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

Documento:

Relatório de Extração de Dados INIDONEOS

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
07/11/2019	1.0	Danilo Cardoso	Inclusão dos dados da base Inidôneos
15/11/2019	1.1	Leticia Valle	Revisão
07/02/2020	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** 5
2. **ORIGEM DOS DADOS EXTRAÍDOS** 5
3. **QUANTITATIVO DE DADOS** 5
4. **MODELAGEM DO BANCO DE DADOS** 5
5. **FLUXOS DE ETL** 6

   Tarefa 1 **- Download do código HTML para captura da URL de download do arquivo de dados** 7

   Tarefa 2 - Download do arquivo de dados 7

   Tarefa 3 **– Tratamento da base** 7

   Tarefa 4 **– Importação da tabela no banco SQL Server** 7

   Tarefa 5 **– Exclusão dos arquivos intermediários** 8
6. **DURAÇÃO DAS ROTINAS ETL** 8
7. **FLUXO DE TRATAMENTO DE ERROS** 9
8. **FLUXO DE AGENDAMENTO DE ROTINAS** 9
9. **ESTIMATIVA DE CRESCIMENTO** 10
10. **AUXÍLIO NOS ESTUDO DE COMPRAS PUBLICAS** 10
11. **EVIDÊNCIA DOS DADOS IMPORTADOS** 10
12. **BIBLIOGRAFIA** 11

1. INTRODUÇÃO

Este relatório tem como objetivo documentar o processo de extração, tratamento carregamento de dados da base de Inabilitados para Função Pública Segundo o TCU (Inidoneos).

A base INIDONEOS apresenta a relação de pessoas físicas inabilitados para exercício de cargo em comissão ou função de confiança na Administração Pública Federal nos termos do art. 60 da Lei nº 8443/92 (LOTCU) pelo TCU com processos transitados em julgado. [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.

Os dados estão disponíveis no site do TCU na URL: http://portal.tcu.gov.br/responsabilizacao-publica/inabilitados-para-funcao-publica/ com um link auxiliar no portal de dados do governo: http://dados.gov.br/dataset/inabilitados-para-funcao-publica-segundo-tcu, sendo este segundo link o endereço utilizado para a extração dos dados nesse trabalho.

3. QUANTITATIVO DE DADOS

A base INIDONEOS possui apenas uma tabela com 6 colunas e 853 registros, na data de entrega desse relatório. Importa salientar que isso não significa que 888 pessoas estão inabilitadas, mas sim que 888 processos foram concluídos inabilitando pessoas, com algumas pessoas tendo mais de um processo sendo movido contra si.

4. MODELAGEM DO BANCO DE DADOS

Após análise da base INIDONEOS, foi realizada a modelagem dos dados para posterior criação do banco e das tabelas. Como a base é pequena e possui apenas 6 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_Inidoneos

Nome da Coluna	Tipo de Dados	Permitir Nulos
CPF	varchar(11)	☑
NOME	varchar(100)	☑
PROCESSO	varchar(14)	☑
DELIBERACAO	varchar(17)	☑
DATA_TRANSINT	varchar(10)	☑
DATA_FINAL	varchar(10)	☑
		☐

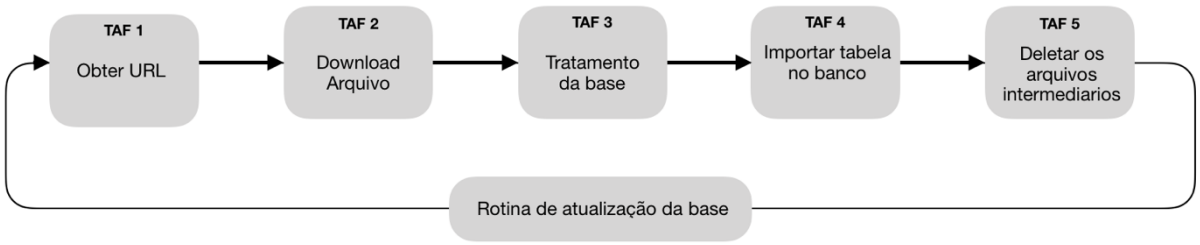
```
CREATE TABLE dbo.Dados_Inidoneos(
CPF varchar(11) NULL,
NOME varchar(100) NULL,
PROCESSO varchar(14) NULL,
DELIBERACAO varchar(17) NULL,
DATA_TRANSINT varchar(10) NULL,
DATA_FINAL varchar(10) NULL
);

CREATE INDEX IDX_DADOS_INIDONEOS ON dbo.Dados_Inidoneos (CPF);
```

5. FLUXOS DE ETL

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

O trabalho de ETL foi desenvolvido na ferramenta Apache Airflow e conta com 5 tarefas diretas, apresentadas no diagrama de blocos a seguir.



