ETL class project - Maestria en IA

Introduction

This project involves creating an ETL pipeline to extract, transform, and load a dataset. You must carefully select a dataset through research that meets these criteria:

- More than 10,000 rows
- Sufficient columns to perform meaningful transformations and extract valuable insights
- Use Python

Choose a dataset that motivates you to explore the problem deeply

Steps

The final project delivery consists of these steps:

1. Data sources: Select one or more data sources (e.g., CSVs, APIs, databases)
2. Data extraction: Use Python to extract the data from the source and store it in a relational database
3. EDA
4. Read the raw data from the staging area database using Python
5. Perform necessary transformations to create value and solve the problem
6. Create a merge task if needed to combine different data sources
7. Load the processed dataset into the database
8. Dashboard: Retrieve data from the ETL pipeline database and create a dashboard using your preferred tool (Power BI, Looker Studio)

> Considerations: The entire project must be in GitHub, including the EDA notebook.
> The ETL pipeline must be automated using Python or an ETL tool (Airflow, Prefect)

What is Expected

Is expected to have the complete pipeline working from the step 1 to the 6, evidences of the data in the database, dashboards and EDA notebooks , GitHub repository

Technologies

We expect you to use in this challenge:

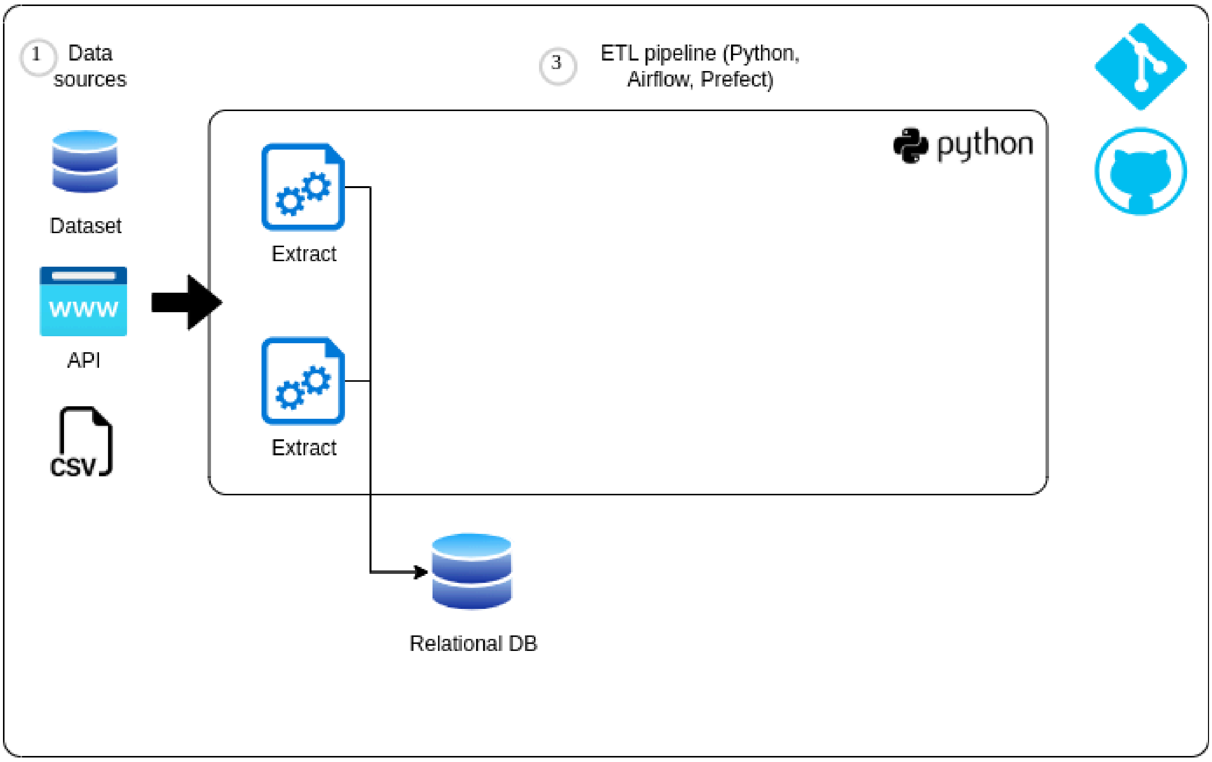
- Python
- Jupiter Notebook
- Database (you choose)
- AirFlow
- CSV files or API consumption
- Visualization tool
- Git/Github

