



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

Documento:

**Relatório de Extração de Dados
SANÇÕES - MG**

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**

Confidencial.

Este documento foi elaborado pela Universidade de Brasília (UnB) para a Enap.
É vedada a cópia e a distribuição deste documento ou de suas partes sem o consentimento, por escrito, da Enap.

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

Confidencial.

Este documento foi elaborado pela Universidade de Brasília (UnB) para a Enap.
É vedada a cópia e a distribuição deste documento ou de suas partes sem o consentimento, por escrito, da Enap.

HISTÓRICO DE REVISÕES

Data	Versão	Autor	Descrição
01/11/2019	1.0	Danilo Cardoso	Inclusão dos dados da base SANÇÕES MG
15/11/2019	1.1	Leticia Valle	Revisão
31/01/2020	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

Confidencial.

Este documento foi elaborado pela Universidade de Brasília (UnB) para a Enap.
É vedada a cópia e a distribuição deste documento ou de suas partes sem o consentimento, por escrito, da Enap.

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
	<i>Tarefa 1 - Coleta dos dados</i>	7
	<i>Tarefa 2 – Tratamento dos dados</i>	7
	<i>Tarefa 3 – Importação da tabela no banco SQL Server</i>	10
	<i>Tarefa 4 – Exclusão dos arquivos intermediários</i>	10
6.	DURAÇÃO DAS ROTINAS ETL.....	11
7.	FLUXO DE TRATAMENTO DE ERROS.....	11
8.	FLUXO DE AGENDAMENTO DE ROTINAS.....	11
9.	ESTIMATIVA DE CRESCIMENTO	11
10.	AUXÍLIO NOS ESTUDO DE COMPRAS PUBLICAS	12
11.	EVIDÊNCIA DOS DADOS IMPORTADOS	12
12.	BIBLIOGRAFIA	13

Confidencial.

Este documento foi elaborado pela Universidade de Brasília (UnB) para a Enap.
É vedada a cópia e a distribuição deste documento ou de suas partes sem o consentimento, por escrito, da Enap.

1. INTRODUÇÃO

Este relatório tem como objetivo documentar o processo de extração, tratamento e carregamento de dados da base de Sanções dadas pelo Governo do Estado de Minas Gerais por meio do CAGEF (Cadastro Geral de Fornecedores).

A base SANÇÕES-MG apresenta a relação de pessoas físicas ou jurídicas que receberam uma sanção do Governo do Estado do Minas Gerais. Esta base é, originalmente, em formato pdf e portanto necessita de um processo de extração posterior para transformar este texto em dados tabulares a ser inseridos no banco de dados.

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 do “Sistema Integrado de Gestão de Aquisições”, na categoria “Fornecedores”, seguida da opção “Consultar Sanções”.

Os dados estão disponíveis no portal do CAGEF, na URL:
<https://www.cagef.mg.gov.br/fornecedor-web/br/gov/prodemge/seplag/fornecedor/publico/>

3. QUANTITATIVO DE DADOS

As informações da base SANCOES-MG estão contidas em um PDF com 7 colunas e 20 páginas, na data de entrega deste relatório. Após tratamento dos dados, a base se converte em uma tabela de 8 colunas e cerca de 394 linhas.

4. MODELAGEM DO BANCO DE DADOS

Após análise da base SANCOES-MG, foi realizada a modelagem dos dados para posterior criação do banco e das tabelas. Como a base é pequena e possui apenas 8 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_SancaoMG			
Nome da Coluna	Tipo de Dados	Permitir Nul...	
CPF	varchar(11)	<input checked="" type="checkbox"/>	
CNPJ	varchar(14)	<input checked="" type="checkbox"/>	
NOME	varchar(100)	<input checked="" type="checkbox"/>	
TIPO_PENALIDADE	varchar(50)	<input checked="" type="checkbox"/>	
DATA_INICIO_PENALIDA...	varchar(10)	<input checked="" type="checkbox"/>	
DATA_FIM_PENALIDADE	varchar(10)	<input checked="" type="checkbox"/>	
DATA_DESPACHO	varchar(10)	<input checked="" type="checkbox"/>	
ORGAO_APLICADOR	varchar(100)	<input checked="" type="checkbox"/>	
		<input type="checkbox"/>	

