Proyecto ETL: Proyecto Predios.

Facultad de Ingeniería, Universidad Autónoma de Occidente

Cali, Valle del Cauca

Presentado por:

Carlos Hidalgo Escobar - 22502395 - [carlos.hidalgo@uao.edu.co](mailto:carlos.hidalgo@uao.edu.co)

### Problem Description

There are numerous difficulties in handling information from the licensed software used by the public agency to project city properties. Poor decision-making and frequent errors occur due to deficiencies in the exported data, which is often extensive and heavy.

### Context

A public agency needs to analyze various variables related to city properties through projections to ensure secure and efficient property tax billing. The goal is to manage the necessary calculations for property analysis and facilitate decision-making by utilizing information extracted from the agency’s licensed software.

As a public city entity, the agency handles large capital figures, making property projections crucial to prevent significant financial losses for both the municipality and taxpayers.

The data extraction phase focuses on creating property tax projections through calculations and analyses comparing previous and current fiscal years (2024-2025).

### Dataset Description

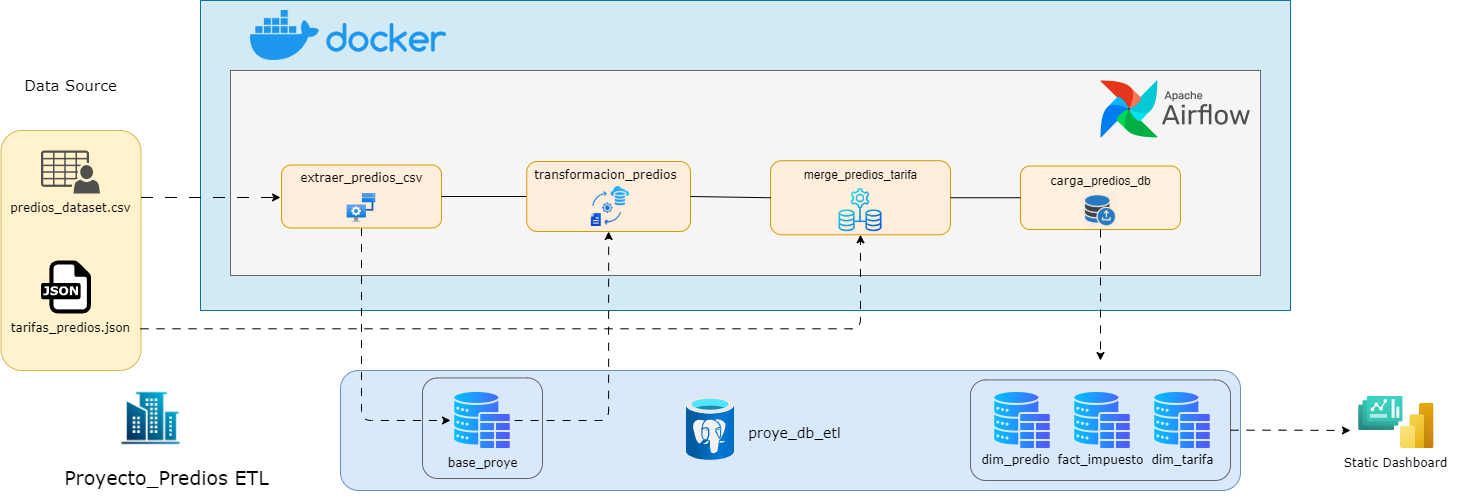
The data used for property projection was obtained through the public agency’s licensed software. Due to confidentiality agreements and data sensitivity, certain dataset columns have been modified to prevent compromising the public entity.

The dataset consists of 100,000 records and 21 columns:

* OBJETO\_NUMERICO: Unique indicator assigned by the licensed software to the property owner.
* TIPOPRED: Property type as defined by regulations.
* AVALPRED\_VIGANT: Previous fiscal year (2024) property valuation.
* USU\_VIGANT: Standardized property indicator for the previous fiscal year.
* ACTIVIDAD\_VIGANT: Standardized property activity for the previous fiscal year.
* ESTRATO\_VIGANT: Socioeconomic stratum for the previous fiscal year.
* AREA\_VIGANT: Property area for the previous fiscal year.
* TERRENO\_VIGANT: Property land value for the previous fiscal year.
* PREDIAL\_VIGANT: Property tax value for the previous fiscal year.
* COMUNA: Municipality subdivision where the property is located.
* BARRIO: Neighborhood where the property is located.
* MANZANA: Block where the property is located.
* TIPO\_PREDIO: Property type used for agency improvements.
* ACTUALIZACION: Classification of the property as rural or urban.
* AVALPRED\_VIGACT: Current fiscal year (2025) property valuation.
* USU\_VIGACT: Standardized property indicator for the current fiscal year.
* ACTIVIDAD\_VIGACT: Standardized property activity for the current fiscal year.
* ESTRATO\_VIGACT: Socioeconomic stratum for the current fiscal year.
* AREA\_VIGACT: Property area for the current fiscal year.
* TERRENO\_VIGACT: Property land value for the current fiscal year.
* CARTERA\_VIGACT: Indicates whether the property has outstanding debts.

