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

Documento:

Relatório de Extração de Dados SANÇÕES - RJ

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

ENAP

Diogo Godinho Ramos Costa
Presidente

Diana Magalhães de Souza Coutinho
Diretor de Pesquisa e Pós-Graduação

Leonardo Monteiro Monasterio
Coordenador Geral de Ciência de Dados

Pedro Masson Sesconetto Souza
Coordenador de Ciência de Dados

EQUIPE TÉCNICA
Cristiano Alves Bezerra
Adalberto Bleme
Wanderson Maia Nascimento

UNIVERSIDADE DE BRASÍLIA

Márcia Abrahão Moura
Reitora

Marileusa Dosolina Chiarello
Diretora do Centro de Apoio ao Desenvolvimento Tecnológico – CDT

Rafael Timóteo de Sousa Júnior
Coordenador do Laboratório de Tecnologias da Tomada de Decisão – LATITUDE

EQUIPE TÉCNICA
Pesquisadores Sêniores
Ugo Silva Dias

EQUIPE TÉCNICA
Leticia Moreira Valle
Eduardo Calandrini Rocha da Costa
Anderson Alves de Oliveira
Andréia Campos Santana
Caio Matheus Campos de Oliveira
Danilo Santos Cardoso
Danilo Santos de Sales
Flávio Sousa da Vitória
Leonardo Pires Simões Vasconcelos
Samyra Lima Pereira

HISTÓRICO DE REVISÕES

Data	Versão	Autor	Descrição
01/11/2019	1.0	Danilo Cardoso	Inclusão dos dados da base SANÇÕES RJ
15/11/2019	1.1	Leticia Valle	Revisão
31/01/2020	1.2	Leticia Valle	Atualização do documento
10/02/2020	1.3	Leticia Valle	Atualização do script de criação de tabelas

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 carregamento de dados da base de Sanções dadas pelo Governo do Estado do Rio de Janeiro por meio do Sistema Integrado de Gestão de Aquisições (Sanções - RJ).

A base SANÇÕES-RJ apresenta a relação de pessoas físicas ou jurídicas que receberam uma sanção do Governo do Estado do Rio de Janeiro. Esta sanção pode ser de um dos quatro tipos: advertência, multa, vigente e decorrida.

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 “Sistema Integrado de Gestão de Aquisições”, na categoria “Fornecedores” e por último na opção “Consultar Sanções”.

Os dados estão disponíveis de maneira direta na URL: http://www.compras.rj.gov.br/Portal-Siga/Sancao/buscar.action#/.

3. QUANTITATIVO DE DADOS

A base SANCOES-RJ possui apenas uma base de dados com uma lista de pessoas físicas e jurídicas sancionadas. Ao inspecionar mais atentamente os itens, é possível observar 12 itens para cada observação. Em formato tabular, é possível observar 6 itens e para realizar a captação dos outros 6, é necessário clicar e averiguar cada registro. A imagem abaixo ilustra o exemplo de um dos registros.

Detalhe da Sanção

Nome/Razão Social:
ZURIEL DE IGUAÇU COMERCIO E REPRESENTAÇÃO LTDA-ME

CPF/CNPJ:	Status da Sanção:	Data de Efetivação:
24.593.578/0001-67	Advertido	04/04/2019
Enquadramento Legal:	**Número do processo:**	**Órgão/Entidade Apenadora:**
Lei Federal Nº 8.666/93, art. 87, Inc. I.	E-26/004/104/2019	CECIERJ - FUND CENTRO CIÊN EDUC SUP DISTÂN DO EST RJ
Prazo:	**Data Início:**	**Data Final:**
Não possui	Não possui	Não possui

Justificativa:
Penalidade de Advertência publicada em DOERJ ANO XLV N.º 064 - PARTE I DE 04/04/2019 PÁG. 31 RETIFICADO PELO DOERJ ANO XLV N.º 171 DE 10/09/2019 PÁG. 18 pelo fato de não ter apresentado documentação de habilitação, referente ao Pregão n.º 19/2018, na qual foi arrematante.

Motivo:
Inexecução total ou parcial do contrato.