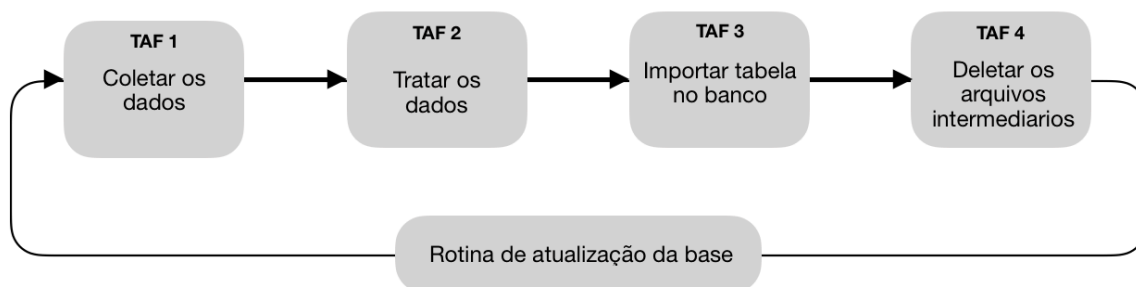
```
CREATE TABLE dbo.Dados_SancaoMG (
    CPF varchar(11) NULL,
    CNPJ varchar(14) NULL,
    NOME varchar(100) NULL,
    TIPO_PENALIDADE varchar(50) NULL,
    DATA_INICIO_PENALIDADE varchar(10) NULL,
    DATA_FIM_PENALIDADE varchar(10) NULL,
    DATA_DESPACHO varchar(10) NULL,
    ORGAO_APLICADOR varchar(100) NULL
);
```

```
CREATE INDEX IDX_DADOS_SANCAOMG ON dbo.Dados_SancaoMG (CPF,CNPJ);
```

5. FLUXOS DE ETL

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

O trabalho de ETL foi desenvolvido na ferramenta Apache Airflow e conta com 4 tarefas diretas, apresentadas no diagrama de blocos a seguir.



Confidencial.

Este documento foi elaborado pela Universidade de Brasília (UnB) para a Enap.
É vedada a cópia e a distribuição deste documento ou de suas partes sem o consentimento, por escrito, da Enap.

Tarefa 1 - Coleta dos dados

Rotina: **get_dados**

Com o auxílio do plugin para airflow em docker da biblioteca selenium [1] é realizada a navegação por meio de software de automação. Primeiro, dirige-se para a página referenciada na seção da origem dos dados, ou seja, a URL <https://www.cagef.mg.gov.br/fornecedor-web/br/gov/prodemge/seplag/fornecedor/publico/>

Nessa URL é necessário fazer algumas seleções. Primeiro, deve-se selecionar dentro do menu principal a opção de consultas, em seguida selecionar os fornecedores impedidos para por fim, apertar o botão de listar todos os fornecedores impedidos, que faz download automaticamente de um arquivo PDF contendo os dados.

Tarefa 2 – Tratamento dos dados

Rotina: **pdf_to_csv**

A partir do arquivo **Relatorio_Fornecedores_Impedidos.pdf** gerado na tarefa 1 e com o auxílio da biblioteca PyPDF2 [2], o arquivo PDF é lido e transformado em um arquivo csv.

Arquivos PDF não tem sinais de quebra de texto, assim, o texto extraído é uma longa frase sem distinção entre linhas ou campos. A única quebra natural do PDF fornecido é entre páginas. Assim, convém criar uma solução que resolve uma página e replicar para as demais. Nas duas imagens a seguir, é possível observar a página do PDF original e qual texto é extraído pela biblioteca do Python.



