ALUNO: FLÁVIO AUGUSTO ALÓ TORRES – 2020030477

BANCO DE DADOS 2: TRABALHO PRÁTICO 1

Primeira parte: desenvolver uma aplicação que consome dados de uma API de Dados

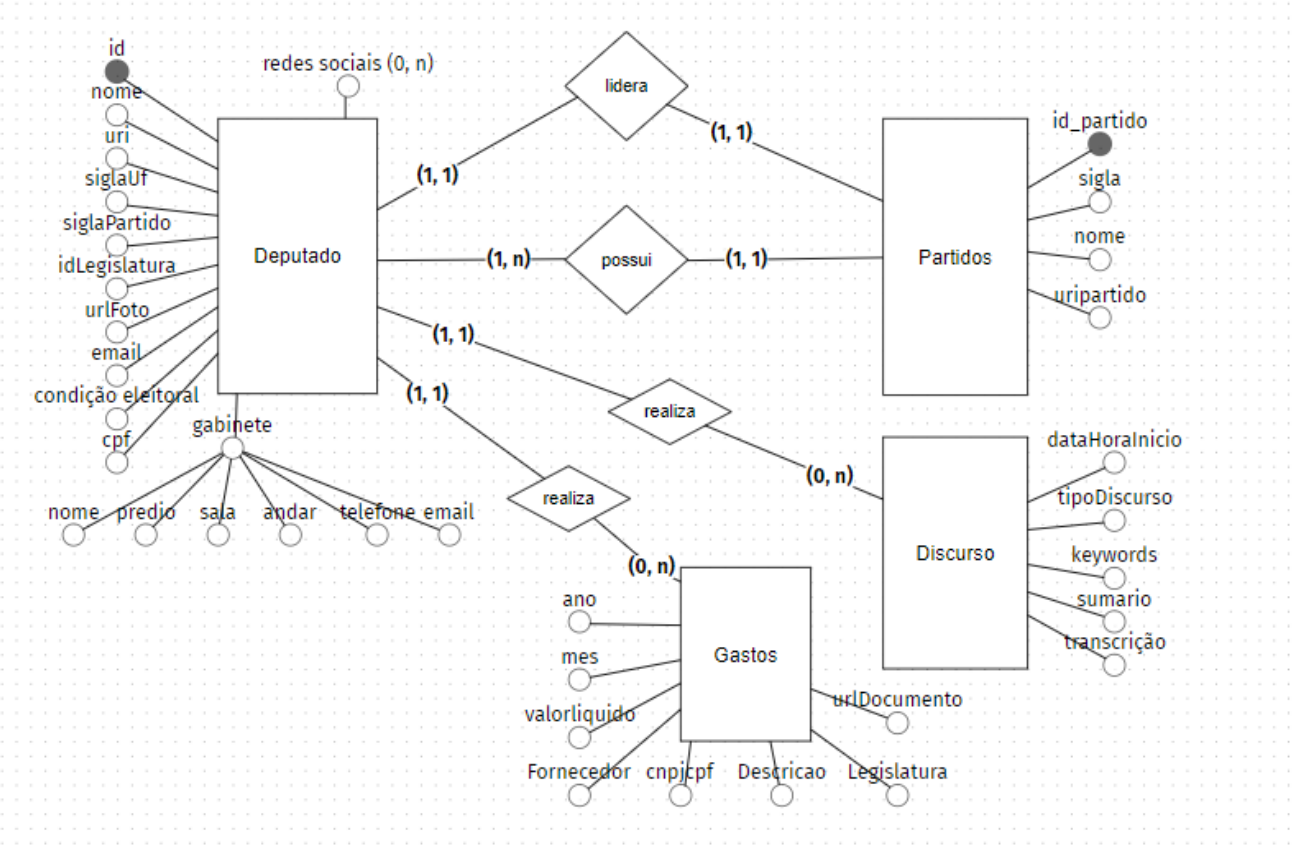
Para o desenvolvimento deste trabalho, eu escolhi consumir dados de uma API da Câmara dos Deputados, do governo. Para obter mais dados sobre a API RESTful utilizada, o link é: https://dadosabertos.camara.leg.br/swagger/api.html#api. Antes de eu começar a utilizar a API mencionada, busquei estudar uma outra API de dados governamentais (de compras e licitações federais), porém conforme fui tentando desenvolver, fiquei confuso em relação a alguns relacionamentos e com a semântica de dados, além das irregularidades e demora em algumas solicitações. Este acontecimento me fez migrar para uma API mais simples.

