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

Documento:

Relatório de Extração de Dados CEIS

Data de Emissão:

22/01/2019

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
06/10/2019	1.1	Leticia Valle	Inclusão dos dados da base CEIS
22/01/2019	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 Empresas Inidôneas e Suspensas (CEIS).

A base CEIS apresenta a relação das empresas e pessoas físicas que sofreram sanções tendo como efeito restrição ao direito de participar de licitações ou de celebrar contratos com a Administração Pública [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) e sanções.

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

3. QUANTITATIVO DE DADOS

A base CEIS possui apenas uma tabela com 22 colunas e 15.439 registros, na data de entrega desse documento.

4. MODELAGEM DO BANCO DE DADOS

Após análise da base CEIS, foi realizada a modelagem dos dados para posterior criação do banco e das tabelas. Para essa base especifica, foi necessário a criação de apenas uma tabela.

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

Dados_CEIS

Nome da Coluna	Tipo de Dados	Permitir Nulos
TIPO_PESSOA	varchar(20)	☑
CPF_CNPJ_SANCIONADO	varchar(14)	☑
NOME	varchar(100)	☑
RAZAO_SOCIAL	varchar(150)	☑
NOME_FANTASIA	varchar(150)	☑
NUMERO_PROCESSO	varchar(50)	☑
TIPO_SANCAO	varchar(20)	☑
DATA_INIIO_SANCAO	varchar(10)	☑
DATA_FINAL_SANCAO	varchar(10)	☑
ORGAO_SANCIONADOR	varchar(150)	☑
UF_ORGAO_SANCIONA...	varchar(2)	☑
ORIGEM_INFORMACOES	varchar(100)	☑
DATA_ORIGEM_INFORM...	varchar(10)	☑
DATA_PUBLICACAO	varchar(10)	☑
PUBLICACAO	varchar(100)	☑
DETALHAMENTO	varchar(500)	☑
ABRAGENCIA_JUDICIAL	varchar(500)	☑
FUNDAMENTACAO_LEG...	varchar(100)	☑
DESCRICAO_FUND_LEGAL	varchar(500)	☑
DATA_TRANSITO_JULGA...	varchar(10)	☑
COMPLEMENTO_ORGAO	varchar(500)	☑
OBSERVACOES	varchar(500)	☑
		☐

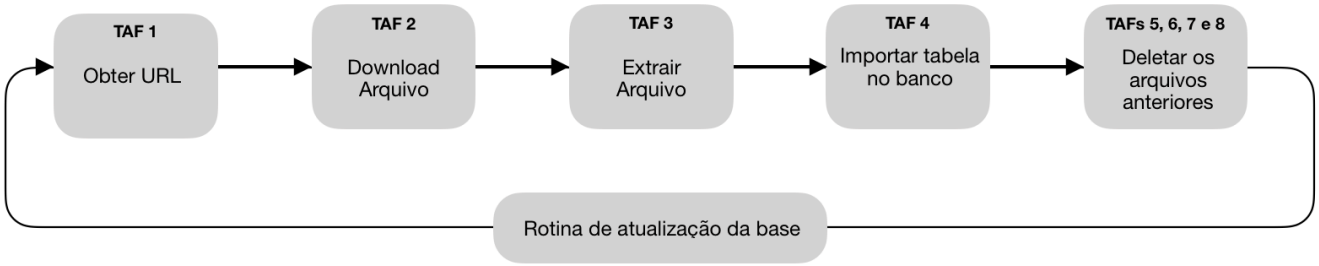
```
CREATE TABLE dbo.Dados_CEIS(
        TIPO_PESSOA varchar(20) ,
        CPF_CNPJ_SANCIONADO varchar(14) ,
        NOME varchar(200) ,
        RAZAO_SOCIAL varchar(150) ,
        NOME_FANTASIA varchar(150) ,
        NUMERO_PROCESSO varchar(50) ,
        TIPO_SANCAO varchar(100) ,
        DATA_INICIO_SANCAO varchar(10) ,
        DATA_FINAL_SANCAO varchar(10) ,
        ORGAO_SANCIONADOR varchar(300) ,
        UF_ORGAO_SANCIONADOR varchar(5) ,
        ORIGEM_INFORMACOES varchar(500) ,
        DATA_ORIGEM_INFORMACOES varchar(10) ,
        DATA_PUBLICACAO varchar(10) ,
        PUBLICACAO varchar(100) ,
        DETALHAMENTO varchar(2500) ,
        ABRAGENCIA_JUDICIAL varchar(500) ,
        FUNDAMENTACAO_LEGAL varchar(100) ,
        DESCRICAO_FUND_LEGAL varchar(2500) ,
        DATA_TRANSITO_JULGADO varchar(10) ,
        COMPLEMENTO_ORGAO varchar(2500) ,
        OBSERVACOES varchar(2500)
);

CREATE INDEX IDX_DADOS_CEIS ON dbo.Dados_CEIS (CPF_CNPJ_SANCIONADO);
```

