Universidad Autónoma de Occidente

Faculty of Engineering and Basic Sciences

Program: Data Engineering and Artificial Intelligence

Course: ETL

Activity: ETL Proyect second delivery

Members:

Dillan Steven Molina

Laura Daniela Astudillo

Maria Paula Pinillo

This document details the implementation of an ETL pipeline designed for the management and analysis of environmental data. It represents the continuation of a previously developed project, to which new functionalities were added to enhance its scope and level of automation.
Among the improvements implemented are the integration of the Colombia Public API (api-colombia.com), used to obtain official and up-to-date information on regions and departments, and the automation of the execution flow using Apache Airflow, which enables a sequential and controlled management of the extraction, transformation, and loading stages of the data.

The document describes the data sources used, the cleaning and transformation processes applied, and the loading of data into a dimensional model designed to support analysis, indicator generation, and the interactive visualization of results.

Refined Objectives

For this delivery, the objectives defined were the following:

- Obtain the total sum of investment in FNCE grouped by each region of the country.
- Obtain the total investment in GEE by region.
- To analyse the temporal evolution of investment in FNCE and GEE within each region.
- Identify which is the most frequent type of investment in each region.

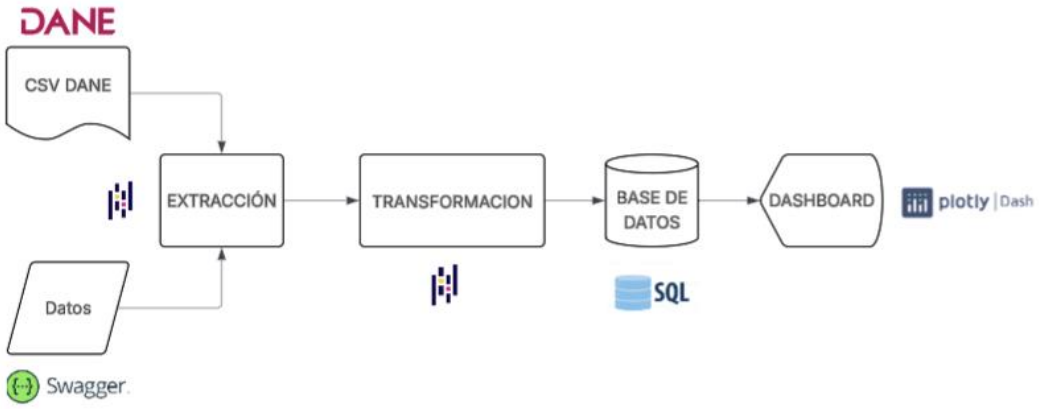
API details and why it was chosen

We selected the Colombia Public API (api-colombia.com) to obtain official and updated information on the regions and departments of the country. This API was chosen because:

- It has no limit on requests, which made it easy to use during testing and integration.
- It provides structured and reliable data on the territorial division of the country.
- It provides exactly the information we needed to populate the geographic dimensions of the Data Warehouse.

Thanks to this integration, we were able to feed the tables of regions and departments, guaranteeing the consistency of the data and enriching the analyses carried out within the project.

Architecture diagram



Description of the ETL pipeline steps

1. **Extraction**

In this phase, data are collected and consolidated from different sources:

- DANE CSV files (EAI) corresponding to the years 2019–2022, with an initial volume of 2,940 columns per set.
- Colombian Public API (api-colombia.com), accessed through Swagger from a Jupyter notebook, using the requests and pandas libraries. This source

provided official information on regions and departments, which was integrated to strengthen the geographical dimension of the analysis.
- XML files with the structure and metadata of the variables, processed to identify the 148 relevant variables of chapter 2 of the EAI.

This combination of sources made it possible to have complete, updated and validated information for the subsequent stages of the pipeline.

2.Transformation

During this stage, the data is cleansed, validated, and enriched in the dataTransformation.py module. The main operations carried out include:

- Data validation: verification of calculations and consistency in the values reported by companies.
- Inconsistency detection and correction: Elimination of erroneous or incomplete records.
- Temporal normalization: standardization of reference dates and periods.

  Calculation of environmental metrics, such as:

  - Energy Efficiency and Clean Power Generation (GEE_final, and FNCE)
  - Tax Credits (Descuento_FNCE, Descuento_GEE)
  - Percentage of environmental investment
- Estimates: projection of tax benefits and potential returns associated with environmental investments.

This phase ensures the quality, consistency, and analytical utility of the data.

3. Load

The transformed data is stored in a SQLite database, structured under a dimensional model implemented in the dataLoad.py module. El modelo incluye:

- Dimension tables: DimEnterprise, DimRegion, DimYear, DimTypeInvestment.
- Fact Table: FactInversiones, which consolidates investments by company, region, year, and type of investment.