### Process

#### Workflow Structure



(1) Workflow image of ETL Property Project

A workflow is defined where the data sources are specified as the files used for the established tasks executed via Python/Airflow in Docker.

Transformed data is stored in PostgreSQL databases, following standard ETL phases: Extraction, Staging, Transformation, Merge, and Load.

Folder Structure in Docker

Apache Airflow is used to automate the workflow by virtualizing it in Docker and creating tasks using Python code.

Docker is chosen for its flexibility and ease of running tools on Linux, as the project runs on Windows. Since Airflow is primarily for Linux environments, Docker serves as the virtual environment enabling Airflow execution on the local machine.



(2) Folder structure image of the ETL Property Project in Docker

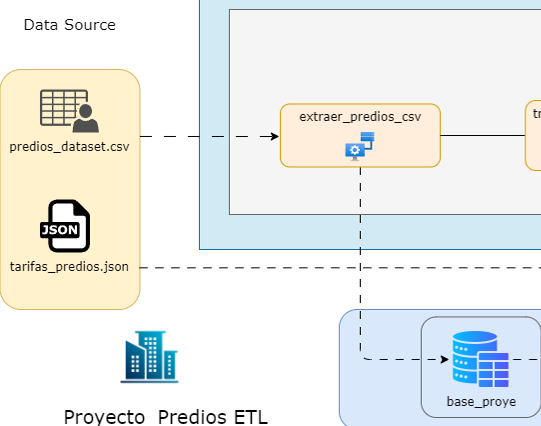
### Tools Used:

* Python
* Jupyter Notebooks
* PostgreSQL
* Power BI
* SQLAlchemy
* Apache Airflow
* Docker

### 

### 

### Extract



(3) Extraction phase image of the ETL Property Project

The extraction phase involves defining and presenting the data to be used in the process.





The first step in the ETL process is establishing a connection to the PostgreSQL server to create the database (proye\_etl\_db) for staging and storing the initial data load in the table (base\_proye).

