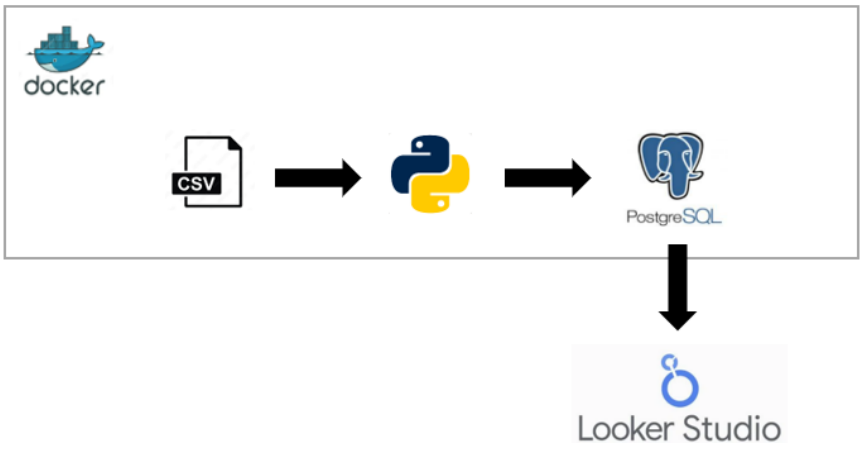
Batch Processing :

