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

Documento:

Relatório de Extração de Dados CEPIM

Data de Emissão:

17/01/2019

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
15/10/2019	1.0	Anderson Alves de Oliveira	Versão inicial do documento
21/10/2019	1.1	Leticia Valle	Inclusão dos dados da base CEPIM
17/01/2019	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 carregamento de dados da base de Cadastro de Entidades Privadas Sem Fins Lucrativos Impedidas (CEPIM).

A base CEPIM apresenta a relação de entidades privadas sem fins lucrativos que estão impedidas de celebrar novos convênios, contratos de repasse ou termos de parceria com a Administração Pública Federal, 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 da transparência na categoria de dados do portal (dados abertos), sanções e entidades sem fins lucrativos impedidas.

A URL do local de origem dos dados é http://www.portaltransparencia.gov.br/download-de-dados/cepim/

3. QUANTITATIVO DE DADOS

A base CEPIM possui apenas uma tabela com 5 colunas e 4738 registros, na data de entrega desse documento.

4. MODELAGEM DO BANCO DE DADOS

Após análise da base CEPIM, foi realizada a modelagem dos dados para posterior criação do banco e das tabelas. Como a base é pequena e possui apenas 5 colunas, 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.

Dados_CEPIM

Nome da Coluna	Tipo de Dados	Permitir Nulos
CNPJ_ENTIDADE	varchar(14)	☑
NOME_ENTIDADE	varchar(250)	☑
NUMERO_CONVENIO	varchar(50)	☑
ORGAO_CONCEDENTE	varchar(250)	☑
MOTIVO_IMPEDIMENTO	varchar(500)	☑
		☐

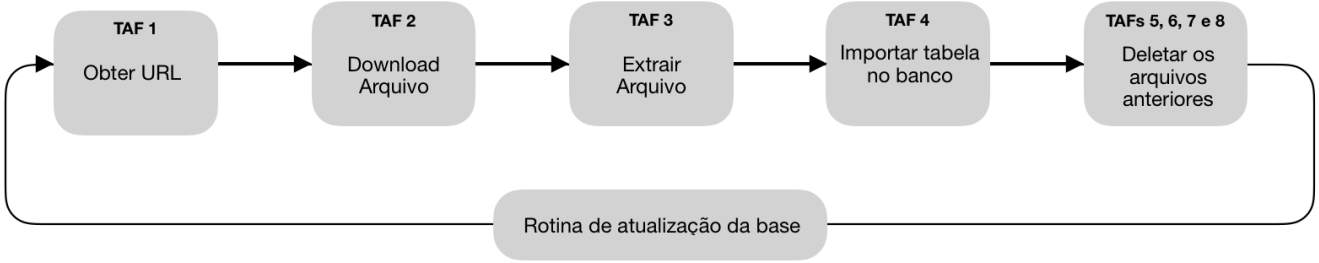
```
CREATE TABLE dbo.Dados_CEPIM
(
CNPJ_ENTIDADE VARCHAR(14),
NOME_ENTIDADE VARCHAR(250),
NUMERO_CONVENIO VARCHAR(50),
ORGAO_CONCEDENTE VARCHAR(250),
MOTIVO_IMPEDIMENTO VARCHAR(500));


CREATE INDEX IDX_DADOS_CEPIM ON dbo.DADOS_CEPIM
(CNPJ_ENTIDADE,NOME_ENTIDADE);
```

5. FLUXOS DE ETL

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

O trabalho de ETL foi desenvolvido na ferramenta Apache Airflow e conta com 8 tarefas diretas, apresentadas no diagrama de bloco a seguir.



