



TDC  
ONLINE



# Inovação com dados em nuvem

TRILHA

#TheDevConf  
Oracle

## Analizando grandes volumes de dados abertos

Daniel Panizzo  
23.03.21 16h00



# Inovação com dados em nuvem

TRILHA

#TheDevConf  
Oracle



Este trabalho está licenciado sob uma Licença Creative Commons Atribuição-Compartilhamento 4.0 Internacional.

Para ver uma cópia desta licença, visite <http://creativecommons.org/licenses/by-sa/4.0/>.



# Inovação com dados em nuvem

**ANALISANDO GRANDES  
VOLUMES DE DADOS  
ABERTOS**





# Analizando grandes volumes de dados abertos

---

Daniel Panizzo  
Engenheiro de Soluções

March, 2021, Version 2.0  
Copyright © 2021, Oracle and/or its affiliates  
Public

# ÍNDICE

---

<b>INTRODUÇÃO</b>	<b>6</b>
<b>LAB.1: COLETAR OS DADOS ABERTOS</b>	<b>7</b>
Obtendo os códigos do Github	7
Coletando os dados da Receita Federal	9
Coletando os dados do IBGE	15
<b>LAB.2: ARMAZENAR OS DADOS ABERTOS</b>	<b>16</b>
Carregando os dados no Object Storage	16
Armazenando os dados no banco de dados	21
<b>LAB.4: ANALISAR OS DADOS ABERTOS</b>	<b>25</b>
Conectando o Analytics com o ADW	25
Importando os metadados	29
Ajustando os metadados	32
Visualizando os dados	33

## INTRODUÇÃO

Neste passo-a-passo, iremos analisar grandes volumes de dados abertos e, para isso, passaremos pelo fluxo geral de um projetos de análise de dados. Coletaremos os dados em sua forma primária nos repositórios disponibilizados por organizações públicas. Em seguida, armazenaremos estes dados em um banco de dados. Uma vez armazenados, iremos prepará-los e padronizá-los para, ao final, analisá-los no Oracle Analytics.

Para executar este passo-a-passo, será necessário que você esteja familiarizado com as seguintes linguagens:

- Python
- SQL

E que você tenha as seguintes ferramentas:

- [Jupyter Notebook ou JupyterLab](#) com Python 3.6 ou maior, instalado na sua máquina.
  - Necessários para executar os notebooks com código Python que automatizam a coleta dos dados. Sugerimos instalação da distribuição [Anaconda](#), que já inclui o Jupyter Notebook, JupyterLab, Python 3.8 e diversas bibliotecas padrões para Cientistas de Dados.
- Uma conta no [OCI - Oracle Cloud Infrastructure](#) com:
  - Um Autonomous Data Warehouse, necessário para o armazenamento e preparação dos dados. Veja [aqui](#) como provisionar um banco de dados autônomo gratuito no OCI.
- [Oracle Analytics Desktop](#) instalado na sua máquina.
  - Necessário para analisar e visualizar os dados.

Durante este passo-a-passo, acessaremos as seguintes origens de dados abertos:

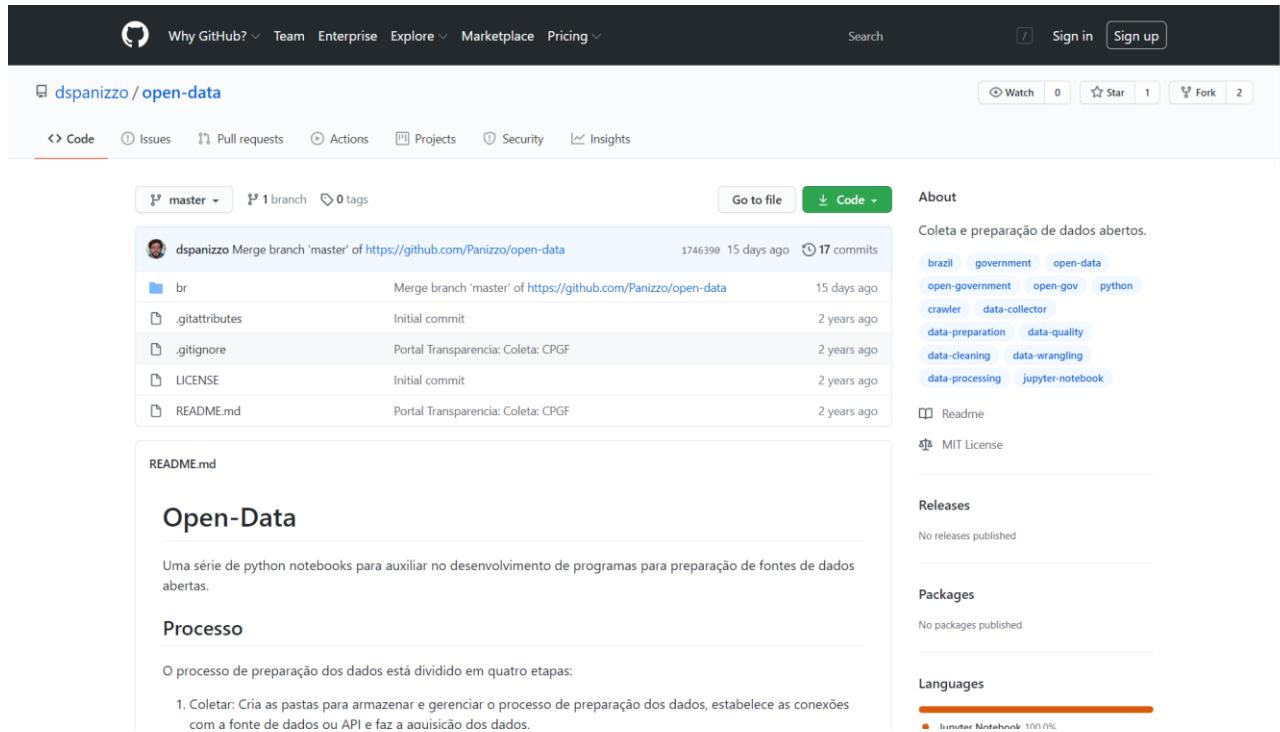
- [Dados públicos do CNPJ da Receita Federal](#)
- [API de Serviço de Dados do Instituto Brasileiro de Geografia e Estatística](#)

Outros scripts necessários para a execução deste passo-a-passo estão contidos dentro da pasta “arquivos-de-apoio” dentro do arquivo ZIP.

## LAB.1: COLETAR OS DADOS ABERTOS

### Obtendo os códigos do Github

1. Acesse o repositório [dspanizzo/opendata](https://github.com/dspanizzo/opendata) no Github.



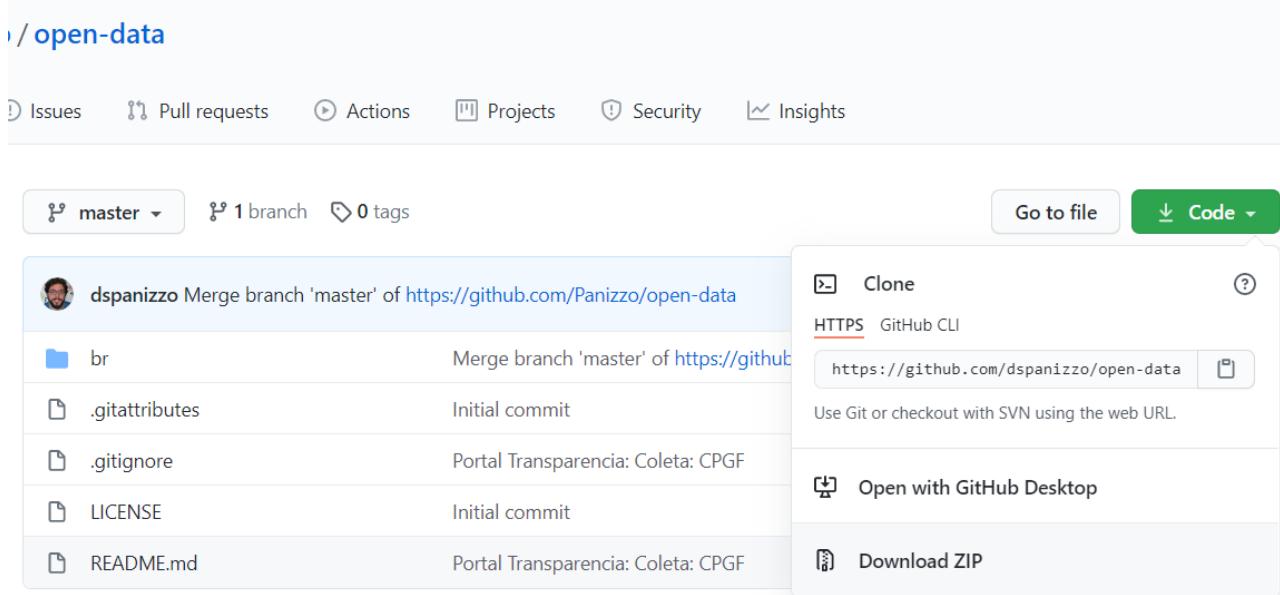
The screenshot shows the GitHub repository page for [dspanizzo/opendata](https://github.com/dspanizzo/opendata). The repository has 17 commits, 1 branch, and 0 tags. The README.md file contains the following content:

```
Open-Data
Uma série de python notebooks para auxiliar no desenvolvimento de programas para preparação de fontes de dados abertas.

Processo
O processo de preparação dos dados está dividido em quatro etapas:
1. Coletar: Cria as pastas para armazenar e gerenciar o processo de preparação dos dados, estabelece as conexões com a fonte de dados ou API e faz a adquisição dos dados.
```

The 'Code' button is highlighted in green, indicating it was clicked.