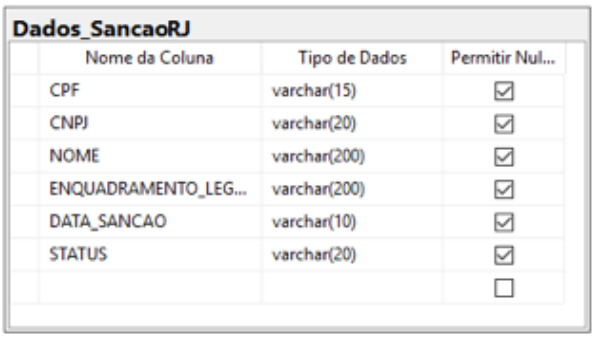
Devido à baixa importância das informações não disponíveis no formato tabular (número do processo, prazo, datas, justificativa e motivo) para os propósitos do Ministério da Economia, foi decidido usar o formato mais simples como forma mais rápida e direta de captar as informações, deixando de lado as informações detalhadas.

Nome/Razão Social	CPF/CNPJ	Enquadramento Legal	Data de Efetivação	Órgão/Entidade Apenadora	Status
ZURIEL DE IGUAÇU COMERCIO E REPRESENTAÇÃO LTDA-ME	24.593.578/0001-67	Lei Federal Nº 8.666/93, art. 87, Inc. I.	04/04/2019	CECIERJ - FUND CENTRO CIÊN EDUC SUP DISTÂN DO EST RJ	Advertido
X-OFFICE SUPPLIES DIST. E SERV INF. LTDA	12.028.146/0001-24	Lei Federal Nº 8.666/93, art. 87, Inc. I.	07/03/2014	SEFAZ - SECRETARIA DE ESTADO DE FAZENDA	Advertido
WNT TRANSPORTES E LOCAÇÕES EIRELI-ME	04.494.163/0001-29	Lei Federal Nº 8.666/93, art. 87, Inc. I.	21/06/2016	FUNPERJ - FUNDO ESPECIAL DA PROCURADORIA GERAL DO ERJ	Advertido

4. MODELAGEM DO BANCO DE DADOS

Após análise da base SANCOES-RJ, foi realizada a modelagem dos dados para posterior criação do banco e das tabelas. Como a base é pequena e possui apenas 6 colunas para cada registro, apenas uma tabela se faz necessária no modelo.

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



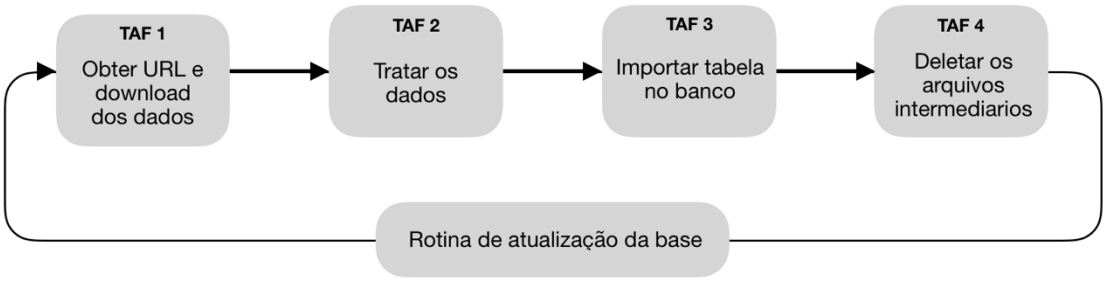

```
CREATE TABLE dbo.Dados_SancaoRJ(
        CPF varchar(15) NULL,
        CNPJ varchar(20) NULL,
        NOME varchar(200) NULL,
        ENQUADRAMENTO_LEGAL varchar(200) NULL,
        DATA_SANCAO varchar (10)NULL,
        STATUS varchar(20) NULL
);

CREATE INDEX IDX_DADOS_SANCAOMG ON dbo.Dados_SancaoRJ (CPF,CNPJ);
```

5. FLUXOS DE ETL

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

O trabalho de ETL foi desenvolvido na ferramenta Apache Airflow e conta com 4 tarefas diretas, apresentadas no diagrama de blocos a seguir.