Tarefa 1 - Download do código HTML para captura da URL de download do arquivo de dados

Rotina: **get_url**

Com o auxilio do plugin para o airflow em docker da biblioteca selenium [2], é realizado o download do código HTML da página que contém o botão de download da base do CEPIM

Com o auxílio da biblioteca beautiful soup [3] é possível procurar, dentro do código HTML da página, o identificador do botão de download do arquivo .zip que contém o .csv da tabela CEPIM. A partir do identificador, conseguimos localizar a URL de download do arquivo e salvamos essa URL em um arquivo .txt nomeado cepim_url.txt

Tarefa 2 - Download do arquivo de dados

Rotina: **get_data_from_url**

À partir do arquivo cepim_url.txt, obtém-se a URL de download do arquivo e com o auxílio da biblioteca *requests* [4], realizamos o download do arquivo .zip e nomeamos de cepim.zip

Tarefa 3 – Extração do arquivo de dados

Rotina: **unzip_file**

Com o arquivo cepim.zip e com o auxílio da biblioteca zipfile [5], realizamos a extração do arquivo .csv do arquivo .zip e o nomeamos o arquivo salvo de cepim_filename.txt.

Tarefa 4 – Importação da tabela no banco SQL Server

Rotina: **copy_to_sqlserver**

Com o arquivo cepim_filename.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.