GOVERNO DO ESTADO DE MINAS GERAIS
SECRETARIA DE ESTADO DE PLANEJAMENTO DE GESTÃO
Sistema Integrado de Administração de Materiais e Serviços - SIAD

Relatórios de Fornecedores Impedidos

		Judicial				DE MINAS GERAIS
130.334.686-91	Aloisio Fonseca Lara	Suspensão por Decisão Judicial	13/10/2016	12/10/2021	08/12/2016	TRIBUNAL DA JUSTICA DO ESTADO DE MINAS GERAIS
029.247.956-56	Altair Sebastião de Souza	Suspensão por Decisão Judicial	04/07/2018	03/07/2028	14/11/2018	TRIBUNAL DA JUSTICA DO ESTADO DE MINAS GERAIS
047.583.556-57	Ana Flávia Alves	Suspensão por até 5 anos	28/06/2018	27/06/2023	15/05/2019	TRIBUNAL REGIONAL ELEITORAL DE MINAS GERAIS
178.424.336-15	Antônio Francelino dos Santos	Suspensão por Decisão Judicial	03/01/2016	02/01/2021	29/08/2017	TRIBUNAL DA JUSTICA DO ESTADO DE MINAS GERAIS
054.922.206-59	Arailton Francisco de Oliveira	Inidoneidade	01/03/2013		01/03/2013	SECRETARIA DE ESTADO DE JUSTICA E SEGURANCA PUBLIC
079.994.766-09	Arailton Francisco de Oliveira Neto	Inidoneidade	01/03/2013		01/03/2013	SECRETARIA DE ESTADO DE JUSTICA E SEGURANCA PUBLIC
02.993.861/0001-43	BERGSON DO BRASIL EIRELI - EPP	Suspensão por até 5 anos	30/09/2015	28/09/2020	30/09/2015	COMPANHIA ENERGÉTICA DE MINAS GERAIS
04.997.313/0001-17	BERMA ENGENHARIA EIRELI	Suspensão por até 2 anos	07/02/2019	06/02/2021	04/09/2019	COMPANHIA DE SANEAMENTO DE MINAS GERAIS
01.856.601/0001-63	BETAMAQ TRATORPECAS LTDA -ME	Inidoneidade	21/12/2007		21/12/2007	SECRETARIA DE ESTADO DE CULTURA E TURISMO
19.167.675/0001-58	BON MENU COMERCIO E ALIMENTACAO LTDA	Inidoneidade	20/10/2018		14/09/2019	SECRETARIA DE ESTADO DE JUSTICA E SEGURANCA PUBLIC
20.676.116/0001-52	BORMAPLAS INDUSTRIA E COMÉRCIO DE ARTEFATOS DE BORRACHA MADEIRA E PLÁSTICO	Suspensão por até 2 anos	23/03/2018	22/03/2020	27/07/2018	COMPANHIA ENERGÉTICA DE MINAS GERAIS
04.826.675/0001-45	BORÁ AGROPECUÁRIA LTDA.	Suspensão por até 2 anos	08/03/2019	07/03/2021	31/08/2019	COMPANHIA ENERGÉTICA DE MINAS GERAIS
10.273.448/0001-32	CARFAG COMERCIO E SERVIÇO DE MANUTENÇÃO LTDA	Inidoneidade	28/07/2017		15/03/2018	FUNDAÇÃO CENTRO DE HEMATOLOGIA E HEMOTERAPIA DE MG
029.176.728-11	CARLITO MOREIRA SILVA	Suspensão por Decisão Judicial	05/12/2014	04/12/2019	24/07/2015	CONTROLADORIA GERAL DO ESTADO DE MINAS GERAIS
089.153.006-10	CARLOS GUIDUCE SOARES	Suspensão por Decisão Judicial	12/04/2015	11/04/2020	08/10/2016	TRIBUNAL DA JUSTICA DO ESTADO DE MINAS GERAIS
66.292.301/0001-44	CASTRO SIMAO ENGENHARIA LTDA	Suspensão por até 2 anos	01/03/2019	28/02/2021	27/07/2019	TRIBUNAL DA JUSTICA DO ESTADO DE MINAS GERAIS
11.862.022/0001-87	CB AGROFLORESTAL LTDA -ME	Suspensão por até 2 anos	15/06/2018	14/06/2020	21/09/2018	EMPRESA DE ASSIST. TECNICA E EXTENSÃO RURAL DE MG