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

Rotina: **get_url**

Com o auxilio da biblioteca requests [2] é realizado o download do código HTML da página do portal de dados do governo que contém o link do recurso.

Com o auxílio da biblioteca beautiful soup [3] é possível procurar, dentro do código HTML da página, o link contido dentro do recurso desejado (na página tem vários recursos, o desejado é aquele que se liga com a base em formato .csv). Encontrado este link estático, a URL de referência é salva em um arquivo temporário denominado "url_inidoneos.txt".

Tarefa 2 - Download do arquivo de dados

Rotina: **get_data_from_url**

À partir do arquivo url_inidoneos.txt, obtém-se a URL de download do arquivo e com o auxílio da biblioteca *requests*, realizamos o download do arquivo com o nome inidoneos.csv

Tarefa 3 – Tratamento da base

Rotina: **edita_base**

Com o arquivo inidoneos.csv e com o auxílio da biblioteca pandas [4], são feitas todas as operações desejadas à base de dados. Neste caso a única alteração feita na fonte foi a transformação do CPF em uma sequencia de números (ex: de 000.000.000-00 para 00000000000), removendo os pontos e traços da variável original. Essa base de dados tratada é então movida para o arquivo inidoneos_final.csv.

Tarefa 4 – Importação da tabela no banco SQL Server

Rotina: **truncate_e_reinsert_to_sqlserver**

Com o arquivo inidoneos_final.csv txt e com o auxílio da biblioteca pyodbc [5], 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).

