Termo de Execução Descentralizada nº 19 – Realização de Estudos em Compras Públicas

Documento:

Relatório de Extração de Dados POLÍTICOS

Data de Emissão:

10/02/2020

Elaborado por:

Escola Nacional de Administração Pública em parceria com Laboratório de Tecnologias da Tomada de Decisão – LATITUDE.UnB

HISTÓRICO DE REVISÕES

Data	Versão	Autor	Descrição
22/11/2019	1.0	Anderson Alves de Oliveira	Inclusão dos dados da base Políticos
25/11/2019	1.1	Leticia Valle	Revisão
10/02/2020	1.2	Leticia Valle	Atualização do documento

Universidade de Brasília – UnB
Campus Universitário Darcy Ribeiro - FT – ENE – Latitude
CEP 70.910-900 – Brasília-DF
Tel.: +55 61 3107-5598 – Fax: +55 61 3107-5590

SUMÁRIO

1. INTRODUÇÃO

Este relatório tem como objetivo documentar o processo de extração, tratamento e carregamento de dados da base de Cadastro de Acordos de Políticos.

O Repositório de dados eleitorais é uma compilação de informações brutas das eleições, desde as de 1945, voltada para pesquisadores, imprensa e pessoas interessadas em analisar os dados de eleitorado, candidaturas, resultados e prestação de contas [1].

Todos os arquivos fornecidos estão em formato TXT e podem ser importados para qualquer programa estatístico, base de dados ou planilha eletrônica. Consultas, filtros e cruzamentos são de responsabilidade do pesquisador. É importante ler o arquivo de instruções e atentar à data de geração do arquivo, para fazer as importações e as consultas de forma adequada.

Estão incompletos os dados de candidatos e de resultados das eleições de 1994 a 2002. Está sendo realizada uma revisão nas fontes de dados e, conforme os trabalhos forem concluídos, os arquivos serão substituídos.

Para a realização do trabalho, foram usadas as ferramentas de ETL Apache Airflow e o banco de dados SQL Server, rodando em um servidor Windows, requisito da equipe do Ministério da Economia.

2. ORIGEM DOS DADOS EXTRAÍDOS

Os dados podem ser encontrados na página do portal do TSE.

A URL do local de origem dos dados é:
http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais

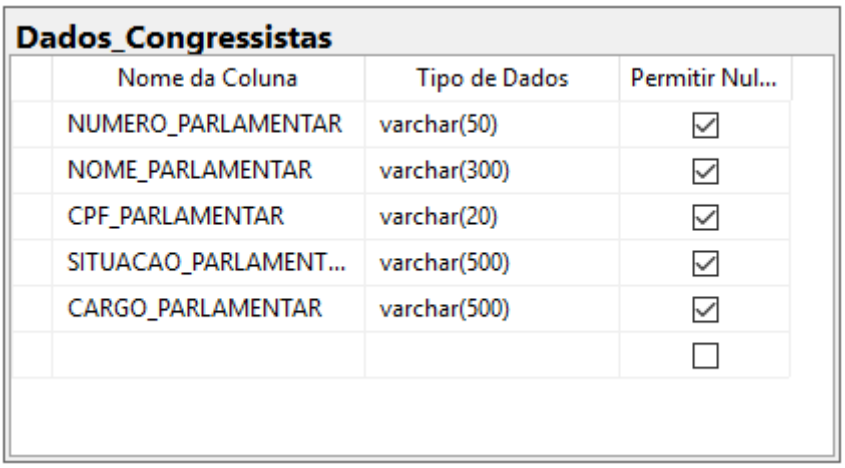
3. QUANTITATIVO DE DADOS

A base de Políticos possui três tabelas com três colunas cada uma das tabelas.

4. MODELAGEM DO BANCO DE DADOS

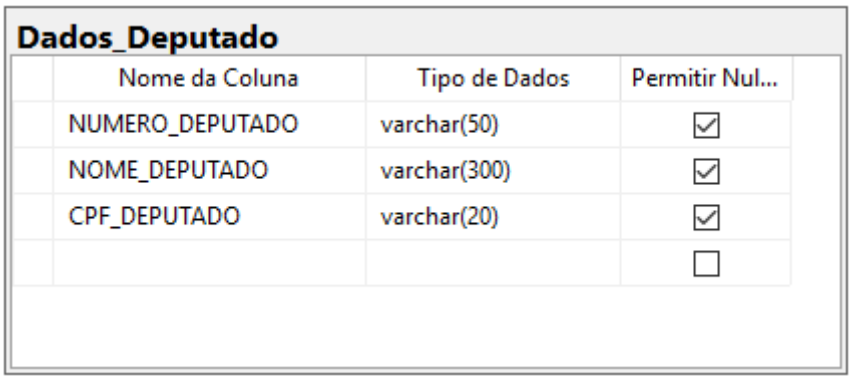
Após análise da base de Políticos, foi realizada a modelagem dos dados para posterior criação do banco e das tabelas. Nesse projeto, somente três tabelas se fazem necessárias no modelo. Como as tabelas são independentes no modelo, elas não possuem nenhum tipo de relacionamento.

