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 – EQUIPAMENTOS REGISTRADOS

Brasília

2025

Thales Rassi Porto de Matos

Gabriel Marques da Rocha

Gabrielle Gutierres

Pedro Klein

Matheus de Morais

Henrique Lessa

Documentação ETL – EQUIPAMENTOS REGISTRADOS

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_EQUIPAMENTO_REGISTRADO_RA.py
Responsável(is): Equipe Radiologia DF

OBJETIVO DO ETL

Extrair do dataset de distribuição geográfica de equipamentos por Região Administrativa (RA) do Distrito Federal a quantidade de cada tipo de equipamento de imagem, tratar e normalizar esses dados, mapear as RAs e tipos de equipamento para suas respectivas dimensões no banco e, por fim, inserir registros consolidados na tabela de fato equipamento_registrado, garantindo integridade referencial com as tabelas regiao_administrativa e tipo_equipamento.

FONTES DE DADOS DE ENTRADA

Arquivos principais:

1. dirty_data_distribuição_geo_equipamentos.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 novas extrações/atualizações dos dados oficiais sobre equipamentos por RA

Observação:

- O caminho do arquivo utilizado é configurado na constante DATASET_EQUIP_RA.
- O ano de referência utilizado para todos os registros inseridos é definido pela constante ANO_PADRAO (no script, ANO_PADRAO = 2025).

TABELAS

1. equipamento_registrado
    - Descrição: tabela que armazena a quantidade de equipamentos de diagnóstico por imagem registrados por tipo de equipamento, Região Administrativa e ano, permitindo análises de distribuição e disponibilidade de equipamentos no Distrito Federal.

2. regiao_administrativa
    - Descrição: tabela utilizada apenas para leitura neste ETL, a partir da qual são obtidos os ids das RAs (id_ra) com base em seus nomes.

3. tipo_equipamento
    - Descrição: tabela utilizada apenas para leitura neste ETL, a partir da qual são obtidos os ids dos tipos de equipamento (id_tipo_equipamento) com base em seus nomes.

FLUXO RESUMIDO DO ETL

Passo 1 – Leitura e tratamento do dataset de equipamentos por RA:

- Verifica se o arquivo definido em DATASET_EQUIP_RA existe no diretório.
    - Verifica se o arquivo definido em DATASET_EQUIP_RA existe no diretório.
    - Caso o arquivo não exista, o script lança FileNotFoundError.
    - Lê o CSV com pandas.read_csv.
    - Valida a existência da coluna "ra" no DataFrame.
    - Valida se todas as colunas esperadas em MAPEAMENTO_COLUNAS_EQUIP existem no CSV; se alguma estiver ausente, lança ValueError.
    - Cria a nova coluna nome_ra a partir de "ra", aplicando limpar_nome_ra, que:
        - converte o valor para string,
        - remove espaços extras (strip()),
        - descarta valores vazios,
        - padroniza o nome para title case.
    - Percorre linha a linha o DataFrame:

        - ignora linhas onde "nome_ra" seja vazio ou inválido,
        - para cada coluna de equipamento definida em MAPEAMENTO_COLUNAS_EQUIP:
            - lê o valor da coluna correspondente no CSV,
            - tenta converter o valor para inteiro,
            - se não conseguir converter (ValueError ou TypeError), ignora o campo,
            - se a quantidade for menor ou igual a zero, ignora o registro,
            - caso contrário, adiciona um registro contendo:
                a. nome_ra
                b. nome_equipamento
                c. quantidade
                d. ano (ANO_PADRAO)
    - Ao final, transforma os registros em um novo DataFrame contendo:
        - nome_ra
        - nome_equipamento
        - quantidade
        - ano

Passo 2 – Identificação das categorias profissionais:

- Define o conjunto de colunas que NÃO são categorias (COLUNAS_NAO_CATEGORIA), como "Ocupações de Nível Superior", "Data", "Total", "Ano/mês compet."
- Para cada coluna de cada dataset:
    - Remove espaços extras (strip).
    - Compara o nome em minúsculas com a lista de colunas não categoria.
    - Ignora colunas vazias ou que estejam na lista de exclusão.
    - Adiciona as demais colunas a um conjunto de categorias.
- Ao final, ordena alfabeticamente as categorias e imprime o resultado no console.

Passo 3 – Carregamento na tabela equipamento_registrado:

- Cria um cursor psycopg2 com DictCursor.
- Monta uma lista de tuplas (linhas) a partir do DataFrame já mapeado:
    - (id_tipo_equipamento, id_ra, ano, quantidade) para cada linha.
    - Se a lista de linhas estiver vazia:
        - imprime “Nenhuma linha para inserir em equipamento_registrado.” e não realiza inserts.
- Caso haja linhas:

    - utiliza psycopg2.extras.execute_values para inserir em lote na tabela equipamento_registrado:
    - INSERT INTO equipamento_registrado (id_tipo_equipamento, id_ra, ano, quantidade) VALUES %s
    - dá commit na transação.
    - imprime no console quantos registros foram inseridos.

Passo 4 – Encerramento:

- Fecha o cursor utilizado para o mapeamento e o cursor de inserção.
- Fecha a conexão com o banco de dados.
- Imprime mensagem final informando que o ETL de equipamento_registrado foi concluído.

MAPEAMENTO DE CAMPOS (ORIGEM → DESTINO):

coluna "ra" do CSV —> regiao_administrativa.nome (usada para localizar regiao_administrativa.id_ra)

colunas de quantidade de equipamentos do CSV (qtd_) —> quantidade (campo quantidade da tabela equipamento_registrado), após conversão para inteiro e filtragem de valores menores ou iguais a zero

colunas de quantidade de equipamentos do CSV (qtd_) —> tipo_equipamento.nome (via MAPEAMENTO_COLUNAS_EQUIP, que associa cada coluna técnica a um nome de tipo de equipamento), posteriormente mapeado para tipo_equipamento.id_tipo_equipamento

constante ANO_PADRAO —> equipamento_registrado.ano (ano de referência para todos os registros inseridos por este ETL)

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 de banco necessárias:

- regiao_administrativa
- tipo_equipamento

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 definido em DATASET_EQUIP_RA esteja no caminho correto.
- Garantir que as tabelas regiao_administrativa, tipo_equipamento existam e estejam com o schema esperado.
- Rodar o comando: python3 ETL_EQUIPAMENTO_REGISTRADO_RA.py

Pré-condições:

- Banco PostgreSQL acessível.
- Tabelas regiao_administrativa, tipo_equipamento e equipamento_registrado criadas.
- Tabela regiao_administrativa com os nomes de RA compatíveis com os nomes limpos gerados por limpar_nome_ra.
- Tabela tipo_equipamento com os nomes compatíveis com os valores definidos em MAPEAMENTO_COLUNAS_EQUIP.
- Arquivo de dados dirty_data_distribuição_geo_equipamentos.csv presente e com estrutura compatível.