Foi utilizado o Driver ODBC [6] 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'
```

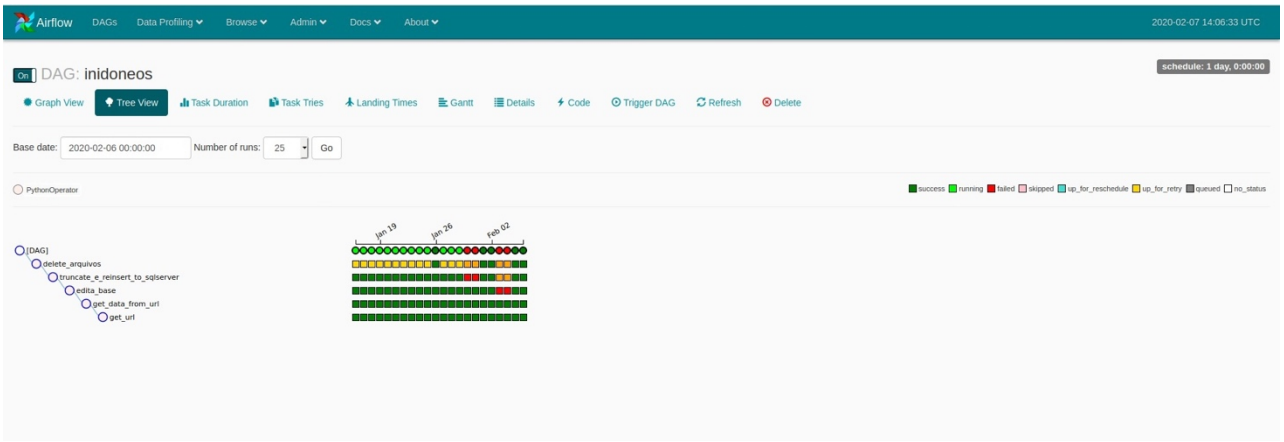
Tarefa 5 – Exclusão dos arquivos intermediários

Rotina: **delete_arquivo**

Com o auxílio da biblioteca *os* [7], deletamos os arquivos baixados e criados ao longo das tarefas intermediarias:

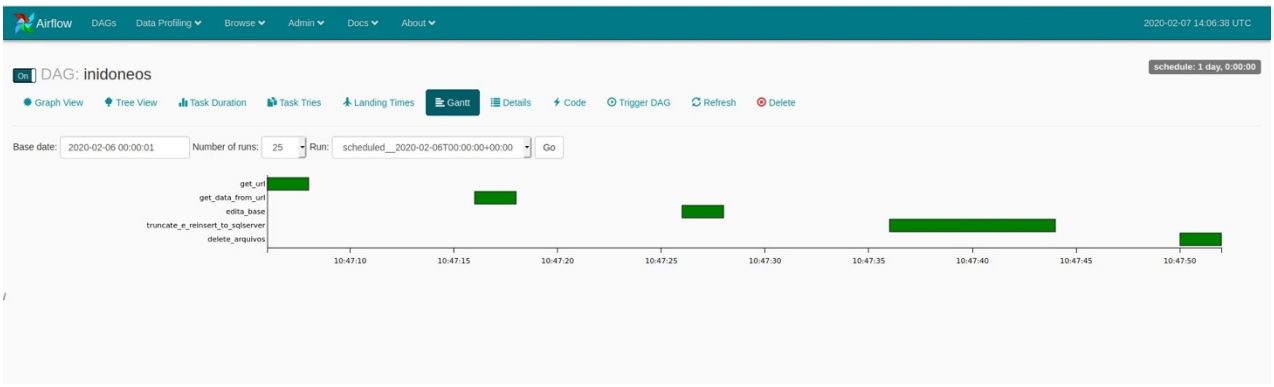
- Tarefa 1: deleta-se o arquivo url_inidoneos.txt
- Tarefa 2: deleta-se o arquivo inidoneos.csv
- Tarefa 3: deleta-se o arquivo inidoneos_final.csv

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 INIDONEOS, o tempo corrido de execução das rotinas foi de cerca de 15 minutos no total, com altíssimo 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 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 resistir mudanças estruturais na página de download do arquivo tais como mais recursos disponíveis ou mudança de ordem.

8. FLUXO DE AGENDAMENTO DE ROTINAS

O TCU não especifica a frequência de atualização da base INIDONEOS, presume-se que seja automático ao haver ocorrências que essas sim, não são frequentes. Olhando os últimos 6 meses em relação ao dia 04/11 aconteceram 8 ocorrências, próximo de uma por mês.

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 com menor frequência sem alterações significativas na base.

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

9. ESTIMATIVA DE CRESCIMENTO

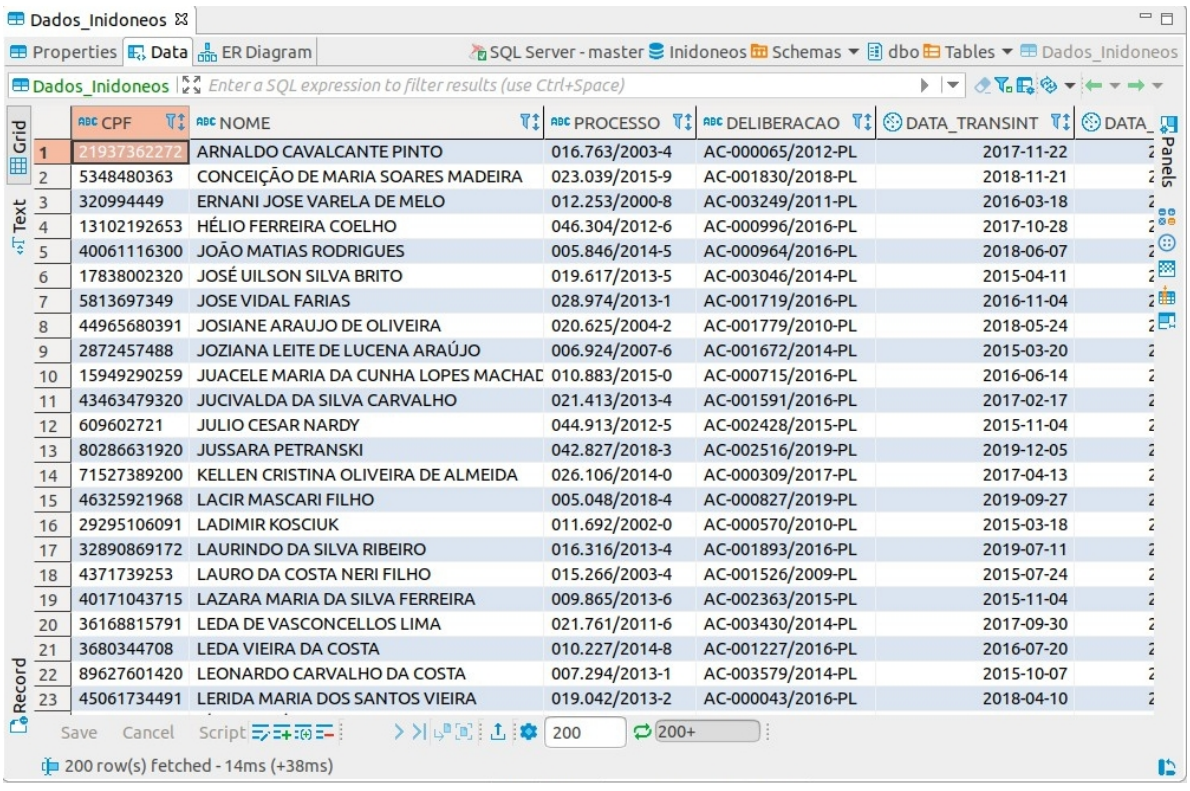
Olhando para o passado recente, com 8 observações em 6 meses, é fato que esta base cresce em ritmo extremamente lento, e dado o tamanho atual de 100 KB dificilmente esta base alcançará 1 MB mesmo em um futuro de longo prazo.

10. AUXÍLIO NOS ESTUDO DE COMPRAS PUBLICAS

Com os dados da base INIDONEOS, 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.



	CPF	NOME	PROCESSO	DELIBERACAO	DATA_TRANSINT	DATA_
1	21937362272	ARNALDO CAVALCANTE PINTO	016.763/2003-4	AC-000065/2012-PL	2017-11-22	2
2	5348480363	CONCEIÇÃO DE MARIA SOARES MADEIRA	023.039/2015-9	AC-001830/2018-PL	2018-11-21	2
3	320994449	ERNANI JOSE VARELA DE MELO	012.253/2000-8	AC-003249/2011-PL	2016-03-18	2
4	13102192653	HÉLIO FERREIRA COELHO	046.304/2012-6	AC-000996/2016-PL	2017-10-28	2
