OLAP Queries with PySpark

In [18]:

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col

# Initialize SparkSession
spark = SparkSession.builder.appName("SpotifyAnalytics").getOrCreate()
```

Example OLAP-style queries with specific column names

1. Aggregating by genre and calculating average popularity

In [19]:

```
# Load the CSV file into a DataFrame
df = spark.read.csv('train.csv', header=True, inferSchema=True)


df_avg_popularity_by_genre = df.groupBy("track_genre").agg({"popularity": "a
df_avg_popularity_by_genre.show()
```

```
+-----------------+------------------+
|      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|
|progressive-house|            46.615|
|            piano|            45.273|
|         mandopop|            45.025|
|       deep-house|            44.808|
|           brazil|             44.67|
|       electronic|            44.325|
|           pagode|            44.298|
|          ambient|            44.191|
|          british| 43.88264794383149|
|            metal|            43.705|
+-----------------+------------------+
only showing top 20 rows
```

2. Finding the top 10 tracks with the highest energy

In [20]:

```
df_top_energy_tracks = df.orderBy(col("energy").desc()).limit(10)
df_top_energy_tracks.show()
```

```
+----------+--------------------+--------------------+--------------------
+--------------------+--------------------+--------------------+----------
----------+------------+--------------------+------+--------+-----+-------
----+------------+----------------+--------+--------+------+--------------
+-----------+
|Unnamed: 0|            track_id|             artists|          album_name
|          track_name|          popularity|         duration_ms|
explicit|danceability|              energy|   key|loudness| mode|speechine
ss|acousticness|instrumentalness|liveness| valence| tempo|time_signature|t
rack_genre|
+----------+--------------------+--------------------+--------------------
+--------------------+--------------------+--------------------+----------
----------+------------+--------------------+------+--------+-----+-------
----+------------+----------------+--------+--------+------+--------------
+-----------+
|     13388|39XjBtONTw3TGoVN5...| Robert Owens;Atjazz|"Black Label #78 ...
|          Snuff Crew| Niedermeier & Wh...|              Si-Tew| Matthias
Heilbro...| John Gazoo"|Hearts And Soul -...|    10|  471050|False|
0.802|       0.753|             2.0|  -7.923|       1|0.0324|       0.0068
5|      0.576|
|     45024|48RME2XQqVXIaH54N...|         Estas Tonne|"""Strings and St...
|       Live in Odeon|        Vienna 2011"|The Song of the G...|
48|      549809|               False| 0.233|   0.825|    4|    -10.095|
0|          0.0415|    0.68|   0.946| 0.565|         0.552|      89.01|
|     76649|7nvdj8uWalevHaOBX...|Giuseppe Verdi;Ri...|         Verdi: Aida
|"Verdi: Aida, Act...| sul crin ti piov...|           Amneris)"|
22|       97093|               False| 0.232|  0.0847|    7|    -22.243|
0|          0.0414|   0.986| 0.00011| 0.343|         0.043|     89.912|
|     16842|2HyJhQXLIyysNuz89...|Wolfgang Amadeus ...|Mozart - All Day ...
|"12 Variations in...| vous dirai-je Ma...| K. 265: 9. Varia...|
15|       58360|               False| 0.435|   0.094|    7|    -25.717|
1|          0.0451|   0.995|   0.915| 0.181|         0.818|    105.188|
|     76547|4aAeGUBWayg0PVqW3...|Giuseppe Verdi;Ri...|         Verdi: Aida
|"Verdi: Aida, Act...|              Ramfis|         Sacerdoti)"|
23|      131706|               False| 0.148| 0.00377|    3|    -40.046|
1|          0.0384|   0.986|   0.447|0.0928|         0.239|     81.078|
|     16173|5aGhOb4iwTbfrgxQG...|Franz Schubert;Mi...|Classical Christm...
|"Ave Maria, ""Ell...|          Op.52 No.6| D.839 (Arr. Joha...|
0|      335666|               False|0.0918|  0.0325|    8|    -26.707|
1|          0.0494|   0.995|   0.864| 0.109|        0.0393|     68.958|
|     76826|16QKb6qf3WlmgsSDV...|Charles Lecocq;Ko...|Ballet Class Musi...
|"Pirouettes in A-...|               Act 1|      Allegro-valse"|
21|      105226|               False| 0.365|   0.109|    9|    -20.241|
1|          0.0511|   0.994|   0.948|0.0968|         0.404|     60.015|
|     16239|1WFKR1UoO2aHEYnCU...|Franz Schubert;Mi...|Weihnachten Klass...
|"Ave Maria, ""Ell...|          Op.52 No.6| D.839 (Arr. Joha...|
0|      335666|               False|0.0918|  0.0325|    8|    -26.707|
1|          0.0494|   0.995|   0.864| 0.109|        0.0393|     68.958|
|     76346|7x56g3SKfKeE4XRg9...|Giuseppe Verdi;Ca...|40 Most Beautiful...
|"Verdi: La travia...|            Violetta|           Gastone)"|
25|      189239|               False| 0.324|   0.149|    5|    -18.318|
1|           0.051|   0.976|1.33e-05| 0.136|         0.154|     95.073|
|     16672|4uaRUpAdVcjF5ME0Q...|Wolfgang Amadeus ...|Mozart: A Night o...
|"12 Variations in...| vous dirai-je Ma...| K. 265: 8. Varia...|
25|       43586|               False| 0.367|   0.175|    0|    -19.716|
1|          0.0428|   0.974|   0.853| 0.232|         0.816|     74.077|
+----------+--------------------+--------------------+--------------------
+--------------------+--------------------+--------------------+----------
----------+------------+--------------------+------+--------+-----+-------
----+------------+----------------+--------+--------+------+--------------
```

```
+-----------+
```

3. Calculating genre distribution of explicit vs. non-explicit tracks

In [21]:

```
df_genre_explicit_distribution = df.groupBy("track_genre", "explicit").count
df_genre_explicit_distribution.show()
```

```
+-----------+--------+-----+
|track_genre|explicit|count|
+-----------+--------+-----+
|          4|   46741|    1|
|death-metal|   False|  749|
| deep-house|   False|  975|
|      dance|    True|  174|
|   cantopop|   False|  998|
|          3|  459360|    6|
|    114.211|      11|    1|
|          4|  226626|    2|
|    country|   False|  970|
|death-metal|    True|  251|
|          4|  320173|    3|
|          3|   52202|    1|
|          4|  262306|    4|
|     74.077|      25|    1|
|    148.759|       6|    2|
|  bluegrass|    True|    5|
|          4|  507146|    1|
|  dancehall|    True|  302|
|          4|  235253|    1|
|          4|   60026|    2|
+-----------+--------+-----+
only showing top 20 rows
```

In [ ]: