Part3: OLAP Queries

Group Members:

Nirdosh Mishra: 2022OG04021

Rishabh Mishra: 2022OG04039

Github link:

https://github.com/nirdoshmishra/BDS_Assignment

Loom Video link:
https://www.loom.com/share/419bdbef10b443d080b06a9639780343?sid=f583c3a4-5d20-4d7a-9ce3-06184e69d716

Configuration: Used PySpark jobs on DataBricks Community Edition.

Query 1: Identify High-Potential Genres for Marketing and Promotion

```
# Cache the DataFrame for better performance
spotify_df.cache()

# Find the top genres based on the average popularity of tracks within each genre
high_potential_genres_query = (
    spotify_df
    .groupBy("track_genre")
    .agg({"popularity": "avg"})
    .orderBy(desc("avg(popularity)"))
    .limit(10)
)
high_potential_genres_query.show()
```

Output: (2) Spark Jobs

high_potential_genres_query:pyspark.sql.dataframe.DataFrame = [track_genre: string, avg(popularity): double]

```
+-----------+------------------+ |track_genre| avg(popularity)| +-----------+-----
-------------+ | pop-film|59.287575150300604| | k-pop| 56.896| | chill| 53.651| |
sad| 52.379| | grunge| 49.594| | indian| 49.539| | anime| 48.772| | emo| 48.128| |
sertanejo| 47.866| | pop| 47.576| +-----------+------------------+
Command took 3.23 seconds -- by nirdoshmishra2003@gmail.com at 25/12/2023, 8:27:06 pm on C2
```

Business value: Got the insight about the tracks which are more popular and listen by the users. These tracks are best suited for the promotional advertisements.

Query2: Identify Top Artists and Genres by Average Popularity

```
top_artists_genres_query = (
    spotify_df
    .groupBy("artists", "track_genre")
    .agg({"popularity": "avg"})
    .orderBy(desc("avg(popularity)"))
    .limit(10)
)
top_artists_genres_query.show()
```

Output: top_artists_genres_query:pyspark.sql.dataframe.DataFrame = [artists: string, track_genre: string ... 1 more field]

```
+--------------------+-----------+---------------+ | artists|track_genre|avg(popularity)| +--------------------+-----------+---------------+ |Sam Smith;Kim Petras| dance| 100.0| |Sam Smith;Kim Petras| pop| 100.0| | Bizarrap;Quevedo| hip-hop| 99.0| | Manuel Turizo| latin| 98.0| | Manuel Turizo| latino| 98.0| | Manuel Turizo| reggae| 98.0| | Manuel Turizo| reggaeton| 98.0| |Bad Bunny;Chencho...| latin| 97.0| |Bad Bunny;Chencho...| reggae| 97.0| |Bad Bunny;Chencho...| latino| 97.0| +--------------------+-----------+---------------+
Command took 3.32 seconds -- by nirdoshmishra2003@gmail.com at 25/12/2023, 8:27:06 pm on C2
```

Business value: It gives us the insight, which artist and tracks are more popular among the users. This insight may be helpful for the business promoter like artist to be approached. Focusing on these artists and genres in marketing and promotions could lead to increased user engagement.

Query3: Identify the Most Common Keys and Modes

Knowing the most common keys and modes in your dataset can help in creating playlists.

```
common_keys_modes_query = (
    spotify_df
    .groupBy("key", "mode")
    .count()
    .orderBy(desc("count"))
)
common_keys_modes_query.show(10)
```

Output: common_keys_modes_query:pyspark.sql.dataframe.DataFrame = [key: string, mode: string ... 1 more field]

```
+---+----+-----+ |key|mode|count| +---+----+-----+ | 0| 1|10179| | 7| 1|10130| | 2| 1| 9052| | 1| 1| 7164| | 9| 1| 6857| | 8| 1| 5436| | 5| 1| 5336| | 11| 0| 5145| | 4| 0| 4724| | 9| 0| 4453| +---+----+-----+ only showing top 10 rows
Command took 1.67 seconds -- by nirdoshmishra2003@gmail.com at 25/12/2023, 8:46:01 pm on C2
```

Business value: It helps us to get insight which key and mode align with popular music trends, contributing to user engagement and satisfaction.

More queries in the code.