A seguir são apresentados o modelo lógico do banco e o script para a criação da tabela.



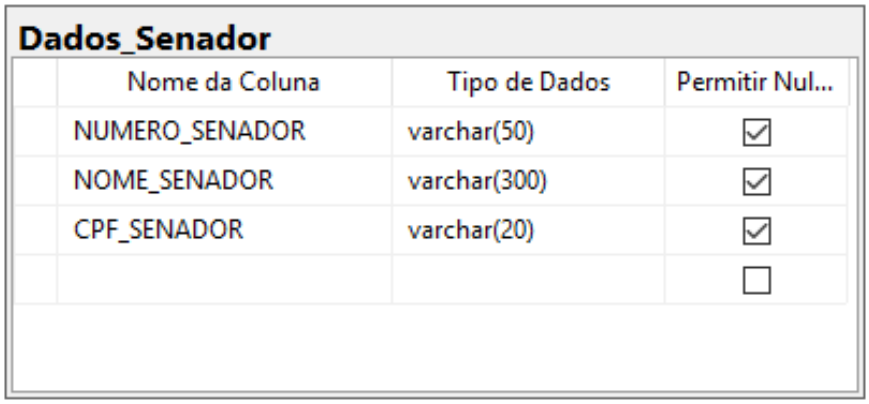
Dados_Congressistas

Nome da Coluna	Tipo de Dados	Permitir Nul...
NUMERO_PARLAMENTAR	varchar(50)	☑
NOME_PARLAMENTAR	varchar(300)	☑
CPF_PARLAMENTAR	varchar(20)	☑
SITUACAO_PARLAMENT...	varchar(500)	☑
CARGO_PARLAMENTAR	varchar(500)	☑
		☐



Dados_Deputado

Nome da Coluna	Tipo de Dados	Permitir Nul...
NUMERO_DEPUTADO	varchar(50)	☑
NOME_DEPUTADO	varchar(300)	☑
CPF_DEPUTADO	varchar(20)	☑
		☐



Dados_Senador

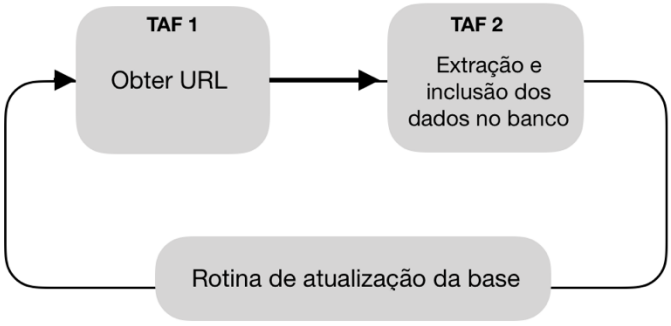
Nome da Coluna	Tipo de Dados	Permitir Nul...
NUMERO_SENADOR	varchar(50)	☑
NOME_SENADOR	varchar(300)	☑
CPF_SENADOR	varchar(20)	☑
		☐

```
CREATE TABLE Dados_Deputado (
NUMERO_DEPUTADO VARCHAR(50),
NOME_DEPUTADO VARCHAR(300),
CPF_DEPUTADO VARCHAR(20));

CREATE INDEX IDX_Dados_Deputado on Dados_Deputado(CPF_DEPUTADO);



CREATE TABLE Dados_Senador (
NUMERO_SENADOR VARCHAR(50),
NOME_SENADOR VARCHAR(300),
CPF_SENADOR VARCHAR(20));

CREATE INDEX IDX_Dados_Senador on Dados_senador(CPF_SENADOR);



CREATE TABLE Dados_Congressistas (
NUMERO_PARLAMENTAR varchar(50),
NOME_PARLAMENTAR varchar(300),
CPF_PARLAMENTAR varchar(20),
SITUACAO_PARLAMENTAR varchar(500),
CARGO_PARLAMENTAR varchar(500))

CREATE INDEX IDX_Dados_Congressistas on
Dados_Congressistas(CPF_PARLAMENTAR);
```

5. FLUXOS DE ETL

A sessão a seguir apresenta um resumo do trabalho de extração, tratamento e carregamento da base de Políticos.

