso contrário:
    - Insere todos os registros em lote (execute_batch)
    - Cada linha resulta em um INSERT na tabela exame_realizado
    - Realiza commit
    - Imprime o total de registros inseridos

Passo 5 – Encerramento

- Fecha o cursor e a conexão.
- Informa no console que o processo foi concluído.

MAPEAMENTO DE CAMPOS (ORIGEM → DESTINO):

"Mes/Ano" → ano (parse)

"Mes/Ano" → mes (parse)

"Exames" → quantidade

Sigla “DF” (fixa no script) → id_uf

Nome "Diagnostico Por Mamografia" (fixo no script) → id_tipo_exame

O script converte e valida cada campo antes da carga.

DEPENDÊNCIAS E PRÉ-REQUISITOS

Bibliotecas Python:

- os
- pandas
- psycopg2
- psycopg2.extras
- python-dotenv (load_dotenv)
- config_db (função get_conn para conexão com o banco)

Tabelas:

- Tabela unidade_da_federacao deve conter sigla = "DF"
- Tabela tipo_exame deve conter nome = "Diagnostico Por Mamografia"
- Tabela exame_realizado deve estar criada com o schema esperado

Requisitos de dados

- Arquivo CSV existente no caminho definido em CAMINHO_EXAMES
- Colunas "Mes/Ano" e "Exames" devem estar presentes e preenchidas adequadamente

COMO EXECUTAR

- Garantir que o ambiente virtual (se houver) esteja ativado.
- Garantir que o arquivo .env esteja configurado com os parâmetros de conexão ao banco.
- Garantir que o arquivo dirty_data_qtd_mamografias_df.csv esteja no caminho definido por CAMINHO_EXAMES.
- Executar: python3 ETL_EXAME_REALIZADO_MAMOGRAFIA_DF.py

Pré-condições:

- Banco PostgreSQL acessível.
- Tabelas unidade_da_federacao, tipo_exame e exame_realizado criadas
- Arquivos de dados presentes e com estrutura compatível.