# IESB

Big Data e Inteligência Analítica

**Projeto Integrador em Big Data e Inteligência Analítica**

Lucas Siqueira Rodrigues

12 de junho de 2025

Brasília - DF
Junho de 2025

# 1 Obtenção da base de dados

## 1.1 Introdução

O trabalho tem como objetivo explorar os dados públicos disponibilizados pela Câmara dos Deputados para analisar e compreender os gastos realizados pelos parlamentares. Para isso, os dados serão obtidos por meio do portal dos dados abertos da câmara[1], armazenados em um banco de dados relacional na nuvem e, posteriormente, analisados.

## 1.2 Motivação

A transparência pública é essencial para garantir a confiança da população nas instituições governamentais. Esse projeto busca realizar um ciclo completo de extração, transformação, armazenamento e análise dos dados de gastos públicos. Além disso, ao construir dashboards dinâmicos espera-se fornecer ferramentas que possam auxiliar na identificação de possíveis irregularidades e na fiscalização das despesas parlamentares.

## 1.3 Script e Banco de Dados

Para o ano de 2022, a obtenção dos dados por meio da API[1] retornou apenas 32 registros.

```python
import io
import json
import zipfile

import httpx
from tqdm import tqdm

class CamaraAPI:
  def __init__(self) -> None:
    self.base_url = "https://dadosabertos.camara.leg.br/api/v2"

  def request(self, endpoint: str) -> dict:
    response = httpx.get(f"{self.base_url}/{endpoint}")
    return response.json()

  def get_deputados(self) -> dict:
    return self.request("deputados").get("dados", {})

  def get_despesas(self, id_: int, year: int = 2022) -> dict:
    return self.request(f"deputados/{id_}/despesas?ano={year}")

despesas = []

api = CamaraAPI()

deputados = api.get_deputados()
for deputado in tqdm(deputados):
  id_ = deputado["id"]
  despesas_deputado = api.get_despesas(id_=id_, year=2022)
  despesas.extend(despesas_deputado["dados"])
print(len(despesas))
```

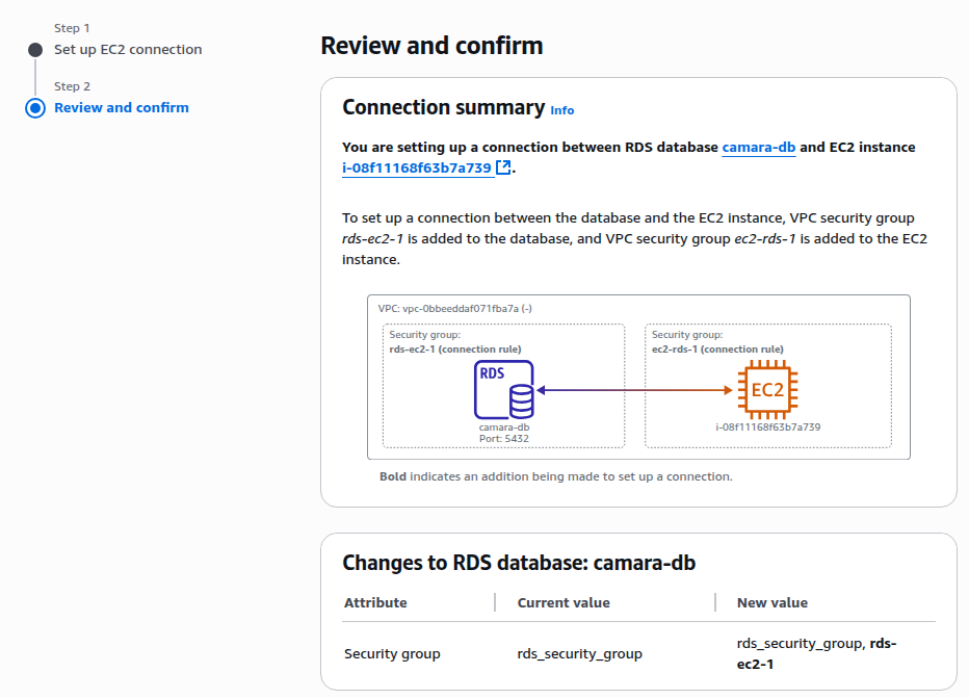


Figura 1: Coleta de dados por meio da API para o ano de 2022.

Por isso, apenas para esse ano a coleta de dados foi por meio de um arquivo no formato JSON, que também é fornecido no portal de dados abertos da câmara por meio da aba "Arquivos".