www.cagef.mg.gov.br/fornecedor-web

Emitido em: 07/11/2019 às 17:36:58

Página 3 de 20

```

1 >>> i=2
2 >>> texto_pagina = pdfReader.getPage(i).extractText()
3 >>> texto_pagina
4 'JudicialDE MINAS GERAIS130.334.686-91Aloisio Fonseca LaraSuspensão por DecisãoJudicial13/10/201612/10/202108/12/2016TRIBUNAL DA JUSTICA DO ESTADODE MINAS GERAIS029.247.956-56Altair Sebastião de SouzaSuspensão por DecisãoJudicial04/07/201803/07/202814/11/2018TRIBUNAL DA JUSTICA DO ESTADODE MINAS GERAIS047.583.556-57Ana Flávia AlvesSuspensão por até 5 anos28/06/201827/06/202315/05/2019TRIBUNAL REGIONAL ELEITORALDE MINAS GERAIS178.424.336-15Antônio Francelino dos SantosSuspensão por DecisãoJudicial03/01/201602/01/202129/08/2017TRIBUNAL DA JUSTICA DO ESTADODE MINAS GERAIS054.922.206-59Arailton Francisco de OliveiraInidoneidade01/03/201301/03/2013SECRETARIA DE ESTADO DE JUSTICA E SEGURANCA PUBLIC079.994.766-09Arailton Francisco de Oliveira NetoInidoneidade01/03/201301/03/2013SECRETARIA DE ESTADO DE JUSTICA E SEGURANCA PUBLIC02.993.861/0001-43BERGSON DO BRASIL EIRELI -EPPSuspensão por até 5 anos30/09/201528/09/202030/09/2015COMPANHIA ENERGÉTICA DE MINASGERAIS04.997.313/0001-17BERMA ENGENHARIA EIRELISuspensão por até 2 anos07/02/201906/02/202104/09/2019COMPANHIA DE SANEAMENTO DE MINAS GERAIS01.856.601/0001-63BETAMAQ TRATORPECAS LTDA-MEInidoneidade21/12/200721/12/2007SECRETARIA DE ESTADO DE CULTURA E TURISMO19.167.675/0001-58BON MENU COMERCIO E ALIMENTACAO LTDAInidoneidade20/10/201814/09/2019SECRETARIA DE ESTADO DE JUSTICA E SEGURANCA PUBLIC20.676.116/0001-52BORMAPLAS INDUSTRIA E COMÉRCIO DE ARTEFATOS DE BORRACHA MADEIRA E PLÁSTICOSuspensão por até 2 anos23/03/201822/03/202027/07/2018COMPANHIA ENERGÉTICA DE MINASGERAIS04.826.675/0001-45BORÁ AGROPECUÁRIA LTDA.Suspensão por até 2 anos08/03/201907/03/202131/08/2019COMPANHIA ENERGÉTICA DE MINASGERAIS10.273.448/0001-32CARFAG COMERCIO E SERVIÇO DE MANUTENÇÃO LTDAInidoneidade28/07/201715/03/2018FUNDACAO CENTRO DE HEMATOLOGIA E HEMOTERAPIA DE MG029.176.728-11CARLITO MOREIRA SILVASuspensão por DecisãoJudicial05/12/201404/12/201924/07/2015CONTROLADORIA GERAL DO ESTADO DE MINAS GERAIS089.153.006-10CARLOS GUIDUCE SOARESSuspensão por DecisãoJudicial12/04/201511/04/202008/10/2016TRIBUNAL DA JUSTICA DO ESTADO DE MINAS GERAIS66.292.301/0001-44CASTRO SIMAO ENGENHARIA LTDAInidoneidade01/03/201928/02/202127/07/2019EMPRESA DE ASSIST. TECNICA E EXTENSÃO RURAL DE MG11.862.022/0001-87CB AGROFLORESTAL LTDA -MESuspensão por até 2 anos15/06/201814/06/202021/09/2018EMPRESA DE ASSIST. TECNICA E EXTENSÃO RURAL DE MG
fornecedor-webEmitido em: 07/11/2019 às 17:36:58Página 3 de 20

```

Existem algumas peculiaridades a serem tratadas:

- ◇ Em algumas páginas o cabeçalho (a parte acima da tabela, GOVERNO DO ESTADO...) é detectado e extraído e em algumas não.
- ◇ Existem situações em que um resquícios do último registro da página anterior persiste na página seguinte (tal como na imagem).
- ◇ O rodapé da página é detectado em todos os casos.