O trabalho de ETL foi desenvolvido na ferramenta Apache Airflow e conta com 2 tarefas diretas, apresentadas no diagrama de bloco a seguir.



Tarefa 1 - Download do código JSON através da URL do API disponibilizado

Rotina: **get_url**

Com o auxílio da biblioteca request [2], é obtido o link do api de acesso aos dados da base Políticos.

A partir do API disponibilizado no site do Portal da transparência, ele disponibiliza um JSON [3] com os dados.

Tarefa 2 - Extração dos dados e importação para a tabela no banco SQL Server

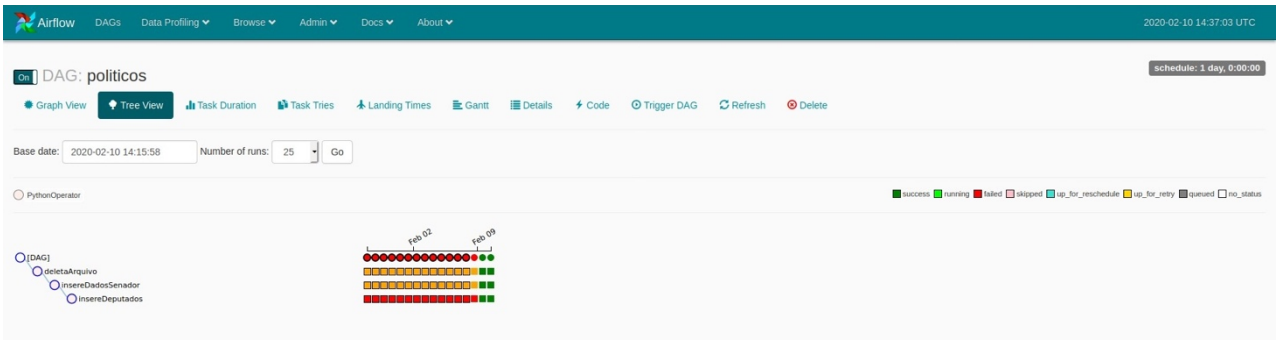
Rotina*:* **copy_to_sqlserver**

Com o arquivo politicos_file.json e com o auxílio da biblioteca pyodbc [4], é possível conectar no banco SQL Server e importar os dados da tabela para o banco.

Foi utilizado o Driver ODBC [5] para SQL Sever para Linux.

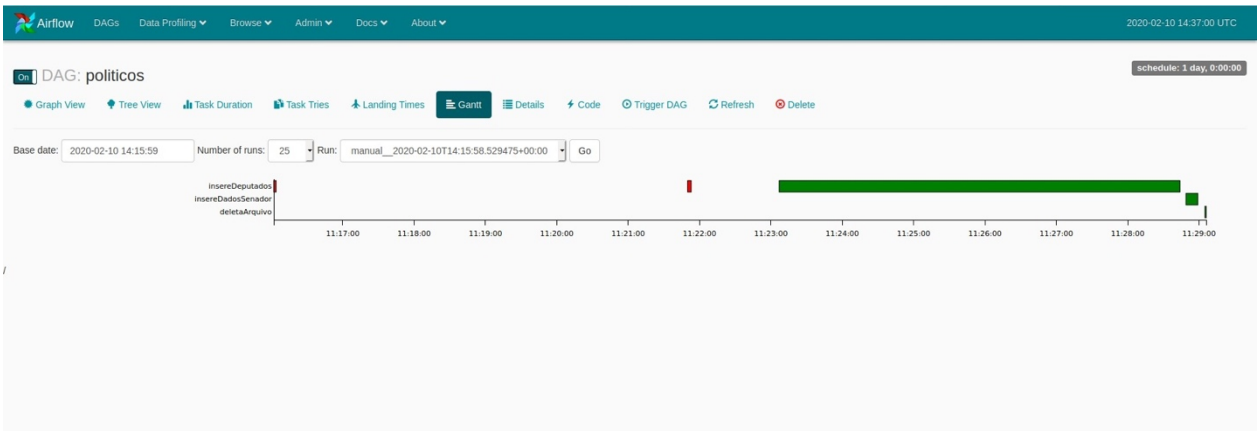
```
conn_string = 'DRIVER={ODBC Driver 17 for SQL Server};SERVER='+
              host+';PORT=1433;DATABASE='+dbname+';UID='+
              username+';PWD='+ password +
              ';UseNTLMv2=yes;TDS_Version=8.0'
```

A imagem a seguir representa as tarefas ETL dentro da ferramenta Apache Airflow.