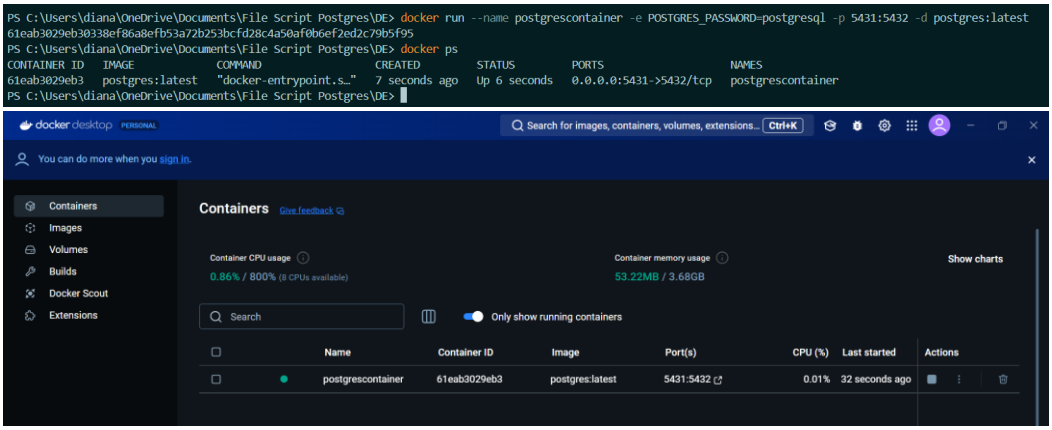
ETL CSV to PostgreSQL in Docker

1. Arstitektur

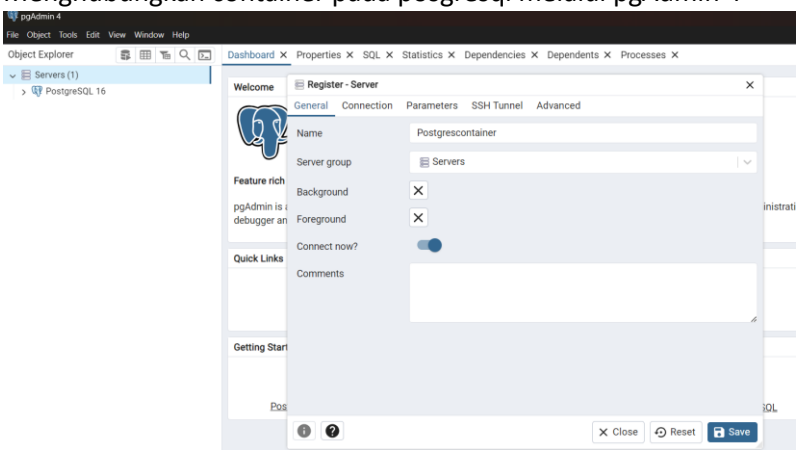


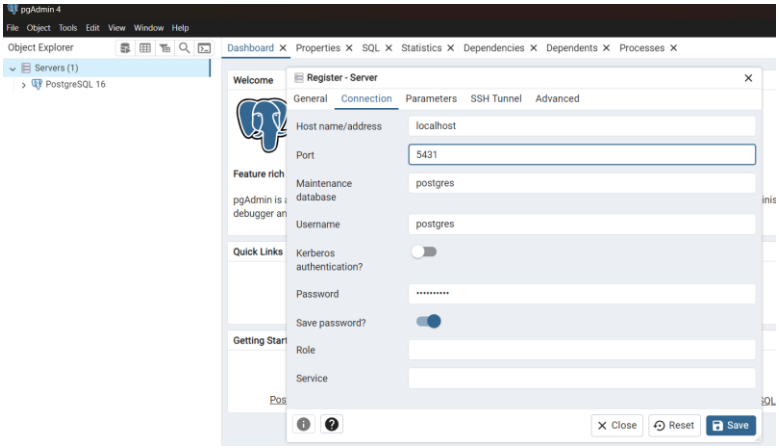
2. Dataset
   https://www.kaggle.com/datasets/ambaliyagati/spotify-dataset-for-playing-around-with-sql/data

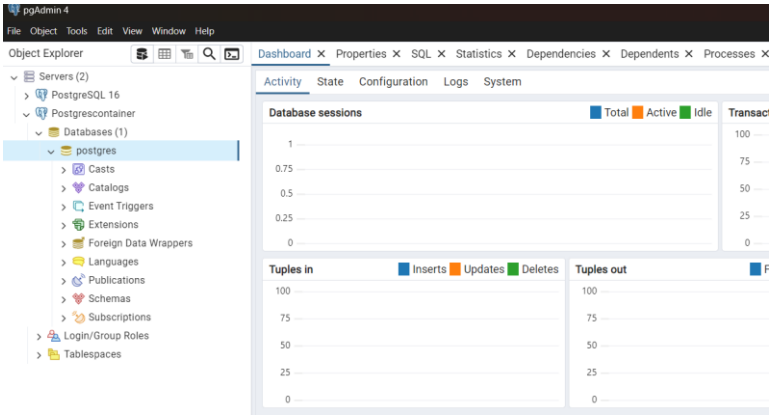
3. Proses
   a. Membuat dan menjalankan container postgresql pada docker



   b. Menghubungkan container pada posgresql melalui pgAdmin 4







c. Extract data csv

```
import os
import pandas as pd

# mengambil lokasi data
file_path = os.path.join(os.path.dirname(__file__), 'dataku', 'spotify_tracks.csv')
#print(f"Path ke file CSV: {file_path}")

# baca data
df = pd.read_csv(file_path)
print(df.head(10))
```

```
PS C:\Users\diana\OneDrive\Documents\File Script Postgres\DE> python -u "c:\Users\diana\OneDrive\Documents\File Script Postgres\DE\Batch\etl_proj
t\test.py"
                     id                                    name     genre  ... popularity duration_ms  explicit
0  7kr3xZk4yb3YSZ4VFtg2Qt                                Acoustic  acoustic  ...         58      172199     False
1  1kJygfS4eoVziBBI93MSYp                                Acoustic  acoustic  ...         57      172202     False
2  6lynns69p4zTCRxmmiSY1x               Here Comes the Sun - Acoustic  acoustic  ...         42      144786     False
3  1RC9slv335IfLce5vt9KTW                             Acoustic #3  acoustic  ...         46      116573     False
4  5o9L8xBuILoVjLECSBi7Vo  My Love Mine All Mine - Acoustic Instrumental  acoustic  ...         33      133922     False
5  742ZnC1OgUGIScdzXEy5ui                                Acoustic  acoustic  ...         14      238146     False
6  3vpJdk93GzerZnlou6Ua0z               Beautiful Things - Acoustic  acoustic  ...          0      201248     False
7  42qGA2116mkpSAaxzQfjEf                               Landslide  acoustic  ...         29      199222     False
8  0OHHcoyrWnkVukc911UeZd                                Acoustic  acoustic  ...         15      129250     False
9  64zEnxASl3epPAIH4MsmSw              Acoustic Energy Vibrations  acoustic  ...         45      118331     False

[10 rows x 8 columns]
```

