CO ThuyLinh / DELLAL Célia / ALMEIDA Thomas / HORD HAMZA
I3 – DE1

**Documentation :** Application of big data project

(Subject 3 : DataViz)

## What it is?

This project is a **Data Visualization Dashboard** designed to analyze user music listening habits and provide actionable insights. It is built to process a dataset of music streams, calculate key performance indicators (KPIs), and present them in an interactive and visually appealing dashboard using **Tableau**.

Data were processed raw from CSV files using an automated ETL pipeline built with Apache Airflow. This pipeline extracts raw music listening data, cleans and standardizes it (e.g., formatting dates, handling missing values), and loads it into a PostgreSQL database hosted on Supabase. From there, SQL queries and a pre-built SQL view (music_data) calculate the KPIs required for visualization in Tableau.

The Tableau dashboard was designed by extracting the dataset from the PostgreSQL database into Tableau Extracts. This approach ensures that the dashboard remains fully functional even without a live database connection, while maintaining high performance and interactivity. The dashboard allows users to easily explore trends, rankings, and behaviors related to music listening patterns, offering insights through dynamic and customizable visualizations.

## The project enables users to:

1. Explore the most popular tracks and albums across different time periods.
2. Analyze weekly trends in music listening behaviors.
3. Identify the top listeners and their contributions to total plays.
4. Perform cross-tabulation analysis, showing listener-artist interactions in a table format.
5. Interact with the dashboard using built-in instructions to filter data dynamically by week, year, or artist for focused analysis.

The dashboard contains clear instructions to guide users on how to interact with the visualizations. For example: Filters are provided for selecting specific years or weeks and legends explain how to filter down into details.

Users can interact with filters, tooltips, and visual elements to explore the data and insights.

## How does it install ?

1. **Compilation process and running instructions**

**To run the ETL pipeline, if needed :**

- A relational database management system (**PostgreSQL**) to store and query data.
- **Python 3.x environment** or equivalent for initial data processing and **Airflow**

**To launch the dashboard**

- The **Tableau** visualization tool (license required).
- A **Tableau account** (to be created if you don't already have one).

### 1.2 Steps to Set Up the Project

1. **Installing and Configuring Tableau:**
    - Download and install Tableau Desktop.
    - Activate your Tableau license by entering a valid product key.
    - Create a Tableau account if you don't already have one.
    - Once Tableau is installed and configured, open the project's .twbx file.
2. **(Optional) If you want to look at to the datasource :**
    - Navigate to the Data Sources sheet in Tableau.
        - Enter the provided access code **"SDavB#_cAf8r8z9"** to establish a connection with the PostgreSQL database.
        - Verify that the connection is active (check the connection status).

### 1.3 Launching the Dashboard

- Open the .twbx file in Tableau Desktop.
- Ensure that:
    - The license key is active.
    - There is no error showing up linked to connection between Tableau and the data.
- To use the dashboard in presentation mode:
    - Press F7 on your keyboard, or
    - Click on the presentation mode icon in the toolbar.
    - This will allow you to interact with the visualizations and explore the KPIs dynamically.
- Navigate to view the visualizations and interact with the KPIs.
- Users can interact with the dashboard by filtering data, exploring trends, and analyzing KPIs for specific timeframes or categories.

## How does it work ?

The ETL pipeline is managed by Apache Airflow. It automates the extraction of raw CSV files from a folder, processes them to clean and structure the data, and loads the results into the flow_in_music_data table in Supabase. The pipeline consists of the following steps:

1. **Extraction**: The pipeline scans a designated folder for all CSV files. Each file is read and loaded into a Pandas DataFrame using the process_csv function.
2. **Transformation**: Data is cleaned to ensure consistency. For instance, invalid dates are replaced with NULL, and columns like singer, album, and title are standardized as strings. Metadata, such as the file name, is added to the dataset.
3. **Loading**: The cleaned data is converted into JSON format and inserted into the flow_in_music_data table in Supabase. To handle large datasets, the data is inserted in batches of 100,000 rows, ensuring efficient and reliable loading.

**SQL View for Tableau**

To make the data accessible and ready for visualization in Tableau, a **SQL view (music_data)** is created in the PostgreSQL database. This view performs additional transformations on the raw data, such as:

1. Replacing invalid or missing dates (NaT) with NULL.
2. Converting the date column to the DATE format for easier manipulation in Tableau.
3. Providing a clean and structured dataset for visualization.

- **SQL Command to Create the View**:

```
CREATE OR REPLACE VIEW music_data AS

SELECT

   id,

   singer,

   title,

   album,

   CASE

      WHEN date = 'NaT' THEN NULL

      ELSE TO_TIMESTAMP(date, 'YYYY-MM-DD HH:MM:SS')::DATE

   END AS date_cleaned

FROM flow_in_music_data;
```

The music_data view simplifies Tableau queries, ensuring consistent and repeatable results across different visualizations.

**Connecting Tableau**

Once the data is ready in Supabase, Tableau is connected to the music_data view. This allows Tableau to dynamically fetch data and calculate the required KPIs for the dashboard. The final dashboard is delivered with data extracts, ensuring that users can interact with preloaded data even without a direct database connection.

In Tableau, users can:

- Filter data by specific weeks or years.
- Analyze trends and rankings through interactive charts and tables.
- Explore KPIs such as:
    - The most listened track of all time.
    - The most listened track for each week.
    - The most listened album of all time.

- The most listened album for each week.
- Cross-tabulation of the number of tracks listened to by user and by artist.
- Ranking of the 10 biggest listeners (all time).
- Weekly ranking of the top 10 listeners.

**Data Visualization:**

- Dynamically displays KPIs via an interactive dashboard in Tableau.
- Provides detailed charts, rankings, and trends that are easy to explore.
- Allows stakeholders to filter data and customize the view based on their needs.

**Use of AI and Data Optimization**

As part of the project, AI was used for the following optimizations:

- Debugging Timeout Issues: When inserting files with more than 100,000 rows into the PostgreSQL database, the process would timeout. AI (via ChatGPT) was used to debug and implement the batch insertion solution, splitting the data into smaller chunks to prevent timeouts and improve the reliability of the ETL process.
- Calculated Fields in Tableau: AI was utilized to simplify and enhance the configuration of calculated fields in Tableau, ensuring optimal performance and easier data manipulation.