DIGITAL INNOVATION ONE

CERTIFICADO DE CONCLUSÃO • DIGITAL INNOVATION ONE • 48560386

Certificamos que

Helio S do Nascimento

em 29 de Agosto de 2021, concluiu o curso

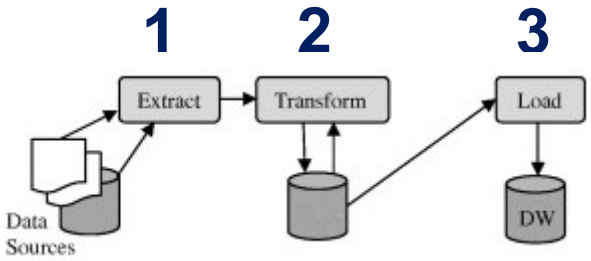
Fundamentos de ETL com Python

com carga horária de 5 horas.

Introdução ao ETL

Fundamentos de ETL com Python

ETL – Definição

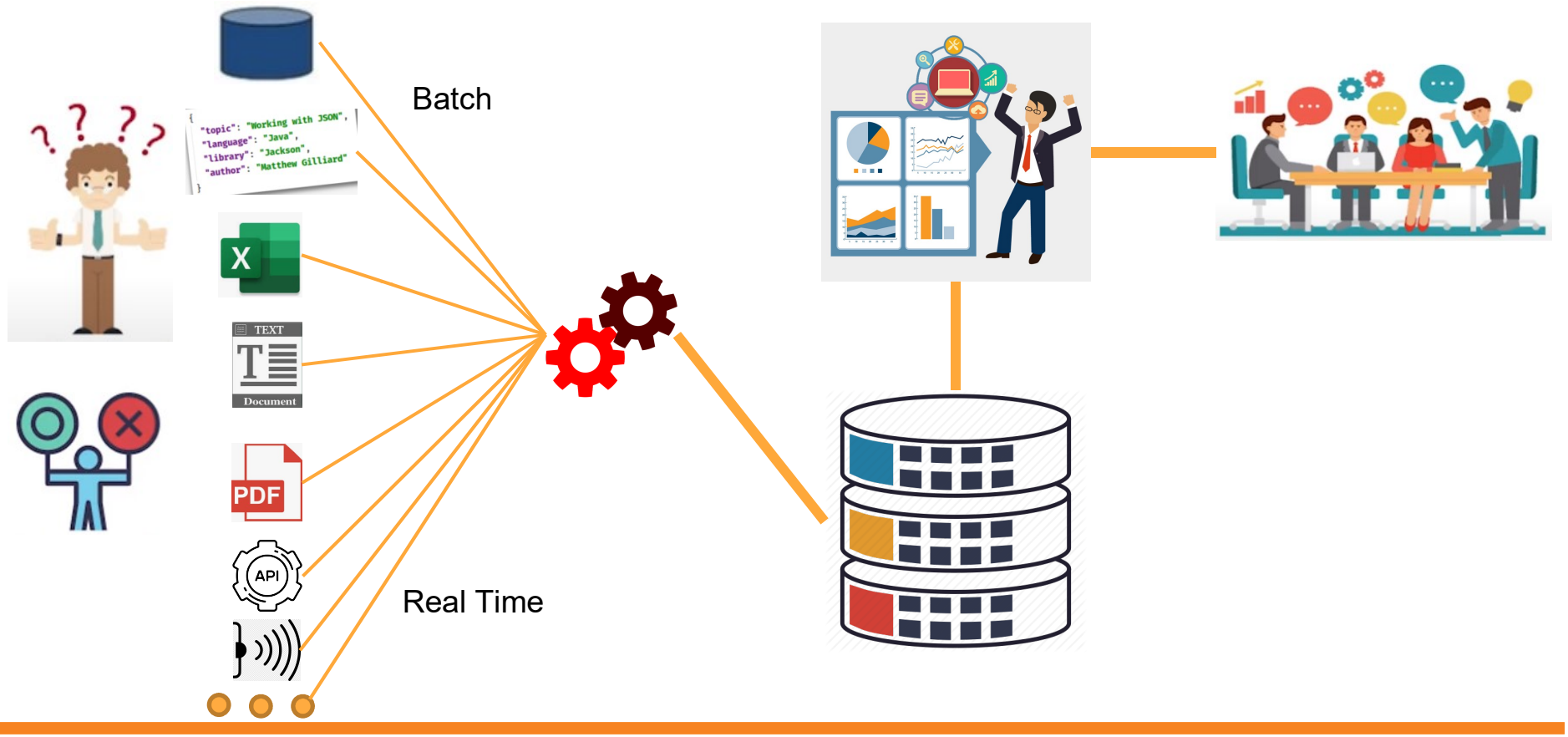


1) **Extract**: os dados são extraídos de diferentes fontes de dados.

