```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
# start_time = time.time()

# end_time = time.time()
# execution_time = end_time - start_time
# print(execution_time)
```

1 :

Installing pyspark module

```
!pip install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.5.1.tar.gz (317.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.0/317.0 MB 3.6 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488491 sha256=c66931c9e14ced449a473c9e92b8e89bb747
  Stored in directory: /root/.cache/pip/wheels/80/1d/60/2c256ed38dddce2fdd93be545214a63e02fbd8d74fb0b7f3a6
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.1
```

Importing the modules

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import count, desc , col, max
import matplotlib.pyplot as plts
import time
```

creating spark session

```
spark = SparkSession.builder.appName('spark_app').getOrCreate()
```

2 :

importing the *Listenings.csv* file:

```
start_time = time.time()

listening_csv_path = '/content/drive/MyDrive/dataset/listenings.csv'
listening_df = spark.read.format('csv').option('inferSchema',True).option('header', True).load(listening_csv_path)

end_time = time.time()
execution_time = end_time - start_time
print(execution_time)
```

```
39.95333695411682
```

let's check the data:

```
start_time = time.time()
listening_df.show()
end_time = time.time()
execution_time = end_time - start_time
print(execution_time)
```

```
+-----------+-------------+--------------------+--------------+--------------------+
|    user_id|         date|               track|        artist|               album|
+-----------+-------------+--------------------+--------------+--------------------+
|000Silenced|1299680100000|           Price Tag|       Jessie J|         Who You Are|
|000Silenced|1299679920000|Price Tag (Acoust...|       Jessie J|           Price Tag|
|000Silenced|1299679440000|Be Mine! (Ballad ...|         Robyn|            Be Mine!|
```

```
|000Silenced|1299679200000|            Acapella|          Kelis|            Acapella|
|000Silenced|1299675660000|   I'm Not Invisible|      The Tease|   I'm Not Invisible|
|000Silenced|1297511400000|Bounce (Feat NORE...|       MSTRKRFT|         Fist of God|
|000Silenced|1294498440000|Don't Stop The Mu...|        Rihanna|Addicted 2 Bassli...|
|000Silenced|1292438340000|               ObZen|      Meshuggah|               ObZen|
|000Silenced|1292437740000|   Yama's Messengers|         Gojira|The Way of All Flesh|
|000Silenced|1292436360000|On the Brink of E...|   Napalm Death|Time Waits For No...|
|000Silenced|1292436360000|On the Brink of E...|   Napalm Death|Time Waits For No...|
|000Silenced|1292435940000|        In Deference|   Napalm Death|      Smear Campaign|
|000Silenced|1292434920000|      Post(?)organic|    Decapitated|Organic Hallucinosis|
|000Silenced|1292434560000|        Mind Feeders|   Dom & Roland| No Strings Attached|
|000Silenced|1292434320000|Necrosadistic War...|Cannibal Corpse|               Kill|
|000Silenced|1292365560000|     Dance All Night|   Dom & Roland|          Chronology|
|000Silenced|1292365260000|          Late Night|   Dom & Roland|          Chronology|
|000Silenced|1292365020000|          Freak Seen|   Dom & Roland|          Chronology|
|000Silenced|1292364720000|Paradrenasite (Hi...|   Dom & Roland|          Chronology|
|000Silenced|1292364300000|               Rhino|   Dom & Roland|          Chronology|
+-----------+-------------+--------------------+---------------+--------------------+
only showing top 20 rows

0.13406062126159668
```

let's delete useless columns:

```
listening_df = listening_df.drop('date')
```

drop the null rows:

```
start_time = time.time()

listening_df = listening_df.na.drop()

end_time = time.time()
execution_time = end_time - start_time
print(execution_time)
```

```
0.02201557159423828
```

let's check the dataset again:

```
start_time = time.time()

listening_df.show()

end_time = time.time()
execution_time = end_time - start_time
print(execution_time)
```

```
+-----------+-------------+--------------------+---------------+--------------------+
|    user_id|         date|               track|         artist|               album|
+-----------+-------------+--------------------+---------------+--------------------+
|000Silenced|1299680100000|           Price Tag|       Jessie J|         Who You Are|
|000Silenced|1299679920000|Price Tag (Acoust...|       Jessie J|           Price Tag|
|000Silenced|1299679440000|Be Mine! (Ballad ...|          Robyn|            Be Mine!|
|000Silenced|1299679200000|            Acapella|          Kelis|            Acapella|
|000Silenced|1299675660000|   I'm Not Invisible|      The Tease|   I'm Not Invisible|
|000Silenced|1297511400000|Bounce (Feat NORE...|       MSTRKRFT|         Fist of God|
|000Silenced|1294498440000|Don't Stop The Mu...|        Rihanna|Addicted 2 Bassli...|
|000Silenced|1292438340000|               ObZen|      Meshuggah|               ObZen|
|000Silenced|1292437740000|   Yama's Messengers|         Gojira|The Way of All Flesh|
|000Silenced|1292436360000|On the Brink of E...|   Napalm Death|Time Waits For No...|
|000Silenced|1292436360000|On the Brink of E...|   Napalm Death|Time Waits For No...|
|000Silenced|1292435940000|        In Deference|   Napalm Death|      Smear Campaign|
|000Silenced|1292434920000|      Post(?)organic|    Decapitated|Organic Hallucinosis|
|000Silenced|1292434560000|        Mind Feeders|   Dom & Roland| No Strings Attached|
|000Silenced|1292434320000|Necrosadistic War...|Cannibal Corpse|               Kill|
|000Silenced|1292365560000|     Dance All Night|   Dom & Roland|          Chronology|
|000Silenced|1292365260000|          Late Night|   Dom & Roland|          Chronology|
|000Silenced|1292365020000|          Freak Seen|   Dom & Roland|          Chronology|
|000Silenced|1292364720000|Paradrenasite (Hi...|   Dom & Roland|          Chronology|
|000Silenced|1292364300000|               Rhino|   Dom & Roland|          Chronology|
+-----------+-------------+--------------------+---------------+--------------------+
only showing top 20 rows