Tarefa 1 - Captação dos dados

Rotina: **get_dados**

Com o auxilio do plugin para airflow em docker da biblioteca selenium [1], é realizada a navegação automática. Primeiro, dirige-se para a página referenciada na seção 2, ou seja, a URL http://www.compras.rj.gov.br/Portal-Siga/Sancao/buscar.action

Nessa URL é necessário fazer algum filtro para poder pesquisar. Não sendo possível acessar todos os dados de uma vez. Assim, para diminuir a quantidade de pesquisas distintas, utiliza-se o filtro de 'Situação'. Assim, são necessárias 4 pesquisas, uma para cada status (Advertido, Decorrido, Multado e Vigente).

Em cada uma das 4 pesquisas, o resultado é uma sequência de páginas com os resultados com certa quantidade de resultados aparecendo em cada página, tal como pesquisas em ferramentas de busca, como Google.

Primeiramente, é alterada a quantidade de registros por página para minimizar a quantidade de trocas de página. Depois de passar de 6 itens por página para 100 itens, utiliza-se a interface do Selenium para utilização dos métodos de interação com a página nativos para selecionar os elementos da página com as classes 'even' e 'odd' (pares e ímpares em português) que são as linhas alternadas da tabela. Em seguida, copia-se as informações desses elementos para uma lista dinâmica em python com todos os dados. Depois de copiar todas as linhas da página, navega-se para a próxima página até a última, onde é finalizada a pesquisa dentro do status e repete-se o processo descrito neste parágrafo para os outros status.

Depois de capturar todas as informações, as informações são passadas para o arquivo temporário **sancoesRJ.csv** que será apagado ao final depois de passado para o banco de dados.

Tarefa 2 - Tratamento dos dados

Rotina: **edita_base**

À partir do arquivo **sancoesRJ.csv** gerado na tarefa 1 e com o auxílio da biblioteca pandas [2], são feitas todas as operações de tratamento na base de dados.

Neste caso, a primeira alteração necessária é transformar a variável “cpf_ou_cnpj” em duas variáveis distintas: “cpf” e “cnpj”, onde cada registro possui apenas uma das duas possibilidades.

A segunda alteração é a remoção de pontos, traços e barras para transformar tanto a variável CPF quanto a variável CNPJ em códigos numéricos.

Depois de feitas estas alterações, esta base de dados é passada para um arquivo temporário sancoesRJfinal.csv com a versão finalizada que irá para o banco de dados.

Tarefa 3 – *Importação da tabela no banco SQL Server*

Rotina: **truncate_e_reinsert_to_sqlserver**

Com o arquivo sancoesRJfinal.csv e com o auxílio da biblioteca pyodbc [3], é possível conectar no banco SQL Server e importar os dados da tabela para o banco, lembrando de limpar o banco de dados anteriormente (apagar os dados de antes para inserir em uma tabela vazia).