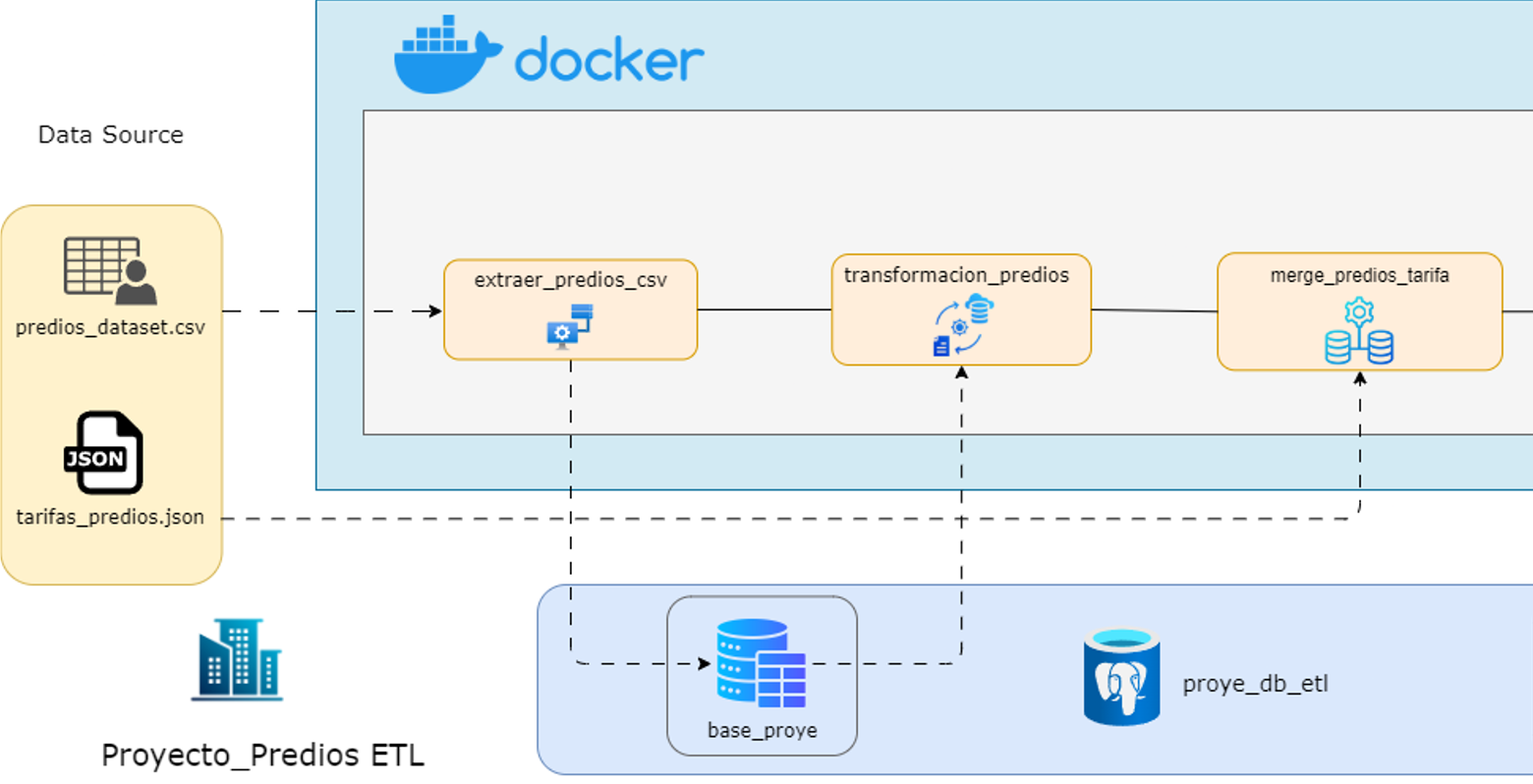
### Transform-Merge

For the transformation-merge phases, the database containing the extracted and staged data (base\_proye) is accessed. Various sub-stages are implemented for better organization and data governance, including data cleaning, standardization, and initial processing.

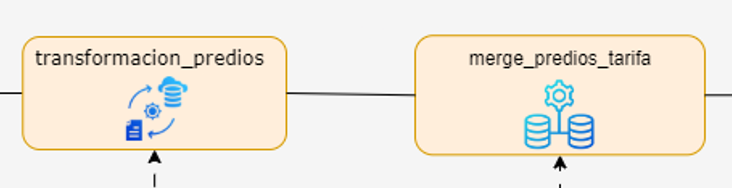
Since the agency's licensed software exports data with formatting inconsistencies, standardization is applied. For example, the OBJETO\_NUMERICO field, a unique identifier for property owners, should contain 20 digits, but the exported data often lacks a digit due to initial numerical format conversions (e.g., 01 → 1).

Data type conversions are also performed to correct any issues from the extraction phase.

The merge phase integrates the agency’s tariff data (in JSON format) with the dataset. Transformations are applied to ensure compatibility and enable property tax calculations.



(4) Transformation-merge phase image of the ETL Property Project

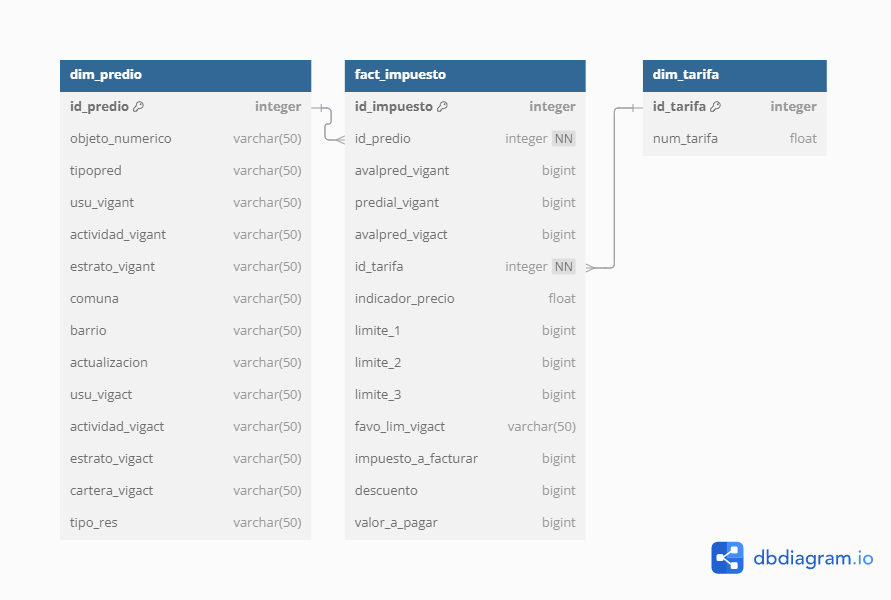


(5) Tasks image for transformation-merge in the ETL Property Project

Transformation and merge tasks communicate via transformed dataframes, passing the final output to the next phase.

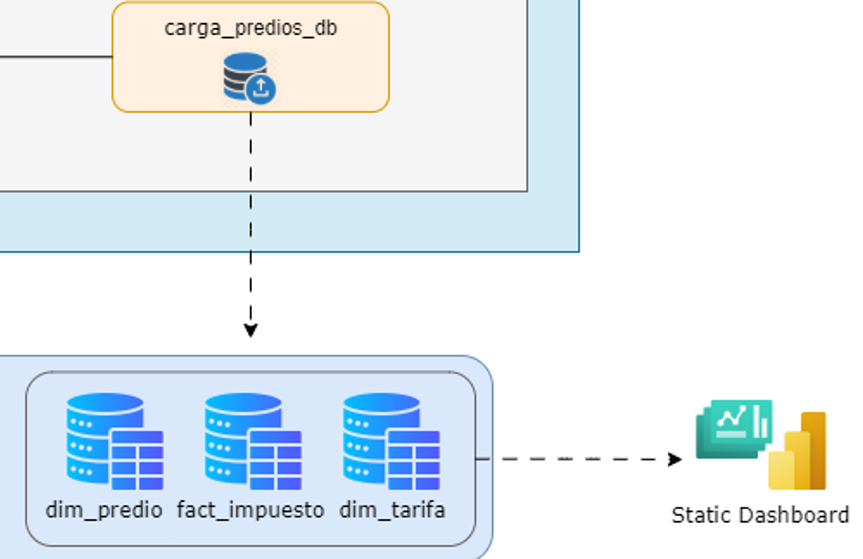
### Load

The Load phase concludes the workflow by inserting the final data into dimensional tables within the database, using the following dimensional model:

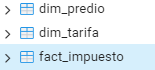


(6) Dimensional model image of the load phase in the ETL Property Project

This model aims to enhance data analysis and visualization in Power BI.

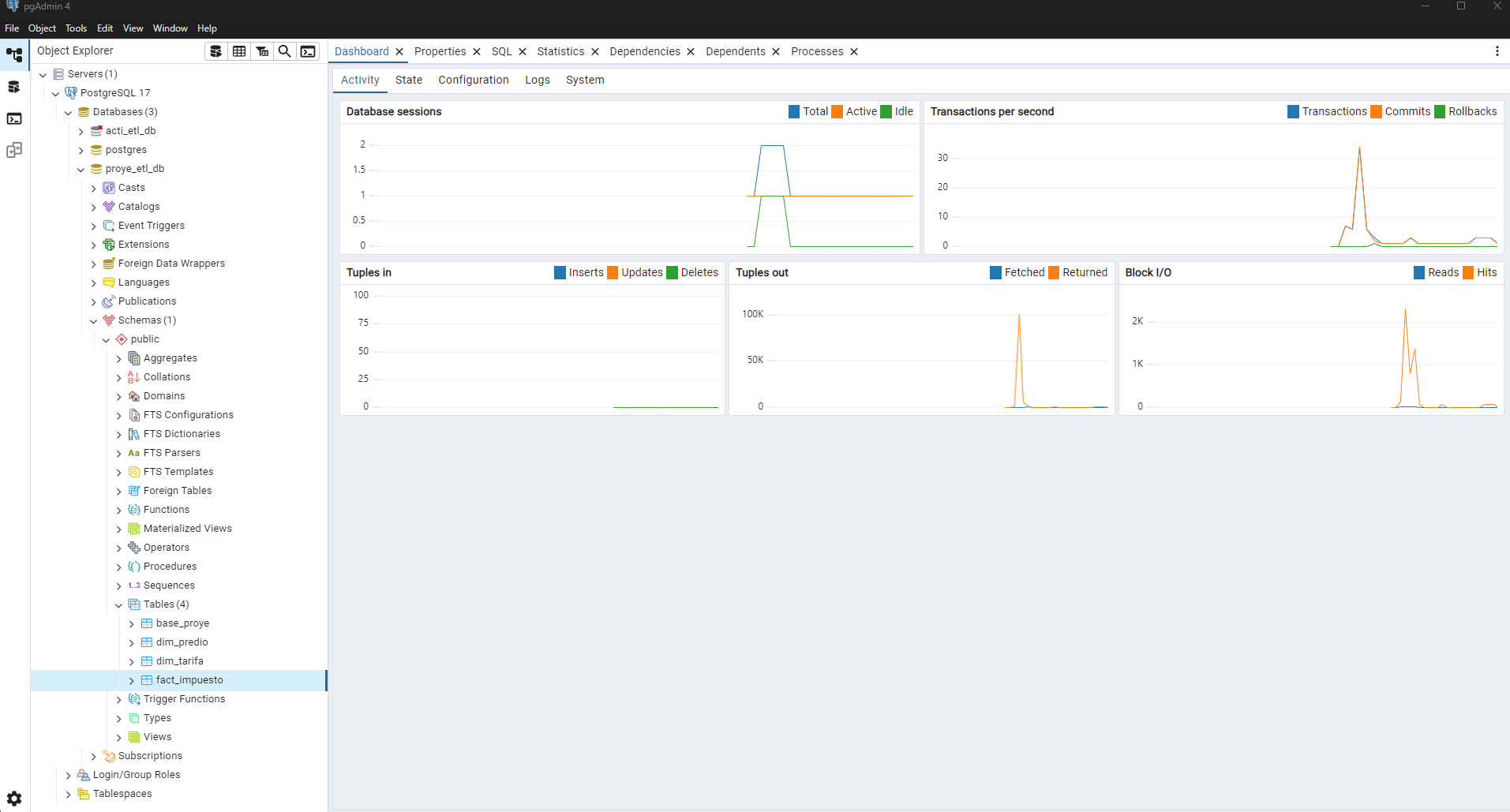


(7) Load phase image of the ETL Property Project

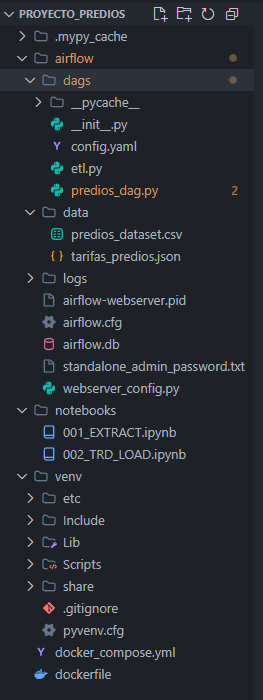


(8) Dimensional-fact table structure