Nadjib KHAMMAR
Abel BOUDAIB
Kaci BOUAOULI
Ingé3 – DE1

Applications of Big Data Project
Data Visualisation
Tune Trends

Summary

Introduction 3

Chapter 1: ETL Process 4

Chapter 2: Data Migration to PostgreSQL 6

Chapter 3: Creating a dashboard with Power BI 9

Introduction

This project aims to study a database from a music streaming application as part of an advanced analysis based on Big Data principles. The primary objective is to build a dynamic dashboard that showcases relevant Key Performance Indicators (KPIs), such as the most played tracks, the most popular albums, and user behavior on the platform.

The project is divided into three main stages:

1. **ETL Process (Extract, Transform, Load):** Data preparation by merging multiple source files, enriching and transforming the data to make it analysis-ready.
2. **Migration to a Relational Database:** Integration of the transformed data into PostgreSQL for more efficient analysis.
3. **Dashboard Creation with Power BI:** Data visualization and presentation of the results in an interactive and visual format.

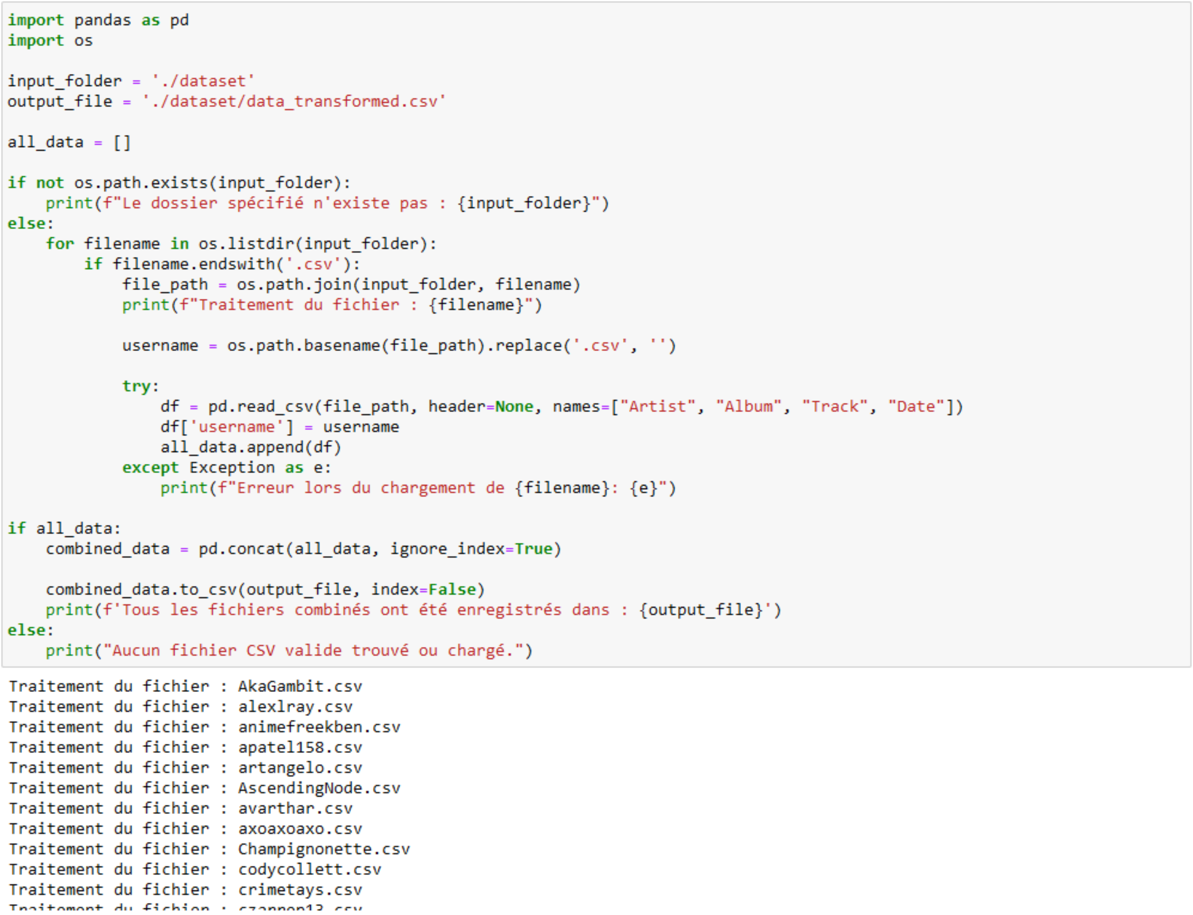
This report details each stage, the challenges encountered, the solutions implemented, and the results achieved.

Chapter 1: ETL Process

The ETL process is a key step in this project, as it transforms raw data from CSV files into a format suitable for analysis tools. Below are the main steps of the implemented ETL pipeline.

1. Merging CSV Files and Adding a Username Column

The data was distributed across multiple CSV files, each representing a user. The first step involved combining these files into a single dataset while adding a **username** column to identify the source of the data.



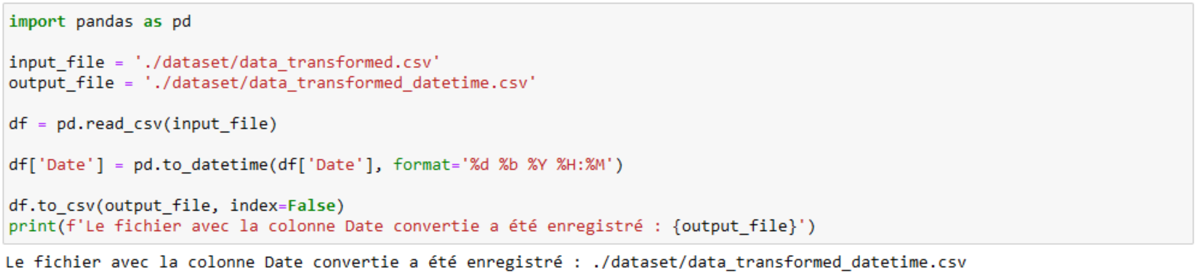
```python
import pandas as pd
import os

input_folder = './dataset'
output_file = './dataset/data_transformed.csv'

all_data = []

if not os.path.exists(input_folder):
    print(f"Le dossier spécifié n'existe pas : {input_folder}")
else:
    for filename in os.listdir(input_folder):
        if filename.endswith('.csv'):
            file_path = os.path.join(input_folder, filename)
            print(f"Traitement du fichier : {filename}")

            username = os.path.basename(file_path).replace('.csv', '')

            try:
                df = pd.read_csv(file_path, header=None, names=["Artist", "Album", "Track", "Date"])
                df['username'] = username
                all_data.append(df)
            except Exception as e:
                print(f"Erreur lors du chargement de {filename}: {e}")

if all_data:
    combined_data = pd.concat(all_data, ignore_index=True)

    combined_data.to_csv(output_file, index=False)
    print(f'Tous les fichiers combinés ont été enregistrés dans : {output_file}')
else:
    print("Aucun fichier CSV valide trouvé ou chargé.")
```

```
Traitement du fichier : AkaGambit.csv
Traitement du fichier : alexlray.csv
Traitement du fichier : animefreekben.csv
Traitement du fichier : apatel158.csv
Traitement du fichier : artangelo.csv
Traitement du fichier : AscendingNode.csv
Traitement du fichier : avarthar.csv
Traitement du fichier : axoaxoaxo.csv
Traitement du fichier : Champignonette.csv
Traitement du fichier : codycollett.csv
Traitement du fichier : crimetays.csv
Traitement du fichier : czannon13.csv
```

- Merging the Files: The script iterates through the folder containing the CSV files, reads each one, and combines them into a single dataset.
- Adding the Username Column: A new username column is added to identify the user corresponding to each source file.
- Saving the Transformed Data: The resulting file is saved under the name data_transformed.csv.

