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

Documento:

Relatório de Extração de Dados SANÇÕES - SP

Data de Emissão:

27/02/2020

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
29/11/2019	1.1	Flávio Vitória	Inclusão dos dados da base de Sanções da BEC
27/02/2020	1.1	Leticia Valle	Revisão

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 do cadastro de fornecedores com sanções existentes na Bolsa Eletrônica de Compras (BEC) do Estado de São Paulo.

A base de sanções da BEC apresenta a relação de entidades privadas com ou sem fins lucrativos que estão impedidas de celebrar novos convênios, contratos de repasse ou termos de parceria com a Administração Pública (Legislativo, Executivo e Judiciário e entidades da Administração Indireta) do Estado de São Paulo, em função de irregularidades não resolvidas em convênios, contratos de repasse ou termos de parceria firmados anteriormente. [1]

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 E-Sanções do Governo de São Paulo, dentro da página da Bolsa Eletrônica de Compras.

A URL do local de origem dos dados é https://www.bec.sp.gov.br/Sancoes_ui/aspx/ConsultaAdministrativaFornecedor.aspx

3. QUANTITATIVO DE DADOS

A base possui uma tabela com 14 colunas e 4660 registros.

4. MODELAGEM DO BANCO DE DADOS

Após análise da base do E-sanções, foi realizada a modelagem dos dados para posterior criação do banco e das tabelas. As três bases existentes no site podem ser entendidas como uma base, em grande parte, parecida, mas com poucas características peculiares em cada uma das sub-divisões. Sendo assim, optou-se por criar uma base única com todas as informações disponíveis no E-Sanções, sendo que os campos só estarão preenchidos com as informações relativas à cada categoria.

Segue o layout da base:

Dados_inadimplentes

Nome da Coluna	Tipo de Dados	Permitir Nul...
TIPO_MEDIDA	varchar(12)	☑
PODER	varchar(20)	☑
SECRETARIA_ORG	varchar(30)	☑
UGE	varchar(50)	☑
NUMERO_PROCESSO	varchar(30)	☑
TIPO_PESSOA	varchar(10)	☑
RAZAO_SOCIAL	varchar(100)	☑
CNPJ_CPF	varchar(14)	☑
TIPO_SANCAO	varchar(40)	☑
PERIOD_SANCAO	varchar(10)	☑
DATA_INICIO	datetime	☑
DATA_TERMINO	datetime	☑
ABRANGENCIA_PENALI...	varchar(200)	☑
VALOR_MULTA	float	☑
		☐

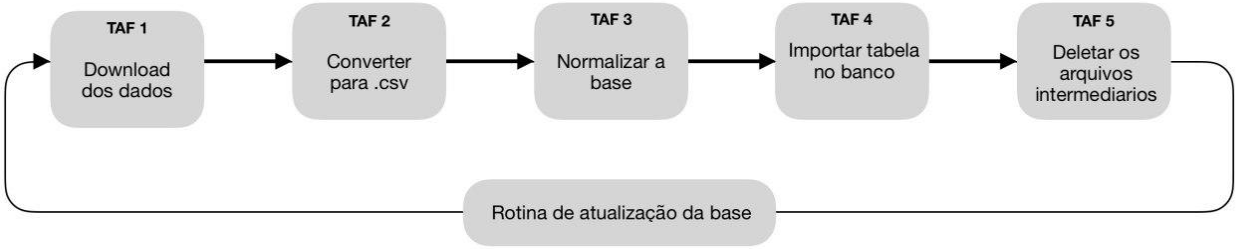
```
CREATE TABLE Dados_inadimplentes (
        TIPO_MEDIDA VARCHAR(25),
        PODER VARCHAR(20),
        SECRETARIA_ORG VARCHAR(50),
        UGE VARCHAR(50),
        NUMERO_PROCESSO VARCHAR(200),
        TIPO_PESSOA VARCHAR(10),
        RAZAO_SOCIAL VARCHAR(200),
        CNPJ_CPF VARCHAR(15),
        TIPO_SANCAO VARCHAR(40),
        PERIOD_SANCAO VARCHAR(20),
        DATA_INICIO VARCHAR(15),
        DATA_TERMINO VARCHAR(15),
        ABRANGENCIA_PENALIDADE VARCHAR(200),
        VALOR_MULTA VARCHAR(20)
);

CREATE INDEX IDX_Dados_Inadimplentes ON Dados_inadimplentes(CNPJ_CPF);
```

5. FLUXOS DE ETL

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

O trabalho de ETL foi desenvolvido na ferramenta Apache Airflow e conta com 6 tarefas diretas, apresentadas no diagrama de bloco a seguir.