Para os anos de 2023 e 2024 a obtenção dos dados foi realizada por meio da API[1] da Câmara dos Deputados, explorando dois principais endpoints:

- `/deputados`: Retorna informações gerais sobre os parlamentares, como seus nomes, partidos, estados e e-mails.
- `/deputados/{id}/despesas`: Fornece detalhes sobre as despesas realizadas pelos parlamentares, incluindo valores, fornecedores, tipos de despesa e datas.

Para organizar os dados de forma eficiente e integrar os dados obtidos por meio da API e por meio do JSON, foi criado um modelo de banco de dados relacional com tabelas normalizadas para representar as informações de deputados, despesas e fornecedores.

Logo após serem obtidos, os dados foram inseridos em um banco de dados MariaDB, que foi criado usando o serviço RDS da AWS.

O programa, ao ser executado, faz a coleta e persiste 224042 registros de despesas.

Todo o código relacionado ao projeto está no Github[3].

## 1.4 Considerações finais

A combinação de tecnologias como Python, AWS e MariaDB foi fundamental para realizar as etapas de coleta, armazenamento e preparação dos dados. A API da Câmara revelou-se limitada em relação à quantidade de dados retornados para o ano de 2022, mas o uso do JSON permitiu superar essa restrição e criar uma base robusta com uma quantidade considerável de dados.

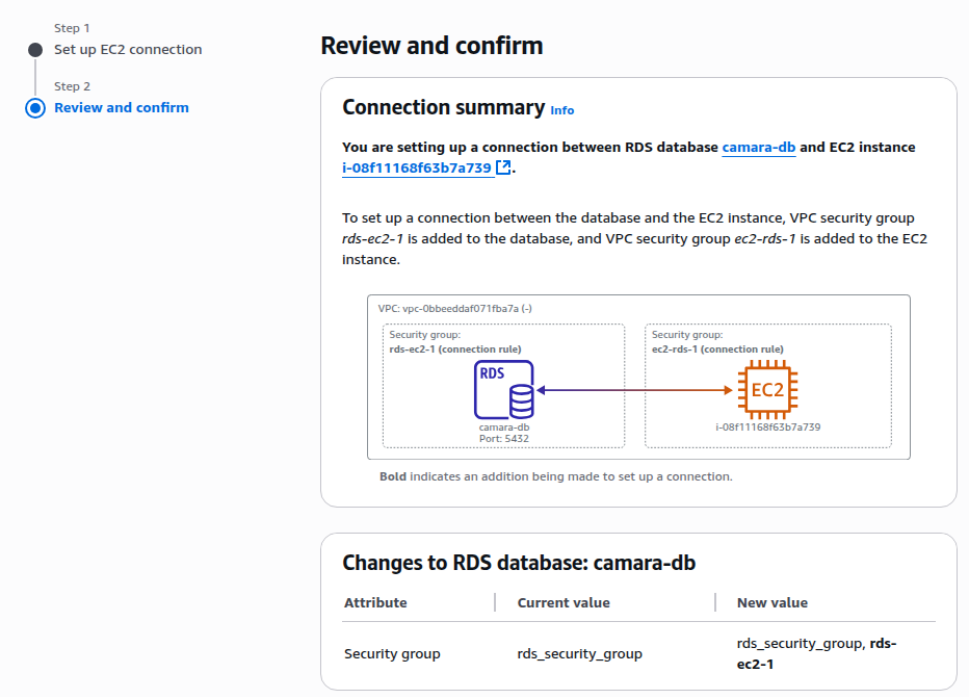


Figura 2: Coleta de dados por meio do arquivo.

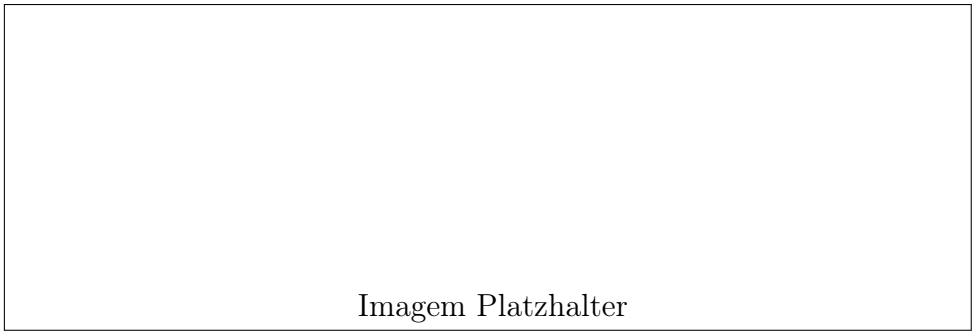


Figura 3: Criação do MariaDB na AWS.