Nessa tarefa, foi utilizado o Driver ODBC [4] para SQL Server 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'
```

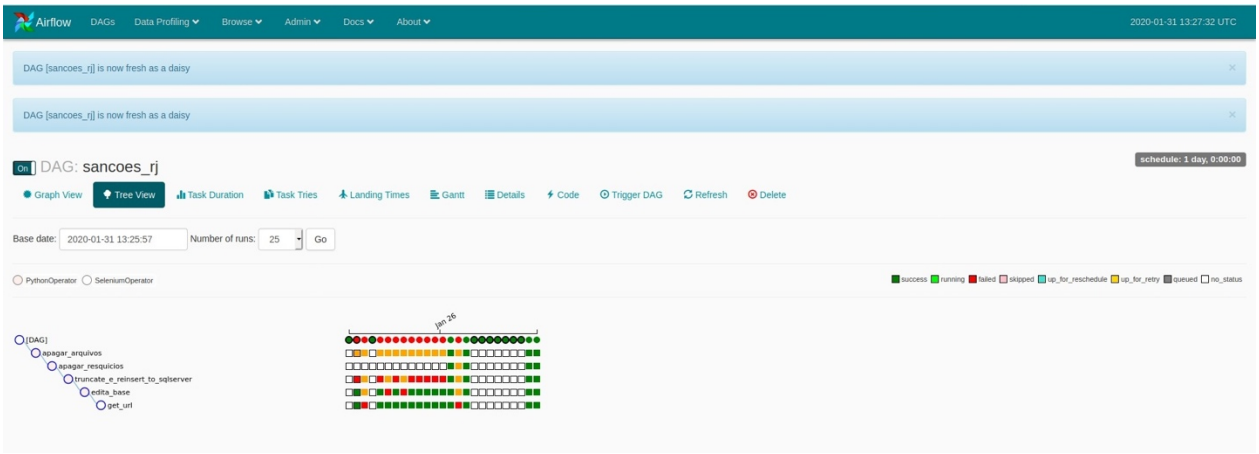
Tarefa 4 – *Exclusão dos arquivos intermediários*

Rotina: **apagar_arquivos**

Com o auxílio da biblioteca *os* [5], deletamos os arquivos baixados e criados ao longo das tarefas intermediarias:

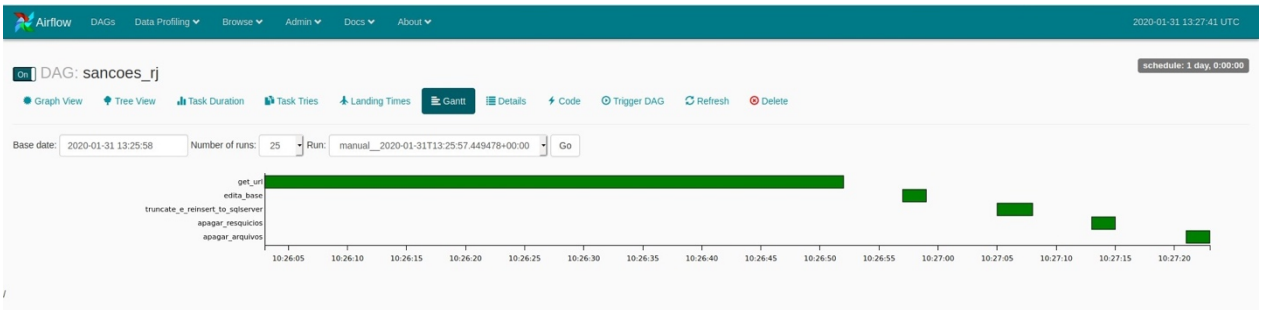
- Tarefa 1: deleta-se o arquivo sancoesRJ.csv
- Tarefa 2: deleta-se o arquivo sancoesRJfinal.csv

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 SANCOES-RJ, o tempo corrido de execução das rotinas foi de cerca de 5 minutos no total, com potencial para melhora com otimização de tempo produtivo no servidor.

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 de parâmetros do sistema, foram tomadas algumas precauções:

◊ A pesquisa para selecionar a quantidade de resultados por página procura sempre a última opção ao invés de tentar trazer sempre 100 resultados por página. Assim, sempre se minimiza a quantidade de páginas a serem pesquisadas mesmo que mais opções de resultados por página sejam adicionadas ou removidas;

◊ São sempre pesquisados todos os status, sem se assumir que serão sempre 4;

◊ A navegação é resistente à quantidade de páginas de resultados.

8. FLUXO DE AGENDAMENTO DE ROTINAS

O SIG não especifica a frequência de atualização da base SANCOES-RJ, presume-se que seja automático ao haver ocorrências e que essas não sejam frequentes. Olhando o ano de 2019, foram observadas 125 ocorrência, aproximadamente uma a cada 3 dias.

Como não foi informada escassez de recursos, não há problema em atualizar esta base diariamente, mas fica a ressalva que se for necessário liberar recursos para processamento de outras bases, esta aqui pode ser atualizada uma vez por semana sem alterações significativas na base.

