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 – RA

Brasília

2025

Thales Rassi Porto de Matos

Gabriel Marques da Rocha

Gabrielle Gutierres

Pedro Klein

Matheus de Morais

Henrique Lessa

Documentação ETL – RA

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_RA.py
Responsável(is): Equipe Radiologia DF

OBJETIVO DO ETL

Extrair todos os nomes de Regiões Administrativas (RAs) do Distrito Federal presentes nos datasets utilizados no projeto (equipamentos e população), padronizar a escrita dos nomes e inserir automaticamente na tabela regiao_administrativa apenas as RAs que ainda não existirem no banco.

Este ETL garante que a dimensão de Regiões Administrativas esteja sincronizada com os datasets brutos e preparada para ser utilizada como chave estrangeira em outras tabelas de fato (equipamentos, população, etc.).

FONTES DE DADOS DE ENTRADA

Arquivo principal:

1. 1. dirty_data_distribuição_geografica_equipamentos_DF.csv
    - Caminho esperado: diretório local do projeto
    - Formato: CSV
    - Separador: vírgula (",")
    - Linha inicial: sem skip (skiprows=0)
    - skipfooter: 0

2. dirty_data_populacao_DF_com_plano_saude_por_RA.csv
    - Caminho esperado: diretório local do projeto
    - Formato: CSV
    - Separador: vírgula (",")
    - skiprows: 1
    - skipfooter: 2

Observação:

- Caso um arquivo não exista, ele é ignorado com aviso no console.
- Caso a coluna de RA configurada não exista no arquivo, o dataset é ignorado com aviso.
- Todos os nomes de RA extraídos são tratados e unificados em uma lista final sem duplicidades.

TABELAS DE DESTINO IMPACTADAS

1. regiao_administrativa
    - Tabela que armazena as Regiões Administrativas do Distrito Federal.

FLUXO RESUMIDO DO ETL

Passo 1 – Extração das RAs nos datasets

- Para cada item em DATASETS_RA:
    - Valida se o arquivo existe; se não existir, exibe aviso e ignora.
    - Lê o CSV com os parâmetros adequados (sep, skiprows, skipfooter).
    - Valida se a coluna indicada em coluna_ra existe.
    - Extrai a coluna de RA:
        - Remove valores nulos
        - Converte para string
        - Aplica strip
        - Remove entradas vazias
        - Converte para formato Title Case
    - Concatena todas as listas de RAs extraídas.
    - Remove duplicidades.
    - Ordena alfabeticamente.
    - Retorna um DataFrame final contendo apenas uma coluna: nome_ra

Passo 2 – Carregamento das RAs na tabela regiao_administrativa

- Busca no banco todas as RAs já existentes para id_uf = 1 (Distrito Federal).
- Constrói um conjunto com os nomes já cadastrados.

- Para cada RA extraída:
- Se ela ainda não existir no banco, adiciona na lista de inserção.
- Se houver novas RAs:
- Insere em lote utilizando:
    - INSERT INTO regiao_administrativa (nome, id_uf) VALUES %s
    - Realiza commit
    - Imprime a quantidade de RAs inseridas
- Caso não haja novas RAs:
    - Imprime “Nenhuma nova RA para inserir.”

Passo 3 – Encerramento

- Fecha o cursor
  Fecha a conexão com o banco
  Imprime mensagem final indicando que o ETL foi concluído

MAPEAMENTO DE CAMPOS (ORIGEM → DESTINO):

Coluna RA dos datasets (ra, Local) → nome

Valor fixo para UF (=1) → id_uf

"Data" → ano, mes

Nome da coluna da categoria → id_categoria (via tabela categoria_profissional)

Valor da célula (numérico ou string) → quantidade (int normalizado)

Sigla "DF" → id_uf (via unidade_da_federacao)

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:

- unidade_da_federacao
- regiao_administrativa

Requisitos de dados

- Arquivos definidos em DATASETS_RA devem existir no diretório esperado
- As colunas indicadas (ra, Local) precisam estar presentes

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 os arquivos listados em DATASETS_RA estejam no diretório correto.
- Executar: python3 ETL_RA.py

Pré-condições:

- Banco PostgreSQL acessível.
- Tabela regiao_administrativa criada
- Tabela unidade_da_federacao possui registro com id_uf = 1 para o DF