6. DURAÇÃO DAS ROTINAS ETL

A duração de importação dos dados depende do tempo de execução de todas as rotinas ETL. A imagem a seguir apresenta o tempo de execução de cada tarefa.



Observa-se que cada tarefa possui um tempo de execução distinto dependendo da complexidade da tarefa. Para a base Políticos, o tempo de execução das rotinas foi de cerca de 8 minutos no total.

7. FLUXO DE TRATAMENTO DE ERROS

Os possíveis erros de execução das tarefas foram tratados à partir da criação de tarefas independentes. Cada tarefa procura pelos arquivos necessários para sua execução e gera como output arquivos que serão usados como entrada de outras tarefas. Dessa forma, caso alguma tarefa não seja cumprida, ou apresente erro, ao reiniciar o sistema, a rotina de tarefas será retomada e os arquivos salvos de tarefas anteriores continuam salvos.

Além disso, afim de evitar futuros erros relacionados à mudança da URL do API que contém o JSON para download, a tarefa inicial do ETL é fazer um filtro usando a URL do API disponibilizado na página, de forma a evitar erros gerados por mudanças estruturais na pagina de download do JSON.

8. FLUXO DE AGENDAMENTO DE ROTINAS

A base de Políticos é atualizada mensalmente. Dessa forma, no código de configuração do Apache Airflow, foi inserida uma rotina de atualização da base a cada 30 dias.

```
dag = DAG(dag_id='politicos, default_args=args, schedule_interval=timedelta(days=30))
```

9. ESTIMATIVA DE CRESCIMENTO

A estimativa de crescimento da base depende do volume atual de dados acrescido da estimativa de volume que vai ser inserido nas atualizações mensais. Como a atualização é feita mensamente e os dados são totalmente substituídos pela inserção dos novos registros do arquivo JSON [2], não houve tempo hábil para estimar qual é o volume incremental da base à cada atualização.

De qualquer forma, o volume total da base é de menos de 1MB, o que implica que mesmo após várias atualizações futuras, a base não deverá passar de 2 MB.