2. Clique no botão verde “Code” e selecione a opção “Download ZIP”.



The screenshot shows the GitHub repository page for [dspanizzo/opendata](https://github.com/dspanizzo/opendata). The 'Code' button is highlighted in green. A context menu is open, showing the following options:

- Clone
- HTTPS GitHub CLI
- Open with GitHub Desktop
- Download ZIP

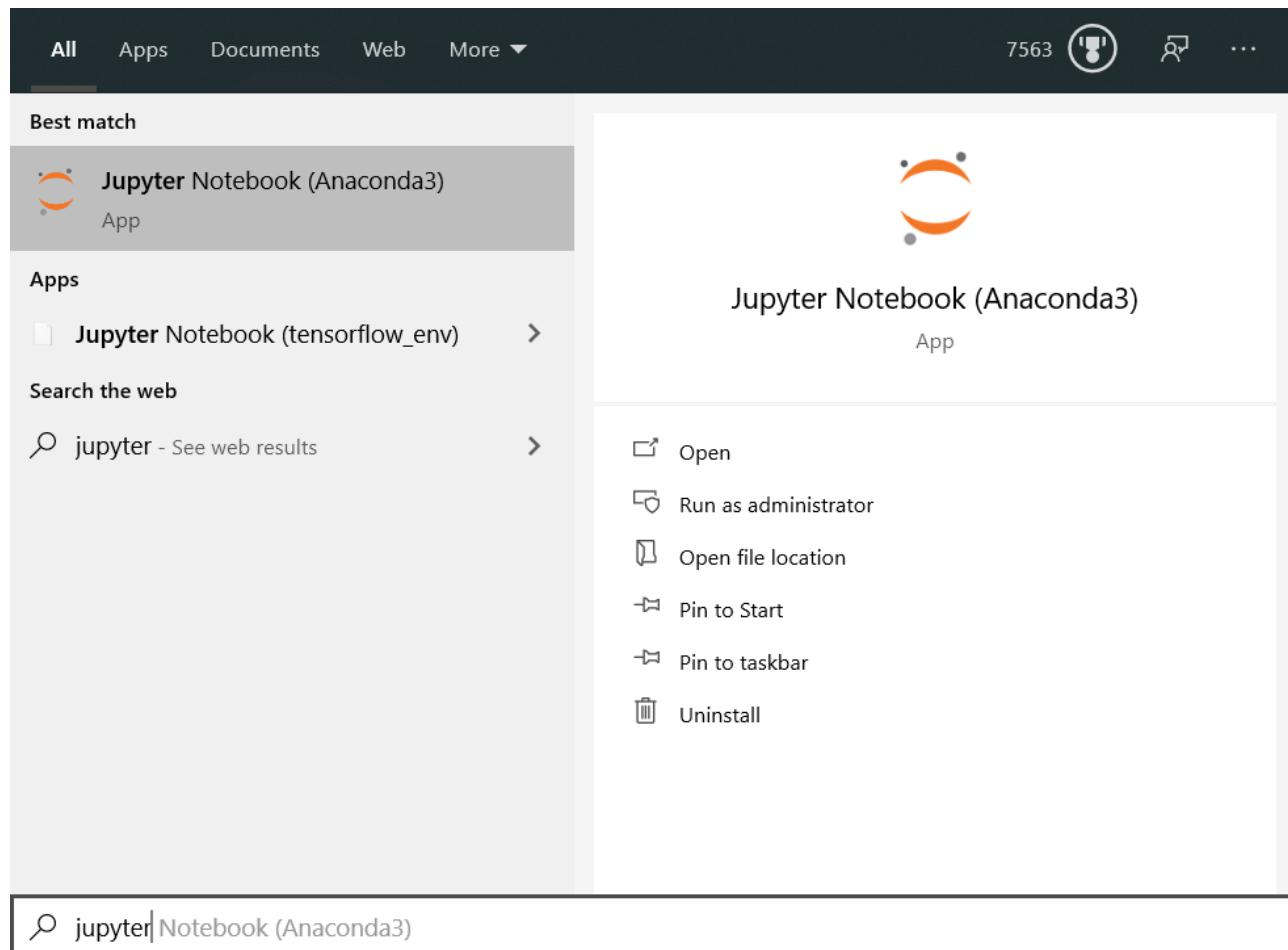
The 'Download ZIP' option is highlighted in green.

3. Localize no seu computador o local de download do arquivo “open-data-master.zip” e copie-o para outra pasta de sua preferência ( Exemplo: C:\Users\dspanizzo\Documents\TDC ). Em seguida, descompacte-o. Você deverá ver os arquivos listados abaixo.

Documents > TDC > open-data-master >

Name	Date modified	Type	Size
br	12/11/2020 18:15	File folder	
.gitattributes	12/11/2020 18:15	GITATTRIBUTES File	1 KB
.gitignore	12/11/2020 18:15	GITIGNORE File	2 KB
LICENSE	12/11/2020 18:15	File	2 KB
README.md	12/11/2020 18:15	MD File	2 KB

4. No seu computador, abra o Jupyter Notebook ou JupyterLab.



## Coletando os dados da Receita Federal

5. Com o Jupyter aberto no seu browser, navegue pelas pastas até o local da pasta “open-data-master” e, a partir daí, navegue para “br -> receita.economia.gov.br -> scripts”. Os notebooks dentro desta pasta são responsáveis por coletar e preparar os dados oriundos dos Dados Abertos do CNPJ da Receita Federal.

The screenshot shows a Jupyter Notebook interface. At the top, there are tabs for 'Files', 'Running', and 'Clusters'. On the right, there are buttons for 'Quit' and 'Logout'. Below the tabs, a message says 'Select items to perform actions on them.' To the right of this message are buttons for 'Upload', 'New', and a refresh icon. A file list table follows, showing the following files:

	Name	Last Modified	File size
<input type="checkbox"/>	0 ..	poucos segundos atrás	
<input type="checkbox"/>	01-data-collector.ipynb	6 meses atrás	11.5 kB
<input type="checkbox"/>	02-data-cleaner.ipynb	um dia atrás	5.33 kB
<input type="checkbox"/>	03-data-organizer.ipynb	5 dias atrás	24.1 kB
<input type="checkbox"/>	04-data-standardizer.ipynb	4 dias atrás	35.3 kB

6. Clique no notebook “01-data-collector.ipynb” para abrí-lo em uma nova guia do browser.

The screenshot shows a Jupyter Notebook interface with a single open cell, In [2]. The cell contains Python code for creating directories. The code is as follows:

```
In [2]: def data_directories():
    """
    Funcao de criacao das pastas para armazenamento dos dados, caso nao existam.

    """

    # Directories para armazenamento dos dados coletados
    directories = ['..\data\01-collected',
                  '..\data\02-cleaned',
                  '..\data\03-organized',
                  '..\data\04-standardized']

    # Log: Mensagem de inicio da criacao dos diretorios
    print(str(datetime.now()) + ': Criacao dos diretorios para armazenamento dos dados iniciada.')

    # Para cada diretorio na lista de diretorios
    for directory in directories:

        # Verifica se o diretorio existe
        if not os.path.exists(directory):

            # Cria o diretorio
            os.makedirs(directory)
```

7. Este é o notebook responsável por coletar os arquivos do repositório da Receita Federal. Execute todas as células de código na sequência em que foram criadas (de cima para baixo) clicando no botão “> Run”. A última célula (com a função “main”) será responsável por chamar as demais funções e executar todo o processo. Caso queira verificar os detalhes, todo o código está comentado para que possa compreender cada etapa executada.

Ao finalizar o processo, a última mensagem exibida pelo log da função “main” deverá ser “Fim do download do arquivo...”.

```
# Prepara o dicionario de variaveis (kwargs = keyworded arguments)
kwargs = {}

# Lista de URLs para download dos arquivos zip
kwargs['url_list'] = [
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_01.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_02.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_03.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_04.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_05.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_06.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_07.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_08.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_09.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_10.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_11.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_12.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_13.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_14.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_15.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_16.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_17.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_18.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_19.zip',
    'http://200.152.38.155/CNPJ/DADOS_ABERTOS_CNPJ_20.zip'
]

# Coleta os arquivos
collect_data(**kwargs)

if __name__ == '__main__':
    main()

2019-08-07 18:54:43.772392: Inicio da criacao dos diretorios para armazenamento dos dados.
2019-08-07 18:54:43.773391: Diretorio ..\data\01-collected criado.
2019-08-07 18:54:43.774388: Diretorio ..\data\02-cleaned criado.
2019-08-07 18:54:44.097553: Diretorio ..\data\03-organized criado.
2019-08-07 18:54:44.098520: Diretorio ..\data\04-standardized criado.
2019-08-07 18:54:44.098520: Fim da criacao dos diretorios para armazenamento dos dados.
2019-08-07 18:54:44.098520: Inicio do download do arquivo ..\data\01-collected\0_CNPJ.zip
2019-08-07 18:54:46.079909: URL encontrada, fazendo o download do arquivo: ..\data\01-collected\0_CNPJ.zip
2019-08-07 21:04:43.443068: Descompactando o arquivo ..\data\01-collected\0_CNPJ.zip
2019-08-07 21:05:20.831825: Organizando os arquivos descompactados de: ..\data\01-collected\0_CNPJ.zip
2019-08-07 21:05:20.831825: Fim do download do arquivo ..\data\01-collected\0_CNPJ.zip
```

8. Retorne para a guia de navegação de diretórios do Jupyter e clique no notebook “02-data-cleaner.ipynb” para abri-lo em uma nova guia do browser.

