Workshop 2: ETL process using airflow

Introduction

This workshop is an exercise on how to build an ETL pipeline using Apache Airflow, the idea is to extract information using three different data sources (API, csv file, database), then do some transformations and merge the transformed data to finally load into google drive as a CSV file and store the data in a DB. As a last step, create a dashboard from the data stored in the DB to visualize the information in the best way you consider.

A few things to consider:

- I ask that you complete this challenge within the timeframe agreed on our conversation.
- **You cannot use tools such as Copilot, Tabnine, Captain Stack, GPT-Code-Clippy, chatGPT, or similar to simplify or generate code to support the challenge. Doing this will be grounds for automatic disqualification.**

Getting Started

Hey, welcome to the **Airflow Data Engineer** code challenge. In this challenge, I am interested in seeing your knowledge about data management and visualizations. I will give you some data, and your final objective is to show me all the ETL process using the three different data sources and some chart visualizations.

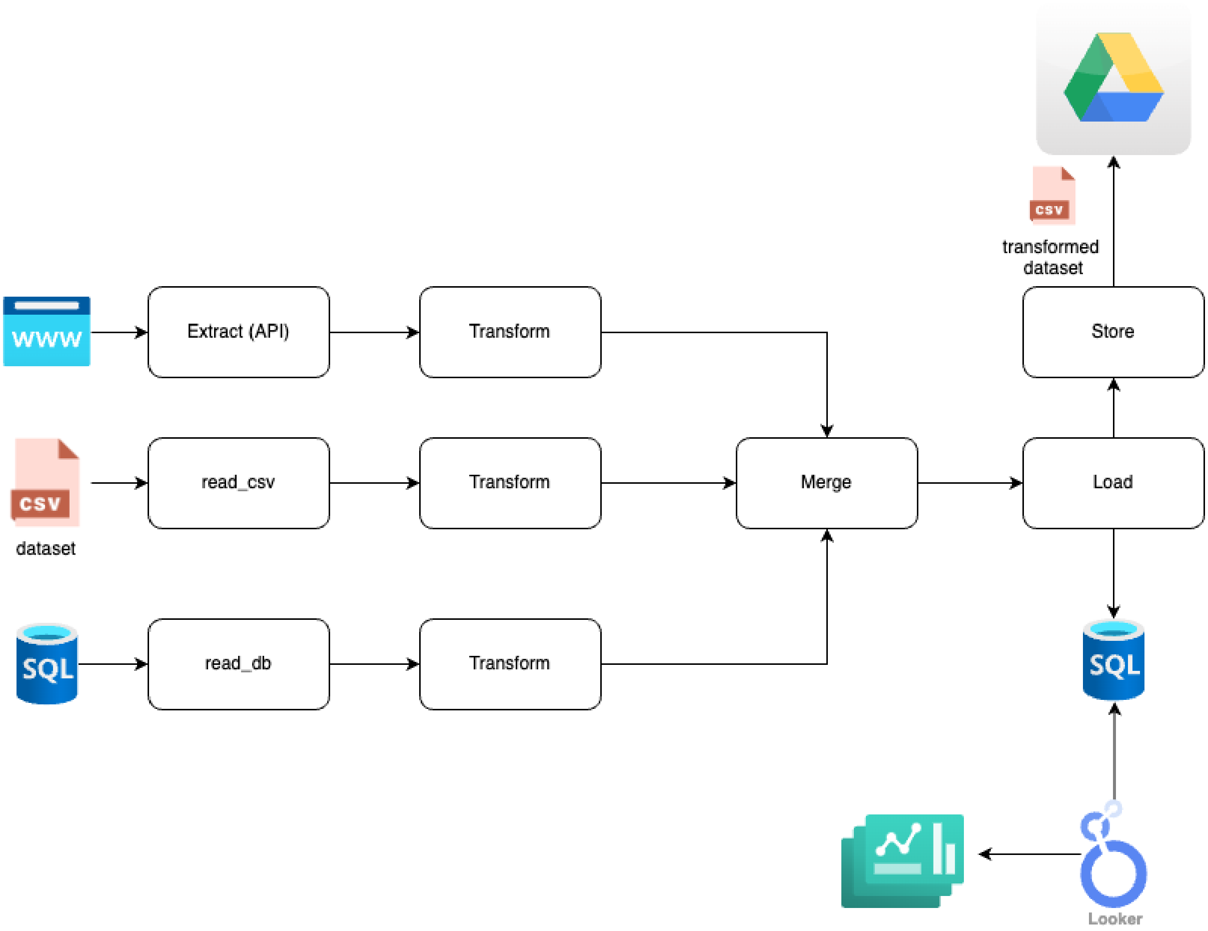
In this workshop we will use the spotify dataset (link in the data section) to be readed in python and airflow, create some transformation and load into a database, on the other hand we will use the grammys dataset to be loaded into a database, then using airflow we will read the data from the database , perform some transformations, merge with the spotify dataset and API dataset and load into the database.

