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 – PROFISSIONAIS REGISTRADOS

Brasília

2025

Thales Rassi Porto de Matos

Gabriel Marques da Rocha

Gabrielle Gutierres

Pedro Klein

Matheus de Morais

Henrique Lessa

Documentação ETL – PROFISSIONAIS 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_PROFISSIONAIS_REGISTRADOS.py
Responsável(is): Equipe Radiologia DF

OBJETIVO DO ETL

Extrair, padronizar e consolidar informações históricas sobre o número de profissionais das categorias ligadas à radiologia (médicos radiologistas, cirurgiões-dentistas radiologistas, auxiliares/técnicos, entre outros) no Distrito Federal, interpretando corretamente o período (ano/mês), associando cada coluna de categoria ao respectivo id_categoria no banco e realizando a carga na tabela profissional_registrado.

Este ETL integra dados de múltiplos datasets, com estruturas distintas, e produz uma base única e padronizada para análises temporais e comparativas entre categorias profissionais.

FONTES DE DADOS DE ENTRADA

Arquivo principal:

1. dirty_data_historico_anual_numero_medicos_radiologistas_e_diagnostico_imagem_SUS - cnes.csv
    - Caminho esperado: diretório local do projeto
    - Formato: CSV
    - Separador: vírgula (";")
    - Encoding: padrão do pandas (utf-8, salvo ajuste no arquivo)
    - Periodicidade de atualização: anual/mensal conforme dados do CNES

2. dirty_data_historico_anual_numero_cirurgioes_dentistas_radiologistas_SUS - denstista_radio_profissinoais.csv
    - Caminho esperado: diretório local do projeto
    - Formato: CSV
    - Separador: vírgula (";")
    - Encoding: padrão do pandas (utf-8, salvo ajuste no arquivo)
    - Periodicidade de atualização: anual/mensal conforme dados do CNES

3. dirty_data_historico_anual_numero_auxiliares_e_tecnicos_em_radiologia_SUS.csv
    - Caminho esperado: diretório local do projeto
    - Formato: CSV
    - Separador: vírgula (";")
    - Encoding: padrão do pandas (utf-8, salvo ajuste no arquivo)
    - Periodicidade de atualização: anual/mensal conforme dados do CNES

Observação:

- As colunas restantes, após ignorar as definidas em colunas_ignorar, correspondem às categorias profissionais.1.

TABELAS DE DESTINO IMPACTADAS

1. profissional_registrado
    - Tabela contendo o número de profissionais registrados por categoria, UF, ano e mês.

FLUXO RESUMIDO DO ETL

Passo 1 – Carregamento de IDs e dicionários de apoio

- Obtém id_uf correspondente à sigla DF.
- Carrega todas as categorias cadastradas em categoria_profissional, produzindo o dicionário: mapa_categorias = { nome_categoria : id_categoria }

Passo 2 – Processamento de cada dataset

- Para cada arquivo de DATASETS_PROFISSIONAIS:
- Verifica se o arquivo existe; caso contrário, lança erro.
- Lê o CSV com pandas.
- Identifica as colunas de categorias como aquelas que não estão em colunas_ignorar.
- Interpreta a coluna de período usando parse_ano_mes, que:

    - separa ano e mês pelo caractere /,
    - converte o mês pela tabela MAPA_MESES (jan., fev., mar. etc.),
    - retorna (ano, mes) ou descarta a linha se o formato estiver inválido.
- Para cada linha válida do dataset:
    - Para cada coluna de categoria:
    - Obtém o id_categoria correspondente pelo nome da coluna.
    - Se a categoria não existir no banco, imprime aviso e ignora.
    - Normaliza a quantidade usando normalizar_quantidade (conversão segura para inteiro).
    - Caso a quantidade seja válida, gera o registro: (id_categoria, id_uf_df, ano, mes, quantidade)

Todos os registros gerados são acumulados em registros_totais.

Passo 3 – Inserção na tabela profissional_registrado

- Após processar todos os datasets, insere todos os registros acumulados:
    - INSERT INTO profissional_registrado (id_categoria, id_uf, ano, mes, quantidade) VALUES (%s, %s, %s, %s, %s);
- A inserção é feita em lote via execute_batch.
- Realiza commit e imprime o total de registros inseridos.

Passo 4 – Encerramento

- Fecha o cursor e a conexão.
- Imprime mensagem final indicando a conclusão do processo.

MAPEAMENTO DE CAMPOS (ORIGEM → DESTINO):

"Ano/mês compet." → ano, mes

"Ocupações de Nível Superior" → ano, mes

"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
- categoria_profissional
- profissional_registrado

Requisitos de dados

- Todos os arquivos definidos em DATASETS_PROFISSIONAIS devem existir.
- Cada coluna de categoria deve estar previamente cadastrada em categoria_profissional (caso contrário, será ignorada).
- As colunas de período devem existir exatamente com o nome indicado.

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.
- Verificar que todos os arquivos listados em DATASETS_PROFISSIONAIS estão no diretório correto.
- Executar: python3 ETL_PROFISSIONAIS_REGISTRADOS.py

Pré-condições:

- Banco PostgreSQL acessível.
- Tabelas unidade_da_federacao, categoria_profissional e profissional_registrado criadas.
- Sigla "DF" existente em unidade_da_federacao.