Name	Last Modified	File size
01-data-collector.ipynb	poucos segundos atrás	11.4 kB
02-data-cleaner.ipynb	3 minutos atrás	5.33 kB
03-data-organizer.ipynb	um dia atrás	24.1 kB
04-data-standardizer.ipynb	5 dias atrás	35.3 kB
	4 dias atrás	

9. Este é o notebook responsável por realizar a limpeza dos dados, eliminando caracteres desconhecidos e codificando o arquivo no padrão UTF-8.

```

In [1]: import os
import codecs
from datetime import datetime

In [2]: def encode_to_utf8(**kwargs):
    """
    Funcao para padronizar a codificacao dos arquivos originais em UTF-8.
    Facilita a futura leitura dos arquivos e elimina caracteres desconhecidos.

    Args:
        original_encode(str): Codificacao original do arquivo.
            Ex.: 'cp1252'
        in_data_path(str): Caminho completo para acesso ao arquivo original.
            Ex.: '..\data\01-collected\2012_Funcoes.csv'
        out_data_path(str): Caminho completo para escrita do arquivo codificado
            em UTF-8.
            Ex.: '..\data\02-cleaned\2012_Funcoes.csv'
    """

    # Log: Mensagem de inicio da codificacao
    print(str(datetime.now()) + ': Inicio da codificacao do arquivo ' + kwargs['in_data_path'])

    # Le arquivo na codificacao original
    with codecs.open(kwargs['in_data_path'], 'rb', encoding = kwargs['original_encode'], errors = 'replace') as file:
        # Substitui os caracteres desconhecidos por interrogacao
        lines = file.read().replace('\x1a', '?').replace('\u0085', '?')

    # Escreve o arquivo codificado em UTF-8
    with codecs.open(kwargs['out_data_path'], 'w', encoding = 'utf-8') as file:
        file.write(lines)

    # Log: Mensagem de termino da codificacao
    print(str(datetime.now()) + ': Fim da codificacao do arquivo ' + kwargs['out_data_path'])

```

10. Mais uma vez, execute todas as células de código na sequência em que foram criadas (de cima para baixo) clicando no botão “> Run”. Ao finalizar o processo, a última mensagem exibida pelo log da função “main” deverá ser “Limpeza dos arquivos finalizada”.

```

# Log: Mensagem de inicio do processo
print(str(datetime.now()) + ': Limpeza dos arquivos iniciada.')

# Para cada arquivo na lista de arquivos originais
for file in kwargs['data_files']:

    # Define o caminho completo de acesso e escrita dos arquivos
    kwargs['data_file'] = file
    kwargs['in_data_path'] = os.path.join(kwargs['in_data_dir'], kwargs['data_file'])
    kwargs['out_data_path'] = os.path.join(kwargs['out_data_dir'], kwargs['data_file'])

    # Codifica o arquivo original em UTF-8
    encode_to_utf8(**kwargs)

# Log: Mensagem de finalizacao do processo
print(str(datetime.now()) + ': Limpeza dos arquivos finalizada.')

```

```

In [4]: def main():
    """
    Limpa os dados do Cadastro Nacional de Pessoas Juridicas (CNPJ).
    Origem: Receita Federal
    """

    ### Limpas os arquivos
    clean_files()

    if __name__ == '__main__':
        main()

2021-03-19 12:10:11.327550: Limpeza dos arquivos iniciada.
2021-03-19 12:10:11.327550: Inicio da codificacao do arquivo ..\data\01-collected\K3241.K032001K.CNPJ.D01120.L00020
2021-03-19 12:11:13.209107: Fim da codificacao do arquivo ..\data\02-cleaned\K3241.K032001K.CNPJ.D01120.L00020
2021-03-19 12:11:13.694536: Limpeza dos arquivos finalizada.

```

11. Retorne para a guia de navegação de diretórios do Jupyter e clique no notebook “03-data-organizer.ipynb” para abri-lo em uma nova guia do browser. Repita as mesmas etapas dos itens anteriores. Este notebook é responsável por estruturar o arquivo em formato tabular (caso seja necessário) e dividir o arquivo por tipo de registro, separando CNPJs, Sócios e CNAEs. Ao finalizar o processo, a última mensagem exibida pelo log da função “main” deverá ser “Organizacao dos arquivos finalizada”.

```

jupyter 03-data-organizer Last Checkpoint: 07/09/2020 (autosaved)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 Logout
main()
2021-03-15 19:07:59.616463: Organizacao dos arquivos iniciada.
2021-03-15 19:07:59.618337: Inicio da organizacao dos tipo de registro 1 do arquivo K3241.K032001K.CNPJ.D01120.L00001
2021-03-15 19:09:55.291739: Arquivo K3241.K032001K.CNPJ.D01120.L00001 organizado por CNPJ.
2021-03-15 19:09:55.291739: Inicio da organizacao dos tipo de registro 2 do arquivo K3241.K032001K.CNPJ.D01120.L00001
2021-03-15 19:11:17.254470: Arquivo K3241.K032001K.CNPJ.D01120.L00001 organizado por SOCIO.
2021-03-15 19:11:17.254470: Inicio da organizacao dos tipo de registro 6 do arquivo K3241.K032001K.CNPJ.D01120.L00001
2021-03-15 19:12:31.222795: Arquivo K3241.K032001K.CNPJ.D01120.L00001 organizado por CNAE.
2021-03-15 19:12:31.222795: Inicio da organizacao dos tipo de registro 1 do arquivo K3241.K032001K.CNPJ.D01120.L00002
2021-03-15 19:14:32.257205: Arquivo K3241.K032001K.CNPJ.D01120.L00002 organizado por CNPJ.
2021-03-15 19:14:32.257205: Inicio da organizacao dos tipo de registro 2 do arquivo K3241.K032001K.CNPJ.D01120.L00002
2021-03-15 19:15:52.564532: Arquivo K3241.K032001K.CNPJ.D01120.L00002 organizado por SOCIO.
2021-03-15 19:15:52.564532: Inicio da organizacao dos tipo de registro 6 do arquivo K3241.K032001K.CNPJ.D01120.L00002
2021-03-15 19:17:07.661136: Arquivo K3241.K032001K.CNPJ.D01120.L00002 organizado por CNAE.
2021-03-15 19:17:07.661136: Inicio da organizacao dos tipo de registro 1 do arquivo K3241.K032001K.CNPJ.D01120.L00003
2021-03-15 19:19:03.503829: Arquivo K3241.K032001K.CNPJ.D01120.L00003 organizado por CNPJ.
2021-03-15 19:19:03.503829: Inicio da organizacao dos tipo de registro 2 do arquivo K3241.K032001K.CNPJ.D01120.L00003
2021-03-15 19:20:24.863576: Arquivo K3241.K032001K.CNPJ.D01120.L00003 organizado por SOCIO.
2021-03-15 19:20:24.863576: Inicio da organizacao dos tipo de registro 6 do arquivo K3241.K032001K.CNPJ.D01120.L00003
2021-03-15 19:21:39.547421: Arquivo K3241.K032001K.CNPJ.D01120.L00003 organizado por CNAE.
2021-03-15 19:21:39.547421: Inicio da organizacao dos tipo de registro 1 do arquivo K3241.K032001K.CNPJ.D01120.L00004
2021-03-15 19:23:40.280568: Arquivo K3241.K032001K.CNPJ.D01120.L00004 organizado por CNPJ.
2021-03-15 19:23:40.280568: Inicio da organizacao dos tipo de registro 2 do arquivo K3241.K032001K.CNPJ.D01120.L00004
2021-03-15 19:25:00.362334: Arquivo K3241.K032001K.CNPJ.D01120.L00004 organizado por SOCIO.
2021-03-15 19:25:00.362334: Inicio da organizacao dos tipo de registro 6 do arquivo K3241.K032001K.CNPJ.D01120.L00004
2021-03-15 19:26:14.936452: Arquivo K3241.K032001K.CNPJ.D01120.L00004 organizado por CNAE.
2021-03-15 19:26:14.936452: Inicio da organizacao dos tipo de registro 1 do arquivo K3241.K032001K.CNPJ.D01120.L00005
2021-03-15 19:28:12.223742: Arquivo K3241.K032001K.CNPJ.D01120.L00005 organizado por CNPJ.
2021-03-15 19:28:12.223742: Inicio da organizacao dos tipo de registro 2 do arquivo K3241.K032001K.CNPJ.D01120.L00005
2021-03-15 19:29:32.873885: Arquivo K3241.K032001K.CNPJ.D01120.L00005 organizado por SOCIO.
2021-03-15 19:29:32.873885: Inicio da organizacao dos tipo de registro 6 do arquivo K3241.K032001K.CNPJ.D01120.L00005
2021-03-15 19:30:49.131692: Arquivo K3241.K032001K.CNPJ.D01120.L00005 organizado por CNAE.
2021-03-15 19:30:49.131692: Organizacao dos arquivos finalizada.

```

12. Retorne para a guia de navegação de diretórios do Jupyter e clique no notebook “04-data-standardizer.ipynb” para abri-lo em uma nova guia do browser. Repita as mesmas etapas dos itens anteriores. Este notebook é responsável por padronizar os dados e metadados do conjunto de dados. Ao finalizar o processo, a última mensagem exibida pelo log da função “main” deverá ser “Padronizacao dos arquivos finalizada”.

The screenshot shows a Jupyter Notebook interface. At the top, there's a toolbar with File, Edit, View, Insert, Cell, Kernel, Widgets, Help, and a Python 3 kernel selection. On the right, there are Trusted and Logout buttons. Below the toolbar, a menu bar has icons for file operations like Open, Save, and Print.

In the main area, a code cell [19] contains the following Python script:

```

# Log: Mensagem de finalizacao do processo
print(str(datetime.now()) + ': Padronizacao dos arquivos finalizada.')