The new information with the names of the regions and information of the departments was also implemented.

4. Pipeline Flow and Automation

The entire flow is executed automatically by Apache Airflow, coordinated by the etl_run_scripts.py script. The order of execution is as follows:

1. Extraction: Fetching data from CSV, XML, and API files.
2. Transformation: cleaning, validation, and calculation of environmental metrics.
3. Loading: Inserting the transformed data into the SQLite dimensional model.
4. Visualization: automatic updating of indicators in the interactive dashboard developed with Dash (Plotly), where KPIs, heat maps and the annual evolution of investments are displayed.

Airflow DAG design

This DAG is titled etl_run_scripts, and its function is to orchestrate the entire ETL flow of the project within Airflow. The flow is composed of five main tasks: debug, extract, run_notebook, transform, and load. Each of them is executed sequentially to ensure that the process of extracting, cleaning, analyzing, and loading data is successful.

Examples of visualizations and insights

A total of 4 KPIs and 8 visualizations were performed. These are aligned with the previously defined objectives.

- For the first graph, the objective was to compare the total level of investment in Non-Conventional Energy Sources (FNCE) and in Efficient Energy Management (EGE) between the different regions of the country, this

graph allows us to observe the priorities regarding investments at the regional level, helping public or private entities to Identify leading areas in energy transition, make decisions on where to concentrate incentives or support policies and evaluate the balance between generation and efficiency at the regional level.

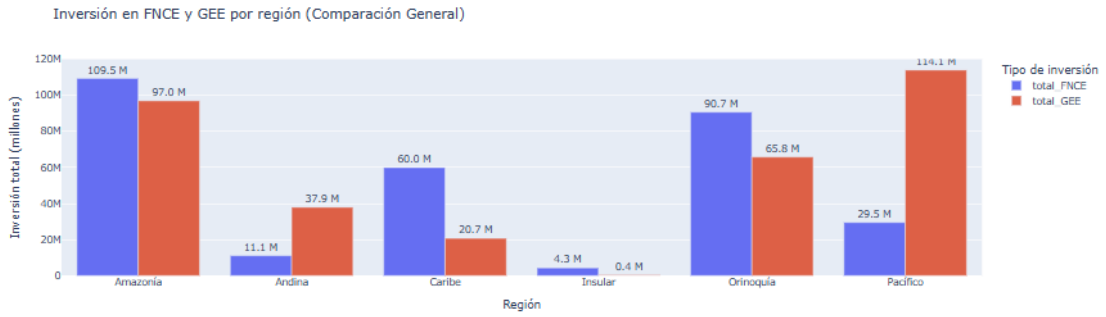


Figure 1. FNCE and GEE Investment by Region

2.

The second graph allows us to determine which is the most frequent or representative type of investment in each region, this focuses not on the total amount invested, instead it allows us to identify what type of expenditure or specific investment predominates, this allows us to understand if the regions are building, consolidating or maintaining their energy infrastructure

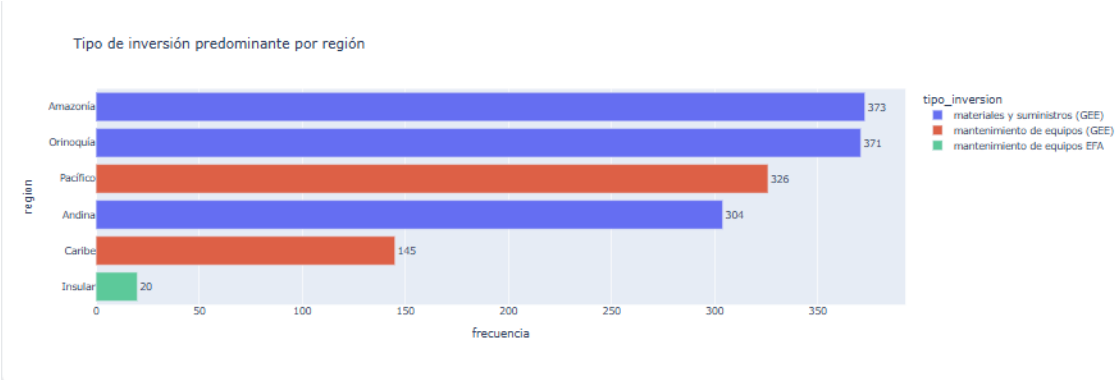


Figure 2. Investment type by region

3. The objective of these graphs is to allow us to analyze the temporal trend of investment in FNCE and GHG in each region, if a region shows a constant increase in FNCE, it shows a growing commitment to the transition to clean energy, on the contrary, if the investment in GEE indicates that the region is strengthening its energy efficiency, optimizing their consumption and reducing waste. Strong

changes could reflect external factors such as economic disruptions or logistical constraints that affected project execution

Figure 3. Evolution of FNCE and GEE investments by region