10. AUXÍLIO NOS ESTUDO DE COMPRAS PUBLICAS

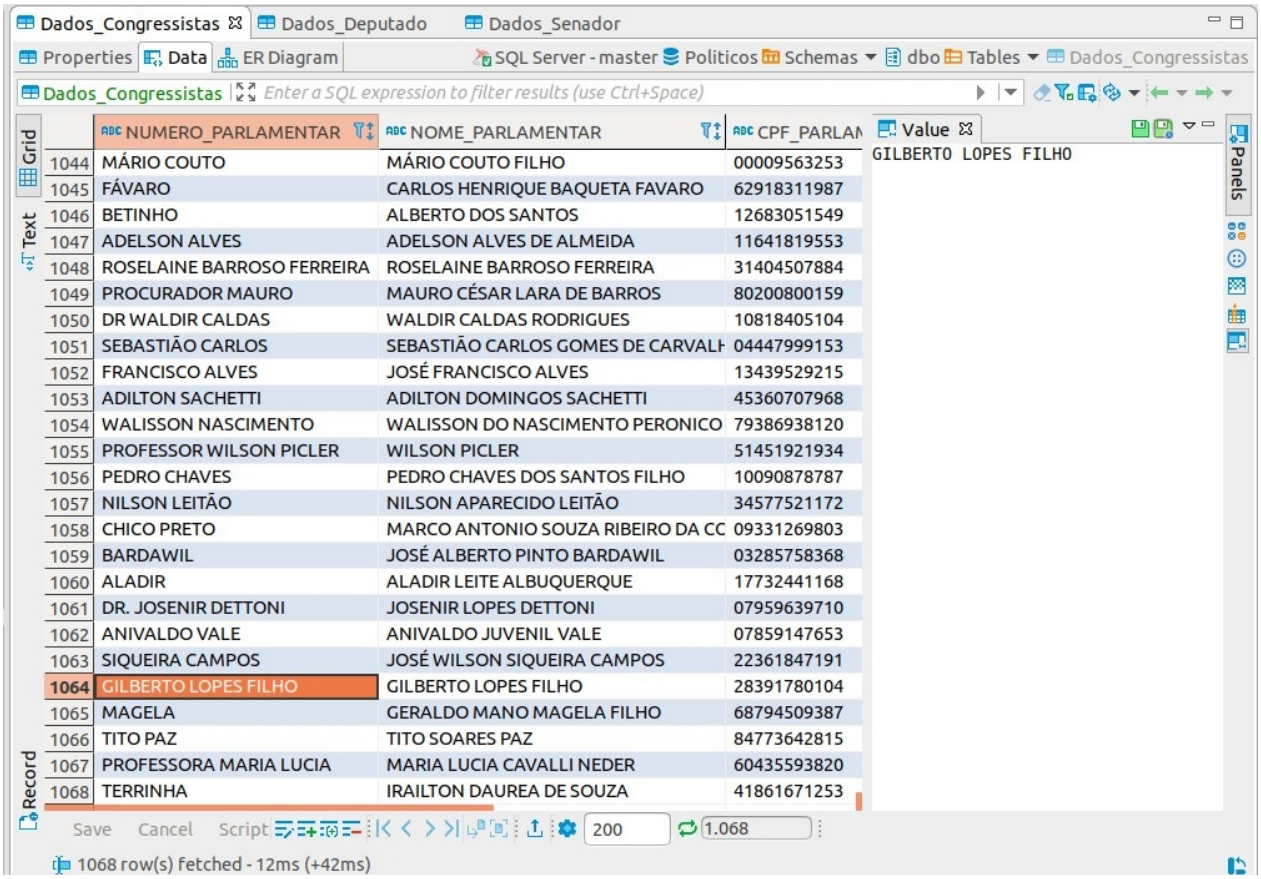
Os dados referentes à base Políticos poderão auxiliar na análise preventiva para os próximos pleitos eleitorais.

Os resultados totalizados são obtidos após o processamento pelo sistema de totalização. Nessa parte, teremos os arquivos de votação nominal e para partido/legenda por município e zona eleitoral.

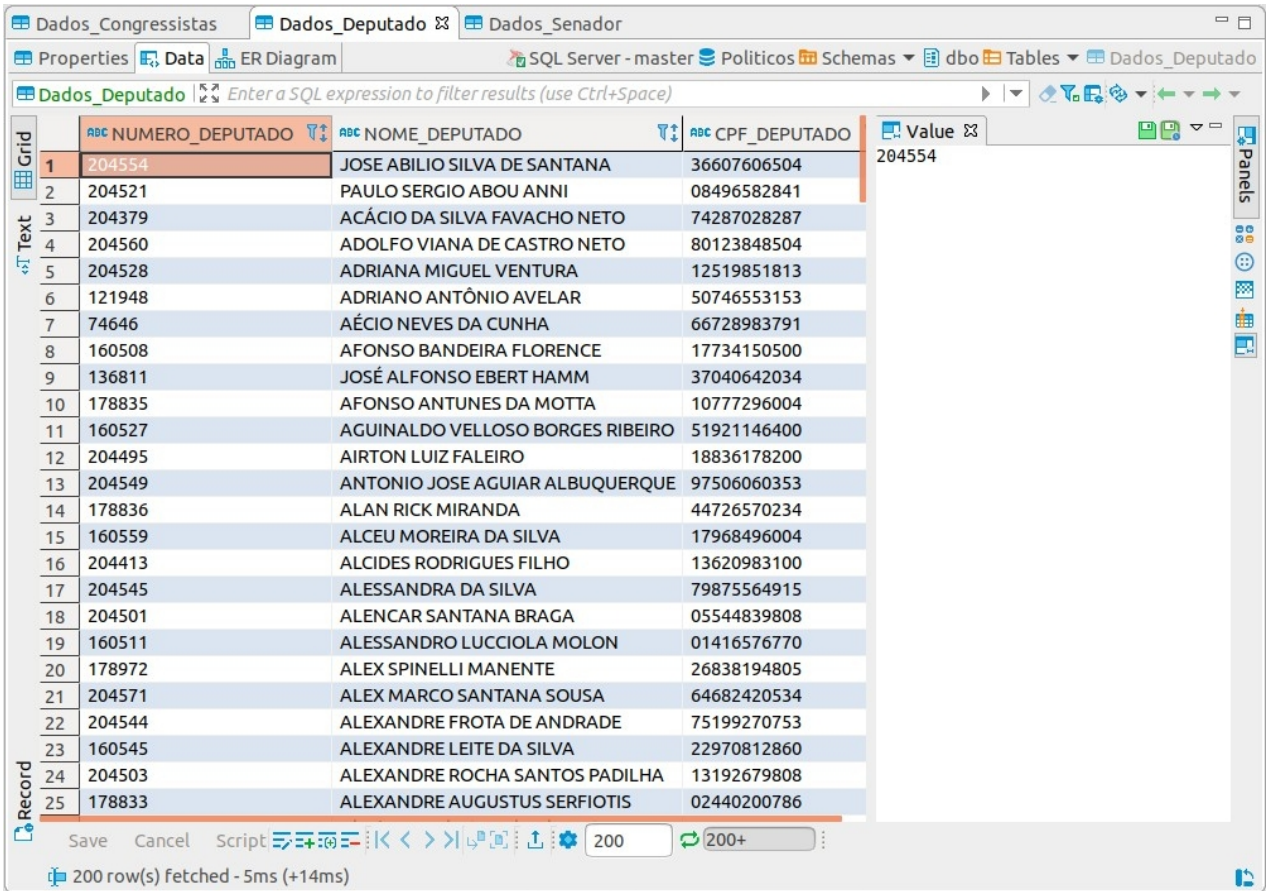
Os boletins de urna são os arquivos de BU das seções eleitorais. É importante ressaltar que, nesses arquivos, constam as informações da forma como saíram da urna eletrônica: sem o processamento do sistema de totalização.