Confidencial.

Este documento foi elaborado pela Universidade de Brasília (UnB) para a Enap.
É vedada a cópia e a distribuição deste documento ou de suas partes sem o consentimento, por escrito, da Enap.

- ◇ É comum que campos distintos se misturem, como no exemplo: “Netolnidoneidade01/03/201301/03/2013/SECRETARIA DE ESTADO”, no qual antes do primeiro espaço tem resquício do nome, a razão da sanção, a data de início, a data de fim e o começo do nome do órgão sem sequer um espaço entre estes campos.
- ◇ Ocasionalmente, existem campos vazios de forma que buscar as informações via ordem de aparecimento é uma regra que necessita de exceções. De maneira mais clara, se órgão é o 7º campo, mas faltar o 4º termo, não adianta procurar o órgão na 7ª posição, mas sim na 6ª.

Para resolver os problemas, primeiro remove-se o irrelevante: cabeçalho e rodapé. Como o cabeçalho e o rodapé são constantes (quando detectados, são sempre a mesma coisa), procura-se sempre os mesmos.

Com apenas o que é relevante, primeiro toma-se vantagem de o primeiro campo da tabela ser um CPF ou CNPJ para inserir um `enter/newline` logo antes de um CPF ou CNPJ para separar as informações de cada linha em um termo separado. Assim, é possível separar as informações relevantes de cada registro sem que uma observação influencie a outra.

Essa separação é feita utilizando expressões regulares. Em python são implementadas utilizando a biblioteca `re` [3]. Expressões regulares são formas de expressão de padrões textuais, com alta capacidade de expressão de baixo nível. Assim, para encontrar os CNPJ procura-se por uma sequência do tipo: `nn.nnn.nnn/nnnn-nn`, onde cada `n` é um dígito numérico. Para se encontrar os CPFs procura-se o padrão `nnn.nnn.nnn-nn`, onde, de mesma forma, `n` são dígitos numéricos. Estes padrões são raros o suficiente para não gerar falsos positivos.

Aproveita-se o fato de adicionar um `enter/newline` antes do CPF/CNPJ para remover o primeiro resultado, afinal o que vier antes do primeiro CPF tem que ser ou resquício da página anterior ou uma linha vazia.

Após esses cuidados, tem-se a separação de cada linha (registro) da tabela, mas apenas foi separado o primeiro campo, o CPF/CNPJ. O próximo passo é separar o campo penalidade. Sabendo que os valores existentes são: “Inidoneidade”, “Suspensão por até 2 anos”, “Suspensão por até 5 anos”, “Suspensão por DecisãoJudicial” e “Suspensão por Decisão”, é possível procurar por cada uma destas expressões. Depois desta separação, já temos os seguintes campos separados: CPF/CNPJ, o nome da entidade sancionada (entre o fim do CPF e o início da penalidade) e a penalidade. Por fim, se quebram as datas com a máscara `nn/nn/nnnn` e todos os campos estão separados.

Após separar todos os campos, é necessário tratar os casos em que existem menos registros que o total. No caso em que são apenas 6 campos, ao invés dos 7 usuais, foi observado que é por que falta a data de fim de vigência. No caso em que são 5 campos, falta o órgão sancionador além da data de vigência. Nesses casos, adicionam-se campos vazios nas posições adequadas movendo os campos que estavam na posição errada para a direita.

Neste momento, todas as informações do PDF foram passadas para uma variável estruturada com todos as informações no seu devido lugar. De um ponto de vista lógico, pode-se começar a tarefa de tratamento de variáveis. Os tratamentos são a remoção de caracteres não-numéricos do CPF/CNPJ, a separação dos mesmos em colunas diferentes,

Confidencial.

Este documento foi elaborado pela Universidade de Brasília (UnB) para a Enap.

É vedada a cópia e a distribuição deste documento ou de suas partes sem o consentimento, por escrito, da Enap.

e o ajuste dos nomes de punição de: "Suspensão por Decisão Judicial" ou "Suspensão por Decisão" para "Suspensão por Decisão Judicial".