```

Below the code cell, the output shows the log message:

```

In [19]: def main():
    """
    Padroniza os dados do Cadastro Nacional de Pessoas Juridicas (CNPJ).
    Origem: Ministerio da Economia - Receita Federal
    """

    ### Padroniza os arquivos
    standardize_files()

if __name__ == '__main__':
    main()

2021-03-16 09:01:14.113240: Padronizacao dos arquivos iniciada.
2021-03-16 09:01:35.018306: Arquivo CNAE_K3241.K032001K.CNPJ.D01120.L00001.csv padronizado.
2021-03-16 09:02:01.559979: Arquivo CNAE_K3241.K032001K.CNPJ.D01120.L00002.csv padronizado.
2021-03-16 09:02:31.235284: Arquivo CNAE_K3241.K032001K.CNPJ.D01120.L00003.csv padronizado.
2021-03-16 09:02:58.909969: Arquivo CNAE_K3241.K032001K.CNPJ.D01120.L00004.csv padronizado.
2021-03-16 09:03:37.204730: Arquivo CNAE_K3241.K032001K.CNPJ.D01120.L00005.csv padronizado.
2021-03-16 09:16:24.962706: Arquivo CNPJ_K3241.K032001K.CNPJ.D01120.L00001.csv padronizado.
2021-03-16 09:29:01.477390: Arquivo CNPJ_K3241.K032001K.CNPJ.D01120.L00002.csv padronizado.
2021-03-16 09:41:49.908996: Arquivo CNPJ_K3241.K032001K.CNPJ.D01120.L00003.csv padronizado.
2021-03-16 09:56:40.919418: Arquivo CNPJ_K3241.K032001K.CNPJ.D01120.L00004.csv padronizado.
2021-03-16 10:09:34.908071: Arquivo CNPJ_K3241.K032001K.CNPJ.D01120.L00005.csv padronizado.
2021-03-16 10:15:27.233176: Arquivo SOCIO_K3241.K032001K.CNPJ.D01120.L00001.csv padronizado.
2021-03-16 10:21:21.818862: Arquivo SOCIO_K3241.K032001K.CNPJ.D01120.L00002.csv padronizado.
2021-03-16 10:27:21.081876: Arquivo SOCIO_K3241.K032001K.CNPJ.D01120.L00003.csv padronizado.
2021-03-16 10:32:36.470243: Arquivo SOCIO_K3241.K032001K.CNPJ.D01120.L00004.csv padronizado.
2021-03-16 10:37:41.271998: Arquivo SOCIO_K3241.K032001K.CNPJ.D01120.L00005.csv padronizado.
2021-03-16 10:37:41.523829: Padronizacao dos arquivos finalizada.

```

13. Ao finalizar a execução dos quatro notebooks, você irá notar que, dentro da pasta “br/tse.jus.br/votacao-candidato-munzona”, temos uma nova pasta chamada “data” com pastas referentes a cada script que executamos.

Computador > Documentos > GitHub > open-data > br > receita.economia.gov.br > data >

Nome	Data de modificação	Tipo	Tamanho
01-collected	19/03/2021 12:09	Pasta de arquivos	
02-cleaned	20/03/2021 09:00	Pasta de arquivos	
03-organized	16/03/2021 14:24	Pasta de arquivos	
04-standardized	17/03/2021 22:40	Pasta de arquivos	

14. Dentro das pastas estão todos os conjuntos de dados que foram preparados durante as execuções dos notebooks. Caso queira, você pode acessar cada pasta e validar como cada notebook preparou os dados em cada etapa.

Nome		Data de modificação	Tipo	Tamanho
CNAE_K3241.K032001K.CNPJ.D01120.L00005		16/03/2021 09:03	Arquivo de Valore...	262.632 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00006		15/03/2021 14:19	Arquivo de Valore...	270.240 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00007		15/03/2021 14:20	Arquivo de Valore...	341.711 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00008		15/03/2021 14:20	Arquivo de Valore...	371.834 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00009		15/03/2021 14:21	Arquivo de Valore...	332.344 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00010		15/03/2021 14:22	Arquivo de Valore...	407.178 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00011		15/03/2021 09:40	Arquivo de Valore...	393.820 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00012		15/03/2021 09:41	Arquivo de Valore...	398.967 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00013		15/03/2021 09:42	Arquivo de Valore...	361.645 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00014		15/03/2021 09:43	Arquivo de Valore...	402.812 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00015		15/03/2021 09:44	Arquivo de Valore...	375.494 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00016		14/03/2021 22:14	Arquivo de Valore...	384.986 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00017		14/03/2021 22:15	Arquivo de Valore...	436.780 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00018		14/03/2021 22:15	Arquivo de Valore...	201.899 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00019		14/03/2021 22:15	Arquivo de Valore...	69.125 KB
CNAE_K3241.K032001K.CNPJ.D01120.L00020		14/03/2021 10:02	Arquivo de Valore...	132.487 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00001		16/03/2021 09:16	Arquivo de Valore...	561.858 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00002		16/03/2021 09:29	Arquivo de Valore...	550.134 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00003		16/03/2021 09:41	Arquivo de Valore...	542.784 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00004		16/03/2021 09:56	Arquivo de Valore...	626.014 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00005		16/03/2021 10:09	Arquivo de Valore...	566.877 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00006		15/03/2021 14:37	Arquivo de Valore...	635.826 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00007		15/03/2021 14:55	Arquivo de Valore...	683.165 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00008		15/03/2021 15:14	Arquivo de Valore...	714.910 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00009		15/03/2021 15:34	Arquivo de Valore...	798.895 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00010		15/03/2021 15:54	Arquivo de Valore...	757.955 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00011		15/03/2021 10:04	Arquivo de Valore...	794.378 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00012		15/03/2021 10:24	Arquivo de Valore...	804.641 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00013		15/03/2021 10:45	Arquivo de Valore...	873.914 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00014		15/03/2021 11:05	Arquivo de Valore...	817.591 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00015		15/03/2021 11:25	Arquivo de Valore...	828.229 KB
CNPJ_K3241.K032001K.CNPJ.D01120.L00016		14/03/2021 22:36	Arquivo de Valore...	819.147 KB

15. Assim, finalizamos a coleta de dados da Receita Federal.

## Coletando os dados do IBGE

1. Com o Jupyter aberto no seu browser, navegue pelas pastas até o local da pasta “open-data-master” e, a partir daí, navegue para “br -> servicodados.ibge.gov.br -> localidades -> scripts”. Os notebooks dentro desta pasta são responsáveis por coletar e preparar os dados oriundos da [API de Serviço de Dados do IBGE](#).

The screenshot shows a Jupyter Notebook interface with a sidebar on the left containing 'Files', 'Running', and 'Clusters'. The main area displays a file list with the following details:

Name	Last Modified	File size
01-data-collector.ipynb	18 dias atrás	8.67 kB
02-data-cleaner.ipynb	18 dias atrás	5.2 kB
03-data-organizer.ipynb	18 dias atrás	9.21 kB
04-data-standardizer.ipynb	18 dias atrás	17.6 kB

2. Assim como na etapa anterior “Coletando dados da Receita Federal”, repita a mesma sequência de execução dos quatro notebooks dentro desta pasta. O resultado final deverá ser um arquivo de municípios do IBGE (Códigos e nomes de municípios, Estados e Regiões) dentro da pasta “br -> servicodados.ibge.gov.br -> localidades -> data -> 04-standardizer”.

The screenshot shows a file explorer interface with the following path: 'cuments > TDC > open-data-master > br > servicodados.ibge.gov.br > localidades > data > 04-standardized'. A table below lists the contents of this folder:

Name	Date modified	Type	Size
20201130_municipios.csv	30/11/2020 07:47	Microsoft Excel Co...	735 KB

3. Repita novamente a sequência de execução dos scripts localizados em ““br -> servicodados.ibge.gov.br -> cnae -> scripts”. O resultado final deverá ser um arquivo da Classificação Nacional de Atividade Econômica dentro da pasta “br -> servicodados.ibge.gov.br -> cnae -> data -> 04-standardizer” .

The screenshot shows a file explorer interface with the following path: 'Computador > Documentos > GitHub > open-data > br > servicodados.ibge.gov.br > cnae > data > 04-standardized'. A table below lists the contents of this folder:

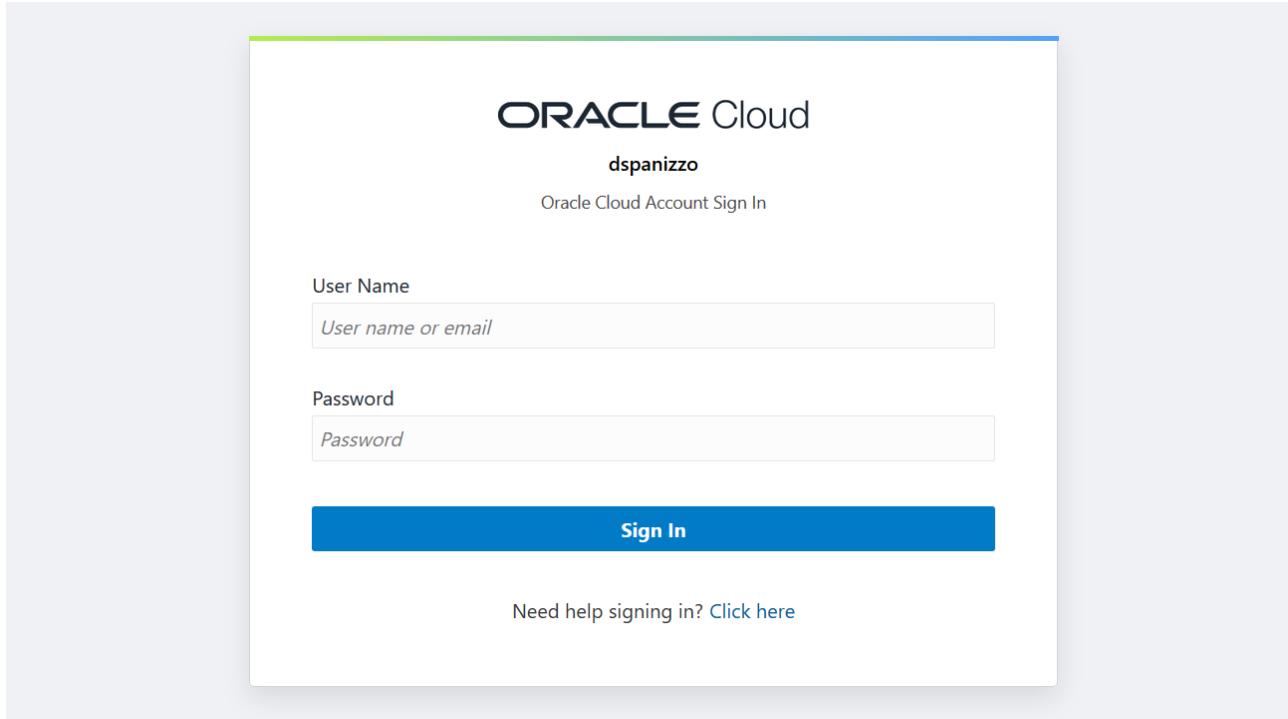
Nome	Data de modificação	Tipo	Tamanho
20210320_cnae	20/03/2021 18:59	Arquivo de Valore...	417 KB

4. Assim, finalizamos a coleta de dados do IBGE.

## LAB.2: ARMAZENAR OS DADOS ABERTOS

### Carregando os dados no Object Storage

1. Com os dados coletados, a próxima etapa será armazenar os dados no banco de dados. Acesse sua conta do Oracle Cloud Infrastructure (OCI) em <https://console.sa-saopaulo-1.oraclecloud.com/>. Caso ainda não tenha uma conta gratuita no OCI, acesse <https://www.oracle.com/br/cloud/free/> para criar uma.