Foi utilizado o Driver ODBC [7] 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'
```

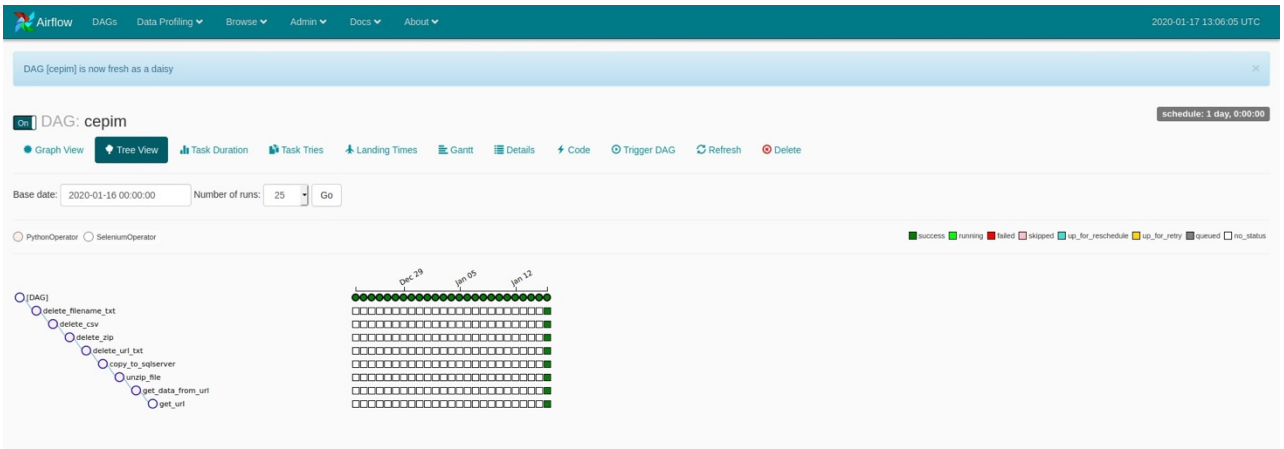
Tarefa 5, 6, 7 e 8 – Exclusão dos arquivos intermediários

Rotina: **delete_arquivo**

Com o auxílio da biblioteca *os* [8], deletamos os arquivos baixados e criados ao longo das tarefas intermediarias:

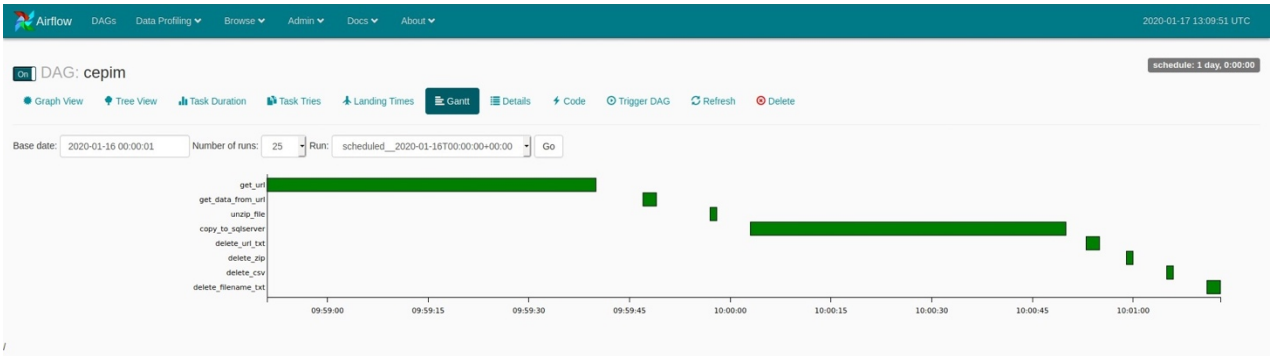
- Tarefa 5: deleta-se o arquivo cepim.zip
- Tarefa 6: deleta-se o arquivo cepim_url.txt
- Tarefa 7: deleta-se o arquivo .csv
- Tarefa 8: deleta-se o arquivo cepim_filename.txt

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 CEPIM, o tempo de execução das rotinas foi de cerca de 3 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 que contém os dados para download, a tarefa inicial do ETL é fazer uma procura dinâmica no código HTML da página, de forma a evitar erros gerados por mudanças estruturais na pagina de download do arquivo.

8. FLUXO DE AGENDAMENTO DE ROTINAS

A base CEPIM é atualizada diariamente. Dessa forma, no código de configuração do Apache Airflow, foi inserida uma rotina de atualização da base a cada 01 dias.

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

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 o arquivo de download é totalmente substituído, não houve tempo hábil para estimar qual é o volume incremental da base à cada atualização.

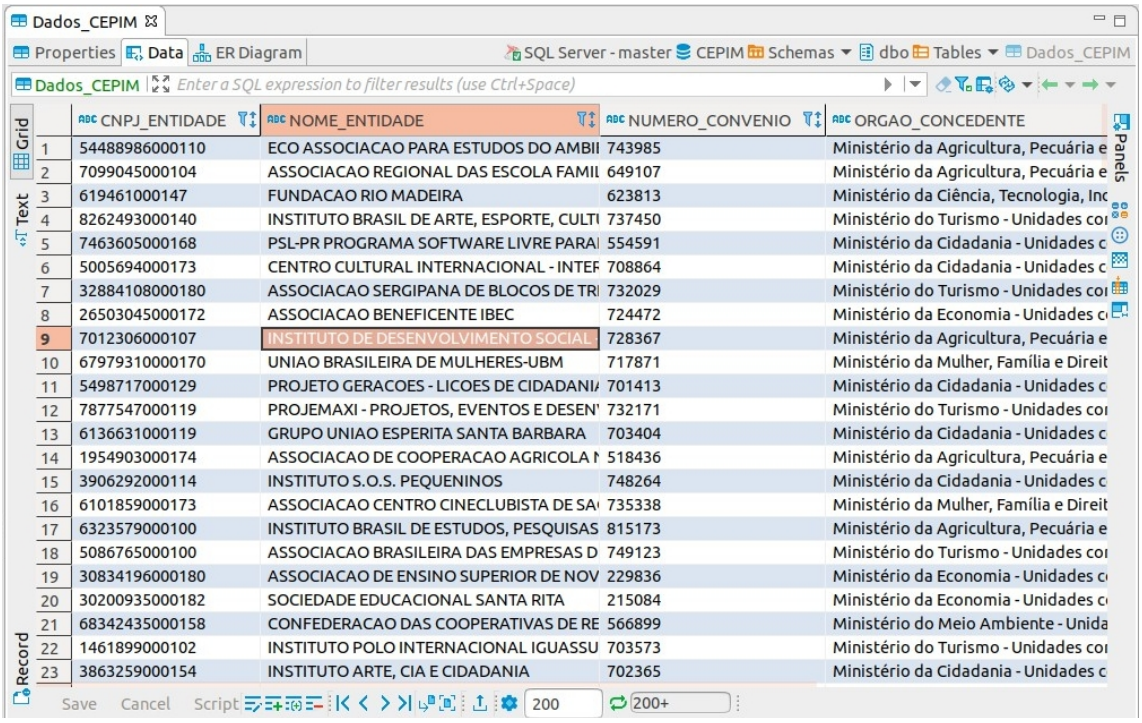
De qualquer forma, o volume total da base é de menos de 1MB, o que implica que mesmo após várias atualizações futuras, a base não devera passar de 1 ou 2 MBs.

10. AUXÍLIO NOS ESTUDO DE COMPRAS PUBLICAS

Com os dados da base CEPIM, é possível realizar o cruzamento entre os dados das empresas privadas sem fins lucrativos inadimplentes 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.



	CNPJ_ENTIDADE	NOME_ENTIDADE	NUMERO_CONVENIO	ORGAO_CONCEDENTE
1	54488986000110	ECO ASSOCIACAO PARA ESTUDOS DO AMBII	743985	Ministério da Agricultura, Pecuária e
2	7099045000104	ASSOCIACAO REGIONAL DAS ESCOLA FAMII	649107	Ministério da Agricultura, Pecuária e
3	619461000147	FUNDACAO RIO MADEIRA	623813	Ministério da Ciência, Tecnologia, Inc
4	8262493000140	INSTITUTO BRASIL DE ARTE, ESPORTE, CULTI	737450	Ministério do Turismo - Unidades cor
5	7463605000168	PSL-PR PROGRAMA SOFTWARE LIVRE PARAI	554591	Ministério da Cidadania - Unidades c
