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

Documento:

Relatório de Extração de Dados LENIÊNCIAS

Data de Emissão:

07/02/2020

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
24/10/2019	1.0	Leonardo Pires Simões Vasconcelos	Versão inicial do documento e inclusão dos dados da base Leniências
25/11/2019	1.1	Leticia Valle	Revisão
07/02/2020	1.1	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** 5
3. **QUANTITATIVO DE DADOS** 5
4. **MODELAGEM DO BANCO DE DADOS** 5
5. **FLUXOS DE ETL** 7

   Tarefa 1 **- Download do código JSON através da URL do API disponibilizado** 7

   Tarefa 2 – Extração dos dados e importação para a tabela no banco SQL Server 7
6. **DURAÇÃO DAS ROTINAS ETL** 8
7. **FLUXO DE TRATAMENTO DE ERROS** 9
8. **FLUXO DE AGENDAMENTO DE ROTINAS** 9
9. **ESTIMATIVA DE CRESCIMENTO** 9
10. **AUXÍLIO NOS ESTUDO DE COMPRAS PUBLICAS** 9
11. **EVIDÊNCIA DOS DADOS IMPORTADOS** 10
12. **BIBLIOGRAFIA** 11

1. INTRODUÇÃO

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

A base Leniência apresenta a relação acordos de leniência com empresas investigadas pela prática de atos lesivos contra a Administração Pública. As empresas podem ter atenuadas ou ficar isentas das respectivas sanções - o que inclui a aplicação de multa e também a pena de inidoneidade (proibição de contratar com o poder público) - desde que colaborem efetivamente com as investigações e o processo administrativo. [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 API de dados, acordos de leniência.

A URL do local de origem dos dados é http://www.transparencia.gov.br/api-de-dados/acordos-leniencia

3. QUANTITATIVO DE DADOS

A base Leniência possui apenas uma tabela com 11 colunas e 11 registros.

4. MODELAGEM DO BANCO DE DADOS

Após análise da base Leniência, foi realizada a modelagem dos dados para posterior criação do banco e das tabelas. Como a base é pequena e possui apenas 11 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_Leniencia

Nome da Coluna	Tipo de Dados	Permitir Nulos
CNPJ	varchar(18)	☑
DATA_FIM_ACORDO	varchar(10)	☑
DATA_INICIO_ACORDO	varchar(10)	☑
ID	int	☑
NOME_EMPRESA	varchar(200)	☑
ORGAO_RESPONSAVEL	varchar(200)	☑
NOME_FANTASIA	varchar(200)	☑
QUANTIDADE	int	☑
RAZAO_SOCIAL	varchar(200)	☑
SITUACAO_ACORDO	varchar(100)	☑
UF_EMPRESA	varchar(2)	☑
		☐

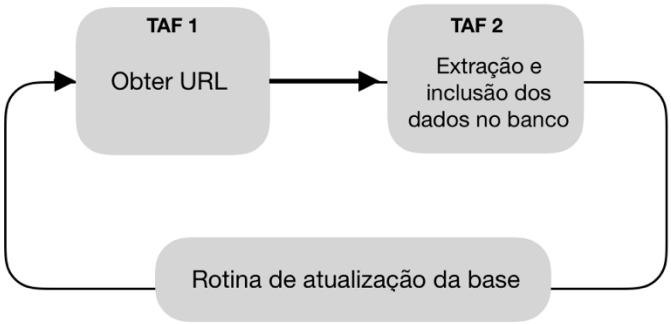
```
CREATE TABLE dbo.Dados_Leniencia
(
   CNPJ VARCHAR(18),
   DATA_FIM_ACORDO VARCHAR(10),
   DATA_INICIO_ACORDO VARCHAR(10),
   ID INT,
   NOME_EMPRESA VARCHAR(200),
   ORGAO_RESPONSAVEL VARCHAR(200),
   NOME_FANTASIA VARCHAR(200),
   QUANTIDADE INT,
   RAZAO_SOCIAL VARCHAR(200),
   SITUACAO_ACORDO VARCHAR(100),
   UF_EMPRESA VARCHAR(2)
);


CREATE INDEX IDX_DadosLeniencia on dbo.Dados_Leniencia(CNPJ);
```