11. EVIDÊNCIA DOS DADOS IMPORTADOS

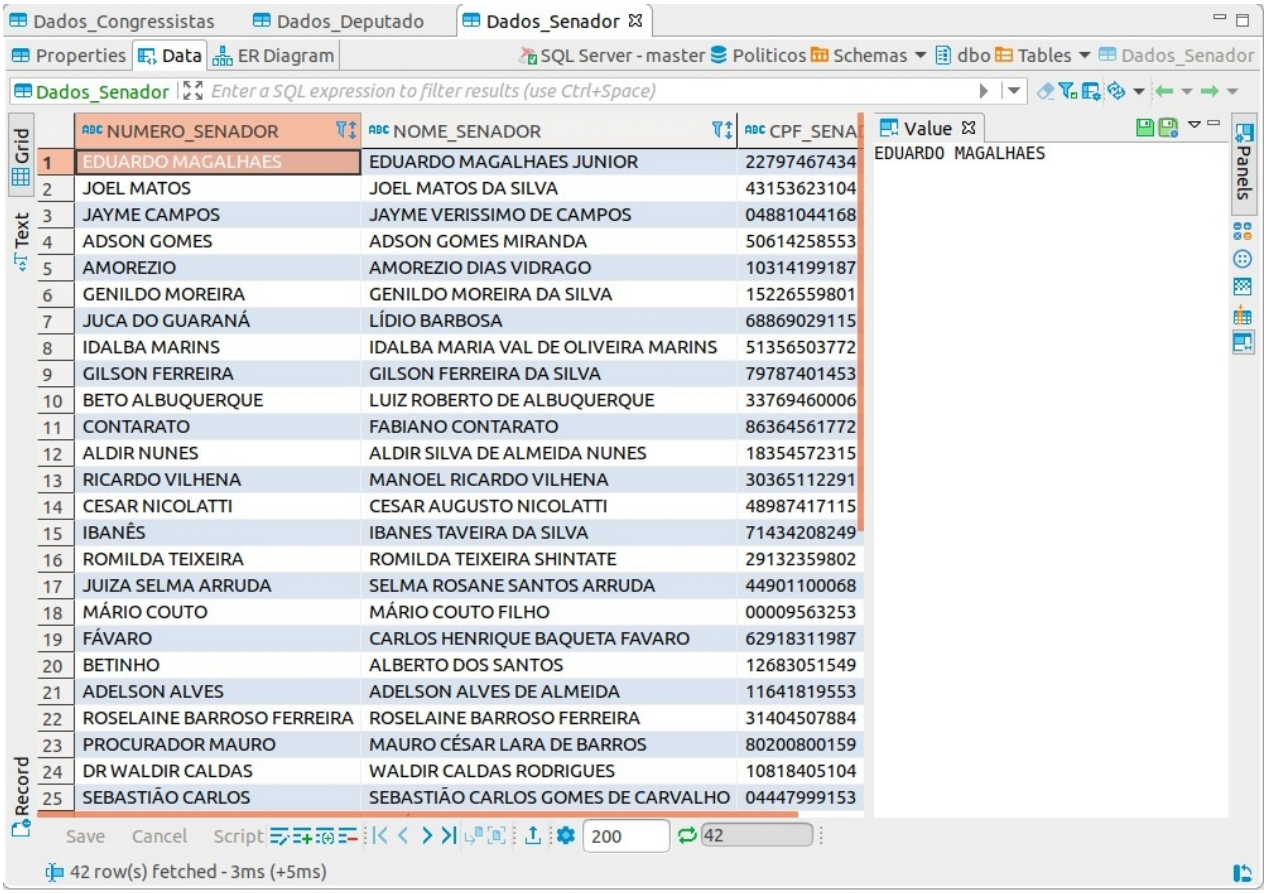
Os dados importados corretamente no SQL Server são apresentados a seguir.