2. Converting the Date Column to Datetime Format

To enable time-based analysis, the Date column is converted to the datetime format. This ensures that the dates can be properly manipulated for temporal analysis.



```
import pandas as pd

input_file = './dataset/data_transformed.csv'
output_file = './dataset/data_transformed_datetime.csv'

df = pd.read_csv(input_file)

df['Date'] = pd.to_datetime(df['Date'], format='%d %b %Y %H:%M')

df.to_csv(output_file, index=False)
print(f'Le fichier avec la colonne Date convertie a été enregistré : {output_file}')
```

```
Le fichier avec la colonne Date convertie a été enregistré : ./dataset/data_transformed_datetime.csv
```

- Loading the Data: The combined file is loaded for transformation.
- Converting to Datetime Format: The Date column is converted into a standard format that is useful for analysis tools.
- Saving the Data: The transformed data is saved into a new file.

The ETL process has successfully transformed raw data into a structured format, ready for migration into PostgreSQL. The data is now cleaned, enriched, and organized, providing a solid foundation for KPI analysis and visualization.

Chapter 2: Data Migration to PostgreSQL

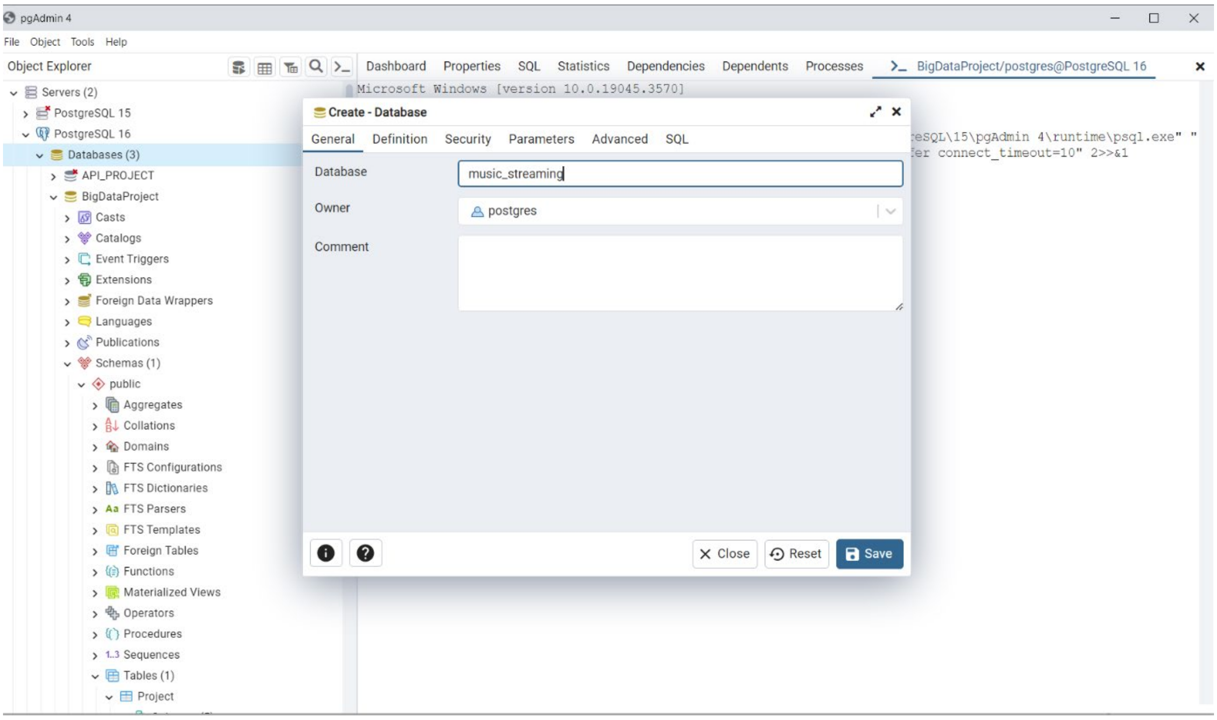
After preparing and transforming the data through the ETL process, the next step is to import it into a relational PostgreSQL database. This migration is crucial for performing efficient analyses and organizing the data in a structured manner.

1. Creating the Database in PostgreSQL

We used **PGAdmin**, a graphical tool for managing PostgreSQL databases. The following steps were followed:

The first step was to launch PGAdmin and connect to the PostgreSQL server.

Next, we created a new database to host the transformed data. We named this database **music_streaming**.



```
Microsoft Windows [version 10.0.19045.3570]
eSQL\15\pgAdmin 4\runtime\psql.exe" "
er connect_timeout=10" 2>>&1
```

Database: music_streaming

Owner: postgres

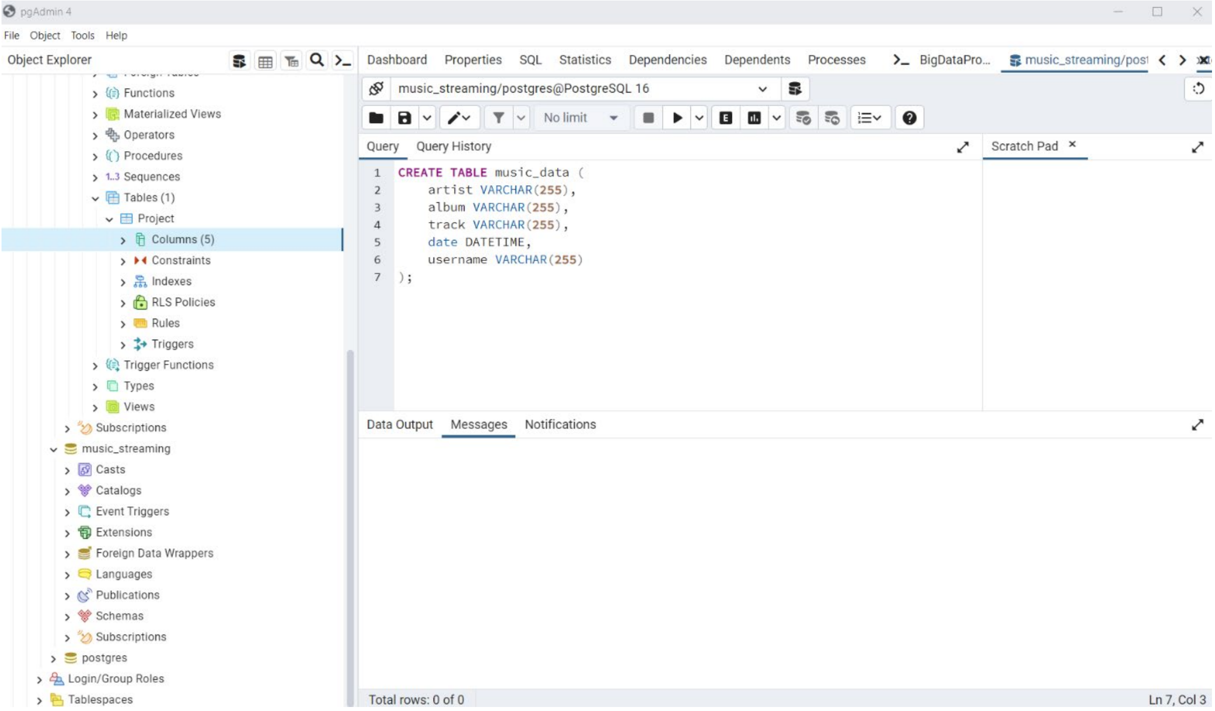
Comment:

2. Creating the Table to Store the Data

Once the database was created, we defined a table to structure the imported data. The table includes the following columns:

- artist (type VARCHAR): The name of the artist.
- album (type VARCHAR): The name of the album.
- track (type VARCHAR): The title of the track.
- date (type DATETIME): The date and time the track was listened to.
- username (type VARCHAR): The identifier of the user.

This table structure ensures that the data is well-organized and ready for analysis.



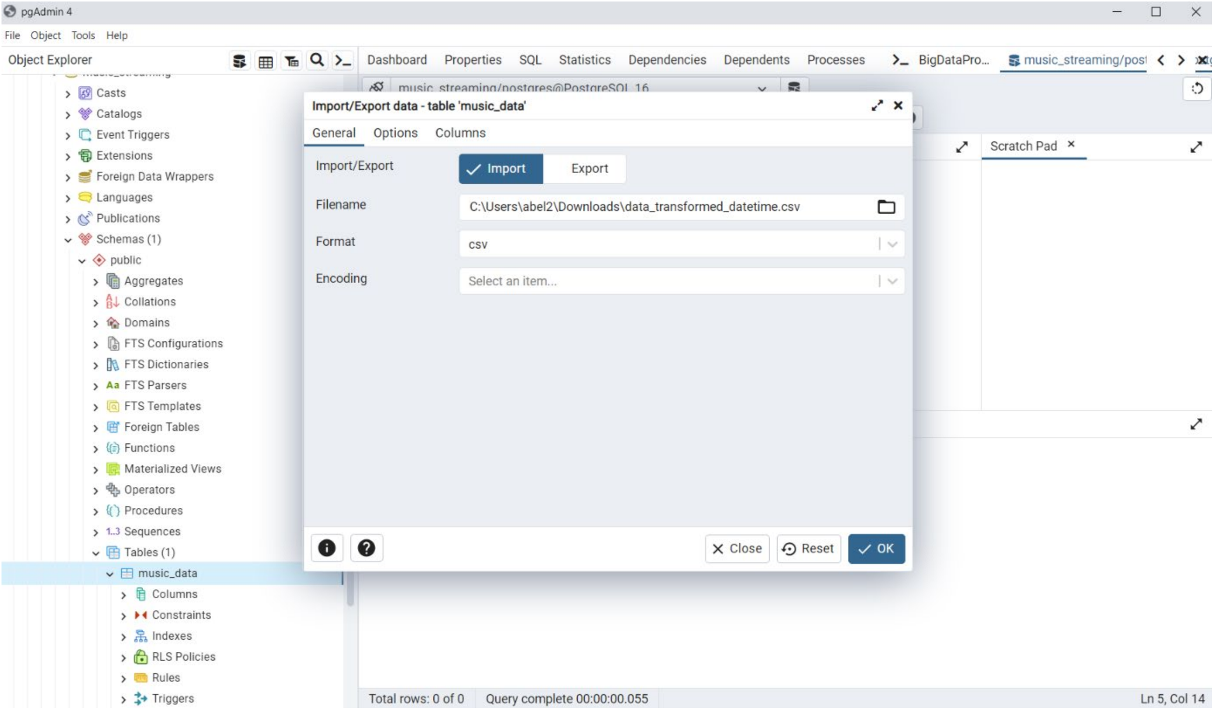
```sql
CREATE TABLE music_data (
    artist VARCHAR(255),
    album VARCHAR(255),
    track VARCHAR(255),
    date DATETIME,
    username VARCHAR(255)
);
```

3. Importing CSV Data into the Table

The CSV file generated at the end of the ETL process was imported into the PostgreSQL table. The following steps were followed:

1. Selecting the "Import" Option: We selected the "Import" option and specified the path to the transformed CSV file (data_transformed_datetime.csv).
2. Configuring the Parameters:
   - Delimiter: Comma (,), used as the column separator.
   - Header: Enabled (the first line contains the column names).
   - Encoding: UTF-8, to ensure proper character encoding.

This process allowed us to successfully import the transformed data into the PostgreSQL database for further analysis.



Import/Export data - table 'music_data'