5. FLUXOS DE ETL

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

O trabalho de ETL foi desenvolvido na ferramenta Apache Airflow e assim como na base CEPIM, 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 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 CEIS.

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 CEIS. A partir do identificador, conseguimos localizar a URL de download do arquivo e salvamos essa URL em um arquivo .txt nomeado ceis_url.txt

Tarefa 2 - Download do arquivo de dados

Rotina: **get_data_from_url**

À partir do arquivo ceis_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 ceis.zip

Tarefa 3 – Extração do arquivo de dados

Rotina: **unzip_file**

Com o arquivo ceis.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 ceis_filename.txt.

Tarefa 4 – Importação da tabela no banco SQL Server

Rotina: **copy_to_sqlserver**

Com o arquivo ceis_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'
```

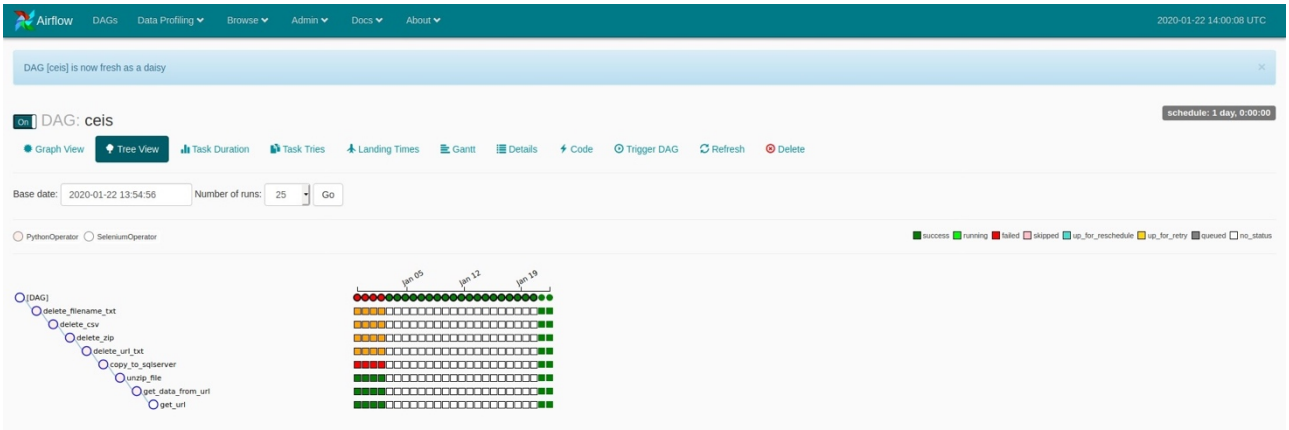
Tarefa 5, 6, 7 e 8 – Exclusão dos arquivos intermediários

Rotina: **delete_arquivo**

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

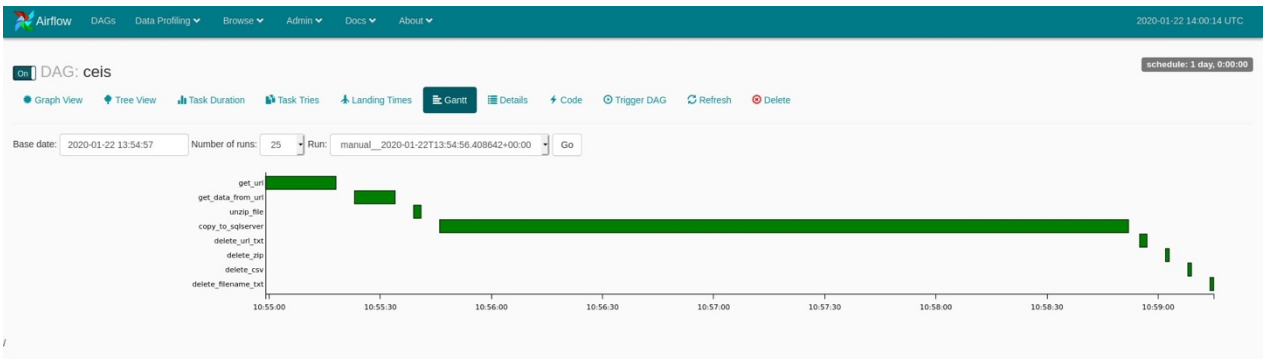
- Tarefa 5: deleta-se o arquivo ceis.zip
- Tarefa 6: deleta-se o arquivo ceis_url.txt
- Tarefa 7: deleta-se o arquivo .csv
- Tarefa 8: deleta-se o arquivo ceis_filename.txt