Phase 1:

Rubrics:

1. Identification of the data problem or objective and dataset selection
2. Data extraction or collection

Diagram:



Deliverables:

Github repository with:

- Relevant code
- Readme file with all the context, instruction to use the repository and considerations
- Gitignore in order to only include relevant files into the respository

Documentation:

- Problem description
- Context
- Dataset description
- Process
- Evidences

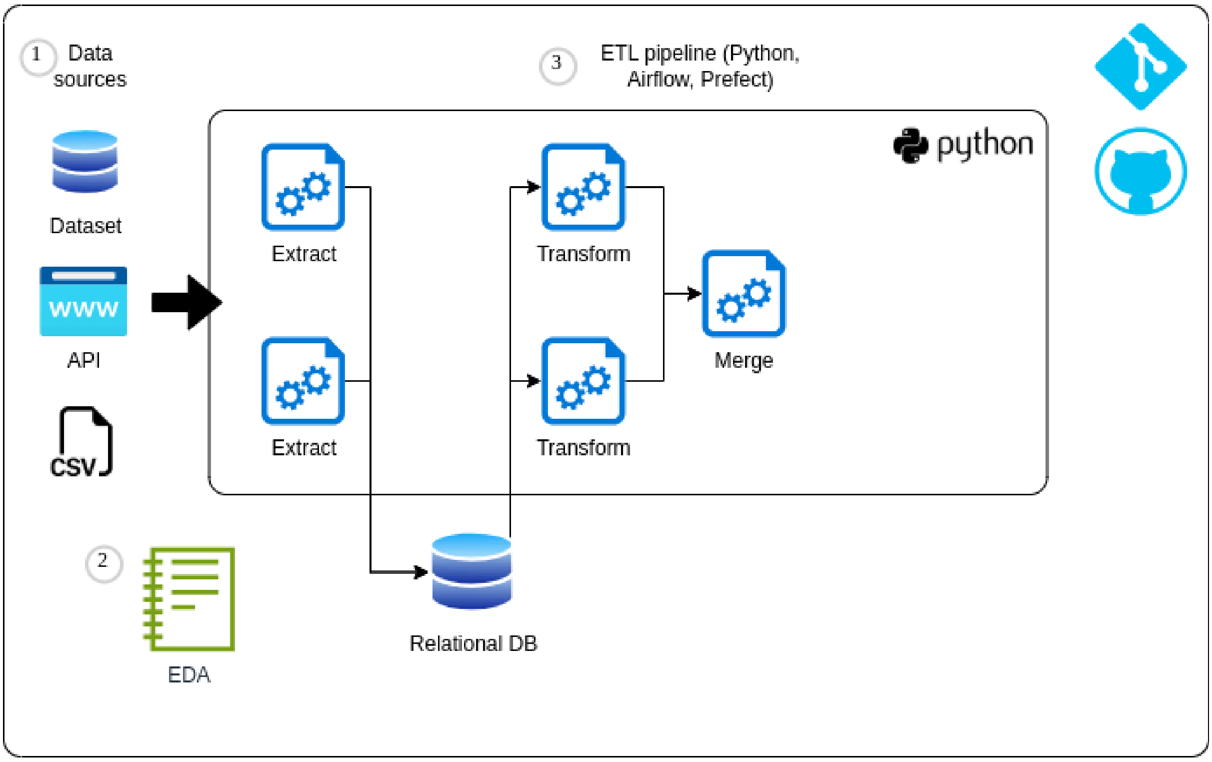
Phase 2:

Rubrics:

3. Data transformation

4. Data pre-analysis and visualization (EDA)

Diagram:



Deliverables:

Github repository with:

- Relevant code
- Readme file with all the context, instruction to use the repository and considerations
- Gitignore in order to only include relevant files into the respository
- EDA jupyter notebook

Documentation:

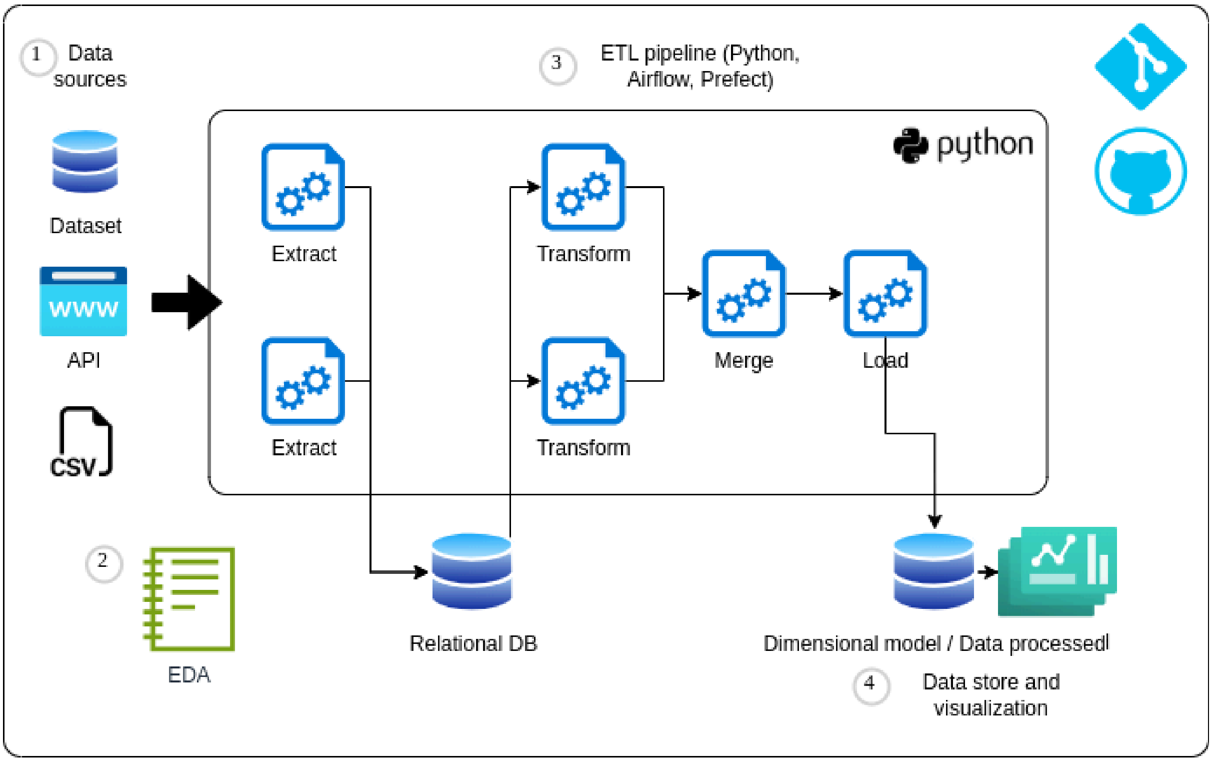
- Problem description
- Context
- Dataset description
- Process
- Evidences

Phase 3:

Rubrics:

3. Data load in a SQL database
4. Presentation and Data story telling

Diagram:



1 Data sources
3 ETL pipeline (Python, Airflow, Prefect)
Dataset
API
CSV
python
Extract
Extract
Transform
Transform
Merge
Load
2
EDA
Relational DB
Dimensional model / Data processedl
4 Data store and visualization

Deliverables:

Github repository with:

- Relevant code
- Readme file with all the context, instruction to use the repository and considerations
- Gitignore in order to only include relevant files into the respository
- EDA jupyter notebook
- Dashboard: in PDF format

Documentation:

- Problem description
- Context
- Dataset description
- Process
- Evidences
- Dashboard

Presentation:

- Presentation to explain the process and what was found (story telling)