Fundamentos ETL com Python

Fundamentos de ETL

Apresentação inicial e objetivos do curso

* Contexto do processo ETL: extração, transformação e carregamento de dados.
* <https://github.com/ftiosso?dio-curso-etl>

Introdução ao ETL - Definição

* Extração: os dados são extraídos de diferentes fontes de dados.
* Transformação: propagados para a área de transformação de dados, onde são transformados e limpos.
* Load: carregados no data warehouse

Por que precisamos?

* Necessidade da obtenção de informações para as tomadas de decisão
* Batch (Pacotes de dados)
* Real time
* Processamento e carregamento de dados

Visão geral sobre ETL



Ferramentas e Pacotes

* [Apache Airflow (Airbnb)](https://airflow.apache.org/)
* [Luigi (Spotify)](http://luigi.readthedocs.io/en/stable)
* [bonobo](https://www.bonobo-project.org/)
* [bubbles](http://bubbles.databrewery.org/)
* [petl (não concebido para pipelines que usam muita memória)](https://petl.readthedocs.io/)
* [pandas](https://pandas.pydata.org/)

Preparação do projeto ETL

Objetivos da aula

* Preparação do ambiente
* Definição do projeto

Instalação e configuração do ambiente - Jupyter no navegador

* <https://jupyter.org/>

Instalação e configuração do ambiente - Jupyter na máquina

* pip install jupyter-lab

Conheça a ferramenta Anaconda

* <https://www.anaconda.com/>

Origem dos dados acessíveis para o projeto

* <https://www2.fab.mil.br/cenipa/>

Visualização dos dados da planilha

Desenvolvimento do projeto ETL - Extração e validação

Objetivos da aula

* Desenvolvimento: Extração, Validação, Limpeza, Transformação, Considerações finais (Carregamento).

Como a extração de dados será realizada

Realizando extração de dados - Parte 1

Realizando extração de dados - Parte 2

Como validar a extração de dados - Parte 1

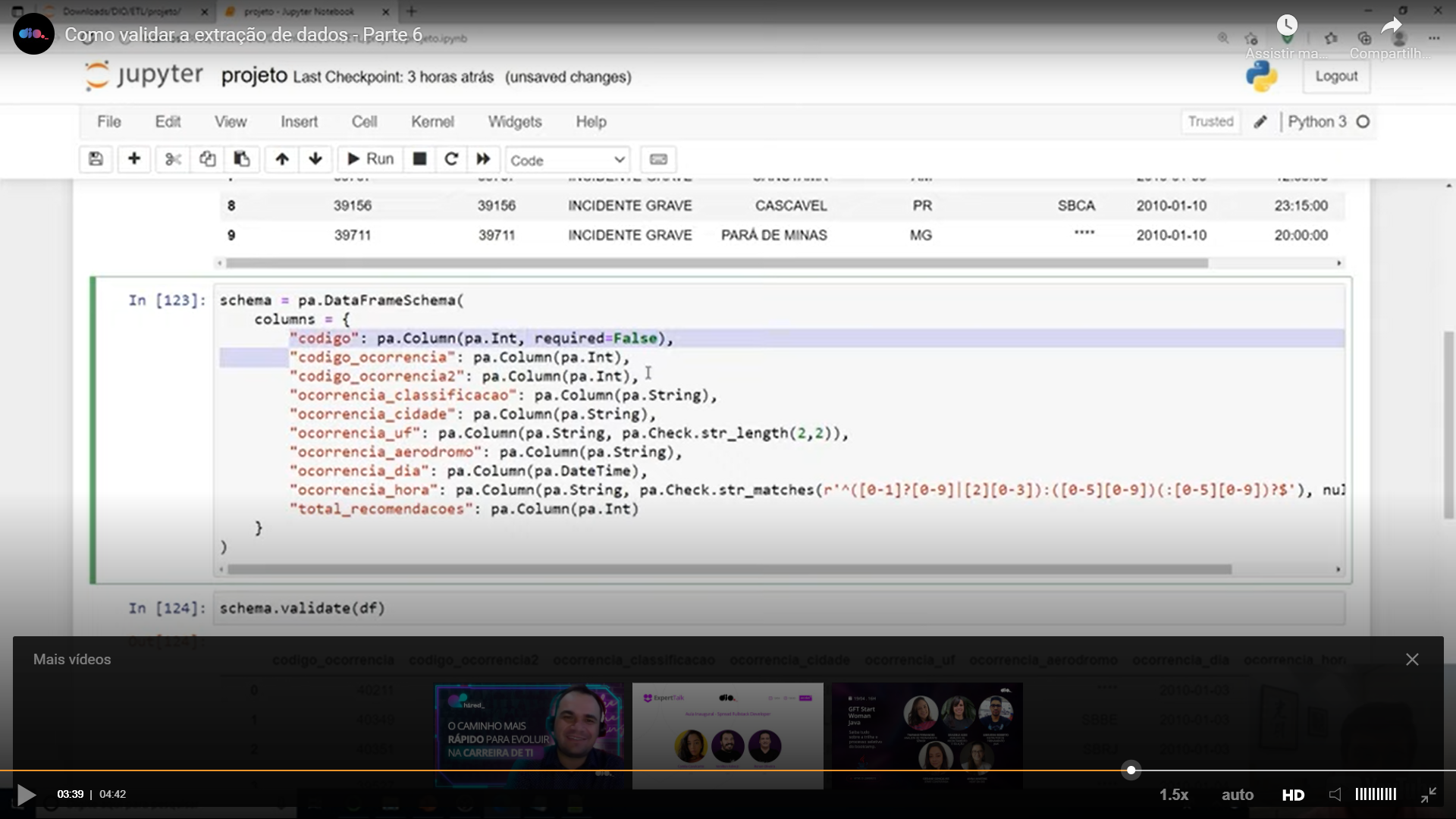
Como validar a extração de dados - Parte 2

Como validar a extração de dados - Parte 3

Como validar a extração de dados - Parte 4

Como validar a extração de dados - Parte 5

Como validar a extração de dados - Parte 6



Desenvolvimento do Projeto ETL - Limpeza

Introdução ao tema de Limpeza

Como a limpeza de dados será realizada

Realizando a limpeza de dados - Parte 1

Realizando a limpeza de dados - Parte 2

Realizando a limpeza de dados - Parte 3

Realizando a limpeza de dados - Parte 4

Desenvolvimento do Projeto ETL - Transformação

Introdução ao tema de Transformação

Realizando a transformação de dados - Parte 1

Realizando a transformação de dados - Parte 2

Realizando a transformação de dados - Parte 3

Realizando a transformação de dados - Parte 4

Realizando a transformação de dados - Parte 5

Realizando a transformação de dados - Parte 6

Realizando a transformação de dados - Parte 7

Realizando a transformação de dados - Parte 8

Realizando a transformação de dados - Parte 9

Realizando a transformação de dados - Parte 10

Realizando a transformação de dados - Parte 11

Realizando a transformação de dados - Parte 12

Realizando a transformação de dados - Parte 13

ETL - Considerações Finais

* 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