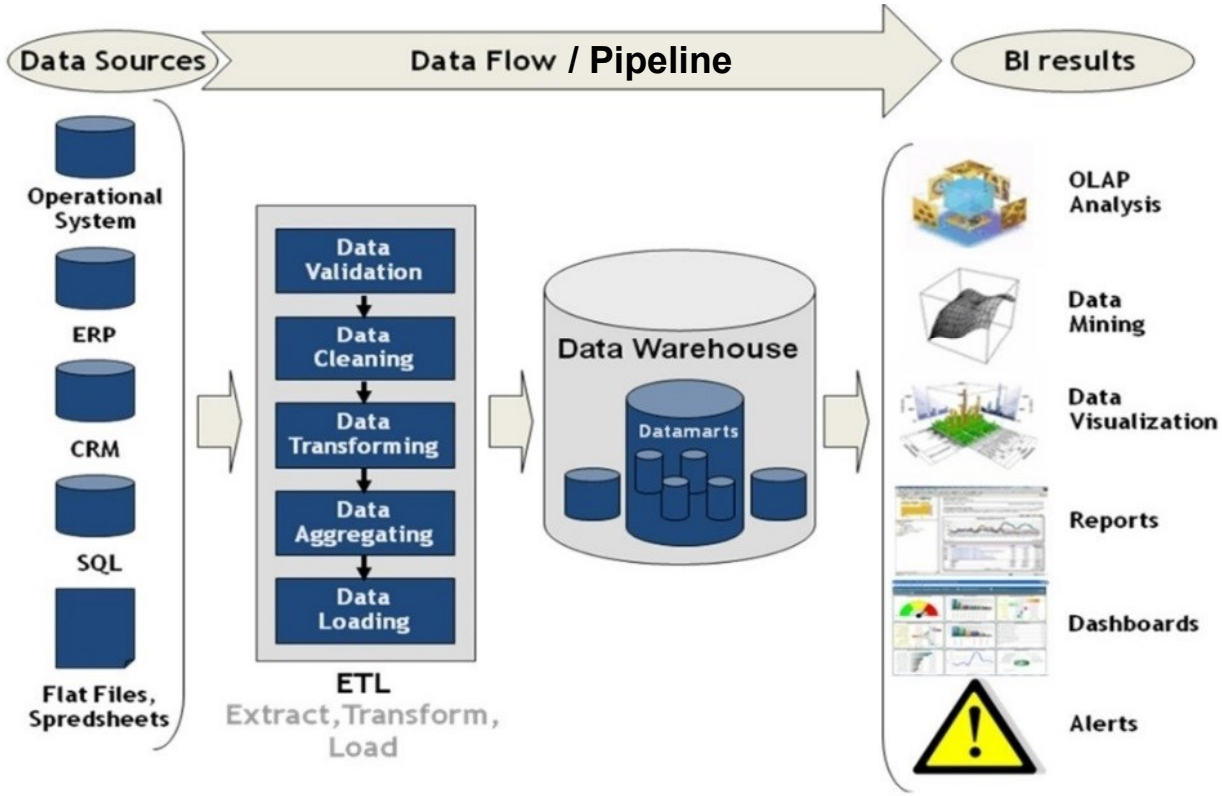
2) **Transform**: Propagados para a área de preparação de dados, onde são transformados e limpos.

3) **Load**: Carregados no data warehouse.

ETL - Por que precisamos?