Depois de todo o tratamento das variáveis, as informações são guardadas no arquivo 'sancoes_mg_dados.txt'.

Tarefa 3 – Importação da tabela no banco SQL Server

Rotina: **truncate_e_reinsert_to_sqlserver**

Com o arquivo sancoes_mg_dados.txt e com o auxílio da biblioteca [pyodbc](#) [4], 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).

Nessa tarefa, foi utilizado o [Driver ODBC](#) [5] 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'
```

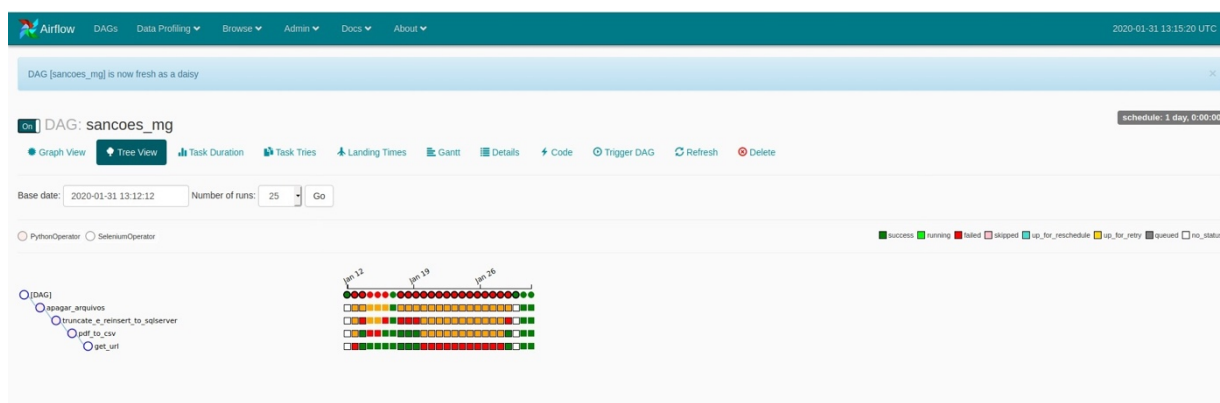
Tarefa 4 – Exclusão dos arquivos intermediários

Rotina: **apagar_arquivos**

Com o auxílio da biblioteca [os](#) [6], apagam-se os arquivos baixados e criados ao longo das tarefas intermediárias:

- Tarefa 1: deleta-se o arquivo Relatorio_Fornecedores_Impedidos.pdf
- Tarefa 2: deleta-se o arquivo sancoes_mg_dados.txt

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

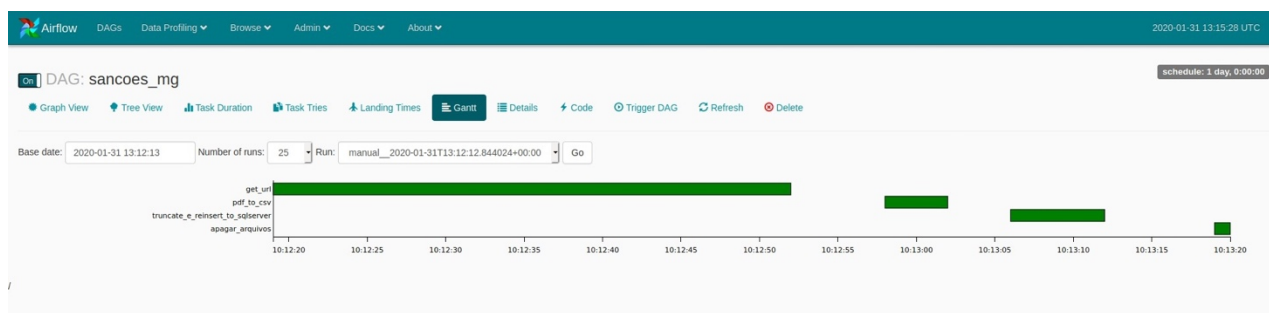


Confidencial.

Este documento foi elaborado pela Universidade de Brasília (UnB) para a Enap.
É vedada a cópia e a distribuição deste documento ou de suas partes sem o consentimento, por escrito, da Enap.

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 SANCOES-MG, o tempo corrido de execução das rotinas foi de cerca de 5 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.

8. FLUXO DE AGENDAMENTO DE ROTINAS

O CAGEF não especifica a frequência de atualização da base SANCOES-MG, presume-se que seja automático ao haver ocorrências e que essas não sejam frequentes. Olhando o ano de 2019, foram observadas 32 ocorrências, aproximadamente uma a cada semana.

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 uma vez a cada 15 dias ou uma vez por mês sem alterações significativas na base.

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