5	40061116300	JOÃO MATIAS RODRIGUES	005.846/2014-5	AC-000964/2016-PL	2018-06-07	2
6	17838002320	JOSÉ UILSON SILVA BRITO	019.617/2013-5	AC-003046/2014-PL	2015-04-11	2
7	5813697349	JOSE VIDAL FARIAS	028.974/2013-1	AC-001719/2016-PL	2016-11-04	2
8	44965680391	JOSIANE ARAUJO DE OLIVEIRA	020.625/2004-2	AC-001779/2010-PL	2018-05-24	2
9	2872457488	JOZIANA LEITE DE LUCENA ARAÚJO	006.924/2007-6	AC-001672/2014-PL	2015-03-20	2
10	15949290259	JUACELE MARIA DA CUNHA LOPES MACHAD	010.883/2015-0	AC-000715/2016-PL	2016-06-14	2
11	43463479320	JUCIVALDA DA SILVA CARVALHO	021.413/2013-4	AC-001591/2016-PL	2017-02-17	2
12	609602721	JULIO CESAR NARDY	044.913/2012-5	AC-002428/2015-PL	2015-11-04	2
13	80286631920	JUSSARA PETRANSKI	042.827/2018-3	AC-002516/2019-PL	2019-12-05	2
14	71527389200	KELLEN CRISTINA OLIVEIRA DE ALMEIDA	026.106/2014-0	AC-000309/2017-PL	2017-04-13	2
15	46325921968	LACIR MASCARI FILHO	005.048/2018-4	AC-000827/2019-PL	2019-09-27	2
16	29295106091	LADIMIR KOSCIUK	011.692/2002-0	AC-000570/2010-PL	2015-03-18	2
17	32890869172	LAURINDO DA SILVA RIBEIRO	016.316/2013-4	AC-001893/2016-PL	2019-07-11	2
18	4371739253	LAURO DA COSTA NERI FILHO	015.266/2003-4	AC-001526/2009-PL	2015-07-24	2
19	40171043715	LAZARA MARIA DA SILVA FERREIRA	009.865/2013-6	AC-002363/2015-PL	2015-11-04	2
20	36168815791	LEDA DE VASCONCELLOS LIMA	021.761/2011-6	AC-003430/2014-PL	2017-09-30	2
21	3680344708	LEDA VIEIRA DA COSTA	010.227/2014-8	AC-001227/2016-PL	2016-07-20	2
22	89627601420	LEONARDO CARVALHO DA COSTA	007.294/2013-1	AC-003579/2014-PL	2015-10-07	2
23	45061734491	LERIDA MARIA DOS SANTOS VIEIRA	019.042/2013-2	AC-000043/2016-PL	2018-04-10	2

12. BIBLIOGRAFIA

1. Inabilitados para função pública. Disponível em: <https://portal.tcu.gov.br/responsabilizacao-publica/inabilitados-para-funcao-publica/>. Acesso em: 04 de novembr 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. Beautiful Soup documentation. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Acesso em: 21 de outubro de 2019.

4. Pandas. Disponível em: <https://pandas.pydata.org/>. Acesso em: 04 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. 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.

7. 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