ETL – Visão Geral



Fonte: https://br.pinterest.com/pin/821836631985166926/

Ferramentas / Pacotes
Para Python

Apache
Airflow

bonobo

pandas

bubbles

petl

Para saber mais

- http://airflow.apache.org/
- https://luigi.readthedocs.io/en/stable
- https://www.bonobo-project.org/
- http://bubbles.databrewery.org/
- https://petl.readthedocs.io/en/stable/
- **https://pandas.pydata.org/**

Projeto ETL – Ambiente

Fundamentos de ETL com Python

Ambiente

ANACONDA®

Instalação pacote Jupyter Lab

Crtl + C => para o processo

Dados usados no curso

Caso o download não inicie, favor clicar no link abaixo:

http://sistema.cenipa.aer.mil.br/cenipa/media/opendata/ocorrencia_2010_2020.csv

*Favor fechar esta aba após a conclusão do download! :)

projeto.ipynb

Código

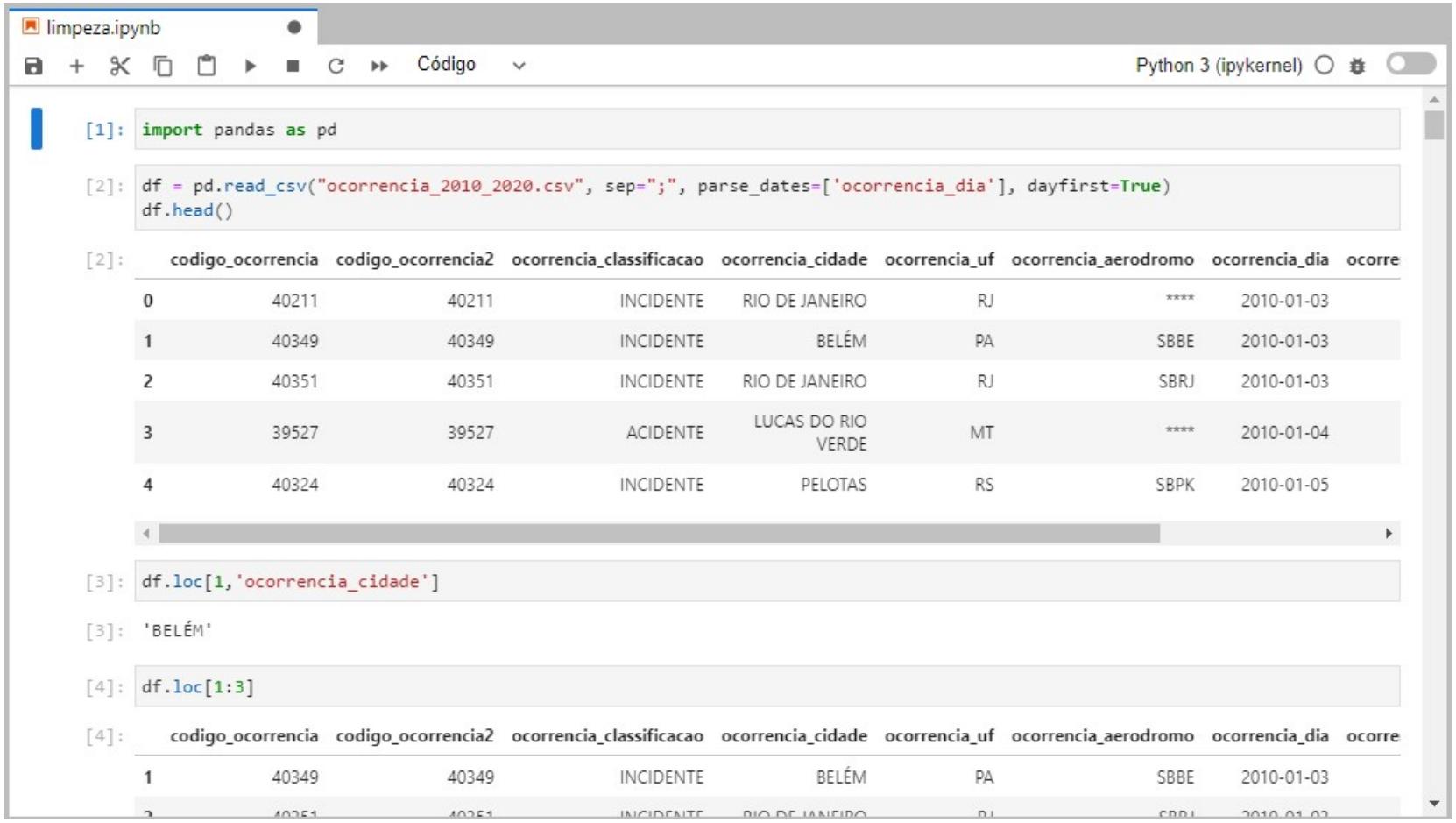
Python 3 (ipykernel)