Lihat keseluruhan kolom yang ada

```
Column data
Index(['id', 'name', 'genre', 'artists', 'album', 'popularity', 'duration_ms',
       'explicit'],
      dtype='object')
```

d. Transform data

Hapus kolom yang tidak perlu

```
# menghapus kolom
hapus_kolom = ['id','explicit']
df.drop(hapus_kolom, axis='columns', inplace=True)
print(df.columns)
```

```
Index(['name', 'genre', 'artists', 'album', 'popularity', 'duration_ms'], dtype='object')
```

simpan hasil transformasi data

```
# simpan data baru ke csv
df.to_csv('dataku/NEWspotify_tracks.csv', index=False)
print("Data berhasil disimpan")
```

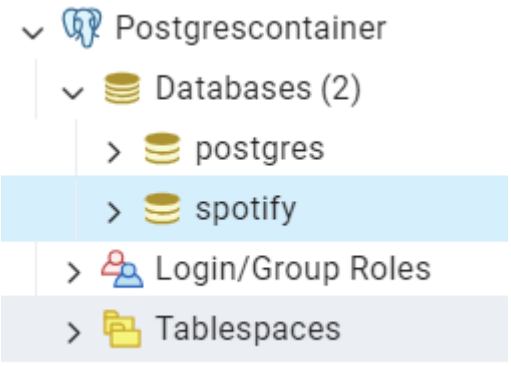
```
Data berhasil disimpan
```

e. Load data ke db postgresql

Buat database tujuan

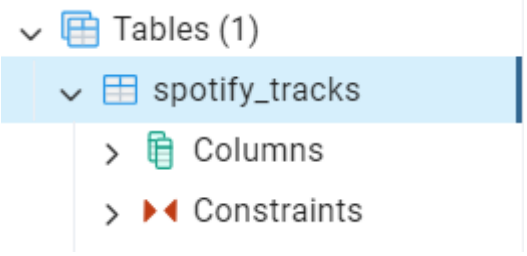
```
PS C:\Users\diana\OneDrive\Documents\File Script Postgres\DE> docker exec -it postgrescontainer psql -U postgres
psql (17.0 (Debian 17.0-1.pgdg120+1))
Type "help" for help.

postgres=#
```

```
postgres=# CREATE DATABASE spotify;
CREATE DATABASE
postgres=# \c spotify
```



Buat table tujuan

```
spotify=# CREATE TABLE spotify_tracks(id SERIAL PRIMARY KEY, name VARCHAR(225), genre VARCHAR(225), artists VARCHAR(225), album VARCHAR(225),
 popularity INT, duration_ms INT);
CREATE TABLE
spotify=#
```



Query Query History

```
1  select* from spotify_tracks
```

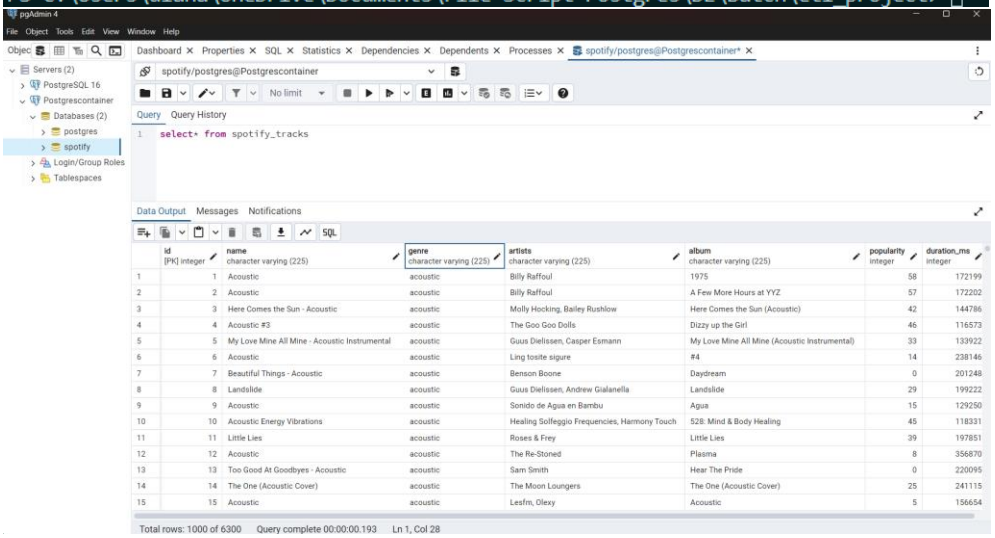
Data Output Messages Notifications

id [PK] integer	name character varying (225)	genre character varying (225)	artists character varying (225)	album character varying (225)	popularity integer	duration_ms integer