5. FLUXOS DE ETL

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

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

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

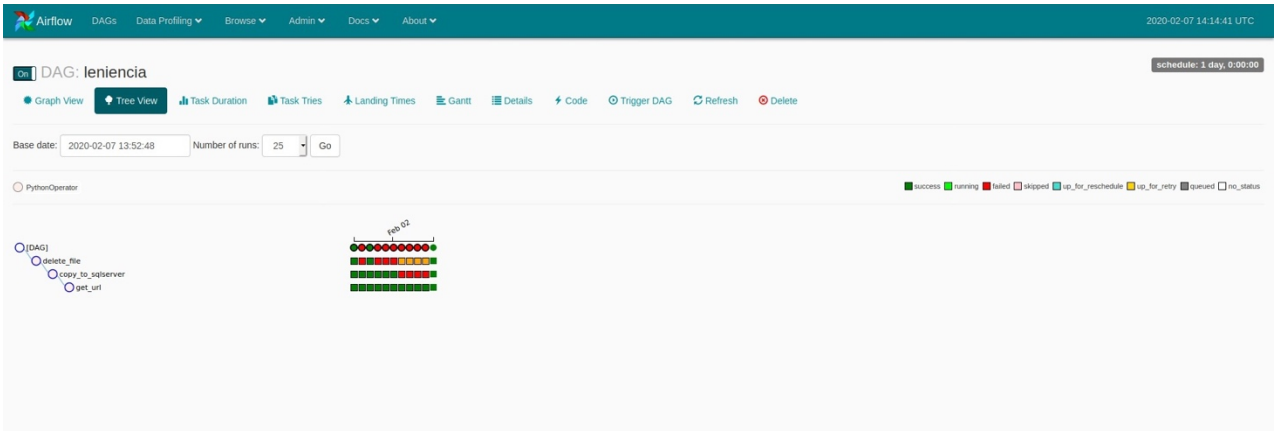
Rotina: **copy_to_sqlserver**

Com o arquivo leniência_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.

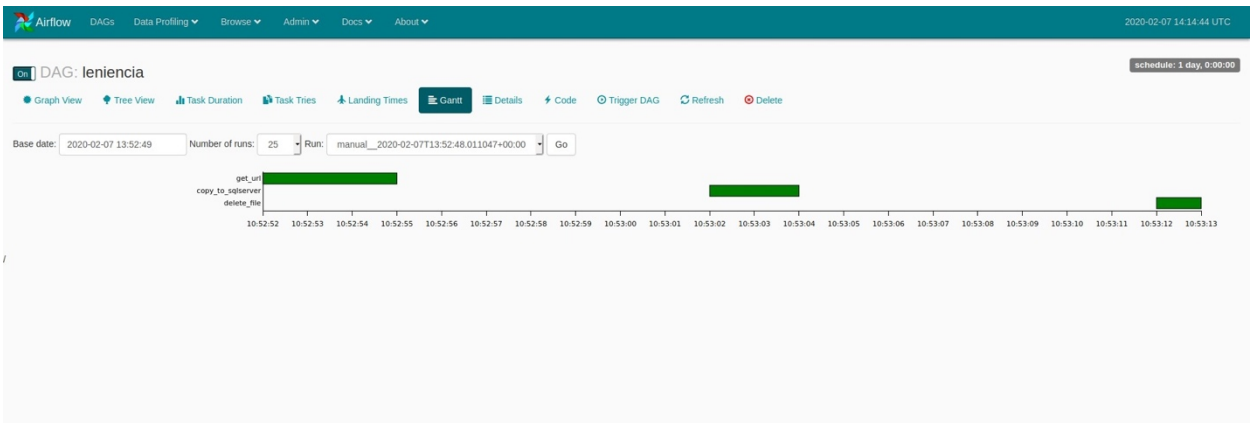
```
conn_string = 'DRIVER={ODBC Driver 17 for SQL Server};SERVER='+
                host+';PORT=1433;DATABASE='+dbname+';UID='+
                username+';PWD='+ password +
                ';UseNTLMv2=yes;TDS_Version=8.0'
```

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 Leniência, o tempo de execução das rotinas foi de cerca de 2 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 Leniência 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='leniencia', 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, não houve tempo hábil para estimar qual é o volume incremental da base à cada atualização.

