



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 – CATEGORIAS PROFISSIONAIS

Brasília

2025

Thales Rassi Porto de Matos

Gabriel Marques da Rocha

Gabrielle Gutierres

Pedro Klein

Matheus de Morais

Henrique Lessa

Documentação ETL – CATEGORIAS PROFISSIONAIS

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	04
OBJETIVO DO ETL	04
FONTES DE DADOS DE ENTRADA	04
TABELAS DE DESTINO IMPACTADAS	05
FLUXO DO ETL	05
MAPEAMENTO DE CAMPOS (ORIGEM → DESTINO)	07
DEPENDÊNCIAS E PRÉ-REQUISITOS	07
COMO EXECUTAR	07

IDENTIFICAÇÃO

Script: ETL_CATEGORIAS_PROFSSIONAIS.py

Responsável(is): Equipe Radiologia DF

OBJETIVO DO ETL

Extrair automaticamente, a partir dos arquivos históricos de profissionais do SUS, todas as categorias profissionais de radiologia presentes nas colunas dos datasets e inseri-las na tabela dimensão categoria_profissional, evitando duplicidades. Este ETL garante que a dimensão de categorias profissionais esteja sempre alinhada com os dados brutos utilizados no projeto Radiologia DF.

FONTES DE DADOS DE ENTRADA

Arquivos principais (datasets de profissionais):

1. 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: conforme novas extrações oficiais do SUS
2. 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: conforme novas extrações oficiais do SUS
3. 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: conforme novas extrações oficiais do SUS
4. dirty_data_historico_anual_numero_medicos_radiologistas_e_diagnosticos_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: conforme novas extrações oficiais do SUS

Observação:

- A lista de arquivos utilizados é configurada na constante DATASETS_PROF.
- Caso algum arquivo não seja encontrado, o script emite um aviso e segue para os demais arquivos.

TABELAS DE DESTINO IMPACTADAS

1. categoria_profissional
 - Descrição: tabela que armazena as categorias profissionais de radiologia utilizadas pelo SUS (ex.: médicos radiologistas, cirurgiões-dentistas radiologistas, técnicos em radiologia, etc.).

FLUXO RESUMIDO DO ETL

Passo 1 – Extração das categorias a partir dos datasets:

- Para cada arquivo listado em DATASETS_PROF:
 - Verifica se o arquivo existe.
 - Se existir, lê o CSV com pandas.read_csv (sep=",").
Imprime no console as colunas encontradas para conferência.

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 categoria_profissional:

- Abre conexão com o banco utilizando get_conn() (config_db).
- Busca todas as categorias já existentes na tabela categoria_profissional (SELECT nome FROM categoria_profissional).
- Normaliza os nomes existentes para minúsculo e remove espaços extras.
- Compara a lista de categorias extraídas dos arquivos com as já existentes no banco. Monta uma lista apenas com as categorias novas (ainda não presentes na tabela).
- Se houver novas categorias, realiza a inserção em lote (execute_values) na tabela categoria_profissional(nome).
- Dá commit na transação e informa no console quantas novas categorias foram inseridas.

Passo 4 – Encerramento:

- Fecha o cursor e a conexão com o banco.
- Imprime mensagem final informando que o ETL de categorias profissionais foi concluído.

MAPEAMENTO DE CAMPOS (ORIGEM → DESTINO):

Este ETL não faz mapeamento linha a linha de registros, e sim mapeamento de “nomes de colunas” dos arquivos para valores na dimensão.

nome da coluna de categoria do CSV —> categoria_profissional.nome

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)

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_PROF estejam no caminho correto.
- Rodar o comando: python3 ETL_CATEGORIAS_PROFISSIONAIS.py

Pré-condições:

- Banco PostgreSQL acessível.
- Tabela categoria_profissional criada.
- Arquivos de dados presentes e com estrutura compatível.