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 CEIS, o tempo de execução das rotinas foi de cerca de 10 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 página de download do arquivo.

8. FLUXO DE AGENDAMENTO DE ROTINAS

A base CEIS 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='ceis', 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 diárias.

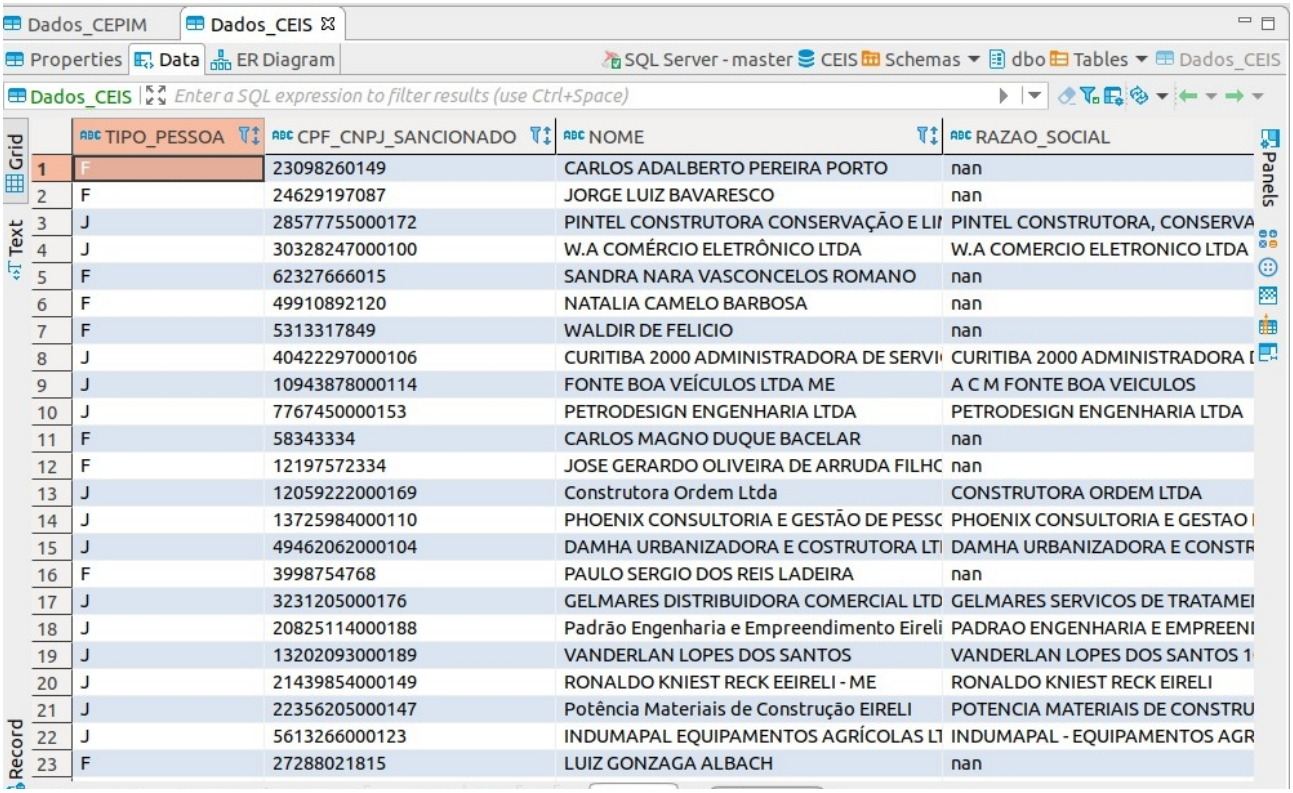
Na entrega desse relatório, o volume total da base é de 24.9MB, o que implica que mesmo após várias atualizações futuras, a base não devera passar de 30 MBs.

10. AUXÍLIO NOS ESTUDO DE COMPRAS PUBLICAS

Com os dados da base CEIS, 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.


Dados_CEPIM | Dados_CEIS

Properties | Data | ER Diagram — SQL Server - master | CEIS | Schemas | dbo | Tables | Dados_CEIS

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

	TIPO_PESSOA	CPF_CNPJ_SANCIONADO	NOME	RAZAO_SOCIAL
1	F	23098260149	CARLOS ADALBERTO PEREIRA PORTO	nan
2	F	24629197087	JORGE LUIZ BAVARESCO	nan
3	J	28577755000172	PINTEL CONSTRUTORA CONSERVAÇÃO E LI	PINTEL CONSTRUTORA, CONSERVA
4	J	30328247000100	W.A COMÉRCIO ELETRÔNICO LTDA	W.A COMERCIO ELETRONICO LTDA
5	F	62327666015	SANDRA NARA VASCONCELOS ROMANO	nan
6	F	49910892120	NATALIA CAMELO BARBOSA	nan
7	F	5313317849	WALDIR DE FELICIO	nan
