MEMORIAL DESCRITIVO:

CONVERSÃO DE BANCO DE DADOS EM CSV PARA SQL

Isadora Garrefa

17/07/2024

1. Introdução

Objetivo do Projeto:

O objetivo deste projeto é converter um arquivo CSV contendo dados de registros válidos de produtos alimentícios extraídos do portal de consultas da ANVISA (de acordo a RDC 27/2010) para um banco de dados SQL realizando um processo ETL com a biblioteca Pandas em Python.

Contextualização e Justificativa:

Com o crescente volume de dados, a necessidade de uma gestão eficiente se torna crucial. Converter dados de um arquivo CSV e utilizar a biblioteca Pandas permite realizar transformações de dados de maneira eficiente para carregá-los em um banco de dados SQL. Um banco de dados SQL facilita a manipulação e análise dessas informações.

Descrição do Arquivo CSV:

- Nome do arquivo: `DADOS_ABERTOS_ALIMENTO.csv`
- Colunas: `NU_CNPJ_EMPRESA, NO_RAZAO_SOCIAL_EMPRESA, NO_PRODUTO, NU_PROCESSO, DS_TIPO_PRODUTO, DS_CATEGORIA_PRODUTO, DT_FINALIZACAO_PROCESSO, NU_REGISTRO_PRODUTO, DT_VENCIMENTO_REGISTRO, ST_SITUACAO_REGISTRO`

2. Metodologia

Ferramentas Utilizadas:

- Excel
- Python 3.12.1
- MySQL Workbench 8.0
- MySQL Server 8.4.1
- VSCode

Procedimentos Adotados:

1. Instalação das ferramentas necessárias.
2. Carregamento e análise dos dados do arquivo CSV (convertido para .xlsx).
3. Transformação dos dados utilizando Pandas.
4. Criação da tabela no banco de dados MySQL.
5. Carregamento dos dados transformados para a tabela SQL.

1. **Instalação**

- *instale o Python 3.12.1*

- *instale o MySQL Server 8.4.1*

- *instale o MySQL Workbench mais recente*

- **Pacotes Python:**

- *pandas*

- *pymysql*

- *openpyxl*

Como utilizar o código e criar um servidor desta base de dados:

- **Etapa 1** Instale os softwares e pacotes necessários

- **Etapa 2** Abra o MySQL Workbench, faça login no seu servidor MySQL e execute o código encontrado na pasta “SQL scripts/criarBancodeDados.sql”

- **Etapa 3** Copie o arquivo de Database/DADOS_ABERTOS_ALIMENTOS.xlsx

- **Etapa 4** Modifique o código para localizar seu arquivo .xlsx e localize as células com a instrução SQL 'INSERT' e descomente-a.

- **Etapa 5** Execute todo o código.

3. Análise de Dados / Data Analysis

Coleta de dados – Data collection
O arquivo CSV utilizado contém informações das empresas, incluindo CNPJ, razão social e número de registro, e nome, categoria, tipo de produto, bem como o número e situação de cada registro.

The CSV file used contains company information, including CNPJ, corporate name and registration number, and name, category, product type, as well as the number and status of each registration.

Análise e Interpretação dos Dados - Analysis
Os dados foram analisados e 3 linhas possuíam dados ausentes (Not a Number) que também geravam inconsistências no formato dos dados. Foi realizado um tratamento desses dados os excluindo.

The data had 3 lines missing data (Not a Number) which also generated inconsistencies in the data format. This data was processed and excluded.