2. No Menu de navegação do OCI, selecione “Armazenamento de Objetos -> Armazenamento do Objetos” entre as opções de “Infraestrutura Básica”.

A screenshot of the Oracle Cloud Infrastructure (OCI) navigation menu. The left sidebar lists various services under 'Infraestrutura Básica', including 'Computação', 'Armazenamento em Blocos', 'Armazenamento de Objetos', 'Armazenamento de Arquivos', 'Rede', 'Oracle Database', 'Visão Geral', 'Autonomous Data Warehouse', 'Autonomous JSON Database', 'Autonomous Transaction Processing', 'Bare Metal, VM e Exadata', 'Exadata Cloud@Customer', 'Banco de Dados', 'MySQL', 'Banco de Dados NoSQL', 'Serviços Relacionados ao Banco de Dados', 'Data Safe', 'Dados e IA', 'Big Data', and 'Assistente Digital'. The 'Armazenamento de Objetos' option is selected and highlighted in blue. The main content area shows a 'Controle' section with several options: 'AUTONOMOUS TRANSACTION PROCESSING' (Criar um banco de dados ATP), 'AUTONOMOUS DATA WAREHOUSE' (Criar um banco de dados ADW), 'GERENCIADOR DE RECURSOS' (Criar uma pilha), 'ARMazenamento de OBJETOS' (Armazenar dados), 'GERENCIADOR DE RECURSOS' (Pré-instalar Kit de Desenvolvimento do Oracle Cloud), and 'PESQUISAR' (Exibir todos os meus recursos). To the right, there are sections for 'Todos os sistemas operacionais', 'Centro de Contas', 'Upgrade', and 'Novidades'.

3. Clique no botão “Criar Bucket”.

Buckets em analytics Compartimento

O Object Storage fornece armazenamento de dados ilimitado, de alto desempenho, durável e seguro. É feito o upload dos dados como objetos armazenados em buckets. [Saiba mais](#)

**Criar Bucket**

Nome	Camada de Armazenamento	Visibilidade	Criado
Nenhum item encontrado.			

Filtros de Tag [adicionar](#) | [limpar](#)

nenhum filtro de tag aplicado

Terminos de Uso e Privacidade | Preferências de Cookies Copyright © 2020. Oracle e/ou suas empresas afiliadas. Todos os direitos reservados.

4. Nas opções de criação do bucket, altere o nome do seu bucket para “opendata” e clique no botão “Criar Bucket”.

**Criar Bucket**

**NOME DO BUCKET**  
opendata

**CAMADA DE ARMAZENAMENTO**

A camada de armazenamento de um bucket só pode ser especificada durante a criação. Depois de definida, não é possível alterar a camada de armazenamento onde reside um bucket.

PADRÃO  
 ARQUIVO COMPACTADO

**EVENTOS DE OBJETO** [?](#)  
 EMITIR EVENTOS DE OBJETO  
 ATIVAR CONTROLE DE VERSÃO DO OBJETO

**CONTROLE DE VERSÃO DO OBJETO** [?](#)  
 ATIVAR CONTROLE DE VERSÃO DO OBJETO

**CRIPTOGRAFIA**

- CRIPTOGRAFAR USANDO CHAVES GERENCIADAS PELA ORACLE  
Deixa todos os assuntos relacionados a criptografia com a Oracle.
- CRIPTOGRAFAR USANDO CHAVES GERENCIADAS PELO CLIENTE  
Exige uma chave válida de um vault ao qual você tem acesso. ([Saiba Mais](#))

**TAGS**

Tag é um sistema de metadados que permite organizar e rastrear recursos na sua tenancy. As tags são compostas de chaves e valores que podem ser anexados a recursos.  
[Saiba mais sobre tags](#)

NAMESPACE DE TAG	CHAVE DE TAG	VALOR
Nenhuma (adicionar uma ta...)		

+ Tag Adicional

Mostrando 0 itens < 1 de 1 >

Terminos de Uso e Privacidade | Preferências de Cookies Copyright © 2020. Oracle e/ou suas empresas afiliadas. Todos os direitos reservados.

5. Clique no bucket “opendata” que acabou de cirar.

The screenshot shows the Oracle Cloud Object Storage Buckets page. On the left, there's a sidebar with 'Armazenamento de Objetos' (Storage) selected. The main area is titled 'Buckets em analytics Compartimento'. It displays a table with one row for the 'opendata' bucket. The table columns are 'Nome' (Name), 'Camada de Armazenamento' (Storage Class), 'Visibilidade' (Visibility), and 'Criado' (Created). The 'opendata' bucket is listed with 'Padrão' as the storage class, 'Privado' (Private) as visibility, and created on 'seg., 30 de nov. de 2020 13:09:01 UTC'. A message at the top says 'Você está usando aproximadamente 0 bytes do limite de 20 GiB de Armazenamento de Objetos e Armazenamento de Arquivos Compactados combinados grátis.' (You are using approximately 0 bytes of the 20 GiB limit for Object Storage and Compressed Files combined free storage.)

6. No seu novo bucket, clique no botão “Fazer Upload”.

The screenshot shows the Oracle Cloud Bucket details page for 'opendata'. At the top, it says 'Detalhes do Bucket'. The bucket icon is a green circle with a white letter 'B'. Below the icon, there are buttons for 'Editar Visibilidade', 'Mover Recurso', 'Recipientar', 'Adicionar Tags', and 'Excluir'. The 'Informações do Bucket' section contains the following details: 'Visibilidade: Privado', 'Namespace: grtzj7pvp01t', 'Camada de Armazenamento: Padrão', 'Contagem aproximada: 0 objetos', 'ETag: eaf8d3c1-2bf2-49cf-a16d-0d8687acb0db', 'OCID: ...26ow5jrq', 'Chave de Criptografia: Chave gerenciada pela Oracle', 'Criado: seg., 30 de nov. de 2020 13:09:01 UTC', 'Compartimento: analytics', 'Tamanho Aproximado: 0 bytes', 'Emitir Eventos de Objeto: Desativado', and 'Controle de Versão do Objeto: Desativado'. The 'Recursos' sidebar on the left has 'Objetos' selected. The 'Objetos' section shows a table with a single row: 'Nome' (Name) is empty, 'Última Modificação' (Last Modified) is 'Nenhum item encontrado.', 'Tamanho' (Size) is empty, and 'Status' is empty. There are buttons for 'Fazer Upload' (Upload) and 'Mais Ações' (More Actions).

7. Na tela de Upload de Objetos, clique no link “selecionar arquivos”.

The screenshot shows the Oracle Cloud Storage interface. On the left, there's a sidebar titled "Arazenamento de Objetos » Detalhes do Bucket" with a large green button labeled "B". Below it are sections for "Informações do Bucket" (Visibility: Private, Namespace: grtzj7ppp01t, etc.) and "Recursos" (Objects, Metrics, Pre-auth Requests, Service Requests). The main area is titled "Fazer Upload de Objetos" and contains fields for "PREFIXO DO NOME DO OBJETO" (OPTIONAL) and a "SELECIONAR ARQUIVOS EM SEU COMPUTADOR" section with a "selecionar arquivos" button. At the bottom, there's a "Mostrar Cabeçalhos e Metadados de Resposta Opcionais" link and a "Fazer Upload" button.

8. Navegue até a pasta onde estão armazenados os arquivos padronizados da Receita Federal (Ex.: C:\Users\dspanizzo\Documents\GitHub\open-data\br\receita.economia.gov.br\data\04-standardized) e selecione todos os arquivos.

The screenshot shows a Windows file explorer window. The path is "Documentos > GitHub > open-data > br > receita.economia.gov.br > data > 04-standardized". The left sidebar shows "Organizar" and "Nova pasta" options. The main area displays a list of files in the "04-cleaned" folder. The columns are "Nome", "Data de modificação", "Tipo", and "Tamanho". The files listed are SOCIO\_K3241.K032001K.CNPJ.D01120.L... files, all modified on 14/03/2021 at various times between 10:23 and 23:36, with sizes ranging from 118.508 KB to 477.300 KB. At the bottom, there are buttons for "Abrir" (Open) and "Cancelar" (Cancel).

9. Clique no botão “Fazer Upload”.

10. Aguarde a finalização do upload de todos os conjuntos de dados.

11. Repita os mesmos passos e faça, também, o upload do arquivo padronizado de municípios do IBGE “YYYYMMDD\_municipios.csv” e da Classificação Nacional de Atividade Econômica “YYYYMMDD\_cnae.csv”.
12. Aguarde a finalização do upload do conjunto de dados.
13. Assim finalizamos o upload dos conjuntos de dados para o Object Storage.

## Armazenando os dados no banco de dados

14. No menu de navegação do OCI, selecione “Autonomous Data Warehouse” entre as opções de “Oracle Database”.

The screenshot shows the OCI console interface. On the left, there's a navigation sidebar with various cloud services like Core Infrastructure, Compute, Block Storage, Object Storage, File Storage, Networking, Oracle Database, and more. Under Oracle Database, 'Autonomous Data Warehouse' is selected. The main content area displays several service cards:

- AUTONOMOUS TRANSACTION PROCESSING**: Create an ATP database (3-5 mins, Always Free Eligible)
- AUTONOMOUS DATA WAREHOUSE**: Create an ADW database (3-5 mins, Always Free Eligible)
- RESOURCE MANAGER**: Create a stack (2-6 mins, Always Free Eligible)
- OBJECT STORAGE**: Store data (2-6 mins, Always Free Eligible)
- RESOURCE MANAGER**: Preinstall Oracle Cloud Development Kit (10-15 mins, Always Free Eligible)
- SEARCH**: View all my resources

Below these cards, there are sections for "Key Concepts and Terminology" and "Get Started with FREE training from Oracle University". The right side of the screen features the "Account Center" with links for User Management, Upgrade, and What's New, along with a news feed and system status information.

