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 – ESPERA EXAME

Brasília

2025

Thales Rassi Porto de Matos

Gabriel Marques da Rocha

Gabrielle Gutierres

Pedro Klein

Matheus de Morais

Henrique Lessa

Documentação ETL – ESPERA EXAME

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_ESPERA_EXAME.py
Responsável(is): Equipe Radiologia DF

OBJETIVO DO ETL

Processar automaticamente todos os arquivos estaduais contendo dados de tempo de espera para realização de mamografias, extrair as informações de quantidade de exames realizados em cada faixa de tempo (0–10 dias, 11–20 dias, 21–30 dias e mais de 30 dias), mapear a sigla da UF e o tipo de exame no banco e inserir ou atualizar esses registros na tabela fato espera_exame.

Este ETL permite comparar tempos de espera entre estados e gera uma base padronizada para análises sobre acesso ao exame de mamografia no Brasil.

FONTES DE DADOS DE ENTRADA

Arquivos principais (um arquivo por UF):
Padrão esperado de nome: mamografia_atendXX.csv

1. dirty_data_distribuição_geo_equipamentos.csv
    - Caminho esperado: diretório local definido por BASE_DIR = "dirty_data_estados_espera_mamografia"
    - Formato: CSV
    - Separador: vírgula (";")
    - Encoding: padrão do pandas (utf-8, salvo ajuste no arquivo)
    - Periodicidade de atualização: conforme atualizações dos datasets estaduais de mamografia

Observação:

- Todos os arquivos são identificados automaticamente via glob.glob("mamografia_atend*.csv") dentro do diretório BASE_DIR.

TABELAS

1. espera_exame
    - Descrição: Tabela que armazena, por UF, tipo de exame e ano, a quantidade de mamografias realizadas dentro de cada faixa de tempo de espera.

FLUXO RESUMIDO DO ETL

Passo 1 – Descoberta e validação dos arquivos

- Define o padrão de busca: mamografia_atend*.csv no diretório BASE_DIR.
- Lista e imprime todos os arquivos encontrados.
- Caso nenhum arquivo seja encontrado, o script encerra.

Passo 2 – Identificação da UF pelo nome do arquivo

- Para cada arquivo:
    - Extrai a UF com base no nome:
    - remove prefixo mamografia_atend
    - remove sufixo .csv
    - valida que restam exatamente 2 caracteres
    - retorna sigla em maiúsculas
- Se o nome estiver fora do padrão, lança erro.

Passo 3 – Leitura e tratamento do arquivo

- Ao ler cada CSV:
- Lê o arquivo com sep=";".
- Remove espaços extras dos nomes de coluna.
- Valida a existência das colunas obrigatórias.
- Seleciona somente as colunas necessárias.
- Renomeia para uma convenção padronizada:
    - ano
    - qtd_0_10
    - qtd_11_20
    - qtd_21_30

    - qtd_30_mais
- Converte todas as quantidades para inteiro, substituindo valores não numéricos por 0.
- Remove linhas onde o ano não é válido (ano nulo).

Passo 4 – Mapeamento de ids

- Para cada UF:
    - Obtém id_uf consultando unidade_da_federacao pela sigla.
    - Obtém id_tipo_exame consultando tipo_exame pelo nome “Diagnostico Por Mamografia”.
    - Caso qualquer id não seja encontrado, lança erro e interrompe o processamento daquela UF.

Passo 5 – Carga na tabela espera_exame

- Constrói lista de valores no formato: (id_uf, id_tipo_exame, ano, qtd_0_10, qtd_11_20, qtd_21_30, qtd_30_mais)
- Se não houver registros, informa e ignora.
- Caso haja registros:
    - Insere usando execute_values com cláusula ON CONFLICT para atualizar dados já existentes.
    - Realiza commit.
    - Informa quantos registros foram inseridos/atualizados para cada UF.

Passo 6 – Encerramento

- Fecha o cursor.
- Fecha a conexão com o banco.
- Imprime mensagem final indicando a conclusão do ETL para todas as UFs.

MAPEAMENTO DE CAMPOS (ORIGEM → DESTINO):

Origem no CSV	Destino na tabela espera_exame
"Ano Resultado"	ano
"0 - 10 dias"	qtd_tempo_espera_0_10
"11 - 20 dias"	qtd_tempo_espera_11_20
"21 - 30 dias"	qtd_tempo_espera_21_30
"> 30 dias"	qtd_tempo_espera_30_mais
Sigla extraída do nome do arquivo	id_uf (via unidade_da_federacao)
Constante NOME_EXAME	id_tipo_exame (via tipo_exame)

DEPENDÊNCIAS E PRÉ-REQUISITOS

Bibliotecas Python:

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

Tabelas de banco necessárias:

- unidade_da_federacao
- tipo_exame

- espera_exame

Requisitos de dados:

- Arquivos com nome no padrão mamografia_atendXX.csv
- Colunas obrigatórias presentes e consistentes
- Sigla da UF cadastrada na tabela unidade_da_federacao
- Tipo de exame “Diagnostico Por Mamografia” cadastrado na tabela de tipo_exame

COMO EXECUTAR

- Ativar ambiente virtual (se houver).
- Confirmar que o arquivo .env contém configurações válidas de banco.
- Confirmar que o diretório definido em BASE_DIR contém os CSVs no padrão esperado.
- Executar: python3 ETL_ESPERA_EXAME_MAMOGRAFIA_UF.py

Pré-condições:

- Banco PostgreSQL acessível.
- Tabelas unidade_da_federacao, tipo_exame e espera_exame criadas.
- Arquivos CSV presentes no diretório configurado e seguindo o padrão.