9. ESTIMATIVA DE CRESCIMENTO

Olhando para o passado recente, com 32 observações em 10 meses, é possível inferir cerca de 50 observações por ano. Considerando que o tamanho atual da base é de cerca

Confidencial.

Este documento foi elaborado pela Universidade de Brasília (UnB) para a Enap.

É vedada a cópia e a distribuição deste documento ou de suas partes sem o consentimento, por escrito, da Enap.

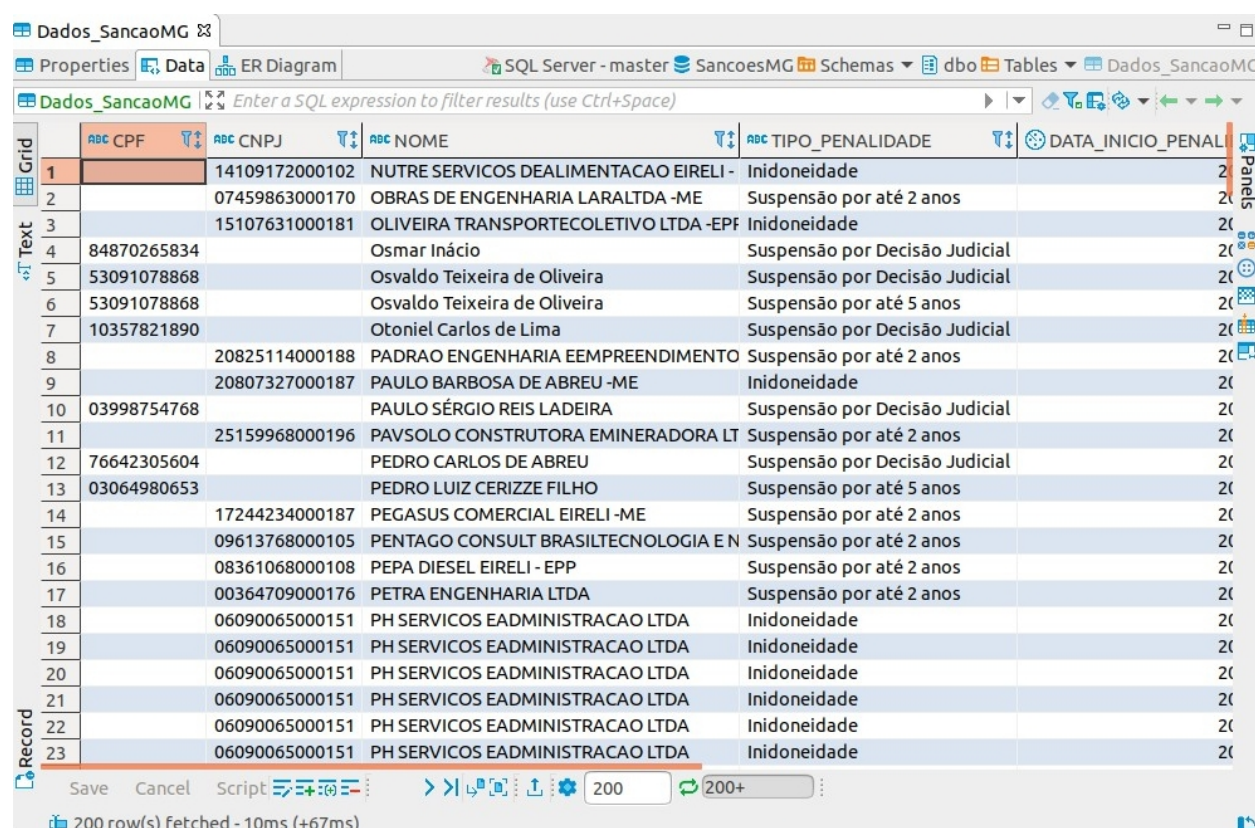
de 400 observações, o tamanho da base levará alguns anos para dobrar de tamanho. Assim, considerando o tamanho atual de cerca de 500KB para o PDF e mais 50 KB para os dados em si, 1 MB é suficiente para atender às necessidades em um futuro de longo prazo

10. AUXÍLIO NOS ESTUDO DE COMPRAS PUBLICAS

Com os dados da base SANCOES-MG, é possível realizar o cruzamento entre os dados das pessoas físicas ou jurídicas sancionadas 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.



Grid	CPF	CNPJ	NOME	TIPO_PENALIDADE	DATA_INICIO_PENAL
1		14109172000102	NUTRE SERVICOS DEALIMENTACAO EIRELI -	Inidoneidade	20
2		07459863000170	OBRAS DE ENGENHARIA LARALTD -ME	Suspensão por até 2 anos	20
3		15107631000181	OLIVEIRA TRANSPORTECOLETIVO LTDA -EPF	Inidoneidade	20
4	84870265834		Osmar Inácio	Suspensão por Decisão Judicial	20
5	53091078868		Osvaldo Teixeira de Oliveira	Suspensão por Decisão Judicial	20
6	53091078868		Osvaldo Teixeira de Oliveira	Suspensão por até 5 anos	20
7	10357821890		Otoniel Carlos de Lima	Suspensão por Decisão Judicial	20
8		20825114000188	PADRAO ENGENHARIA EEMPREENDIMENTO	Suspensão por até 2 anos	20
9		20807327000187	PAULO BARBOSA DE ABREU -ME	Inidoneidade	20
10	03998754768		PAULO SÉRGIO REIS LADEIRA	Suspensão por Decisão Judicial	20