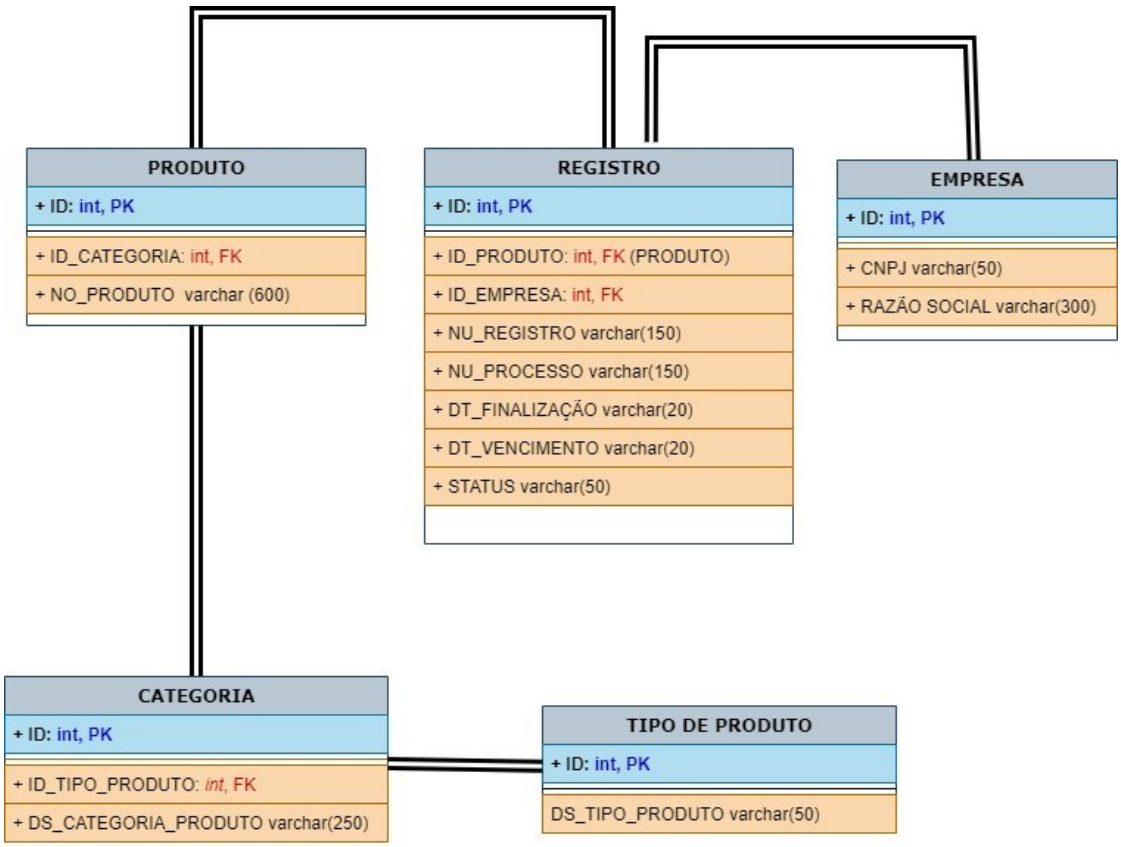
Considerações sobre a Qualidade dos Dados – Data quality
Foram encontrados alguns registros com acentuações equivocas, além de diferenciação entre letras maiúsculas e minúsculas, e números de CNPJ fora do formato esperado.

Some records were found with incorrect accents, in addition to differentiation between upper- and lower-case letters, and CNPJ (Company registration number) numbers outside the expected format.

Procedimento ETL (EXTRACT, TRANSFORM, LOAD)
O script Python completo utilizado no projeto foi incluído na pasta "scripts/criarBancodeDados.sql".

4. Diagrama de Entidade-Relacionamento

Descrição geral:
O diagrama desenvolvido para este projeto foi baseado nos dados fornecidos pelo arquivo CSV. Tem por objetivo representar visualmente a estrutura lógica do banco de dados, mostrando as entidades, seus atributos e os relacionamentos entre elas de forma clara e intuitiva**.**
The ERD developed for this project was based on the data provided by the CSV file. The purpose is to visually represent the logical structure of the database, showing the entities, their attributes and the relationships between them in a clear and intuitive way.