Tarefa 1 – Download dos dados

Rotina: get_url

Essa rotina tem como objetivo navegar pela página da BEC, com o auxílio da biblioteca selenium [2], realizando o download das informações a partir da opção selecionada pelo usuário. Essa rotina implementa um controle de fluxo evitando que possíveis quedas da página, ou eventuais demoras excessivamente longas da mesma, consigam ser contornadas e que o download continue, a partir de uma informação de estado mantida pelo programa e retornada a cada ocasião de queda.

Essa rotina implementa, também, um extenso controle de exceções prevendo os vários tipos de erros que possam acontecer na página, seja por queda na conexão com o servidor ou com queda no desempenho do servidor da BEC. Essa rotina é a responsável por fazer o trabalho de navegar na página da BEC, selecionar fornecedor a fornecedor, entrar na página relativa do mesmo, selecionar o tipo de informação desejada pelo usuário, baixar o arquivo .xls/.xml contendo o conjunto das informações do mesmo, e voltar à tela anterior. Nessa rotina, a página apresenta instabilidade no seu funcionamento, seja ou por tempos longos de resposta ou por conta do conteúdo exibido de uma forma extensa (271 páginas, cada uma com 15 fornecedores, até a data), e essas instabilidades foram tomadas em consideração na geração do código, e tratadas em sua particularidade. Para garantir a continuidade do processo de download, a rotina garante um controle de fluxo, guardando a informação do estado do download e recuperando o mesmo caso haja uma queda na atividade**.**

Tarefa 3 – Conversão dos dados em csv

Rotina: **html_to_csv**

Os arquivos baixados são movidos para uma pasta no mesmo diretório no qual o código está sendo executado. O site da BEC informa que o arquivo baixado é um .xls (Excel), mas, examinando mais cautelosamente, percebe-se que é um arquivo .xml. Dessa forma, com o auxílio da biblioteca beautiful soup [4] é possível procurar, dentro do arquivo, as informações de interesse, existentes dentro do código .xml da mesma.

Assim sendo, os .xml são analisados e concatenados, sendo que, aqueles artigos que não apresentam informação são descartados nessa fase e as informações

desnecessárias ou redundantes são retiradas do arquivo final. Essa rotina também imprime um tipo da medida tomada em cada registro.

Tarefa 4 – Normalização dos dados

Rotina: normalize_data

Já que as tabelas até aqui seguem suas configurações originais, e sabendo que as 3 informações serão armazenadas em uma mesma tabela, é necessário completar com brancos (‘’) os campos de cada tabela que contem informação que não faz sentido nas outras., para fazer essa normalização necessária à coexistência de todas as informações em um mesmo banco.

Tarefa 5 – Inserção da tabela no banco SQL Server

Rotina: **copy_to_sql_server**

Com o arquivo Multas.txt/SancoesRestritivas.txt/Advertencias.txt e com o auxílio da biblioteca pyodbc [6], é possível conectar no banco SQL Server e importar os dados da tabela para o banco.

Utilizando a url 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 foi feito o download do pacote de instalação para Linux do driver de conexão.