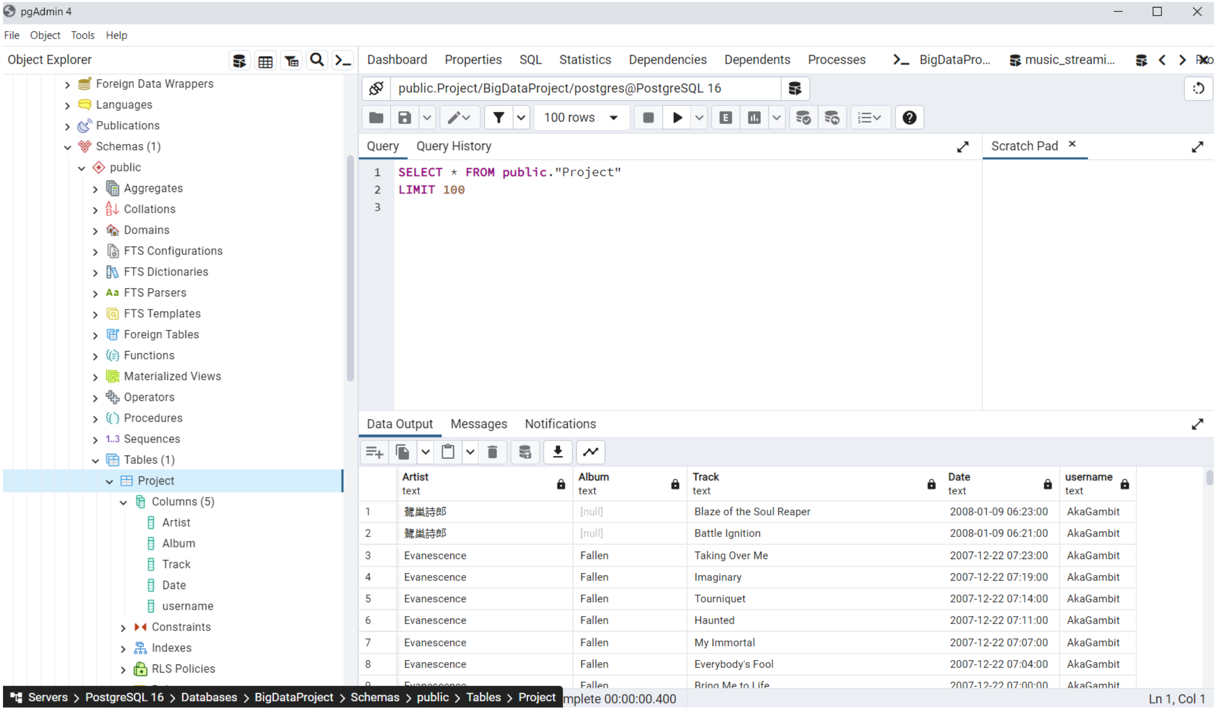
General · Options · Columns

- Import/Export: Import
- Filename: C:\Users\abel2\Downloads\data_transformed_datetime.csv
- Format: csv
- Encoding: Select an item...

Total rows: 0 of 0 · Query complete 00:00:00.055 · Ln 5, Col 14

4. Verifying the Imported Data

To ensure that the data was correctly imported, we executed a simple query:



```
SELECT * FROM public."Project"
LIMIT 100
```

	Artist text	Album text	Track text	Date text	username text
1	鷺巣詩郎	[null]	Blaze of the Soul Reaper	2008-01-09 06:23:00	AkaGambit
2	鷺巣詩郎	[null]	Battle Ignition	2008-01-09 06:21:00	AkaGambit
3	Evanescence	Fallen	Taking Over Me	2007-12-22 07:23:00	AkaGambit
4	Evanescence	Fallen	Imaginary	2007-12-22 07:19:00	AkaGambit
5	Evanescence	Fallen	Tourniquet	2007-12-22 07:14:00	AkaGambit
6	Evanescence	Fallen	Haunted	2007-12-22 07:11:00	AkaGambit
7	Evanescence	Fallen	My Immortal	2007-12-22 07:07:00	AkaGambit
8	Evanescence	Fallen	Everybody's Fool	2007-12-22 07:04:00	AkaGambit
9	Evanescence	Fallen	Bring Me to Life	2007-12-22 07:00:00	AkaGambit

Servers > PostgreSQL 16 > Databases > BigDataProject > Schemas > public > Tables > Project · mplete 00:00:00.400 · Ln 1, Col 1

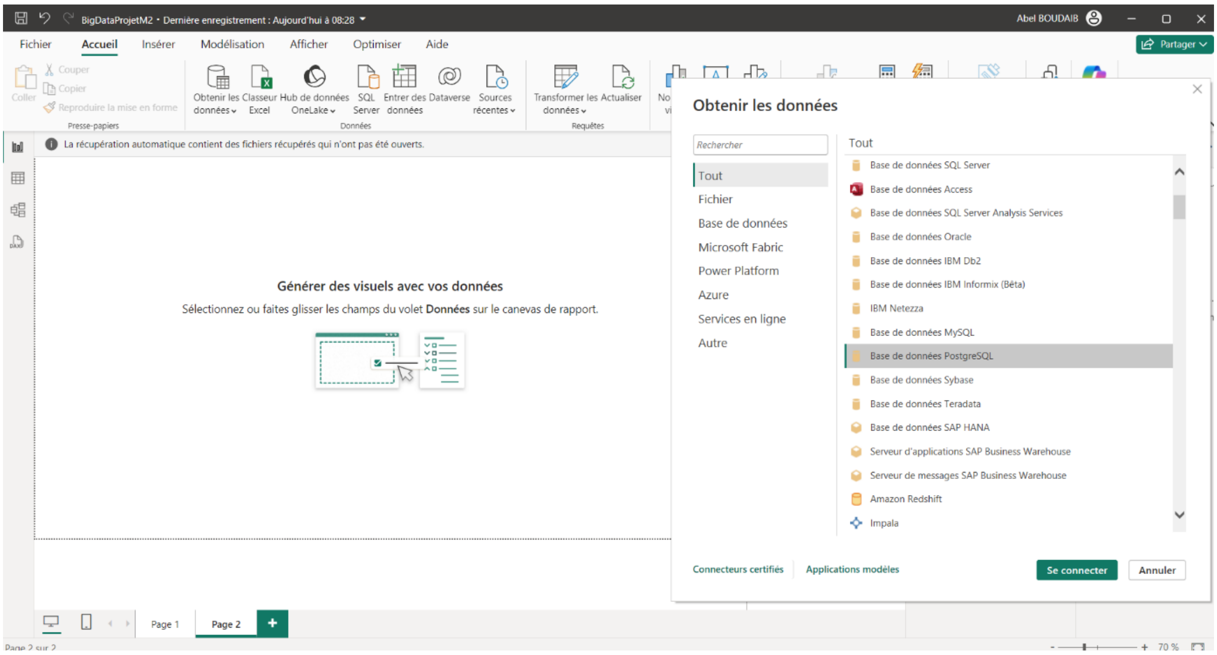
This query allows you to view the first 100 rows of the table and verify that the data matches the contents of the CSV file.

Chapter 3: Creating a dashboard with Power BI

In this section, we detail the steps to connect PostgreSQL data to Power BI, calculate KPIs, and create interactive visualizations in the form of charts and tables.

1. Connecting PostgreSQL data to Power BI

After launching Power BI, we connected to the PostgreSQL database. This allows Power BI to load data from the database in real time.