```
dag = DAG(dag_id='sancoesRJ', default_args=args, schedule_interval=timedelta(days=1))
```

9. ESTIMATIVA DE CRESCIMENTO

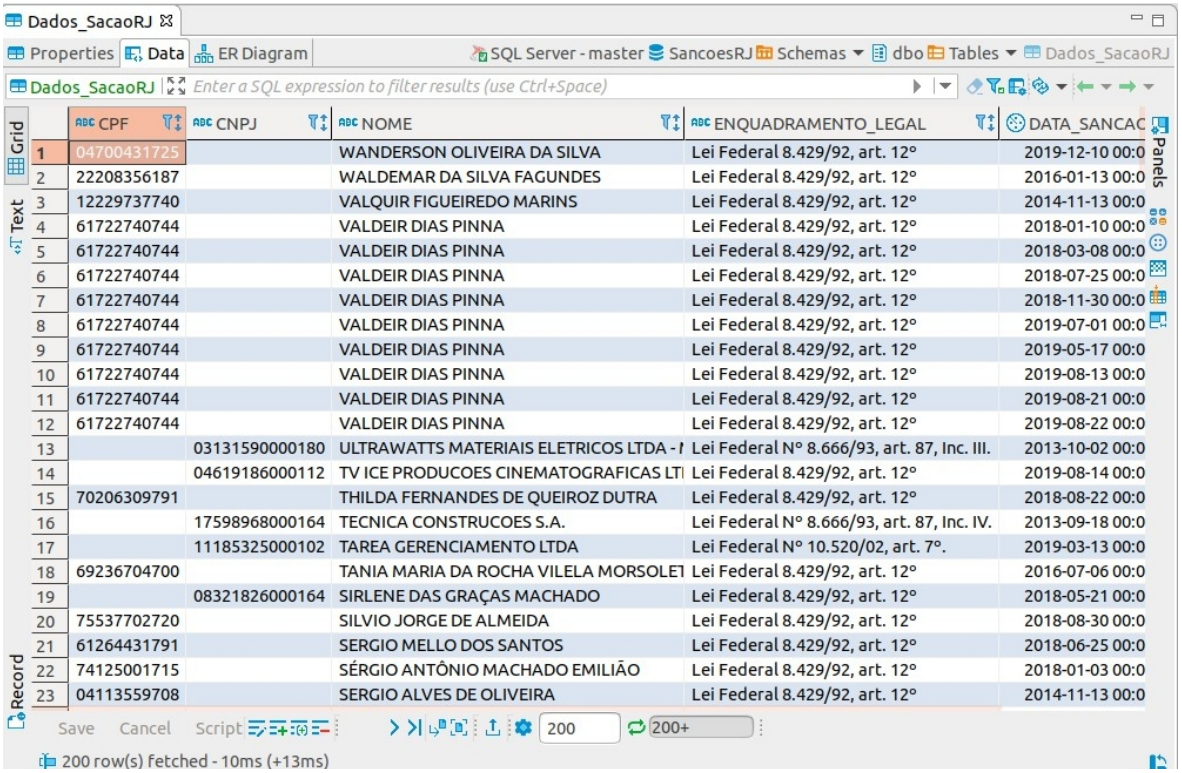
Olhando para o passado recente, com 125 observações em 11 meses, é possível inferir uma média de cerca de 150 observações por ano. Considerando que o tamanho atual da base é de aproximadamente 1000 observações, o tamanho da base levará alguns anos para dobrar de tamanho. Assim, considerando o tamanho atual de cerca de 150KB, dificilmente esta base alcançará 1 MB mesmo em um futuro de longo prazo.

10. AUXÍLIO NOS ESTUDO DE COMPRAS PUBLICAS

Com os dados da base SANCOES-RJ, é possível realizar o cruzamento entre os dados das pessoas físicas ou jurídicas sancionadas e os novos convênios, contratos de repasse ou termos de parceria com a administração publica federal, afim de evitar a participação de tais empresas nesses pleitos.

11. EVIDÊNCIA DOS DADOS IMPORTADOS

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



Dados_SacaoRJ

Properties | Data | ER Diagram | SQL Server - master | SancoesRJ | Schemas | dbo | Tables | Dados_SacaoRJ

Dados_SacaoRJ | Enter a SQL expression to filter results (use Ctrl+Space)

	CPF	CNPJ	NOME	ENQUADRAMENTO_LEGAL	DATA_SANCAC
1	04700431725		WANDERSON OLIVEIRA DA SILVA	Lei Federal 8.429/92, art. 12º	2019-12-10 00:0
2	22208356187		WALDEMAR DA SILVA FAGUNDES	Lei Federal 8.429/92, art. 12º	2016-01-13 00:0
3	12229737740		VALQUIR FIGUEIREDO MARINS	Lei Federal 8.429/92, art. 12º	2014-11-13 00:0
4	61722740744		VALDEIR DIAS PINNA	Lei Federal 8.429/92, art. 12º	2018-01-10 00:0
5	61722740744		VALDEIR DIAS PINNA	Lei Federal 8.429/92, art. 12º	2018-03-08 00:0
6	61722740744		VALDEIR DIAS PINNA	Lei Federal 8.429/92, art. 12º	2018-07-25 00:0
7	61722740744		VALDEIR DIAS PINNA	Lei Federal 8.429/92, art. 12º	2018-11-30 00:0
8	61722740744		VALDEIR DIAS PINNA	Lei Federal 8.429/92, art. 12º	2019-07-01 00:0
9	61722740744		VALDEIR DIAS PINNA	Lei Federal 8.429/92, art. 12º	2019-05-17 00:0