8	J	40422297000106	CURITIBA 2000 ADMINISTRADORA DE SERVI	CURITIBA 2000 ADMINISTRADORA D
9	J	10943878000114	FONTE BOA VEÍCULOS LTDA ME	A C M FONTE BOA VEICULOS
10	J	7767450000153	PETRODESIGN ENGENHARIA LTDA	PETRODESIGN ENGENHARIA LTDA
11	F	58343334	CARLOS MAGNO DUQUE BACELAR	nan
12	F	12197572334	JOSE GERARDO OLIVEIRA DE ARRUDA FILHO	nan
13	J	12059222000169	Construtora Ordem Ltda	CONSTRUTORA ORDEM LTDA
14	J	13725984000110	PHOENIX CONSULTORIA E GESTÃO DE PESSO	PHOENIX CONSULTORIA E GESTAO
15	J	49462062000104	DAMHA URBANIZADORA E COSTRUTORA LTI	DAMHA URBANIZADORA E CONSTR
16	F	3998754768	PAULO SERGIO DOS REIS LADEIRA	nan
17	J	3231205000176	GELMARES DISTRIBUIDORA COMERCIAL LTD	GELMARES SERVICOS DE TRATAMEI
18	J	20825114000188	Padrão Engenharia e Empreendimento Eireli	PADRAO ENGENHARIA E EMPREENI
19	J	13202093000189	VANDERLAN LOPES DOS SANTOS	VANDERLAN LOPES DOS SANTOS 1
20	J	21439854000149	RONALDO KNIEST RECK EEIRELI - ME	RONALDO KNIEST RECK EIRELI
21	J	22356205000147	Potência Materiais de Construção EIRELI	POTENCIA MATERIAIS DE CONSTRU
22	J	5613266000123	INDUMAPAL EQUIPAMENTOS AGRÍCOLAS LT	INDUMAPAL - EQUIPAMENTOS AGR
23	F	27288021815	LUIZ GONZAGA ALBACH	nan

12. BIBLIOGRAFIA

1. Portal da transparência. Página interna CEIS. Disponível em: <http://www.portaltransparencia.gov.br/pagina-interna/603245-ceis >. 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