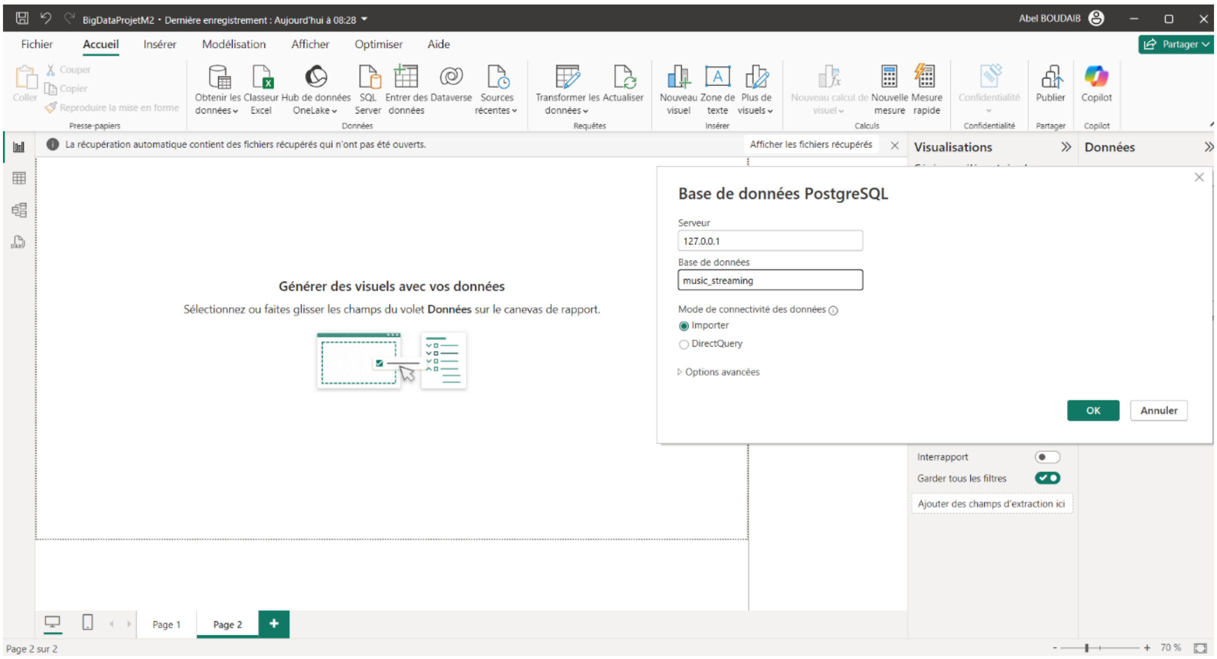
BigDataProjetM2 • Dernière enregistrement : Aujourd'hui à 08:28

La récupération automatique contient des fichiers récupérés qui n'ont pas été ouverts.

Générer des visuels avec vos données

Sélectionnez ou faites glisser les champs du volet **Données** sur le canevas de rapport.

Obtenir les données



BigDataProjetM2 • Dernière enregistrement : Aujourd'hui à 08:28

La récupération automatique contient des fichiers récupérés qui n'ont pas été ouverts.

Générer des visuels avec vos données

Sélectionnez ou faites glisser les champs du volet **Données** sur le canevas de rapport.

Base de données PostgreSQL

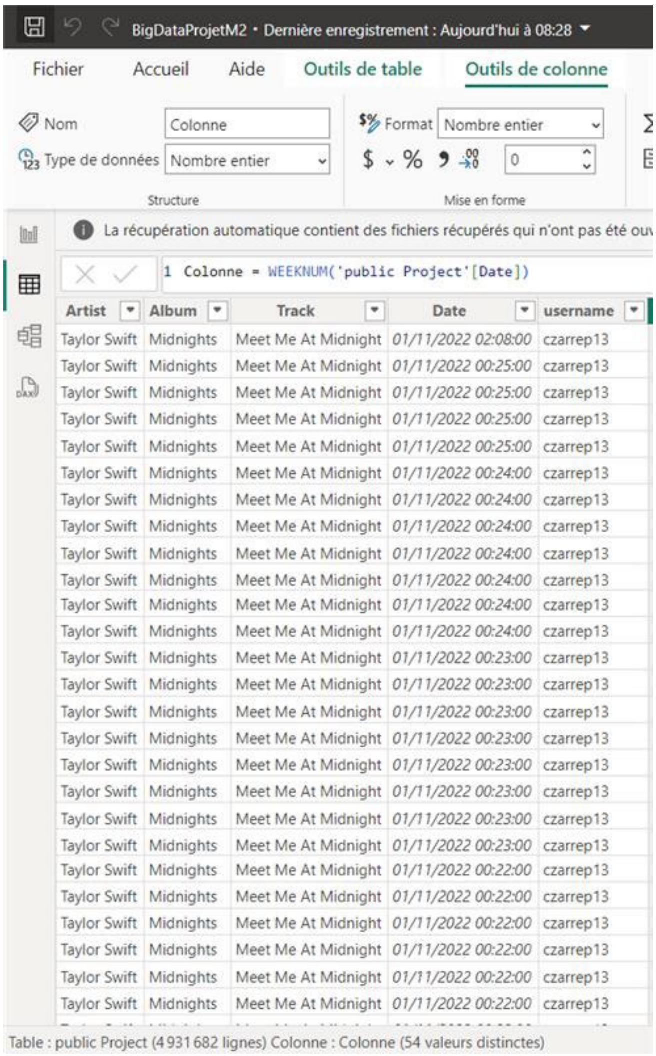
Serveur: 127.0.0.1

Base de données: music_streaming

Mode de connectivité des données

● Importer

○ DirectQuery



Nom: Colonne

Type de données: Nombre entier

Format: Nombre entier

$ % , 0

La récupération automatique contient des fichiers récupérés qui n'ont pas été ouv

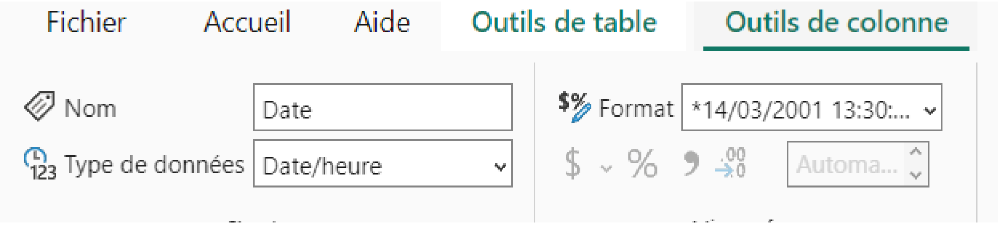
```
1 Colonne = WEEKNUM('public Project'[Date])
```

Artist	Album	Track	Date	username
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 02:08:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:25:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:25:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:25:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:25:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13

Table : public Project (4 931 682 lignes) Colonne : Colonne (54 valeurs distinctes)

2. Creating new columns for future analysis

For the next analysis that we need for our dashboard, we need to create other columns in order to make our dashboard more complete. First, let's set the type of the column "Date" as Date in power BI. That will help us to build a hierarchy of Date in our model.



Nom: Date

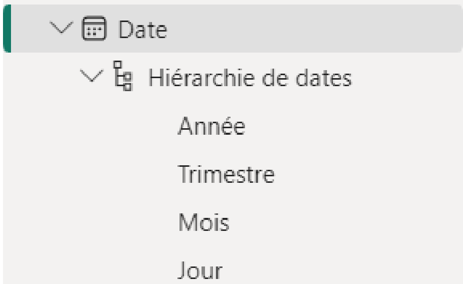
Type de données: Date/heure

Format: *14/03/2001 13:30:...

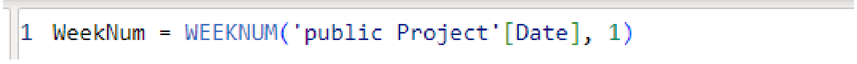
$ % , Automa...

When setting the type of our column, we can see that power BI set automatically a hierarchy in our dataset to split our date. Usually, Power BI split the date into 4 features: Year, Trimester, Months, Day.

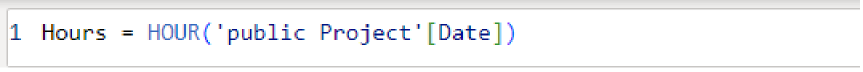
Here's the result:



But for our research we need more values to make the Date more accurate. So, we decided to count the week number of each year and the hours of each date for our dashboard using the DAX tool of power BI.