image of the ETL Property Project

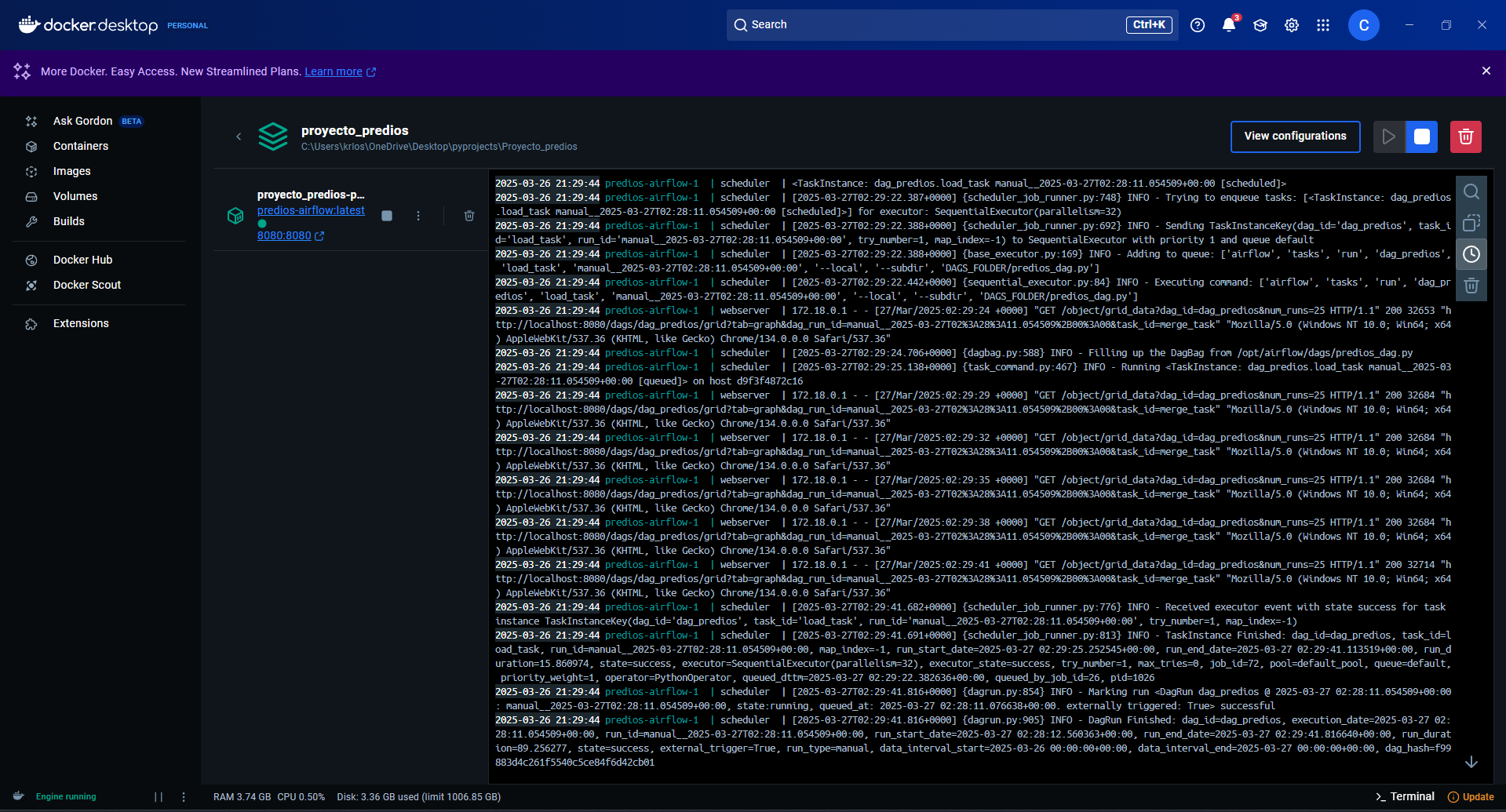
### Evidences



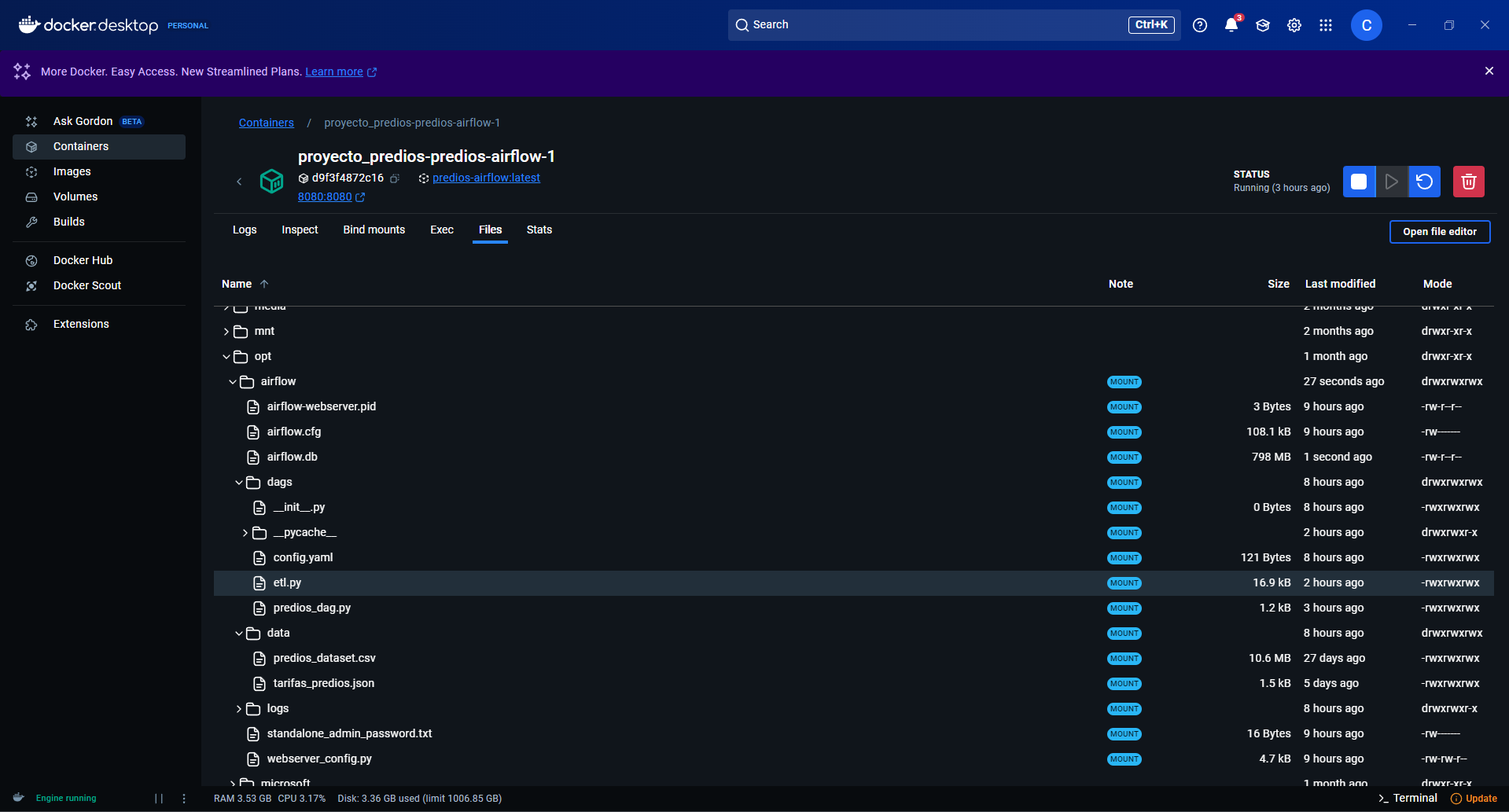
(1) PostgreSQL Workplace pgAdmin image