the technologies we expect to evaluate are described in the technologies section.

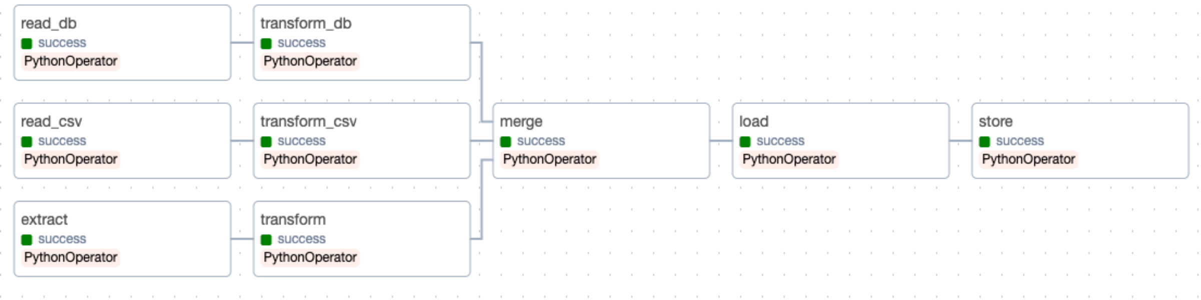
What is Expected

I expect that you get the CSV files and the API and create an ETL data pipeline to extract the information, transform, merge, and store (in a DB and in google drive as a CSV file). Also, you will display those data from the database in chart visualizations; remember, the data should be stored in a database, and your reports must come from the database, not the CSV file.

The following figure shows a diagram of the project:



The next figure is a possible visualization of how the project may looks like in apache airflow



Data:

Dataset to be readed as CSV: Spotify dataset

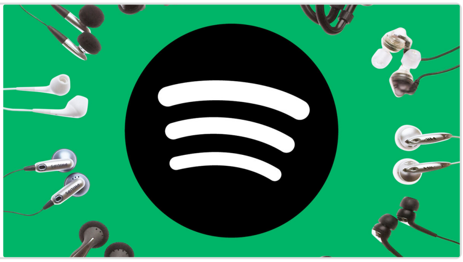
Dataset to be loaded into the initial database: Grammys Dataset

Note: The original information about the datasets can be found in:

🎹 Spotify Tracks Dataset

A dataset of Spotify songs with different genres and their audio features

https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset



Grammy Awards

Grammy Awards, 1958 - 2019

https://www.kaggle.com/datasets/unanimad/grammy-awards

Visualizations:

Find a way to merge the three data sources in order to create an impressive dashboard with useful information

Technologies

We expect you to use in this challenge:

- Python
- Jupiter Notebook

- Database (you choose)
- Apache airflow
- CSV files
- Visualization tool (Looker, powerBI, Tableau)

G-drive folder

Folder in google drive with all information:

wokshop_002 - Google Drive

https://drive.google.com/drive/folders/10RUGzwbzOgovhH6BhSMhaEbNS5cZD1CO?usp=drive_link