[1]:
```
import pandas as pd
import pandera as pa
```

[2]:
```
df = pd.read_csv("ocorrencia_2010_2020.csv", sep=";", parse_dates=['ocorrencia_dia'], dayfirst=True)
df.head(10)
```

[2]:

	codigo_ocorrencia	codigo_ocorrencia2	ocorrencia_classificacao	ocorrencia_cidade	ocorrencia_uf	ocorrencia_aerodromo	ocorrencia_dia	ocorre
0	40211	40211	INCIDENTE	RIO DE JANEIRO	RJ	****	2010-01-03	
1	40349	40349	INCIDENTE	BELÉM	PA	SBBE	2010-01-03	
2	40351	40351	INCIDENTE	RIO DE JANEIRO	RJ	SBRJ	2010-01-03	
3	39527	39527	ACIDENTE	LUCAS DO RIO VERDE	MT	****	2010-01-04	
4	40324	40324	INCIDENTE	PELOTAS	RS	SBPK	2010-01-05	
5	39807	39807	INCIDENTE	SALVADOR	BA	****	2010-01-06	
6	40215	40215	INCIDENTE	COARI	AM	SBUY	2010-01-07	
7	39707	39707	INCIDENTE GRAVE	CANUTAMA	AM	****	2010-01-09	
8	39156	39156	INCIDENTE GRAVE	CASCAVEL	PR	SBCA	2010-01-10	
9	39711	39711	INCIDENTE GRAVE	PARÁ DE MINAS	MG	****	2010-01-10	



	codigo_ocorrencia	codigo_ocorrencia2	ocorrencia_classificacao	ocorrencia_cidade	ocorrencia_uf	ocorrencia_aerodromo	ocorrencia_dia	ocorre
0	40211	40211	INCIDENTE	RIO DE JANEIRO	RJ	****	2010-01-03	
1	40349	40349	INCIDENTE	BELÉM	PA	SBBE	2010-01-03	
2	40351	40351	INCIDENTE	RIO DE JANEIRO	RJ	SBRJ	2010-01-03	
3	39527	39527	ACIDENTE	LUCAS DO RIO VERDE	MT	****	2010-01-04	
4	40324	40324	INCIDENTE	PELOTAS	RS	SBPK	2010-01-05	

	codigo_ocorrencia	codigo_ocorrencia2	ocorrencia_classificacao	ocorrencia_cidade	ocorrencia_uf	ocorrencia_aerodromo	ocorrencia_dia	ocorre
1	40349	40349	INCIDENTE	BELÉM	PA	SBBE	2010-01-03	

transformacao.ipynb

Código

Python 3 (ipykernel)