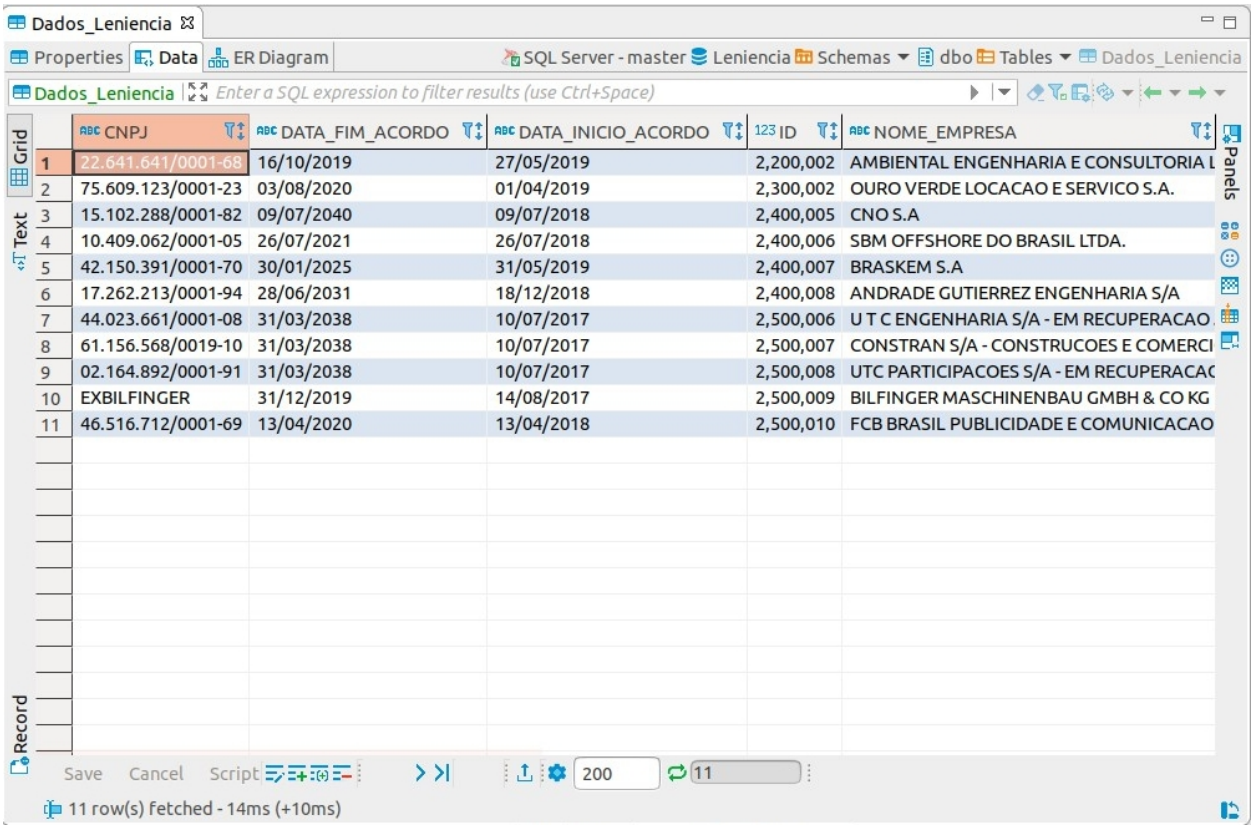
De qualquer forma, o volume total da base é de menos de 500Kb, o que implica que mesmo após várias atualizações futuras, a base não deverá passar de 1 MB.

10. AUXÍLIO NOS ESTUDO DE COMPRAS PUBLICAS

Com os dados da base Leniência, é possível realizar o cruzamento entre os dados das empresas que tiveram acordos de leniência, conforme previsto na Lei Anticorrupção cujo a Controladoria-Geral da União detém a competência exclusiva, com a administração pública federal, afim de verificar se existem algum impedimento ou isenção de sua participação em contratos ou convênios nesses pleitos.

11. EVIDÊNCIA DOS DADOS IMPORTADOS

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



	CNPJ	DATA_FIM_ACORDO	DATA_INICIO_ACORDO	ID	NOME_EMPRESA
1	22.641.641/0001-68	16/10/2019	27/05/2019	2,200,002	AMBIENTAL ENGENHARIA E CONSULTORIA L
2	75.609.123/0001-23	03/08/2020	01/04/2019	2,300,002	OURO VERDE LOCACAO E SERVICO S.A.
3	15.102.288/0001-82	09/07/2040	09/07/2018	2,400,005	CNO S.A
4	10.409.062/0001-05	26/07/2021	26/07/2018	2,400,006	SBM OFFSHORE DO BRASIL LTDA.
5	42.150.391/0001-70	30/01/2025	31/05/2019	2,400,007	BRASKEM S.A
6	17.262.213/0001-94	28/06/2031	18/12/2018	2,400,008	ANDRADE GUTIERREZ ENGENHARIA S/A
7	44.023.661/0001-08	31/03/2038	10/07/2017	2,500,006	U T C ENGENHARIA S/A - EM RECUPERACAO
8	61.156.568/0019-10	31/03/2038	10/07/2017	2,500,007	CONSTRAN S/A - CONSTRUCOES E COMERCI
9	02.164.892/0001-91	31/03/2038	10/07/2017	2,500,008	UTC PARTICIPACOES S/A - EM RECUPERACAC
10	EXBILFINGER	31/12/2019	14/08/2017	2,500,009	BILFINGER MASCHINENBAU GMBH & CO KG
11	46.516.712/0001-69	13/04/2020	13/04/2018	2,500,010	FCB BRASIL PUBLICIDADE E COMUNICACAO

12. BIBLIOGRAFIA

1. Portal da Controladoria Geral da União. Página de Acordos de Leniência. Disponível em: <http://www.cgu.gov.br/assuntos/responsabilizacao-de-empresas/lei-anticorrupcao/acordo-leniencia>. 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. 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.

Escola Nacional de Administração Pública

Laboratório de Tecnologias da Tomada de Decisão – LATITUDE

www.enap.gov.br – www.redes.unb.br