	NUMERO_PARLAMENTAR	NOME_PARLAMENTAR	CPF_PARLAM
1044	MÁRIO COUTO	MÁRIO COUTO FILHO	00009563253
1045	FÁVARO	CARLOS HENRIQUE BAQUETA FAVARO	62918311987
1046	BETINHO	ALBERTO DOS SANTOS	12683051549
1047	ADELSON ALVES	ADELSON ALVES DE ALMEIDA	11641819553
1048	ROSELAINE BARROSO FERREIRA	ROSELAINE BARROSO FERREIRA	31404507884
1049	PROCURADOR MAURO	MAURO CÉSAR LARA DE BARROS	80200800159
1050	DR WALDIR CALDAS	WALDIR CALDAS RODRIGUES	10818405104
1051	SEBASTIÃO CARLOS	SEBASTIÃO CARLOS GOMES DE CARVALH	04447999153
1052	FRANCISCO ALVES	JOSÉ FRANCISCO ALVES	13439529215
1053	ADILTON SACHETTI	ADILTON DOMINGOS SACHETTI	45360707968
1054	WALISSON NASCIMENTO	WALISSON DO NASCIMENTO PERONICO	79386938120
1055	PROFESSOR WILSON PICLER	WILSON PICLER	51451921934
1056	PEDRO CHAVES	PEDRO CHAVES DOS SANTOS FILHO	10090878787
1057	NILSON LEITÃO	NILSON APARECIDO LEITÃO	34577521172
1058	CHICO PRETO	MARCO ANTONIO SOUZA RIBEIRO DA CC	09331269803
1059	BARDAWIL	JOSÉ ALBERTO PINTO BARDAWIL	03285758368
1060	ALADIR	ALADIR LEITE ALBUQUERQUE	17732441168
1061	DR. JOSENIR DETTONI	JOSENIR LOPES DETTONI	07959639710
1062	ANIVALDO VALE	ANIVALDO JUVENIL VALE	07859147653
1063	SIQUEIRA CAMPOS	JOSÉ WILSON SIQUEIRA CAMPOS	22361847191
1064	GILBERTO LOPES FILHO	GILBERTO LOPES FILHO	28391780104
1065	MAGELA	GERALDO MANO MAGELA FILHO	68794509387
1066	TITO PAZ	TITO SOARES PAZ	84773642815
1067	PROFESSORA MARIA LUCIA	MARIA LUCIA CAVALLI NEDER	60435593820
1068	TERRINHA	IRAILTON DAUREA DE SOUZA	41861671253