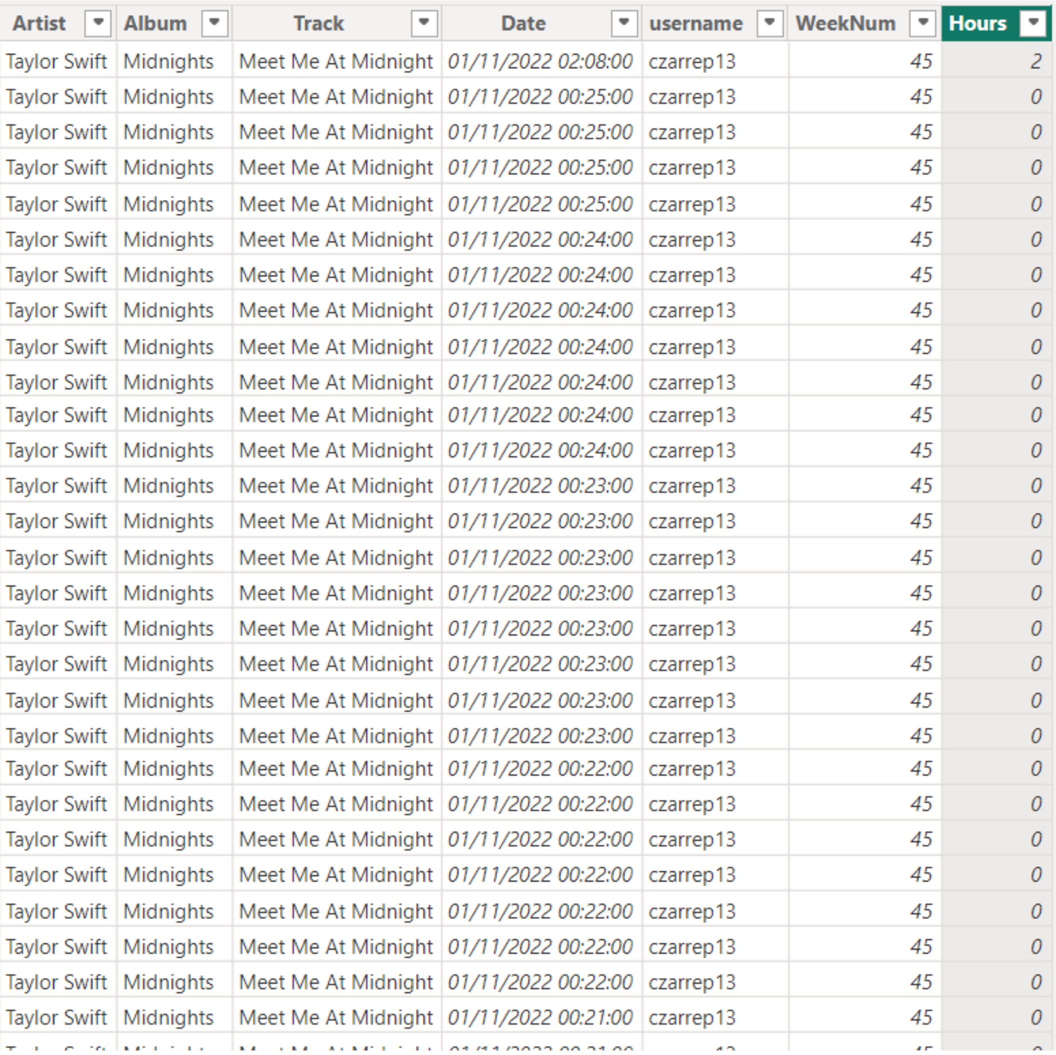


```
WeekNum = WEEKNUM('public Project'[Date], 1)
```



```
Hours = HOUR('public Project'[Date])
```

Using this tool, we can now add new columns for our dataset in order to start our analysis.



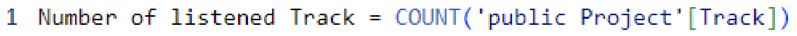
Artist	Album	Track	Date	username	WeekNum	Hours
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 02:08:00	czarrep13	45	2
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:25:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:25:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:25:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:25:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:24:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:23:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:22:00	czarrep13	45	0
Taylor Swift	Midnights	Meet Me At Midnight	01/11/2022 00:21:00	czarrep13	45	0

3. Calculating metrics

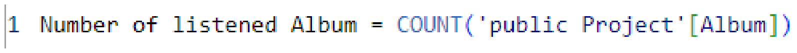
For our dashboard, we need also other data valuable for our research. Power BI has an amazing tool named metrics with which you can use DAX syntax to compute data and save it in your model.

In order to to complete our dashboard, we will need these metrics:

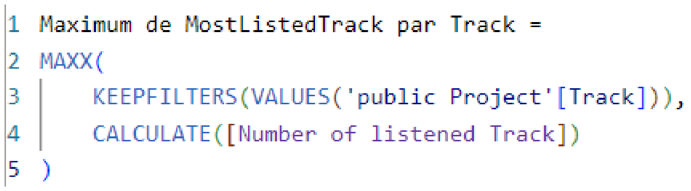
- Number of listened Track
- Number of listened Album
- Maximum of listened Track
- Maximum of listened Album
- Unique Track listened



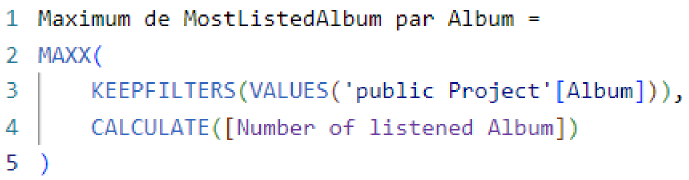
```
Number of listened Track = COUNT('public Project'[Track])
```



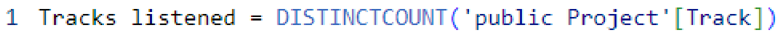
```
Number of listened Album = COUNT('public Project'[Album])
```



```
Maximum de MostListedTrack par Track =
MAXX(
    KEEPFILTERS(VALUES('public Project'[Track])),
    CALCULATE([Number of listened Track])
)
```



```
Maximum de MostListedAlbum par Album =
MAXX(
    KEEPFILTERS(VALUES('public Project'[Album])),
    CALCULATE([Number of listened Album])
)
```



```
Tracks listened = DISTINCTCOUNT('public Project'[Track])
```

4. Setting the Dashboard

Our dashboard is divided into 4 pages:

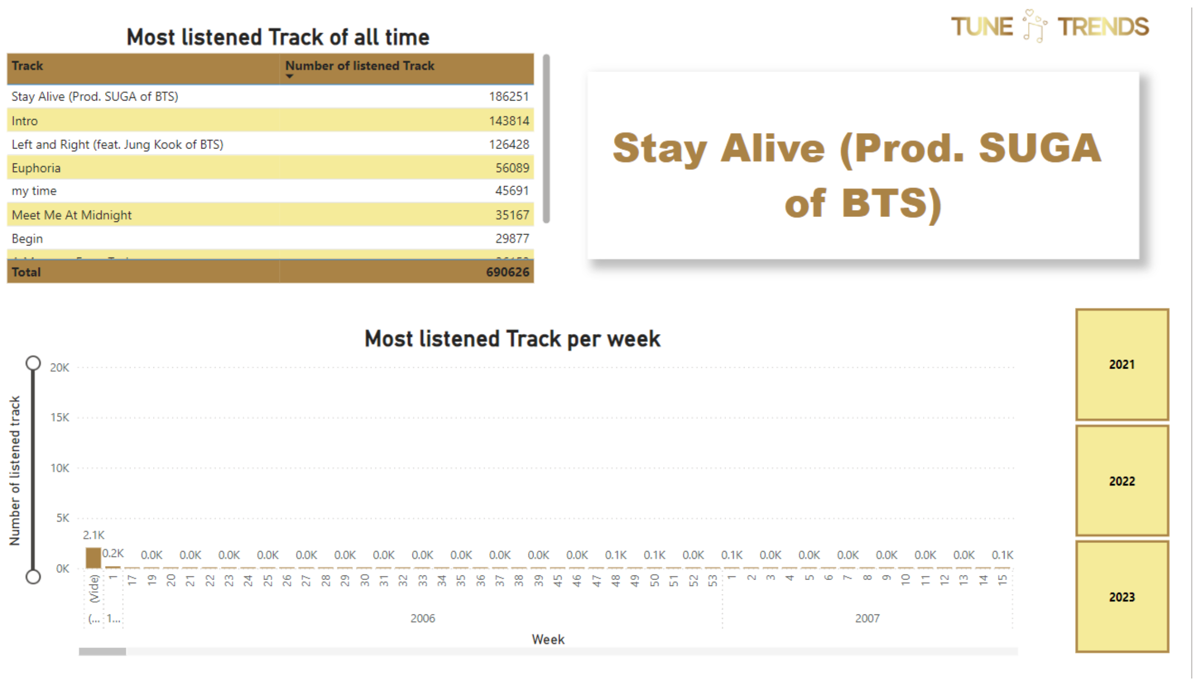
- A page dedicated to the track
- A page dedicated to the album
- A page dedicated to the user
- A page dedicated to more analysis

Track page

For our first page, we will make 3 graphs: A table, an histogram and a card that will be updated everytime we click on a value of the graphs.

The table listed all the track of our dataset and it is ordered by the number of listened.

For our histogram, the X line is the number of week for each years. We can also choose what year you want to focus on. The Y axis is the maximum value of listened track.



Most listened Track of all time

Track	Number of listened Track
Stay Alive (Prod. SUGA of BTS)	186251
Intro	143814
Left and Right (feat. Jung Kook of BTS)	126428
Euphoria	56089
my time	45691
Meet Me At Midnight	35167
Begin	29877
Total	**690626**

Stay Alive (Prod. SUGA of BTS)

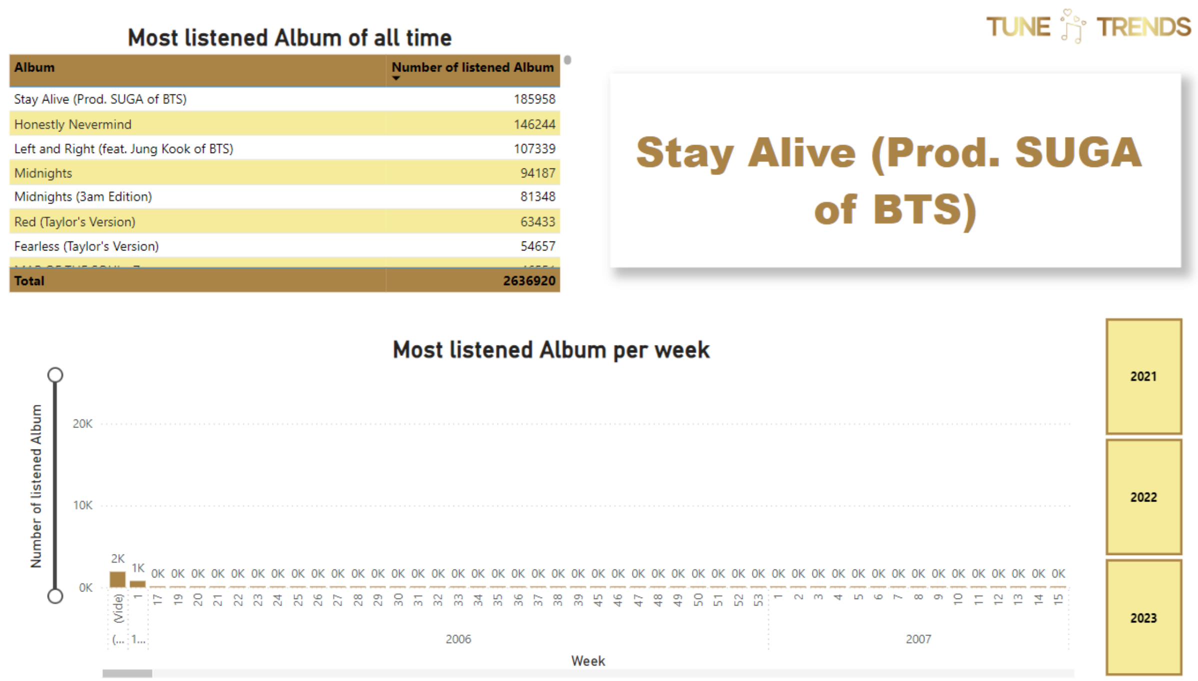
TUNE TRENDS
Most listened Track per week
Number of listened track
20K
15K
10K
5K
0K
2.1K
0.2K
0.0K 0.0K 0.0K 0.0K 0.0K 0.0K 0.0K 0.0K 0.0K 0.0K 0.0K 0.0K 0.1K 0.1K 0.0K 0.1K 0.0K 0.0K 0.0K 0.0K 0.0K 0.0K 0.1K
(Vide) 1 17 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 45 46 47 48 49 50 51 52 53 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
(... 1...
2006
2007
Week
2021
2022
2023

Album page

For our second page, we will make 3 graphs: A table, an histogram and a card that will be updated everytime we click on a value of the graphs.

The table listed all the album of our dataset and it is ordered by the number of listened.

For our histogram, the X line is the number of week for each years. We can also choose what year you want to focus on. The Y axis is the maximum value of listened Album.