O Sistema Gerenciador de Bancos de Dados utilizado foi o PostgreSQL. Como os dados abertos ao público, o funcionamento das requisições da API utilizada não necessitam do uso de uma chave individual. Para obter-se os dados, as requisições são feitas via HTTP GET, onde podemos forneces os parâmetros necessários para a busca através da URL.

O primeiro passo do desenvolvimento foi pegar os dados gerais referentes aos atuais deputados através de uma requisição. Depois de tratar e inserir esses dados básicos na tabela deputados do banco de dados, o próximo passo foi pegar as informações adicionais dos deputados (ainda terminando de preencher a tabela deputado). Isso é possível através de requisições na API utilizando como parâmetro de busca o id de cada deputado. Vale ressaltar que um novo tipo de dado foi criado dentro do atributo gabinete, para poder guardar as informações de onde é possível encontrar o deputado.

O passo seguinte foi a criação de uma nova tabela para guardar o conteúdo multivalorado do atributo das redes sociais do deputado (ver o modelo entidade relacionamento abaixo). Além disso, um deputado pode realizar discursos, que são guardados na tabela discurso. A API fornecia alguns dados adicionais, mas após observar os dados, percebi que muitos atributos acabavam sendo um pouco irrelevantes ou nem estavam preenchidos na maioria dos casos. Esses atributos não foram utilizados.

O modelo entidade relacionamento (MER) concebido foi:

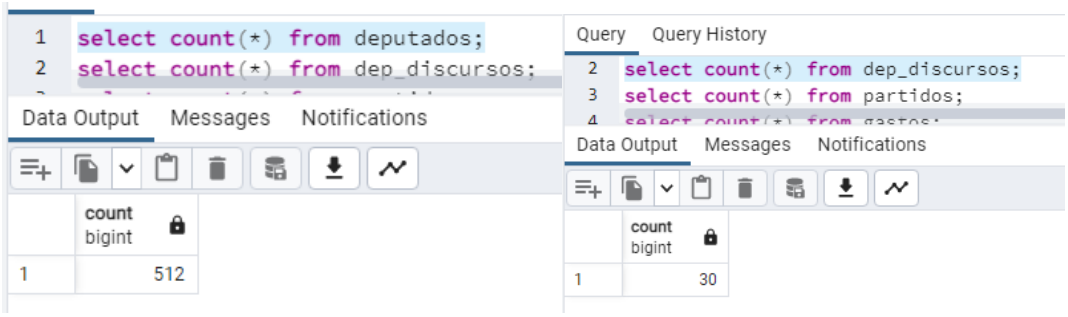


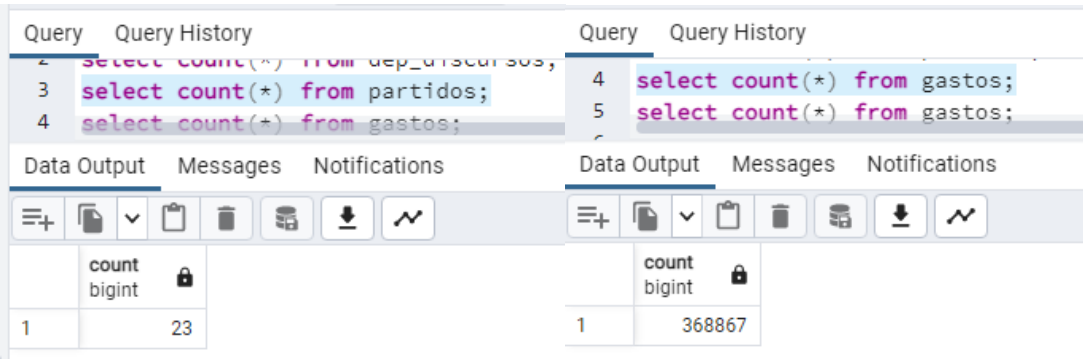
Na questão de permissões, foram criados dois tipos de usuário:

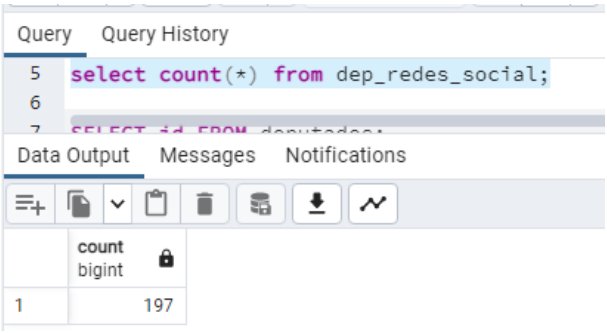
- Um usuário padrão, que tem acesso a ler todos os dados do banco
- Um usuário administrador, com permissão de inserir dados no banco (usuário e senha estão no arquivo SQL)

Não criei índices. As consultas estão retornando em um tempo imperceptível.

O print com o count das tabelas pode ser observado abaixo:





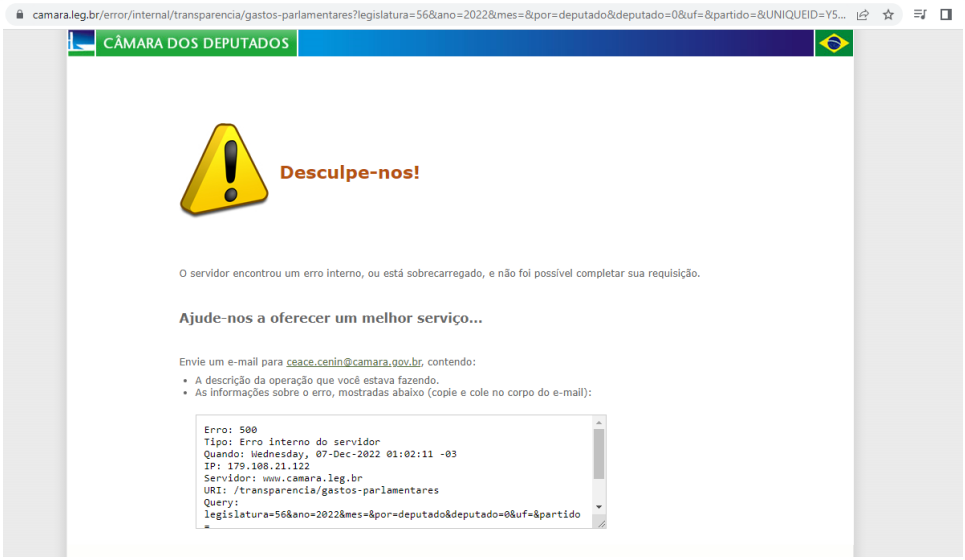


O arquivo SQL para a criação do banco se encontra na pasta zipada, junto podem ser encontrados um backup do banco de dados, além de uma pasta com dois arquivos python do código do programa.