10	61722740744		VALDEIR DIAS PINNA	Lei Federal 8.429/92, art. 12º	2019-08-13 00:0
11	61722740744		VALDEIR DIAS PINNA	Lei Federal 8.429/92, art. 12º	2019-08-21 00:0
12	61722740744		VALDEIR DIAS PINNA	Lei Federal 8.429/92, art. 12º	2019-08-22 00:0
13		03131590000180	ULTRAWATTS MATERIAIS ELETRICOS LTDA - I	Lei Federal Nº 8.666/93, art. 87, Inc. III.	2013-10-02 00:0
14		04619186000112	TV ICE PRODUCOES CINEMATOGRAFICAS LTI	Lei Federal 8.429/92, art. 12º	2019-08-14 00:0
15	70206309791		THILDA FERNANDES DE QUEIROZ DUTRA	Lei Federal 8.429/92, art. 12º	2018-08-22 00:0
16		17598968000164	TECNICA CONSTRUCOES S.A.	Lei Federal Nº 8.666/93, art. 87, Inc. IV.	2013-09-18 00:0
17		11185325000102	TAREA GERENCIAMENTO LTDA	Lei Federal Nº 10.520/02, art. 7º.	2019-03-13 00:0
18	69236704700		TANIA MARIA DA ROCHA VILELA MORSOLET	Lei Federal 8.429/92, art. 12º	2016-07-06 00:0
19		08321826000164	SIRLENE DAS GRAÇAS MACHADO	Lei Federal 8.429/92, art. 12º	2018-05-21 00:0
20	75537702720		SILVIO JORGE DE ALMEIDA	Lei Federal 8.429/92, art. 12º	2018-08-30 00:0
21	61264431791		SERGIO MELLO DOS SANTOS	Lei Federal 8.429/92, art. 12º	2018-06-25 00:0
22	74125001715		SÉRGIO ANTÔNIO MACHADO EMILIÃO	Lei Federal 8.429/92, art. 12º	2018-01-03 00:0
23	04113559708		SERGIO ALVES DE OLIVEIRA	Lei Federal 8.429/92, art. 12º	2014-11-13 00:0

Save | Cancel | Script | 200 | 200+

200 row(s) fetched - 10ms (+13ms)

12. BIBLIOGRAFIA

1. What is Selenium?. Disponível em: < https://www.seleniumhq.org/>. Acesso em: 21 de outubro de 2019.

2. Pandas. Disponível em: <https://pandas.pydata.org/>. Acesso em: 04 de outubro de 2019

3. 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.

4. Microsoft ODBC Driver para SQL Server em Linux. <https://docs.microsoft.com/en-us/sql/connect/odbc/linux-mac/installing-the-microsoft-odbc-driver-for-sql-server?view=sql-server-ver15#microsoft-odbc-driver-131-for-sql-server> 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