6	5005694000173	CENTRO CULTURAL INTERNACIONAL - INTEF	708864	Ministério da Cidadania - Unidades c
7	32884108000180	ASSOCIACAO SERGIPANA DE BLOCOS DE TRI	732029	Ministério do Turismo - Unidades cor
8	26503045000172	ASSOCIACAO BENEFICENTE IBEC	724472	Ministério da Economia - Unidades c
9	7012306000107	INSTITUTO DE DESENVOLVIMENTO SOCIAL	728367	Ministério da Agricultura, Pecuária e
10	67979310000170	UNIAO BRASILEIRA DE MULHERES-UBM	717871	Ministério da Mulher, Família e Direit
11	5498717000129	PROJETO GERACOES - LICOES DE CIDADANI/	701413	Ministério da Cidadania - Unidades c
12	7877547000119	PROJEMAXI - PROJETOS, EVENTOS E DESEN'	732171	Ministério do Turismo - Unidades cor
13	6136631000119	GRUPO UNIAO ESPERITA SANTA BARBARA	703404	Ministério da Cidadania - Unidades c
14	1954903000174	ASSOCIACAO DE COOPERACAO AGRICOLA I	518436	Ministério da Agricultura, Pecuária e
15	3906292000114	INSTITUTO S.O.S. PEQUENINOS	748264	Ministério da Cidadania - Unidades c
16	6101859000173	ASSOCIACAO CENTRO CINECLUBISTA DE SA(	735338	Ministério da Mulher, Família e Direit
17	6323579000100	INSTITUTO BRASIL DE ESTUDOS, PESQUISAS	815173	Ministério da Agricultura, Pecuária e
18	5086765000100	ASSOCIACAO BRASILEIRA DAS EMPRESAS D	749123	Ministério do Turismo - Unidades cor
19	30834196000180	ASSOCIACAO DE ENSINO SUPERIOR DE NOV	229836	Ministério da Economia - Unidades c
20	30200935000182	SOCIEDADE EDUCACIONAL SANTA RITA	215084	Ministério da Economia - Unidades c
21	68342435000158	CONFEDERACAO DAS COOPERATIVAS DE RE	566899	Ministério do Meio Ambiente - Unida
22	1461899000102	INSTITUTO POLO INTERNACIONAL IGUASSU	703573	Ministério do Turismo - Unidades cor
23	3863259000154	INSTITUTO ARTE, CIA E CIDADANIA	702365	Ministério da Cidadania - Unidades c

12. BIBLIOGRAFIA

1. Portal da transparência. Página interna CEPIM. Disponível em: <http://www.portaltransparencia.gov.br/pagina-interna/603243-cepim>. 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. Beautiful Soup documentation. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Acesso em: 21 de outubro de 2019.

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

5. Zipfile documentation. Disponível em: <https://docs.python.org/3/library/zipfile.html>. Acesso em: 21 de outubro de 2019.

6. 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.

7. 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.

8. 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