Most listened Album of all time

Album	Number of listened Album
Stay Alive (Prod. SUGA of BTS)	185958
Honestly Nevermind	146244
Left and Right (feat. Jung Kook of BTS)	107339
Midnights	94187
Midnights (3am Edition)	81348
Red (Taylor's Version)	63433
Fearless (Taylor's Version)	54657
Total	**2636920**

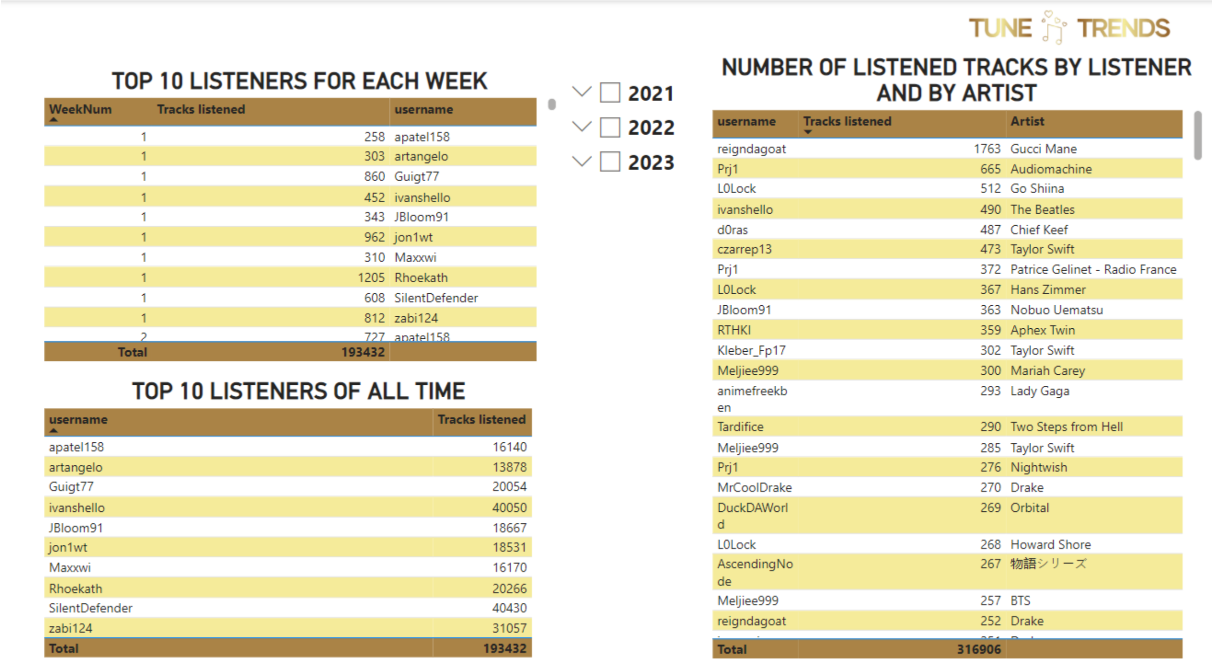
Stay Alive (Prod. SUGA of BTS)

TUNE TRENDS
Most listened Album per week
Number of listened Album
20K
10K
0K
2K
1K
0K
(Vide) 1 17 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 45 46 47 48 49 50 51 52 53 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
(... 1...
2006
2007
Week
2021
2022
2023

User page

For our user page, we will make 3 tables:

- Top 10 listeners of all time
- Number of listened track by listener and by artist
- Top 10 listeners for each week



TUNE TRENDS

TOP 10 LISTENERS FOR EACH WEEK

WeekNum	Tracks listened	username
1	258	apatel158
1	303	artangelo
1	860	Guigt77
1	452	ivanshello
1	343	JBloom91
1	962	jon1wt
1	310	Maxxwi
1	1205	Rhoekath
1	608	SilentDefender
1	812	zabi124
2	727	apatel158
Total	193432	

□ 2021
□ 2022
□ 2023

TOP 10 LISTENERS OF ALL TIME

username	Tracks listened
apatel158	16140
artangelo	13878
Guigt77	20054
ivanshello	40050
JBloom91	18667
jon1wt	18531
Maxxwi	16170
Rhoekath	20266
SilentDefender	40430
zabi124	31057
Total	193432

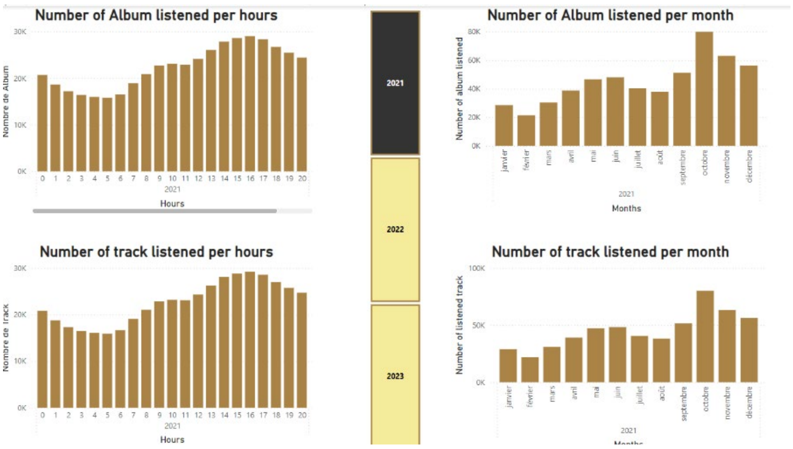
NUMBER OF LISTENED TRACKS BY LISTENER AND BY ARTIST

username	Tracks listened	Artist
reigndagoat	1763	Gucci Mane
Prj1	665	Audiomachine
L0Lock	512	Go Shiina
ivanshello	490	The Beatles
d0ras	487	Chief Keef
czarrep13	473	Taylor Swift
Prj1	372	Patrice Gelinet - Radio France
L0Lock	367	Hans Zimmer
JBloom91	363	Nobuo Uematsu
RTHKI	359	Aphex Twin
Kleber_Fp17	302	Taylor Swift
Meljiee999	300	Mariah Carey
animefreekben	293	Lady Gaga
Tardifice	290	Two Steps from Hell
Meljiee999	285	Taylor Swift
Prj1	276	Nightwish
MrCoolDrake	270	Drake
DuckDAWorld	269	Orbital
L0Lock	268	Howard Shore
AscendingNode	267	物語シリーズ
Meljiee999	257	BTS
reigndagoat	252	Drake
Total	316906	

Trend page

For our this page, we are interested in what time and month we have the most listened album and tracks.

Like the other pages, we can select what year we want to see.



Conclusion

The Applications of Big Data course has provided us with an invaluable opportunity to work on a comprehensive project, encompassing every critical stage of the data pipeline. From exploring raw data and performing ETL (Extract, Transform, Load) processes to integrating data into a PostgreSQL database for structured storage and analysis, we have gained hands-on experience with essential tools and methodologies. Additionally, designing a dynamic dashboard using Power BI allowed us to showcase actionable insights through effective visualizations, reinforcing the importance of storytelling with data.

This project has not only deepened our understanding of Big Data concepts but also strengthened our technical skills in data preparation, database management, and visualization. It has highlighted the challenges of working with real-world data and the satisfaction of deriving meaningful insights.

As a potential future enhancement, we would like to push the boundaries of this project by incorporating a real-time data management system. By leveraging tools such as Apache Airflow, we aim to automate ETL tasks and streamline the process of pushing data into the database. This addition would bring greater efficiency, scalability, and adaptability to our pipeline, further aligning the project with industry best practices.