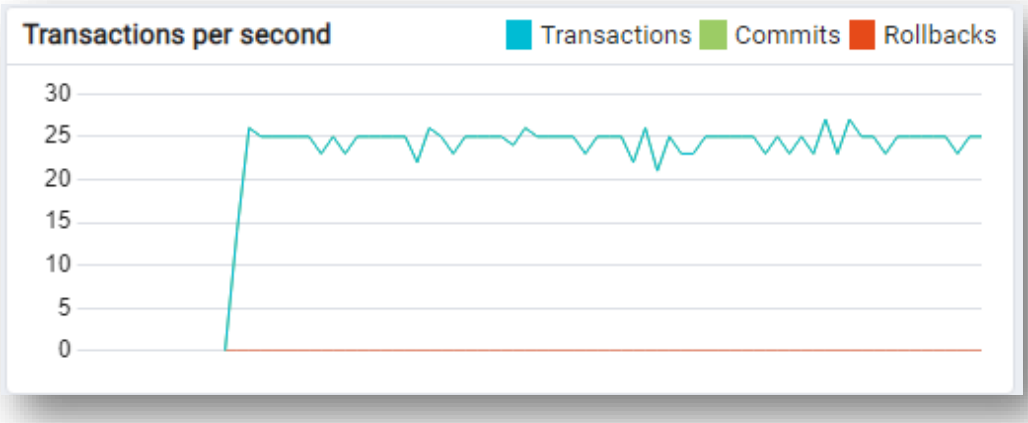
O vídeo do programa funcionando pode ser visto no link: [Trabalho Banco de Dados 2 - consumindo dados da API da câmara]

O github do trabalho pode ser encontrado aqui: [https://github.com/flavio055063/Api-dados-deputados-app-bd.git] (peço perdão se houver algo de errado no git pois é a primeira vez que uso).

Infelizmente, ainda é possível que ocorra algumas falhas no código python na hora de inserir os dados consultados da API (mas creio que já resolveram o problema). Não sei o que aconteceu, mas, até alguns dias atrás, a URL de requisição de detalhes sobre os deputados parou de funcionar, por causa de um erro interno no servidor da API.



Em questão de desempenho, enquanto a aplicação estava inserindo uma quantidade enorme de tuplas, uma a uma. Do jeito que está o software está funcionando, toda a vez que a função de inserir tupla é chamada, uma nova conexão é aberta e logo após fechada. O PostgreSQL manteu uma média de 25 transações por segundo (sem usar o processamento paralelo).



Tentando inserir os valores dos gastos de cada deputado, meu computador i5 de 4° geração com 16 GB de RAM DDR3 já levou mais de 80 minutos, conseguindo inserir apenas 53813. Isso dá uma média de 673 tuplas por minuto de funcionamento.

```
# Abrindo em formato de arquivo os gastos, porque a API está com algum problema

with open("gastos2022.json", encoding='utf-8') as gastos_json:
    objetos_gastos = json.load(gastos_json)

dados_gastos = objetos_gastos['dados']
df_gastos = pd.DataFrame(dados_gastos)
df_gastos.info()
# como existem gastos que nao pertencem a um deputado, e sim a liderança, excluirei os
# dados que nao possuirem idDeputado... o dataframe identificou o idDeputado como float,
# entao terei que converter para int.
df_gastos['idDeputado'] = pd.to_numeric(df_gastos['idDeputado'], errors='coerce')
df_gastos = df_gastos.dropna(subset=['idDeputado'])
df_gastos['idDeputado'] = df_gastos['idDeputado'].apply(int)


#inserindo os dados relevantes de gastos do dataframe no PostgreSQL
for i in df_gastos.index:
    sql = """
            INSERT into public.gastos (id_deputado, legislatura, descricao, fornecedor, cnpjCPF, valorLiquido, mes, ano, url_documento)
            values('%s','%s','%s','%s','%s','%s', '%s', '%s', '%s');
            """ % (df_gastos['idDeputado'][i], df_gastos['legislatura'][i], df_gastos['descricao'][i], df_gastos['fornecedor'][i], df_gastos['cnpjCPF'][i], df_
    inserir_db(sql)
```

81m 55.5s Python

Para que as inserções ocorressem mais rápido, tive que adaptar a parte do código que insere os gastos na tabela, abrindo apenas uma conexão, inserindo tudo e por fim encerrando a conexão. Nesse método, todas as tuplas foram inseridas em menos de 5 minutos

Para fazer o código funcionar, execute o *main.py* dentro da pasta Modelo MVC.