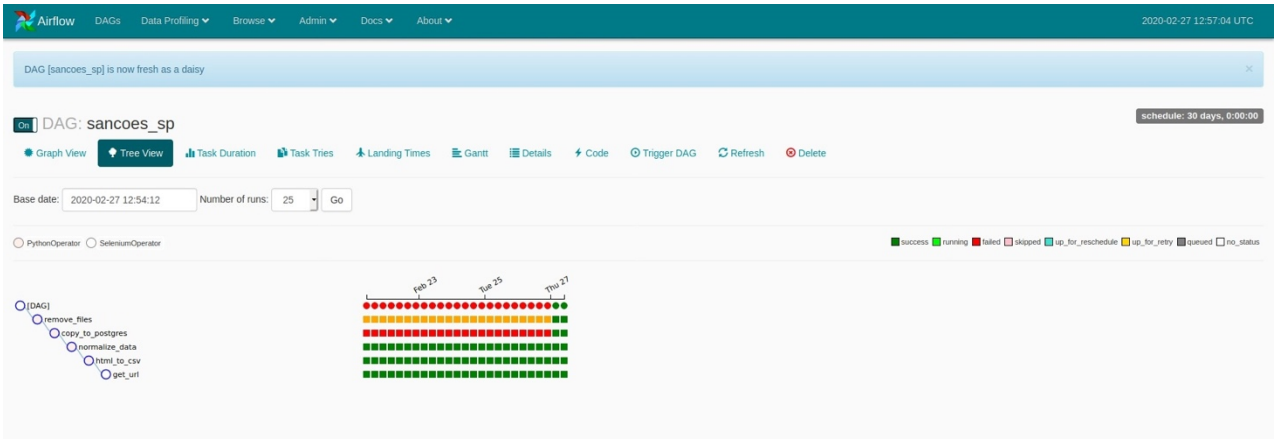
```
conn_string = 'DRIVER={ODBC Driver 17 for SQL Server};SERVER='+
            host+';PORT=1433;DATABASE='+dbname+';UID='+
            username+';PWD='+ password +
            ';UseNTLMv2=yes;TDS_Version=8.0'
```

Tarefa 6 – Exclusão dos arquivos originais

Rotina: **remove_files**

Com o auxílio da biblioteca *os* [7], deletamos os arquivos baixados e criados ao longo das tarefas intermediarias. O arquivo final tem, em média, 1MB, dos quase, em média, 20MB de dados baixados.

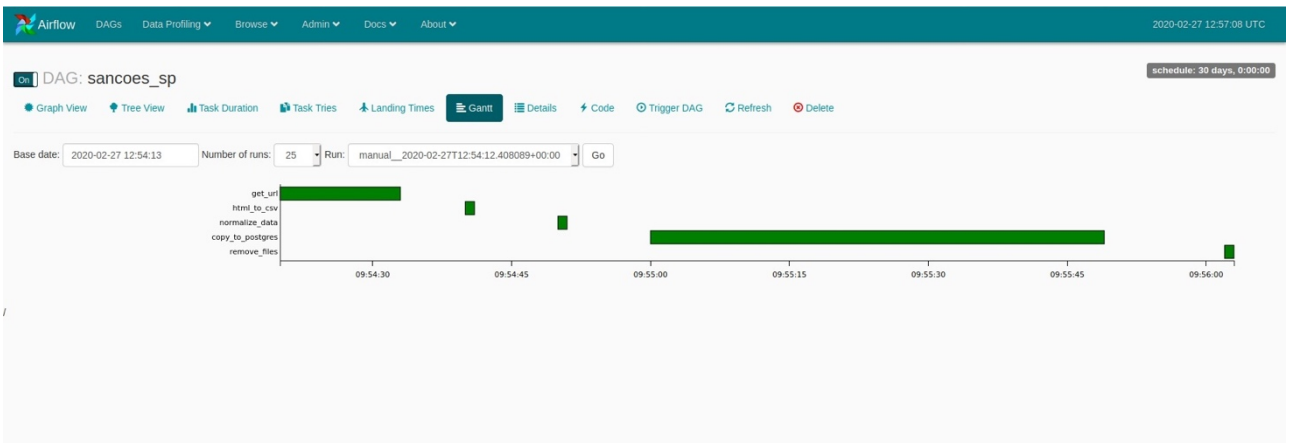
A imagem a seguir representa as tarefas ETL dentro da ferramenta Apache Airflow.



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.

7. FLUXO DE TRATAMENTO DE ERROS

Os possíveis erros de execução das tarefas foram tratados a 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. Aqui, também, o script armazena, tanto em tempo de execução como em disco, um arquivo .tmp indicando o último arquivo baixado com sucesso.