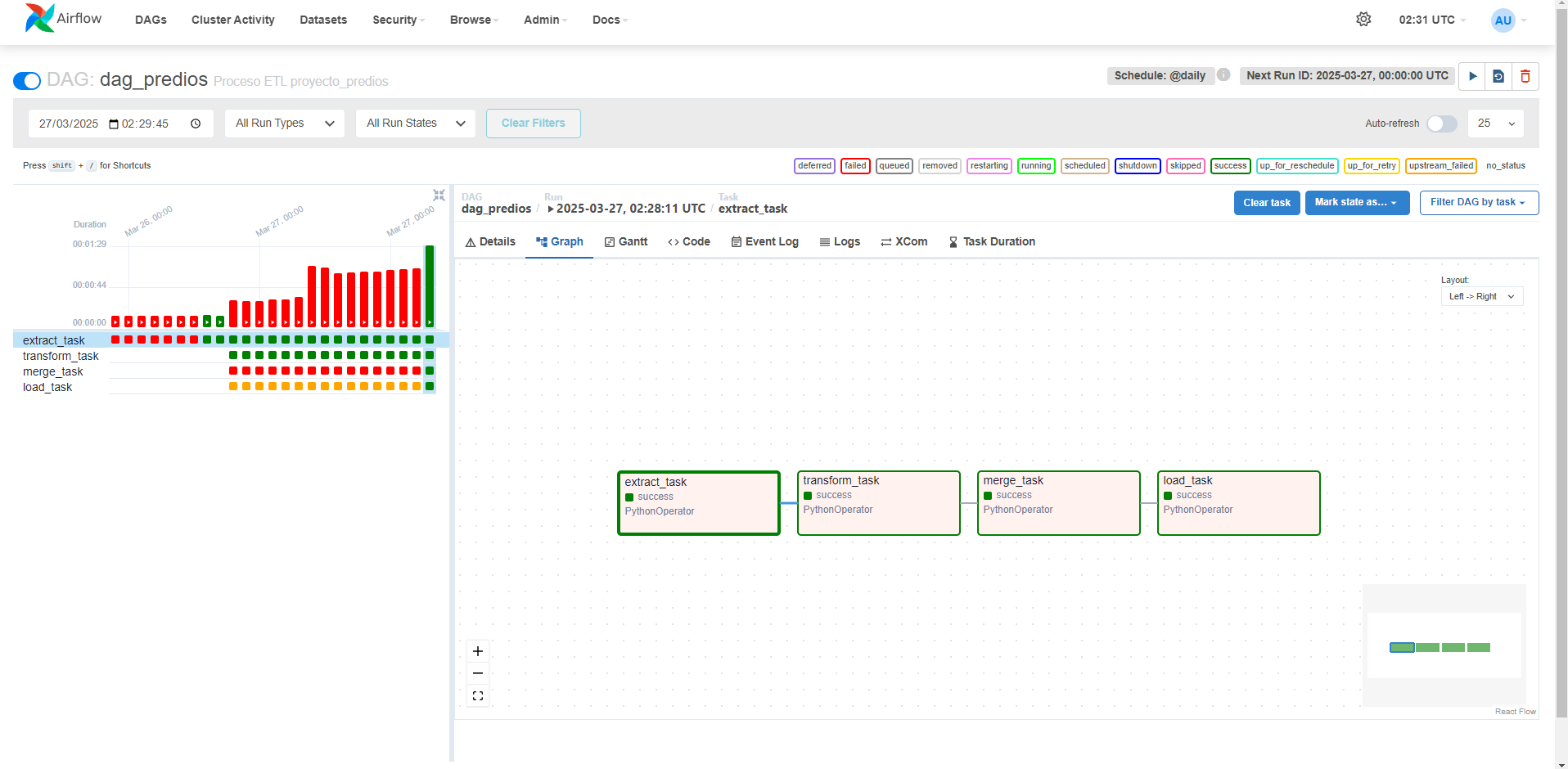
(2) Folder structure image in VSCode



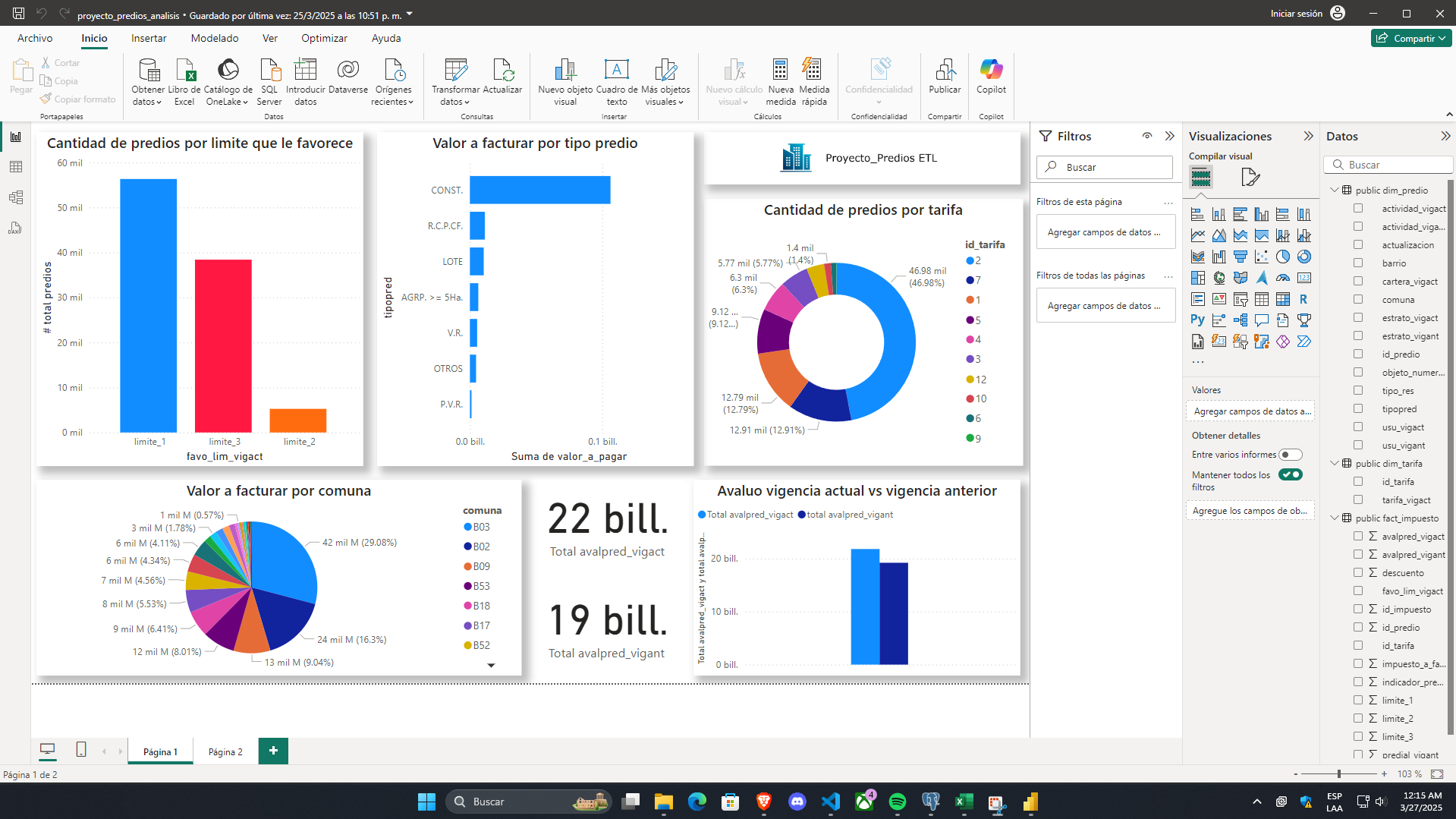
(3) Docker container image for the ETL Property Project