# 2 Relatório Analítico

## 2.1 Introdução

A análise de dados públicos desempenha um papel crucial na promoção da transparência governamental e no combate a irregularidades. Utilizando uma plataforma de BI, como o Power BI, é possível transformar grandes volumes de dados em informações compreensíveis e acessíveis para a população e órgãos de fiscalização.

## 2.2 Demonstração

O Power BI foi integrado ao banco de dados PostgreSQL, permitindo consultas em tempo real e construção de dashboards dinâmicos.

Aqui fazemos uma conexão direta com o banco, e os dados do dashboard podem ser atualizados em tempo real.

Foram construídas 3 tabelas, e para facilitar o processo de construção de gráficos uma query foi feita, realizando o join das tabelas.

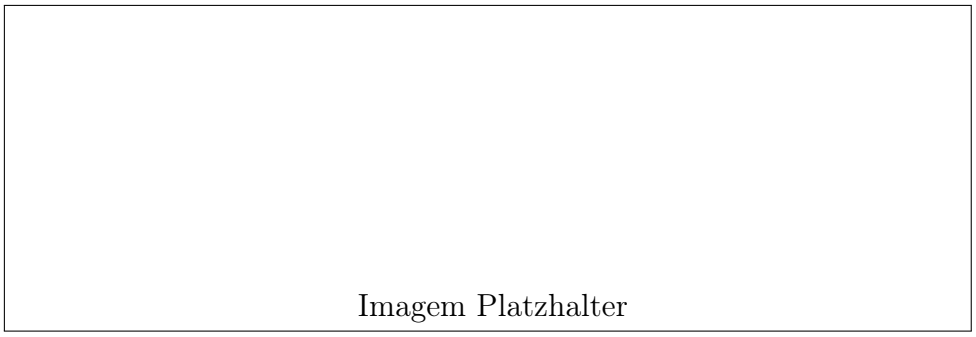


Figura 4: Base de dados em processo de criação.

E finalmente temos a nossa base de dados conectada com o PowerBI.

Temos diversas colunas interessantes que podemos usar na construção dos nossos gráficos:

- cnpj_cpf_despesa
- mes
- descricao
- descricao_especificacao
- valor_documento
- nome_deputado
- uf
- sigla_partido
- nome_fornecedor

## 2.3 Relatórios e tabelas

Foram construídas algumas páginas contendo alguns gráficos.

Primeiramente, podemos ver que o valor total de despesas no ano de 2022 foi de 221,4 milhões de reais. Há vários registros na coluna nome_deputado com o nome LIDERANÇA DO CIDADANIA, que possui o maior valor de despesas, totalizando R$ 912.660,00, seguido da Joenia Wapichana com o valor R$ 565.630,00 e do Jesus Sérgio com o valor R$ 549.970,00. A atividade com o maior gasto foi de Divulgação da Atividade Parlamentar, com R$ 52 milhões gastos, seguida por Passagem Aérea com R$ 48 milhões e Locação ou Fretamento do veículos automotores com R$ 29 milhões. Ao separar os gastos pelo fornecedor, temos o seguinte dado: Podemos notar que há uma grande discrepância de gastos com o fornecedor GOL quando comparado com outros fornecedores. Em outra página do dashboard, temos gráficos de gastos separados por Partido, UF e Mês do Ano. O partido que mais gastou foi o PL, totalizando R$ 32 milhões. O estado que mais gastou foi São Paulo, com R$ 26 milhões. E o mês com maior gasto foi o mês de Dezembro, com um gasto total de R$ 24 milhões.

## 2.4 Considerações finais

O PowerBI é uma ferramenta poderosa para análise de dados, por meio dela podemos construir gráficos dinâmicos que auxiliam em muito a análise e conseguimos tirar diversos insights. Outra grande vantagem é poder criar gráficos que retornam os dados diretamente do banco de dados, assim podemos ter gráficos atualizados e em tempo real.

# 3 Machine Learning

## 3.1 Introdução

O machine learning pode ser usado para encontrar padrões nos dados. Nessa análise, não vi nenhum tipo de machine learning que poderia nos trazer algum tipo de informação sobre os dados.

# 4 Vídeo

O vídeo de 10 minutos mostrando todo o projeto foi gravado e disponibilizado por meio do google drive através do link: link.

## 4.1 Dicionário de dados

## 4.2 Considerações finais

# Referências


[1] Portal de Dados Abertos da Câmara dos Deputados: https://dadosabertos.camara.leg.br/swagger/api.html

[2] dbdiagram: https://dbdiagram.io/

[3] Repositório do Projeto no Github: https://github.com/lucassiro/pi