15. Selecione o seu Autonomous Data Warehouse (ADW). Obs.: Caso ainda não tenha criado um ADW, siga [estes passos](#) para criar um. Também será necessário ter uma credencial de acesso entre o Object Storage e o ADW configurado, veja [aqui](#) como criar esta credencial.

The screenshot shows the OCI console with the "Autonomous Database" section selected in the navigation bar. The main content area displays a table titled "Autonomous Databases in analytics Compartment".

Create Autonomous Database							
Display Name	State	Dedicated	OCPUs	Storage (TB)	Workload Type	Autonomous Data Guard	Created
Open Data	Always Free	Available	No	1	0.02	Data Warehouse	Disabled

At the bottom of the table, it says "Displaying 1 Autonomous Database < 1 of 1 >".

16. No seu ADW, selecione a aba “Tools” (“Ferramentas”) e clique no botão “Database Tools” (“SQL Developer Web”).

The screenshot shows the Oracle Cloud interface for an Autonomous Database. At the top, there's a navigation bar with 'ORACLE Cloud' and a search bar. Below it, the path 'Overview » Autonomous Database » Autonomous Database Details' is shown. A large green banner on the left says 'ADW' and 'AVAILABLE'. The main content area has a tab bar with 'Autonomous Database Information', 'Tools' (which is selected), and 'Tags'. Under 'Tools', there are four sections: 'Database Actions', 'Oracle Application Express', 'Oracle ML User Administration', and 'SODA Drivers'. Each section contains a brief description and a 'View Details' button.

17. Acesse o SQL Developer Web com o usuário e senha do ADMIN. Definida durante a criação do seu ADW.

The screenshot shows a sign-in form for 'Database Actions'. It has fields for 'Username' (containing 'ADMIN') and 'Password' (containing a series of dots). Below the password field is a blue 'Sign in' button.

18. No ambiente do SQL Developer Web, é possível executar algumas instruções SQL para importar os dados armazenados no Object Storage para o ADW. Faremos este processo utilizando o recurso de tabelas externas.

The screenshot shows the Oracle Database Actions interface with the title bar "ORACLE Database Actions — Planilha SQL". The left sidebar includes "Navegador" and "Planilhas" tabs, and dropdowns for "ADMIN" and "Tabelas". A search bar "Pesquisar..." is also present. The main workspace is titled "[Planilha]\*" and contains a single cell with the number "1". Below the workspace are tabs for "Resultado da Consulta", "Saída do Script", "Saída DBMS", "Plano de Explicação", "Rastreamento automático", "Histórico SQL", and "Carga de Dados". The status bar at the bottom shows "2 △ 0 ⚡ 0 | 14:10:46 - Chamada REST resolvida com sucesso."

19. Copie e cole o script “01\_TB\_EXT\_CNPJ\_CNAE.sql” na área de trabalho do SQL Developer Web. Substitua os devidos parâmetros (como o nome da sua credencial de acesso ao Object Storage e o namespace do seu bucket) e execute o script SQL. Este script irá criar uma tabela externa que acessa os arquivos CSV armazenados no Object Storage.

The screenshot shows the Oracle Database Actions interface with the title bar "ORACLE Database Actions | SQL". The left sidebar lists various tables such as TB\_ESPECTRO\_POLITICO\_BR, TB\_EXT\_ESPECTRO\_POLITICO\_PARTIDARICK, TB\_EXT\_MUNICIPIO\_IBGE, TB\_EXT\_VOTACAO\_CANDIDATO\_MUI, TB\_LOCALIDADE, TB\_MUNICIPIO\_IBGE, TB\_MUNICIPIO\_TSE, and TB\_VOTACAO\_CANDIDATO\_MUNZOL. The main workspace contains a PL/SQL script:

```

1 BEGIN
2   DBMS_CLOUD.CREATE_EXTERNAL_TABLE(
3     table_name => 'TB_EXT_CNPJ_CNAE',
4     credential_name => 'OBJ_STORE_CRED',
5     file_url_list => 'https://objectstorage.sa-saoaulo-1.oraclecloud.com/n/grtzj7pvp01t/b/opendata/o/CNAE_*',
6     format => json object(
7       'dateformat' value 'yyyy-mm-dd',
8       'delimiter' value '\n',
9       'record delimiter' value '\r\n\r\n',
10      'skipheaders' value '1'
11    ),
12    column_list => tp_registro VARCHAR2(250)
13      ,in_forma_envio VARCHAR2(250)
14      ,tp_atualizacao VARCHAR2(250)
15      ,nr_cnpj VARCHAR2(250)
16      ,cd_cidade_secundaria VARCHAR2(250)
17      ,dt_referencia VARCHAR2(250)
18      ,nm_fonte_dados VARCHAR2(250)
19      ,nm_arquivo_dados VARCHAR2(250));
20  END;
21 /

```

The status bar at the bottom shows "PL/SQL procedure successfully completed. Elapsed: 00:00:00.169".

20. Copie e cole o script “02\_TB\_VOTACAO\_CANDIDATO\_MUNZONA.sql” na área de trabalho do SQL Developer Web. Este arquivo irá copiar os dados da tabela externa que aponta para o Object Storage para uma nova tabela no ADW.

The screenshot shows the Oracle Database Actions | SQL Developer interface. In the top navigation bar, it says "ORACLE Database Actions | SQL" and "ADMIN". On the left, there's a sidebar titled "Navegador" with "Planilhas" selected. It lists several tables under "Tabelas": TB\_ESPECTRO\_POLITICO\_BR, TB\_ESPECTRO\_POLITICO\_PARTIDARK, TB\_EXT\_CNPJ\_CNAE, TB\_EXT\_ESPECTRO\_POLITICO\_PARTIC, TB\_EXT\_MUNICIPIO\_IBGE, TB\_EXT\_VOTACAO\_CANDIDATO\_MUI, TB\_LOCALIDADE, TB\_MUNICIPIO\_IBGE, TB\_MUNICIPIO\_TSE, and TB\_VOTACAO\_CANDIDATO\_MUNZOI. Below the sidebar is a code editor window with the following SQL script:

```
1 CREATE TABLE TB_CNPJ_CNAE AS
2 SELECT
3     TP_REGISTRO AS TP_REGISTRO
4     , IN_FORMA_ENVIO AS IN_FORMA_ENVIO
5     , TP_ATUALIZACAO AS TP_ATUALIZACAO
6     , NR_CNPJ AS NR_CNPJ
7     , CD_CNAE_SECUNDARIA AS CD_CNAE_SECUNDARIA
8     , TO_CHAR(DT_REFERENCIA
9         , 'YYYY-MM-DD') AS DT_REFERENCIA
10    , NM_FONTE_DADOS AS NM_FONTE_DADOS
11    , NM_ARQUIVO_DADOS AS NM_ARQUIVO_DADOS
12
13
14 FROM
15     TB_EXT_CNPJ_CNAE
16 ;
```

Below the code editor is a results panel titled "Resultado da Consulta". It shows the message "PL/SQL procedure successfully completed." and "Elapsed: 00:00:00.169". It also shows the message "Table TB\_CNPJ\_CNAE created." and "Elapsed: 00:00:04.013".

21. Repita o mesmo processo com os scripts

- “03\_TB\_EXT\_CNPJ.sql” e “04\_TB\_CNPJ.sql”;
- “05\_TB\_EXT\_CNPJ\_SOCIO.sql” e “06\_TB\_CNPJ\_SOCIO.sql”
- “07\_TB\_EXT\_CNAE.sql” e “08\_TB\_CNAE.sql”

Assim teremos todos os dados armazenados dentro do banco de dados.

## LAB.4: ANALISAR OS DADOS ABERTOS

### Conectando o Analytics com o ADW

- No sua ambiente do seu ADW, clique no botão “Conxão do DB”.

The screenshot shows the Oracle Cloud interface with the navigation bar "ORACLE Cloud" and "Procurar recursos, serviços e documentação". The location is "Visão Geral » Autonomous Database » Detalhes do Autonomous Database". The region is "Brazil East (Sao Paulo)". The main content area displays an "Open Data" card for an "ADW" instance, which is "DISPONÍVEL". The card has tabs for "Informações do Autonomous Database", "Ferramentas", and "Tags". The "Informações Gerais" section contains detailed information about the database, including its name ("opendata"), type ("Data Warehouse"), compartment ("dsanipzo (raiz)/dev/analytics"), OCID ("...3typ7q"), creation date ("qua., 7 de out. de 2020 11:27:16 UTC"), number of OCPUs (1), storage (20 GB), license type (included), version (19c), automatic dimensioning status (disabled), life cycle state (available), instance type (free upgrade to paid), and mode (read/write). The "Infraestrutura" section includes "Autonomous Data Guard" (disabled), network access (allow secure access from anywhere), and scheduled maintenance (next on Saturday, December 5, 2020, from 4:00:00 UTC to 8:00:00 UTC). Other sections like "Fazer Backup" and "Rede" also provide specific details.

- Nas opções de conexão, clique no botão “Download da Wallet”.

The screenshot shows the Oracle Cloud interface with the navigation bar "ORACLE Cloud" and "Procurar recursos, serviços e documentação". The location is "Visão Geral » Autonomous Database » Detalhes do Autonomous Database". The region is "Brazil East (Sao Paulo)". A modal dialog box titled "Fazer Download das Credenciais do Cliente (Wallet)" is open. It instructs the user to select the wallet type and click "Fazer Download da Wallet". The dropdown menu shows "Wallet da Instância". Below the dialog, the "Informações Gerais" section of the ADW instance is visible, showing the same details as the previous screenshot. The "Fazer Backup" and "Rede" sections are also partially visible at the bottom.

