COLETANDO E TRANSFORMANDO DADOS ABERTOS

Pandas

DuckDB + dbt

Streamlit

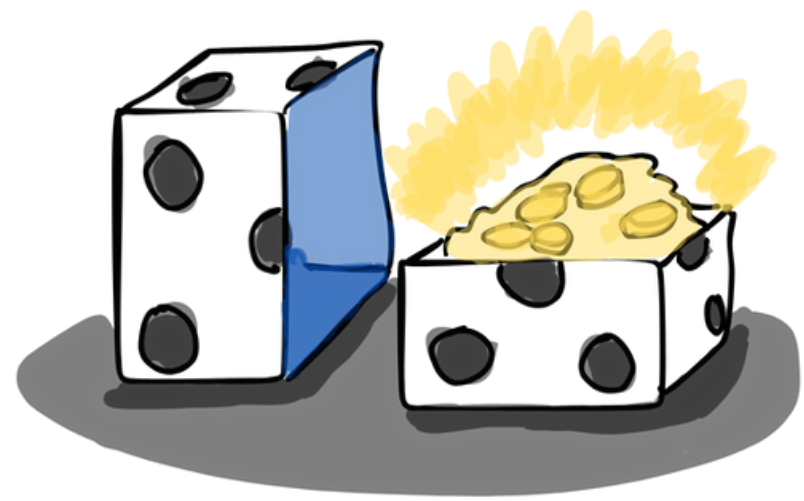
DADOS ABERTOS

Este projeto tem os seguintes objetivos:

1. Criar um processo automático de **download** dos dados de Cadastro Nacional da Pessoa Jurídica (**CNPJ**), disponíveis no site: https://dados.gov.br/dados/conjuntos-dados/cadastro-nacional-da-pessoa-juridica---cnpj

2. Criar um banco de dados funcional em **DuckDB**

3. Criar uma aplicação para **consulta dos dados coletados**



KUNG FU PANDAS

Os dados estão separados em diversos arquivos no formato **CSV**

Como existem arquivos com muitas linhas que podem não caber na memória utilizamos a biblioteca pandas para consumir as bases em '**chunks**' e salvá-las em arquivos do tipo **parquet**.

Foi escolhida o Pandas por também suportar o **encoding** necessário para leitura correta das informações

LÁ VEM O PATO

Foi utilizado o **dbt** para documentação, teste e criação da base de dados

A base de dados foi construída em **DuckDB** devido à **facilidade e agilidade**, em minutos é possível transformar todos os arquivos parquet em uma base funcional!

É possível utilizar o **DBeaver** para conectar na base recém criada e verifiar seus dados

CONSULTANDO OS DADOS

Foi criado um aplicativo Web com o **Streamlit** para a **consulta** completa de um CNPJ sem a necessidade de interação direta com a base de dados.

Com isso você não precisa mais de internet para consultar um CNPJ!

As bases são atualizadas mensalmente, portanto, é necessário **rodar este processo uma vez por mês** para garantir que os dados são os mais atualizados.

Confira o projeto completo no GitHub!