11		25159968000196	PAVSOLO CONSTRUTORA EMINERADORA LT	Suspensão por até 2 anos	20
12	76642305604		PEDRO CARLOS DE ABREU	Suspensão por Decisão Judicial	20
13	03064980653		PEDRO LUIZ CERIZZE FILHO	Suspensão por até 5 anos	20
14		17244234000187	PEGASUS COMERCIAL EIRELI -ME	Suspensão por até 2 anos	20
15		09613768000105	PENTAGO CONSULT BRASILTECNOLOGIA E N	Suspensão por até 2 anos	20
16		08361068000108	PEPA DIESEL EIRELI - EPP	Suspensão por até 2 anos	20
17		00364709000176	PETRA ENGENHARIA LTDA	Suspensão por até 2 anos	20
18	06090065000151		PH SERVICOS EADMINISTRACAO LTDA	Inidoneidade	20
19	06090065000151		PH SERVICOS EADMINISTRACAO LTDA	Inidoneidade	20
20	06090065000151		PH SERVICOS EADMINISTRACAO LTDA	Inidoneidade	20
21	06090065000151		PH SERVICOS EADMINISTRACAO LTDA	Inidoneidade	20
22	06090065000151		PH SERVICOS EADMINISTRACAO LTDA	Inidoneidade	20
23	06090065000151		PH SERVICOS EADMINISTRACAO LTDA	Inidoneidade	20

Confidencial.

Este documento foi elaborado pela Universidade de Brasília (UnB) para a Enap.
É vedada a cópia e a distribuição deste documento ou de suas partes sem o consentimento, por escrito, da Enap.

12. BIBLIOGRAFIA

1. What is Selenium?. Disponível em: < <https://www.seleniumhq.org/>>. Acesso em: 21 de outubro de 2019.
2. PyPDF3 Documentation. Disponível em: < <https://pythonhosted.org/PyPDF2/> >. Acesso em 13 de novembro de 2019.
3. Regular Expressions Operations. Disponível em: < <https://docs.python.org/3/library/re.html> >. Acesso em 13 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.
6. 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



Confidencial.

Este documento foi elaborado pela Universidade de Brasília (UnB) para a Enap.
É vedada a cópia e a distribuição deste documento ou de suas partes sem o consentimento, por escrito, da Enap.