	NUMERO_DEPUTADO	NOME_DEPUTADO	CPF_DEPUTADO
1	204554	JOSE ABILIO SILVA DE SANTANA	36607606504
2	204521	PAULO SERGIO ABOU ANNI	08496582841
3	204379	ACÁCIO DA SILVA FAVACHO NETO	74287028287
4	204560	ADOLFO VIANA DE CASTRO NETO	80123848504
5	204528	ADRIANA MIGUEL VENTURA	12519851813
6	121948	ADRIANO ANTÔNIO AVELAR	50746553153
7	74646	AÉCIO NEVES DA CUNHA	66728983791
8	160508	AFONSO BANDEIRA FLORENCE	17734150500
9	136811	JOSÉ ALFONSO EBERT HAMM	37040642034
10	178835	AFONSO ANTUNES DA MOTTA	10777296004
11	160527	AGUINALDO VELLOSO BORGES RIBEIRO	51921146400
12	204495	AIRTON LUIZ FALEIRO	18836178200
13	204549	ANTONIO JOSE AGUIAR ALBUQUERQUE	97506060353
14	178836	ALAN RICK MIRANDA	44726570234
15	160559	ALCEU MOREIRA DA SILVA	17968496004
16	204413	ALCIDES RODRIGUES FILHO	13620983100
17	204545	ALESSANDRA DA SILVA	79875564915
18	204501	ALENCAR SANTANA BRAGA	05544839808
19	160511	ALESSANDRO LUCCIOLA MOLON	01416576770
20	178972	ALEX SPINELLI MANENTE	26838194805
21	204571	ALEX MARCO SANTANA SOUSA	64682420534
22	204544	ALEXANDRE FROTA DE ANDRADE	75199270753
23	160545	ALEXANDRE LEITE DA SILVA	22970812860
24	204503	ALEXANDRE ROCHA SANTOS PADILHA	13192679808
25	178833	ALEXANDRE AUGUSTUS SERFIOTIS	02440200786



	NUMERO_SENADOR	NOME_SENADOR	CPF_SENA
1	EDUARDO MAGALHAES	EDUARDO MAGALHAES JUNIOR	22797467434
2	JOEL MATOS	JOEL MATOS DA SILVA	43153623104
3	JAYME CAMPOS	JAYME VERISSIMO DE CAMPOS	04881044168
4	ADSON GOMES	ADSON GOMES MIRANDA	50614258553
5	AMOREZIO	AMOREZIO DIAS VIDRAGO	10314199187
6	GENILDO MOREIRA	GENILDO MOREIRA DA SILVA	15226559801
7	JUCA DO GUARANÁ	LÍDIO BARBOSA	68869029115
8	IDALBA MARINS	IDALBA MARIA VAL DE OLIVEIRA MARINS	51356503772
9	GILSON FERREIRA	GILSON FERREIRA DA SILVA	79787401453
10	BETO ALBUQUERQUE	LUIZ ROBERTO DE ALBUQUERQUE	33769460006
11	CONTARATO	FABIANO CONTARATO	86364561772
12	ALDIR NUNES	ALDIR SILVA DE ALMEIDA NUNES	18354572315
13	RICARDO VILHENA	MANOEL RICARDO VILHENA	30365112291
14	CESAR NICOLATTI	CESAR AUGUSTO NICOLATTI	48987417115
15	IBANÊS	IBANES TAVEIRA DA SILVA	71434208249
16	ROMILDA TEIXEIRA	ROMILDA TEIXEIRA SHINTATE	29132359802
17	JUIZA SELMA ARRUDA	SELMA ROSANE SANTOS ARRUDA	44901100068
18	MÁRIO COUTO	MÁRIO COUTO FILHO	00009563253
19	FÁVARO	CARLOS HENRIQUE BAQUETA FAVARO	62918311987
20	BETINHO	ALBERTO DOS SANTOS	12683051549
21	ADELSON ALVES	ADELSON ALVES DE ALMEIDA	11641819553
22	ROSELAINE BARROSO FERREIRA	ROSELAINE BARROSO FERREIRA	31404507884
23	PROCURADOR MAURO	MAURO CÉSAR LARA DE BARROS	80200800159
24	DR WALDIR CALDAS	WALDIR CALDAS RODRIGUES	10818405104
25	SEBASTIÃO CARLOS	SEBASTIÃO CARLOS GOMES DE CARVALHO	04447999153

12. BIBLIOGRAFIA

1. Repositorio de dados eleitorais. Disponível em: < http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais>. Acesso em: 24 de outubro de 2019.

2. Request library documentation. Disponível em: <https://docs.python.org/3/library/urllib.request.html>. Acesso em: 21 de outubro de 2019.

3. JSON documentation. Disponível em: < http://www.json.org/>. Acesso em: 06 de novembro de 2019.

4. Driver SQL Python. Disponível em: < https://docs.microsoft.com/pt-br/sql/connect/python/pyodbc/python-sql-driver-pyodbc?view=sql-server-ver15>. Acesso em: 21 de outubro de 2019.

5. Miscellaneous operating system interfaces. Disponível em: <https://docs.python.org/3.4/library/os.html>. Acesso em: 21 de outubro de 2019.

Escola Nacional de Administração Pública

Laboratório de Tecnologias da Tomada de Decisão – LATITUDE

www.enap.gov.br – www.redes.unb.br