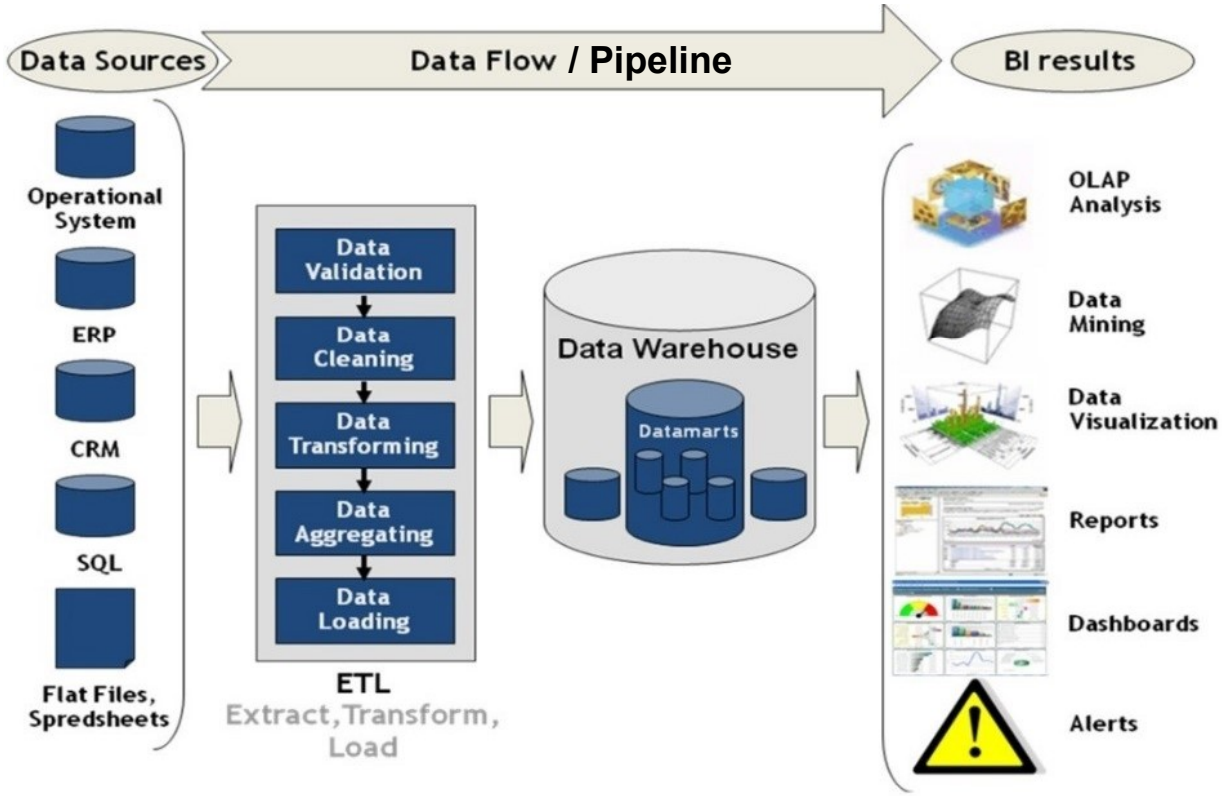
[1]:
```
import pandas as pd
import pandera as pa
```

[2]:
```
valores_ausentes = ['**','###!','####','****','*****','NULL']
df = pd.read_csv("ocorrencia_2010_2020.csv", sep=";", parse_dates=['ocorrencia_dia'], dayfirst=True, na_values=valores_aus
df.head(10)
```

[2]:

	codigo_ocorrencia	codigo_ocorrencia2	ocorrencia_classificacao	ocorrencia_cidade	ocorrencia_uf	ocorrencia_aerodromo	ocorrencia_dia	ocorre
0	40211	40211	INCIDENTE	RIO DE JANEIRO	RJ	NaN	2010-01-03	
1	40349	40349	INCIDENTE	BELÉM	PA	SBBE	2010-01-03	
2	40351	40351	INCIDENTE	RIO DE JANEIRO	RJ	SBRJ	2010-01-03	
3	39527	39527	ACIDENTE	LUCAS DO RIO VERDE	MT	NaN	2010-01-04	
4	40324	40324	INCIDENTE	PELOTAS	RS	SBPK	2010-01-05	
5	39807	39807	INCIDENTE	SALVADOR	BA	NaN	2010-01-06	
6	40215	40215	INCIDENTE	COARI	AM	SBUY	2010-01-07	
7	39707	39707	INCIDENTE GRAVE	CANUTAMA	AM	NaN	2010-01-09	
8	39156	39156	INCIDENTE GRAVE	CASCAVEL	PR	SBCA	2010-01-10	
9	39711	39711	INCIDENTE GRAVE	PARÁ DE MINAS	MG	NaN	2010-01-10	

ETL – Considerações finais



Fonte: https://br.pinterest.com/pin/821836631985166926/

Para saber mais

- **Leitura de arquivo csv (read_csv)**

  https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html

- **Validação: Pandera**

  https://pandera.readthedocs.io/en/stable/

- **Limpeza: Valores ausentes**

  https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html

- **Transformação: Variações de filtros (tempo execução)**

  https://medium.com/data-hackers/a-maneira-eficiente-de-filtrar-um-data-frame-pandas-4158a4e37c10

Fim