8. FLUXO DE AGENDAMENTO DE ROTINAS

A base de sanções da BEC é atualizada sob demanda. Para captar as informações em um intervalo razoável, no código de configuração do Apache Airflow, foi inserida uma rotina de atualização da base a cada 30 dias.

```
dag = DAG(dag_id='BEC', 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 sob demanda e o arquivo de download é incrementado, não houve tempo hábil para estimar qual é o volume incremental da base à cada atualização.

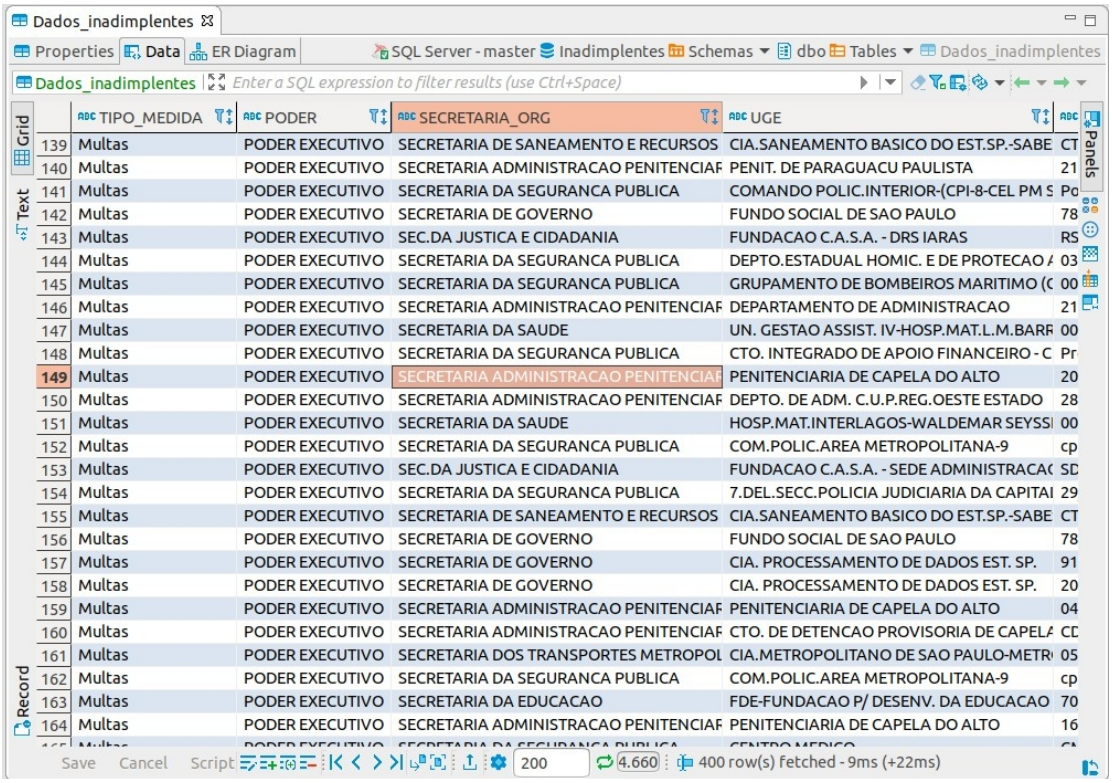
De qualquer forma, o volume total da base é por volta de 30MB, o que implica que mesmo após varias atualizações futuras, a base não deverá passar de 35 MBs.

10. AUXÍLIO NOS ESTUDOS DE COMPRAS PUBLICAS

Com os dados da base de sanções na BEC, é possível realizar o cruzamento entre os dados das empresas privadas com ou sem fins lucrativos inadimplentes com o Governo de São Paulo (uma parte significativa do volume de fornecedores nacional) 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.



	TIPO_MEDIDA	PODER	SECRETARIA_ORG	UGE	
139	Multas	PODER EXECUTIVO	SECRETARIA DE SANEAMENTO E RECURSOS	CIA.SANEAMENTO BASICO DO EST.SP.-SABE	CT
140	Multas	PODER EXECUTIVO	SECRETARIA ADMINISTRACAO PENITENCIAF	PENIT. DE PARAGUACU PAULISTA	21
141	Multas	PODER EXECUTIVO	SECRETARIA DA SEGURANCA PUBLICA	COMANDO POLIC.INTERIOR-(CPI-8-CEL PM S	Po
142	Multas	PODER EXECUTIVO	SECRETARIA DE GOVERNO	FUNDO SOCIAL DE SAO PAULO	78
143	Multas	PODER EXECUTIVO	SEC.DA JUSTICA E CIDADANIA	FUNDACAO C.A.S.A. - DRS IARAS	RS
144	Multas	PODER EXECUTIVO	SECRETARIA DA SEGURANCA PUBLICA	DEPTO.ESTADUAL HOMIC. E DE PROTECAO A	03
145	Multas	PODER EXECUTIVO	SECRETARIA DA SEGURANCA PUBLICA	GRUPAMENTO DE BOMBEIROS MARITIMO (C	00
146	Multas	PODER EXECUTIVO	SECRETARIA ADMINISTRACAO PENITENCIAF	DEPARTAMENTO DE ADMINISTRACAO	21
147	Multas	PODER EXECUTIVO	SECRETARIA DA SAUDE	UN. GESTAO ASSIST. IV-HOSP.MAT.L.M.BARR	00
148	Multas	PODER EXECUTIVO	SECRETARIA DA SEGURANCA PUBLICA	CTO. INTEGRADO DE APOIO FINANCEIRO - C	Pr
149	Multas	PODER EXECUTIVO	SECRETARIA ADMINISTRACAO PENITENCIAF	PENITENCIARIA DE CAPELA DO ALTO	20