(4) Folder structure image inside Docker container



(5) Apache Airflow DAG tasks image

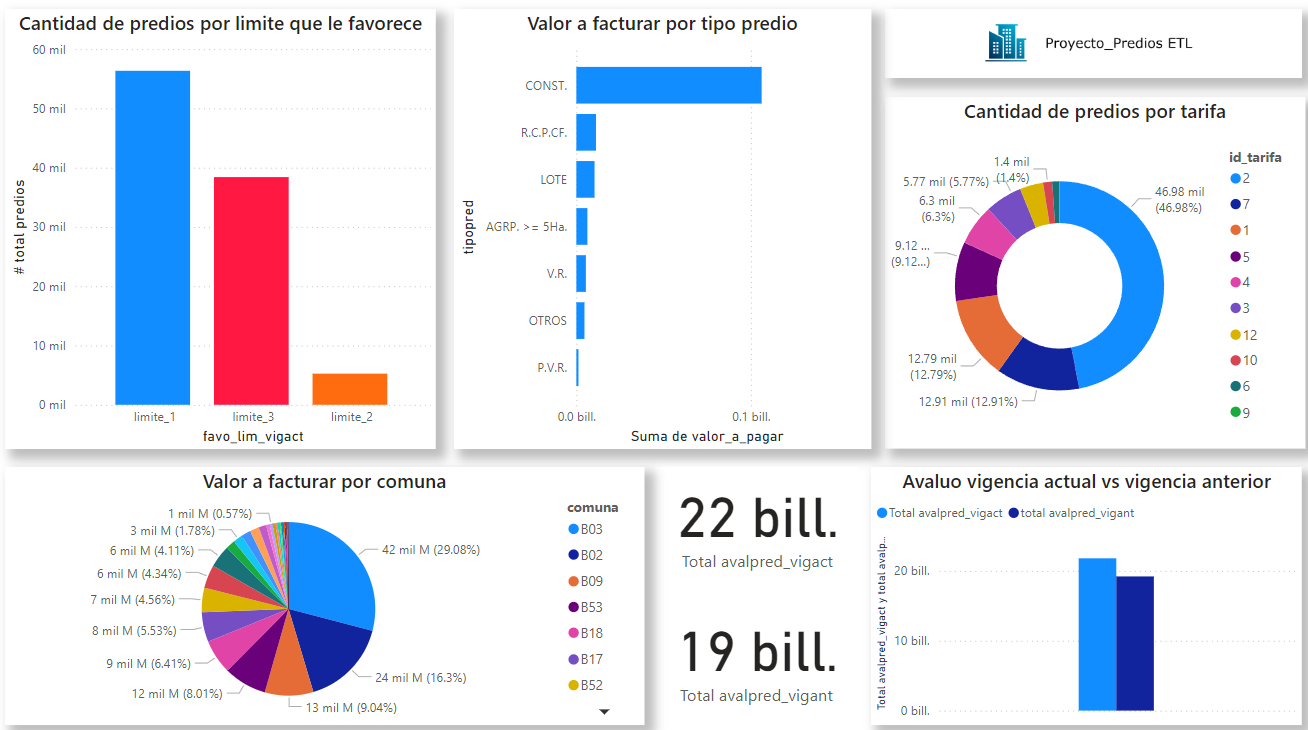


(6) Power BI dashboard creation image

### Dashboard

The dashboard was built in Power BI, connecting directly to the final database from the Load process.

This allows for an analysis of key variables related to the 2025 property tax projections.



(10) Dashboard page 1 image in Power BI



(11) Dashboard page 2 image in Power BI

Since advanced data analysis skills are currently lacking, the dashboard primarily displays metrics and values that highlight differences in property valuations between previous and current fiscal years.

Additionally, the project emphasizes the significance of handling vast amounts of data for property projections and decision-making, given the substantial capital managed in the city’s real estate sector.