databricks mongoDb OLAP queries

(https://databricks.com)

```
# BDS Assignment - vivek dahiya 2022OG04043
# Loom Video Link https://www.loom.com/share/e03002e1df0c40968457a071582a6a93?sid=0d7e9e59-0110-4fdb-8fa2-250ec88036ad
# Git https://github.com/InsightfulSage/BDS-Assignment

# Install Libraries..
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("MyApp").getOrCreate()

# Connect to Mongo DB
uri = "mongodb+srv://vivek29dahiya:mMnSweoIsEBNlyHg@vivekdahiya.arn7gly.mongodb.net/?retryWrites=true&w=majority"
spark = (
    SparkSession.builder.appName("MongoDB Spark Connector")
    .config(
        "spark.mongodb.input.uri",
        "mongodb+srv://vivek29dahiya:mMnSweoIsEBNlyHg@vivekdahiya.arn7gly.mongodb.net/?retryWrites=true&w=majority/myDatabase.spotify"
    )
    .config(
        "spark.mongodb.output.uri",
        "mongodb+srv://vivek29dahiya:mMnSweoIsEBNlyHg@vivekdahiya.arn7gly.mongodb.net/?retryWrites=true&w=majority/myDatabase.spotify"
    )
    .getOrCreate()
)

# Load data from Mongo Db
df = (
    spark.read.format("mongo")
    .option(
        "uri",
        "mongodb+srv://vivek29dahiya:mMnSweoIsEBNlyHg@vivekdahiya.arn7gly.mongodb.net/myDatabase.spotify?retryWrites=true&w=majority",
    )
    .load()
)

display(df.limit(5))
```

Table

	_id	acousticness	album_name	artists
1	{"oid": "65881bce107e381a2cd76ba2"}	0.0322	Comedy	Gen Hoshino
2	{"oid": "65881bce107e381a2cd76ba3"}	0.924	Ghost (Acoustic)	Ben Woodward
3	{"oid": "65881bce107e381a2cd76ba4"}	0.21	To Begin Again	Ingrid Michaelson;ZAYN
4	{"oid": "65881bce107e381a2cd76ba5"}	0.905	Crazy Rich Asians (Original Motion Picture Soundtrack)	Kina Grannis
5	{"oid": "65881bce107e381a2cd76ba6"}	0.469	Hold On	Chord Overstreet