3. Informe a senha do ADMIN e clique em “Fazer Download” para baixar o arquivo ZIP.

The screenshot shows the Oracle Cloud interface for an Autonomous Database. A modal window titled "Download da Wallet" is open, prompting for a password. Below the modal, the database details are visible, including its name (ADW), status (DISPONÍVEL), and various configuration parameters like OCPUs, storage, and license type. To the right of the modal, there are sections for "Fazer Backup", "Rede", and "Manutenção". At the bottom left, a download link for "Wallet\_opendata.zip" is shown, and at the bottom right, there's a "Exibir todos" button.

4. Em posse das credenciais. Abra o seu Oracle Analytics Desktop

The screenshot shows the Oracle Analytics Desktop search interface. On the left, a search bar contains the query "Analytics". Below it, a list of search results includes "Analytics - See web results", "analytics google", "analytics twitter", "analytics youtube", "analytics vidhya", "analytics uber eats", and "analytics brasil". On the right, a detailed view of the "Oracle Analytics Desktop 12c" app is shown, featuring its icon, name, status as an "App", an "Open" button, and a "Recent" section listing several Excel files: "IBGE\_Localidades.xlsx", "200730\_base\_receitas\_municipais.xlsx", "Transparencia\_ComprasNET\_Contratos\_Oracle\_v02.xlsx", and "ComprasNET\_UASGs.xlsx".

5. Na Home do Oracle Analytics, clique no botão “Create” e, logo em seguida, na opção “Connection”.

The screenshot shows the Oracle Analytics Home interface. At the top right, there is a 'Create' button with a dropdown menu. The menu is open, showing options: Project, Data Set, Data Flow, Sequence, and Connection. Below the menu, there is another 'Create' button with a script icon. The main area displays a grid of project cards, including 'OD - Receita Municipal', 'IBGE - Localidades', 'OD - Receitas Municipais...', 'CONTRATOS ORACLE', 'COMPRASNET - UASG', 'CONTRATOS ORACLE', 'HealthInsurance\_PT-BR...', 'HealthCare', and 'Vanilla Advanced Analytics'. Below the grid, there are sections for 'Projects' and 'Data Sets', each containing several project cards. A 'View all' link is located at the top right of the 'Projects' section.

6. Na tela de criar conexão, selecione a opção “Autonomous Data Warehouse”.

The screenshot shows the 'Create Connection' dialog box overlying the Oracle Analytics Home interface. The dialog has a title 'Select Connection Type' and a search bar. It displays a grid of connection types, each with an icon and a label: Oracle Applications, Oracle Autonomous Data Warehouse, Oracle Autonomous Transaction Processing, Oracle Database, Oracle Essbase, Oracle NetSuite, Oracle Service Cloud, Oracle Talent Acquisition, Actian Ingres, Actian Matrix, Actian Vector, Amazon Aurora, Amazon EMR, Amazon Redshift, and Apache Drill. At the bottom right of the dialog, there are 'Cancel' and 'Next Step' buttons. The background of the dialog shows the same project cards and sections as the Home page.

7. Na opção “Client Credentials”, informe o local onde armazenou o arquivo ZIP da Wallet do ADW. Em seguida, informe um nome para a conexão, usuário e senha do ADMIN. Clique em “Save”.

The screenshot shows the Oracle Analytics web application. The top navigation bar includes the Oracle Analytics logo, a menu icon, and a search bar labeled "Search Everything". Below the navigation is a horizontal menu with tabs: "Projects", "Data", "Recent Data Sets", and "Machine".

The main content area is titled "Create Connection". It features a "Cloud with Database" icon and the text "Oracle Autonomous Transaction Processing". The connection configuration form contains the following fields:

- \* Connection Name: dspanizzo - open data
- Description: Dados abertos
- \* Service Name: opendata\_high
- \* Client Credentials: cwallet.sso (with a "Select..." button)
- \* Username: ADMIN
- \* Password: [REDACTED]

At the bottom right of the dialog are "Cancel" and "Save" buttons.

On the left side of the main content area, there is a sidebar titled "Projects" with two items: "OD - Receita Municipal" and "CONTRATOS ORACLE".

On the right side, there are several preview cards for different analytical models:

- OD - Receitas Municipais...
- OD - Receitas Municipais...
- HealthCare
- Vanilla Advanced Analytics
- View all
- Vanilla Advanced Analytics
- Vanilla Advanced Analytics
- CONFEA - Big Data

Each card displays a small thumbnail image of its respective dashboard.

8. Com a conexão criada, selecione no menu a opção “Dados”.

**Analytics**

**ORACLE Analytics**

**Criar** **Opções**

**Home**

**Catálogo**

**Dados**

**Aprendizado por Máquina**

**Jobs**

**Console**

**Academia**

**Pesquisar Tudo**

**Projetos e Relatórios** | **Dados** | **Conjuntos de Dados Recentes** | **Projetos Favoritos** | **Aprendizado por Máquina**

**Projetos**

**Exibir todos os**

**Analytics Live - Brazil Pol...** **Workshop** **TSE - POLITICAL SPEC...** **Ceará Global** **IBGE - Mapas da API** **CGU - HISTÓRICO** **CGU - CONTRATOS**

**Painéis de Controle e Relatórios**

**Dashboard - Monit...** **\_portal - Home**

**Conjuntos de Dados**

**Exibir todos os**

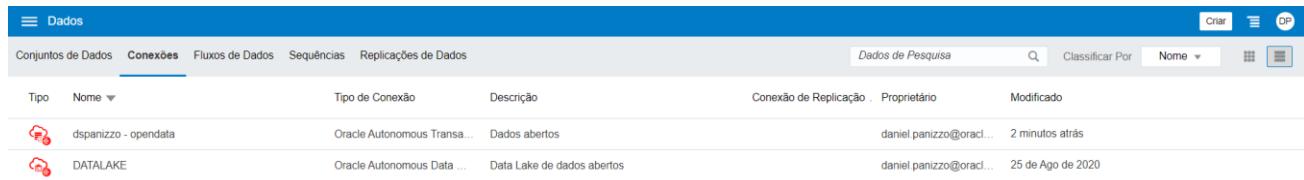
**Ceará Global - Vis...** **Ceará Global - Exp...** **CNAE\_COD** **Ceará Global - Po...** **MD\_MIDIAS\_CEARA** **Ceará Global - Eta...** **Ceará Global - Eta...**  
**Ceará Global - Eta...** **Ceará Global - Eta...** **Ceará Global - Eta...** **Ceará Global - Eta...** **Ceará Global - Vis...** **Ceará Global - CO...** **Ceará Global - Po...**  
**Espectro Político P...** **TB\_PERFIL\_ELEI...** **SH - CUSTOMERS** **CGU - CONTRATOS** **TSE - ESPECTRO...** **BR - LOCALIDADES** **IBGE - LOCALIDA...**

**Fluxos de Dados**

**Ceará Global - Pre...** **Ceará Global - 02 -...** **Ceará Global - 06 -...** **Ceará Global - 04 -...** **Ceará Global - 03 -...** **Ceará Global - 09 -...** **Ceará Global - 05 -...**  
**» Ceará Global - 01 -...** **» Ceará Global - 08 -...** **» KAGGLE - TITANI...** **» KAGGLE - TITANI...** **» TSE-IBGE - PERFI...** **» TSE - SEQUENCI...** **» TSE - MULTIDIME...**  
**» TSE-IBGE - VOTA...** **» SH - SALES PREP...** **» SH - DATA PREPA...**

**javascript:void(0)**

9. Na área de “Dados”, selecione a aba “Conexões”.

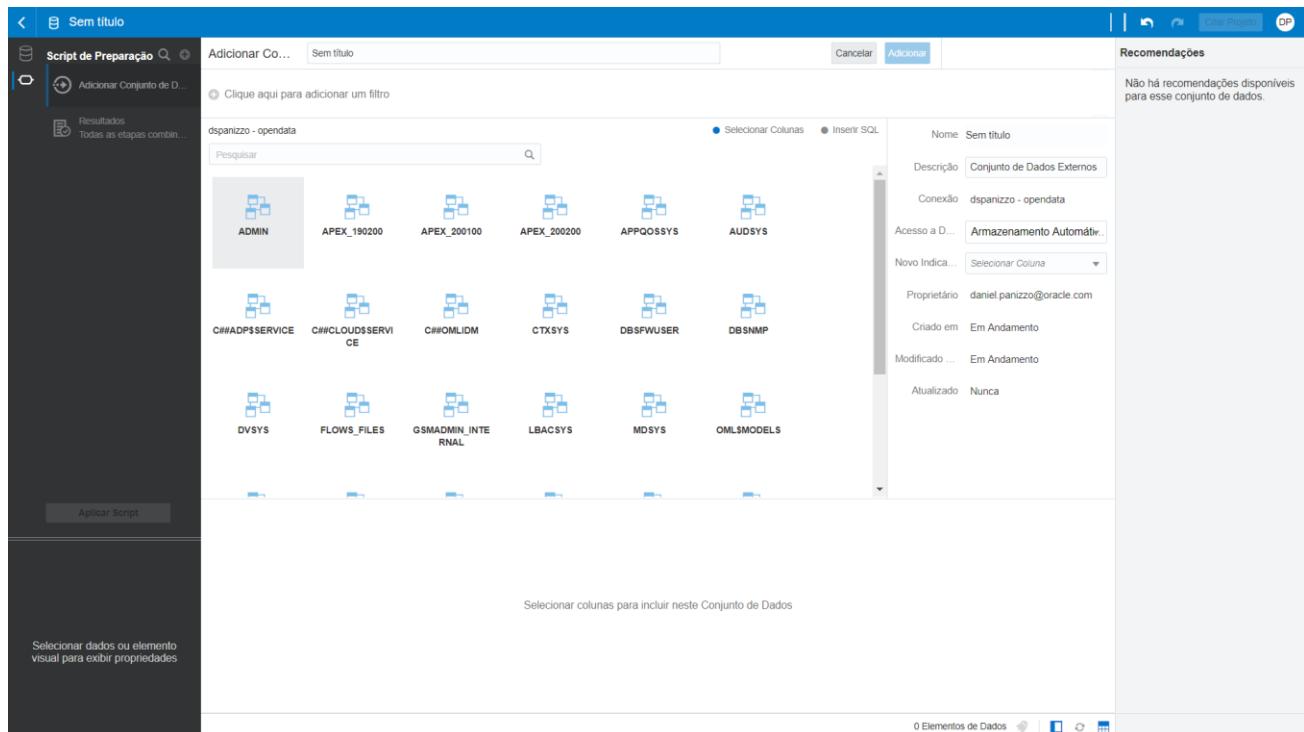


A screenshot of the Oracle Autonomous Data interface showing the 'Conexões' (Connections) tab selected. The table lists two connections:

Tipo	Nome	Tipo de Conexão	Descrição	Conexão de Replicação	Proprietário	Modificado
🔗	dspanizzo - opendata	Oracle Autonomous Transa...	Dados abertos		daniel.panizzo@oracle...	2 minutos atrás
☁️	DATALAKE	Oracle Autonomous Data ...	Data Lake de dados abertos		daniel.panizzo@oracle...	25 de Ago de 2020

## Importando os metadados

10. A conexão que acabamos de criar deve aparecer como uma das opções. Clique na nossa conexão “opendata” e a tela importação de metadados deve aparecer. Selecione o esquema “ADMIN”.