Input data ke postgresql

```
# Iterasi baris DataFrame dan memasukkan ke tabel PostgreSQL
for index, row in df.iterrows():
    cur.execute("INSERT INTO spotify_tracks (name, genre, artists, album, popularity, duration_ms) VALUES (%s,%s,%s,%s,%s,%s)",
                (row['name'],row['genre'], row['artists'], row['album'], row['popularity'], row['duration_ms']))

conn.commit()
cur.close()
conn.close()

print("Data berhasil dimuat ke PostgreSQL")
```

```
Data berhasil dimuat ke PostgreSQL
PS C:\Users\diana\OneDrive\Documents\File Script Postgres\DE\Batch\etl_project>
```



f. Analisis data menggunakan Looker studio

Melakukan analisis popularitas lagu spotify berdasarkan genre dan artisnya. Makin tinggi nilai popularitas maka makin baik.

Berdasarkan analisis, dapat diambil kesimpulan bahwa nilai popularitas suatu lagu dipengaruhi oleh genre dari lagu tersebut. Genre paling popular saat ini adalah genre ‘Rock’ dengan rata-rata nilai popularitas dengan genre tersebut adalah 60,34%.

https://lookerstudio.google.com/reporting/bd1a9c25-2f5f-4cec-b089-342df5c198cd

Spotify Popularity Insights

Analisis Popularitas Lagu Berdasarkan Genre dan Artis

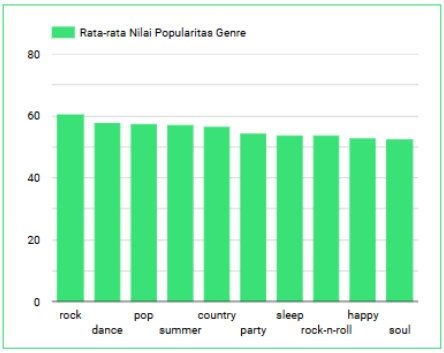
Top 10 Berdasarkan Nilai Popularitas Tertinggi

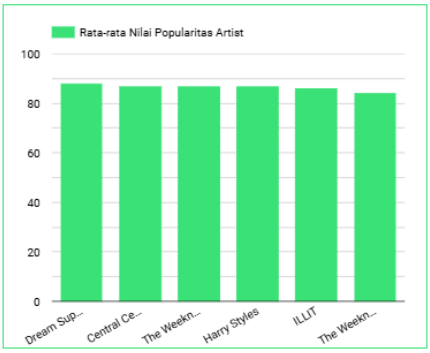
	Artis	Lagu	Album	Genre	Nilai Popula...
1.	Eminem	Houdini	Houdini	edm	90
2.	Dream Supplier, Baby Sleeps, Background White Noise	Clean Baby Sleep White Noise (Loopable)	Best White Noise For Sleeping Baby	sleep	88
3.	The Weeknd, JENNIE, Lily-Rose Depp	One Of The Girls (with JENNIE, Lily Rose Depp)	The Idol Episode 4 (Music from the HBO Original Series)	j-idol	87
4.	Harry Styles	As It Was	Harry's House	house	87
5.	Central Cee, Lil Baby	BAND4BAND (feat. Lil Baby)	BAND4BAND (feat. Lil Baby)	r-n-b	87
6.	ILLIT	Magnetic	SUPER REAL ME	k-pop	86
7.	Benson Boone	Beautiful Things	Fireworks & Rollerblades	r-n-b	86
8.	ILLIT	Magnetic	SUPER REAL ME	j-pop	86
9.	Benson Boone	Slow It Down	Fireworks & Rollerblades	r-n-b	85
10.	Zach Bryan	Pink Skies	Pink Skies	punk	85

Genre Berdasarkan Dengan Nilai Rata-Rata Popularitas

	Genre	Rata-rata Nilai Popularitas
1.	rock	60,34
2.	dance	57,52
3.	pop	57,32
4.	summer	57,06
5.	country	56,5
6.	party	54,22
7.	sleep	53,64
8.	rock-n-roll	53,54
9.	happy	52,82

1 - 100 / 126





Total Lagu Per-Genre
6.300

Nilai Popularitas Tertinggi
90

Nilai Popularitas Terendah
0

Rata-Rata Nilai Popularitas
30,75