5 rows

```
import pandas as pd
import numpy as np
df.printSchema()
```

```
root
 |-- _id: struct (nullable = true)
 |    |-- oid: string (nullable = true)
 |-- acousticness: double (nullable = true)
 |-- album_name: string (nullable = true)
 |-- artists: string (nullable = true)
 |-- danceability: double (nullable = true)
 |-- duration_ms: integer (nullable = true)
 |-- energy: double (nullable = true)
 |-- explicit: boolean (nullable = true)
```

```
 |-- instrumentalness: double (nullable = true)
 |-- key: integer (nullable = true)
 |-- liveness: double (nullable = true)
 |-- loudness: double (nullable = true)
 |-- mode: integer (nullable = true)
 |-- number: integer (nullable = true)
 |-- popularity: integer (nullable = true)
 |-- speechiness: double (nullable = true)
 |-- tempo: double (nullable = true)
 |-- time_signature: integer (nullable = true)
```

```
  df.describe().show()
```

```
+-------+------------------+---------------------------+---------+------------------+------------------+--------------------+---
---------------+-----------------+-------------------+------------------+------------------+-----------------+-
-----------------+------------------+------------------+----------+--------------------+-------------+------------------+
|summary|       acousticness|                 album_name|  artists|       danceability|       duration_ms|              energy|  i
nstrumentalness|              key|           liveness|          loudness|              mode|            number|       popularity|
speechiness|             tempo|     time_signature|track_genre|            track_id|    track_name|            valence|
+-------+------------------+---------------------------+---------+------------------+------------------+--------------------+---
---------------+-----------------+-------------------+------------------+------------------+-----------------+-
-----------------+------------------+------------------+----------+--------------------+-------------+------------------+
|  count|            113986|                     113986|   113986|            113986|            113986|              113986|
113986|           113986|             113986|            113986|            113986|           113986|
113986|            113986|             113986|     113986|              113986|        113986|             113986|
|   mean|0.31493895007035494|                        NaN|      NaN| 0.5667923999438398|228024.92559612583| 0.6413529245451236|0.1
560573085745602|5.309353780288808| 0.2135498105907735|-8.259232809292476|0.6375695260821505| 56996.95359956486|33.23840647097012|
0.08465498219079524| 122.1443550436048| 3.9040233011071535|       NULL|               NULL|           NaN|0.47403706885934116|
| stddev| 0.3325314866928782|                        NaN|      NaN|0.17354992568338184|  107301.378249364|0.25152659618372847|0.3
095692747805911|3.560016352152802|0.19038555522653905| 5.029580189728978|0.4807043298439006|32909.925217888194|22.30638959860173|
0.1057377015966203|29.976770081929978|0.43264610810578213|       NULL|               NULL|           NaN| 0.2592540834338674|
|    min|               0.0|        ! ! ! ! ! Whisper...|   !nvite|               0.0|                 0|                 0.0|
0.0|              0|               0.0|          -49.531|                 0|                 0|                 0|
0.0|              0.0|                  0|   acoustic|0000vdREvCVMxbQTk...|!I'll Be Back!|               0.0|
```

```
  # Find Artists with most songs
  from pyspark.sql import functions as F
  print('Artists with mosr songs...')
  top_df  = df.groupBy("artists").count().orderBy(F.desc('count'))
  top_df.show(10)
```

```
Artists with mosr songs...
+---------------+-----+
|        artists|count|
+---------------+-----+
|    The Beatles|  279|
|   George Jones|  271|
|  Stevie Wonder|  236|
|    Linkin Park|  224|
|Ella Fitzgerald|  222|
|  Prateek Kuhad|  217|
|           Feid|  202|
|    Chuck Berry|  190|
|Håkan Hellström|  183|
|    OneRepublic|  181|
+---------------+-----+
only showing top 10 rows
```

```
  # Most Popular Albums
  from pyspark.sql import functions as F
  top_df  = df.groupBy('album_name').agg(F.count('*').alias('Songs'), F.avg('popularity').alias('averagePopularity')).orderBy(F.desc('av
  print('Most Popular Albums..')
  top_df.show(10)
```

```
Most Popular Albums..
+--------------------+-----+-----------------+
|          album_name|Songs|averagePopularity|
```

```
+--------------------+-----+-----------------+
|Unholy (feat. Kim...|    2|            100.0|
|Quevedo: Bzrp Mus...|    1|             99.0|
|          La Bachata|    4|             98.0|
|I Ain’t Worried (...|    3|             96.0|
|   Indigo (Extended)|    2|             96.0|
|            PROVENZA|    2|             93.0|
|         RENAISSANCE|    1|             93.0|
|   Super Freaky Girl|    2|             92.0|
|Left and Right (F...|    2|             92.0|
|Calm Down (with S...|    1|             92.0|
+--------------------+-----+-----------------+
only showing top 10 rows
```

```
  # Popularity of artists with most songs
  top_df = df.groupBy('artists').agg(F.count('*').alias('Songs'), F.avg('popularity').alias('averagePopularity')).orderBy(F.desc('Songs'
  print('Most Popular Artists..')
  top_df.show(10)
```

```
Most Popular Artists..
+---------------+-----+------------------+
|        artists|Songs| averagePopularity|
+---------------+-----+------------------+
|    The Beatles|  279|61.007168458781365|
|   George Jones|  271|16.505535055350553|
|  Stevie Wonder|  236|1.0635593220338984|
|    Linkin Park|  224| 56.07142857142857|
|Ella Fitzgerald|  222|0.7342342342342343|
|  Prateek Kuhad|  217| 46.33179723502304|
|           Feid|  202|10.084158415841584|
|    Chuck Berry|  190| 7.873684210526315|
|Håkan Hellström|  183| 31.48633879781421|
|    OneRepublic|  181|30.861878453038674|
+---------------+-----+------------------+
only showing top 10 rows
```

```
TOP 10 dance able songs
+--------------------+--------------------+--------------------+--------------+
|          track_name|             artists|          album_name|   track_genre|
+--------------------+--------------------+--------------------+--------------+
|            Sol Clap|             Quantic| The Best of Quantic|      trip-hop|
|       Medicaid Baby|   That Girl Lay Lay|Tha Cheat Code Re...|          kids|
|         Inspiration|        Delano Smith|          An Odyssey|detroit-techno|
|      Daily Routines|     Oliver Schories|Fields Without Fe...|minimal-techno|
|             Bitches|             dj funk|Dance Mania: Ghet...| chicago-house|
|Featuring Mixx Ma...|     Mixx Master Lee|The Mississippi C...|          kids|
|  Dancing in My Room|            347aidan|  Dancing in My Room|           sad|
|   Plastik Fantastik|   Felix Da Housecat|         He Was King| chicago-house|
|The Underground -...|DJ Pierre;My Digi...| Mono:Disko, Vol. 12| chicago-house|
|The Soccer Song (...|           CoComelon|CoComelon Kids Hi...|      children|
+--------------------+--------------------+--------------------+--------------+
```