A screenshot of the Oracle Autonomous Data interface showing the 'Script de Preparação' (Preparation Script) screen. The 'dspanizzo - opendata' connection is selected. In the center, a grid of database schemas is displayed, with 'ADMIN' highlighted. On the right, detailed properties for the selected schema are shown:

- Nome: Sem título
- Descrição: Conjunto de Dados Externos
- Conexão: dspanizzo - opendata
- Acesso a D...: Armazenamento Automático
- Novo Indica...: Selecionar Coluna
- Proprietário: daniel.panizzo@oracle.com
- Criado em: Em Andamento
- Modificado ...: Em Andamento
- Atualizado: Nunca

The bottom left shows a note: "Selecionar dados ou elemento visual para exibir propriedades".

11. As tabelas que criamos no banco de dados devem aparecer como opções. Selecione a tabela “TB\_CNPJ”.

The screenshot shows the Oracle Data Integrator interface. On the left, there's a sidebar with buttons for 'Pesquisar', 'Adicionar Conjunto de Dados...', and 'Resultados'. The main area has a title bar 'Sem título' with tabs for 'Adicionar Co...' and 'Sem título'. A message says 'Clique aqui para adicionar um filtro'. Below it, a dropdown shows 'dspanizzo - opendata > ADMIN' with 'TB\_C' selected. To the right, there's a 'Seleções' section with 'Selecionar Colunas' selected. On the far right, dataset properties are listed: Nome: Sem título, Descrição: Conjunto de Dados Externos, Conexão: dspanizzo - opendata, Acesso a Dados: Armazenamento Automático no Cache, Novo Indicador de Dados: Selecionar Coluna, Proprietário: daniel.panizzo@oracle.com, Criado em: Em Andamento, Modificado em: Em Andamento, and Atualizado: Nunca. At the bottom, a note says 'Selecionar colunas para incluir neste Conjunto de Dados'.

12. As opções de colunas para serem importadas serão exibidas. Clique no botão “Adicionar Tudo”.

This screenshot continues from the previous one. The title bar now says 'TB\_CNPJ'. The 'Seleções' section shows '48(48)' items selected. The properties on the right remain the same. At the bottom, there's a large 'Obter Dados de Visualização' button with a circular arrow icon. The note 'Selecionar colunas para incluir neste Conjunto de Dados' is still present.

13. Do lado direito, nas propriedades da tabela, altere a opção de “Acesso a Dados” para “Ativo”.

The screenshot shows the Oracle Data Integrator interface for managing tables. On the left, there's a sidebar with options like 'Adicionar Conjunto de Dados...', 'Resultados', and 'Aplicar Script'. The main area shows a tree view of selected columns under 'TB\_CNPJ' and 'ADMIN'. On the right, the properties panel is open for 'TB\_CNPJ'. Key settings include:

- Nome:** TB\_CNPJ
- Descrição:** Conjunto de Dados Externos
- Coneção:** dspanizzo - opendata
- Acesso a Dados:** Ativo (checkbox checked)
- Novo Indicador de Dados:** Ativo (checkbox checked)
- Proprietário:** dspanizzo - opendata
- Criado em:** Em Andamento
- Modificado em:** Em Andamento
- Atualizado:** Nunca

14. Na parte superior, clique no botão “Adicionar”.

This screenshot shows the Oracle Data Integrator interface for the TB\_CNPJ\_TDC table. The layout is similar to the previous one, with a sidebar, a central tree view of columns, and a properties panel on the right. The properties panel for 'TB\_CNPJ\_TDC' includes:

- Nome:** TB\_CNPJ\_TDC
- Descrição:** Conjunto de Dados Externos
- Coneção:** dspanizzo - opendata
- Acesso a Dados:** Ativo (checkbox checked)
- Novo Indicador de Dados:** Selecionar Coluna (dropdown menu open)
- Proprietário:** daniel.panizzo@oracle.com
- Criado em:** Em Andamento
- Modificado em:** Em Andamento
- Atualizado:** Nunca

## Ajustando os metadados

15. A ferramenta irá para a etapa de preparação dos dados.

The screenshot shows the Oracle Data Visualization interface in preparation mode. The main area displays a large dataset with columns: NR\_ANO\_ELECAO, CD\_MUNICIPIO, DS\_ESPECTRO, QT\_VOTOS, and VL\_RANK\_ES... . The data grid contains numerous rows of election results. To the left, there's a sidebar with a script editor and a recommendation panel titled 'Recomendações (8)' containing various data extraction options for the 'NR\_ANO\_ELECAO' column. At the bottom, there are buttons for 'Aplicar Script' and 'Cancelar'.

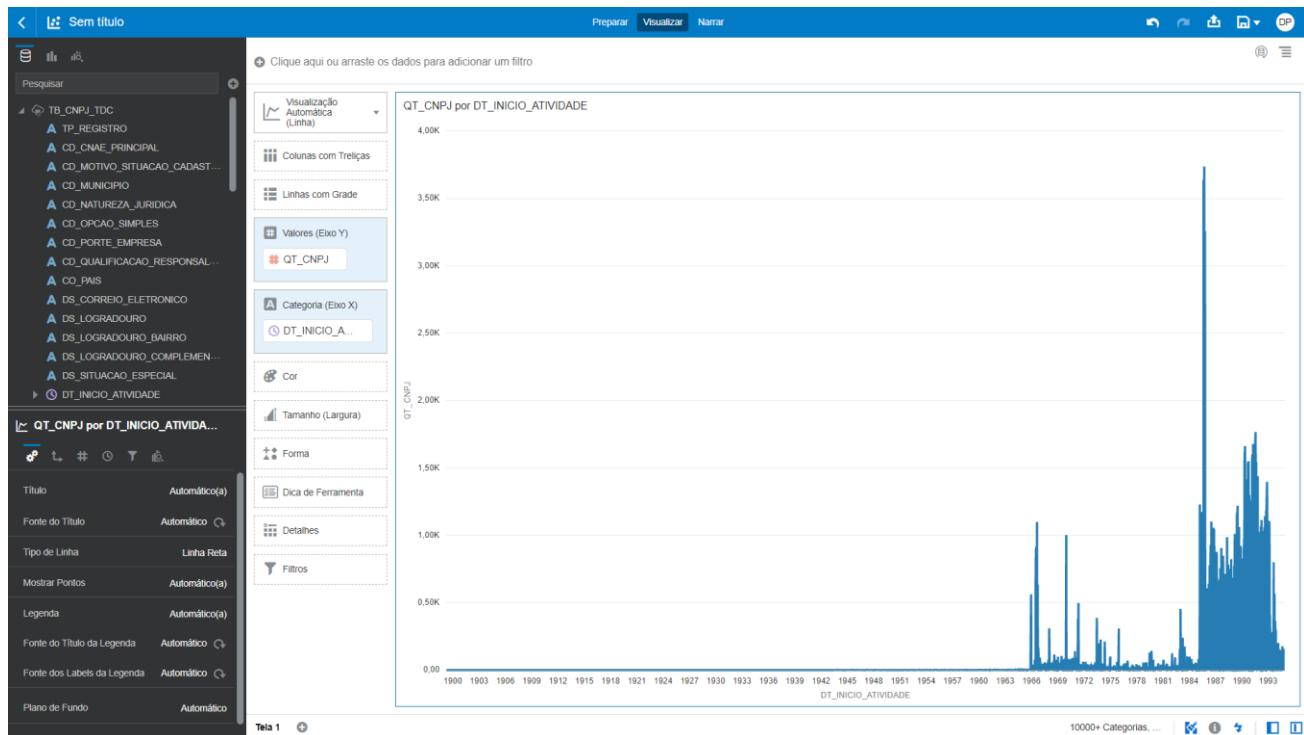
16. No navegador da ferramenta, clique no botão “Criar Projeto”. A etapa de visualização de dados irá iniciar.

The screenshot shows the Oracle Data Visualization project creation screen. On the left, a sidebar lists datasets: 'TB\_CNPJ\_TDC' and 'TP\_REGISTRO'. Under 'TP\_REGISTRO', there are sections for 'Nome' (set to 'TP\_REGISTRO'), 'Tratar Como' (set to 'Atributo'), 'Tipo de Dados' (set to 'Texto'), and 'Agregação' (set to 'Nenhum(a)'). The main workspace is titled 'Sem título' and contains a message 'Clique aqui ou arraste os dados para adicionar um filtro'. Below this, there are sections for 'Selecionar Visualização para Ver os Detalhes' and 'Soltar Visualizações ou Dados Aqui'. At the bottom, there are navigation icons for 'Tela 1', 'Home', 'Search', and 'Help'.

## Visualizando os dados

17. Na área de “Dados” (menu esquerdo), segurando CTRL, selecione as colunas DT\_INICIO\_ATIVIDADE e QT\_CNPJ. Arraste as colunas para a área de visualização de dados.

18. Agora já podemos explorar os detalhes da criação de empresas ao longo do tempo. Agora a criatividade da exploração é com você!



19. Assim finalizamos nossa etapa de análise de dados.