Principais componentes

- **Entidades (entities):** `'Tipo de produto', 'Categoria', 'Produto', 'Registro', 'EMPRESA'`
- **Relacionamentos (relationships):** `'tipo de produto-categoria', 'empresa-registro', 'categoria-produto', 'produto-registro'.`

- **Descrição das Entidades**

    - **Entidade: Tipo de produto**
        - **Atributos:**

- `ID` (Chave Primária)
- `DS_TIPO_PRODUTO`

Descrição: Representa os tipos de produto cadastrados, por exemplo, nesse caso, há apenas um tipo de produto "Alimento".

    - **Entidade: Categoria**
        - **Atributos:**

- `ID` (Chave Primária)
- `ID_TIPO_PRODUTO` (chave estrangeira)
- `DS_CATEGORIA_PRODUTO`

Descrição: Armazena informações sobre as categorias de tipo de produtos cadastradas.

    - **Entidade: Produto**
        - **Atributos:**

- `ID` (Chave Primária)
- `ID_CATEGORIA` (chave estrangeira)
- `NO_PRODUTO`

Descrição: Armazena os nomes de produtos listados dentro das categorias

    - **Entidade: Empresa**
        - **Atributos:**

- `ID` (Chave Primária)
- `CNPJ`
- `RAZÃO SOCIAL`

Descrição: Contém as informações sobre as empresas cadastradas

    - **Entidade: Registro**
        - **Atributos:**

- `ID` (Chave Primária)
- `ID_PRODUTO` (chave estrangeira)

- ID_EMPRESA (chave estrangeira)
- NU_REGISTRO
- NU_PROCESSO
- DT_FINALIZAÇÃO
- DT_VENCIMENTO
- STATUS

Descrição: Cada registro é criado para um produto, uma empresa pode ter muitos produtos registrados. Detalha a situação e vencimentos dos registros em dia e hora.

- **Descrição dos Relacionamentos**

  Relacionamento: Tipo de produto-Categoria

  Tipo: Um-para-Muitos

  - **Entidades Envolvidas:** 'Tipo de produto' e 'Categoria"
  - **Descrição:** Há um tipo de produto (Alimento) que abrange várias categorias

  Relacionamento: Categoria-Produto

  Tipo: Um-para-Muitos

- **Entidade Envolvidas: 'Categoria' e 'Produto'**
- **Descrição:** Uma categoria de alimentos possui vários produtos cadastrados, e um produto cadastrado pertence apenas a uma categoria de alimento

  Relacionamento: Empresa-Registro

  Tipo: Um-para-Muitos

- **Entidade Envolvidas: 'Empresa' e 'Registro'**
- **Descrição:** Uma empresa possui vários registros, e cada registro corresponde a uma única empresa.

  Relacionamento: Produto-Registro

  Tipo: Um-para-Muitos

- **Entidade Envolvidas: 'Produto' e 'Registro'**
- **Descrição:** Um produto para vários registros, duas empresas distintas podem cadastrar o mesmo produto tendo números de registro diferentes.

  Os relacionamentos definidos no diagrama ER são cruciais para manter a integridade referencial e garantir que todas as interações entre empresas, produtos e registros sejam corretamente registradas e acessíveis.

5. Conclusão

Resumo dos Principais Pontos do Projeto:

- A conversão do arquivo CSV para o MySQL Workbench com tratamento dos dados foi realizada com sucesso.
- O processo de importação foi eficiente e os dados foram validados corretamente.

Lições Aprendidas:

- Importância da verificação prévia dos dados para evitar problemas de importação.
- Utilidade das ferramentas gráficas e de linha de comando para diferentes necessidades.
- Garantir que todos os aspectos do sistema sejam capturados e relacionados adequadamente.

6. Anexos

- Link para banco de dados CSV: https://data.amerigeoss.org/id/dataset/alimentos-registrados-no-brasil