150	Multas	PODER EXECUTIVO	SECRETARIA ADMINISTRACAO PENITENCIAF	DEPTO. DE ADM. C.U.P.REG.OESTE ESTADO	28
151	Multas	PODER EXECUTIVO	SECRETARIA DA SAUDE	HOSP.MAT.INTERLAGOS-WALDEMAR SEYSSI	00
152	Multas	PODER EXECUTIVO	SECRETARIA DA SEGURANCA PUBLICA	COM.POLIC.AREA METROPOLITANA-9	cp
153	Multas	PODER EXECUTIVO	SEC.DA JUSTICA E CIDADANIA	FUNDACAO C.A.S.A. - SEDE ADMINISTRACAC	SD
154	Multas	PODER EXECUTIVO	SECRETARIA DA SEGURANCA PUBLICA	7.DEL.SECC.POLICIA JUDICIARIA DA CAPITAI	29
155	Multas	PODER EXECUTIVO	SECRETARIA DE SANEAMENTO E RECURSOS	CIA.SANEAMENTO BASICO DO EST.SP.-SABE	CT
156	Multas	PODER EXECUTIVO	SECRETARIA DE GOVERNO	FUNDO SOCIAL DE SAO PAULO	78
157	Multas	PODER EXECUTIVO	SECRETARIA DE GOVERNO	CIA. PROCESSAMENTO DE DADOS EST. SP.	91
158	Multas	PODER EXECUTIVO	SECRETARIA DE GOVERNO	CIA. PROCESSAMENTO DE DADOS EST. SP.	20
159	Multas	PODER EXECUTIVO	SECRETARIA ADMINISTRACAO PENITENCIAF	PENITENCIARIA DE CAPELA DO ALTO	04
160	Multas	PODER EXECUTIVO	SECRETARIA ADMINISTRACAO PENITENCIAF	CTO. DE DETENCAO PROVISORIA DE CAPELA	CD
161	Multas	PODER EXECUTIVO	SECRETARIA DOS TRANSPORTES METROPOL	CIA.METROPOLITANO DE SAO PAULO-METR	05
162	Multas	PODER EXECUTIVO	SECRETARIA DA SEGURANCA PUBLICA	COM.POLIC.AREA METROPOLITANA-9	cp
163	Multas	PODER EXECUTIVO	SECRETARIA DA EDUCACAO	FDE-FUNDACAO P/ DESENV. DA EDUCACAO	70
164	Multas	PODER EXECUTIVO	SECRETARIA ADMINISTRACAO PENITENCIAF	PENITENCIARIA DE CAPELA DO ALTO	16

1. BIBLIOGRAFIA

1. Portal E-Sanções. Disponível em: <https://www.bec.sp.gov.br/FAQ_UI/FAQ.aspx>. Acesso em: 21 de outubro de 2019.

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

3. Geckodriver. Disponível em: < https://github.com/mozilla/geckodriver>. Acesso em: 21 de outubro de 2019.

4. Beautiful Soup documentation. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Acesso em: 21 de outubro de 2019.

5. 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.

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

7. Argparse — Parser for command-line options, arguments and sub-commands. Disponível em <https://docs.python.org/3/library/argparse.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