0.17213749885559082
```

let's see the schema:

```
listening_df.printSchema()
```

```
root
 |-- user_id: string (nullable = true)
 |-- track: string (nullable = true)
 |-- artist: string (nullable = true)
 |-- album: string (nullable = true)
```

let's see the shape of our dataframe:

```
start_time = time.time()

shape = (listening_df.count() , len(listening_df.columns))
print(shape)

end_time = time.time()
execution_time = end_time - start_time
print(execution_time)
```

```
(13758905, 5)
47.91310524940491
```

3:

select two columns: track and artist

```
start_time = time.time()

q0 = listening_df.select('artist' , 'track')
q0.show()

end_time = time.time()
execution_time = end_time - start_time
print(execution_time)
```

```
+---------------+--------------------+
|         artist|               track|
+---------------+--------------------+
|       Jessie J|           Price Tag|
|       Jessie J|Price Tag (Acoust...|
|          Robyn|Be Mine! (Ballad ...|
|          Kelis|            Acapella|
|      The Tease|   I'm Not Invisible|
|       MSTRKRFT|Bounce (Feat NORE...|
|        Rihanna|Don't Stop The Mu...|
|      Meshuggah|               ObZen|
|         Gojira|   Yama's Messengers|
|   Napalm Death|On the Brink of E...|
|   Napalm Death|On the Brink of E...|
|   Napalm Death|        In Deference|
|    Decapitated|      Post(?)organic|
|   Dom & Roland|        Mind Feeders|
|Cannibal Corpse|Necrosadistic War...|
|   Dom & Roland|     Dance All Night|
|   Dom & Roland|          Late Night|
|   Dom & Roland|          Freak Seen|
|   Dom & Roland|Paradrenasite (Hi...|
|   Dom & Roland|               Rhino|
+---------------+--------------------+
only showing top 20 rows

0.3125152587890625
```

Let's find all of the records of those users who have listened to ***Rihanna***

```
start_time = time.time()

q1 = listening_df.select('*').filter(listening_df.artist == 'Rihanna')
q1.show()

end_time = time.time()
execution_time = end_time - start_time
print(execution_time)
```

```
+-----------+-------------+--------------------+-------+--------------------+
|    user_id|         date|               track| artist|               album|
+-----------+-------------+--------------------+-------+--------------------+
|000Silenced|1294498440000|Don't Stop The Mu...|Rihanna|Addicted 2 Bassli...|
|000Silenced|1285438440000|           Disturbia|Rihanna|Good Girl Gone Ba...|
|00williamsl|1361485800000|   Hatin On The Club|Rihanna|              Random|
```

```
|00williamsl|1361485800000|   Hatin On The Club|Rihanna|              Random|
|00williamsl|1361048640000|         Complicated|Rihanna|                Loud|
|00williamsl|1360439280000|What's My Name (f...|Rihanna|                Loud|
|00williamsl|1360434480000|Kanye West feat R...|Rihanna|                Loud|
|    0502008|1440985800000|Only Girl (In the...|Rihanna|                Loud|
|      0rdos|1319599320000|Pon De Replay (Re...|Rihanna|    Music of the Sun|
|      0rdos|1319599080000|          Now I Know|Rihanna|    Music of the Sun|
|      0rdos|1319598780000|There's a Thug in...|Rihanna|    Music of the Sun|
|      0rdos|1319598600000|                Rush|Rihanna|    Music of the Sun|
|      0rdos|1319598420000|             Let Me|Rihanna|    Music of the Sun|
|      0rdos|1319598180000|    Music of the Sun|Rihanna|    Music of the Sun|
|      0rdos|1319597940000|     Willing to Wait|Rihanna|    Music of the Sun|
|      0rdos|1319597640000|       The Last Time|Rihanna|    Music of the Sun|
|      0rdos|1319596860000|If It's Lovin' Th...|Rihanna|    Music of the Sun|
|      0rdos|1319596680000|     Here I Go Again|Rihanna|    Music of the Sun|
|      0rdos|1319596380000|       Pon de Replay|Rihanna|    Music of the Sun|
|      0rdos|1319596140000|                 Cry|Rihanna|  Good Girl Gone Bad|
+-----------+-------------+--------------------+-------+--------------------+
only showing top 20 rows

0.21714329719543457
```

Let's find top 10 users who are fan of ***Rihanna***

```
start_time = time.time()

q2 = listening_df.select('user_id').filter(listening_df.artist == 'Rihanna').groupby('user_id').agg(count('user_id').alias('count')).or
q2.show()

end_time = time.time()
execution_time = end_time - start_time
print(execution_time)
```

```
+---------------+-----+
|        user_id|count|
+---------------+-----+
|        thiessu|  179|
|    eyessetkyle|  166|
|           adxx|  164|
|missnumberthree|  156|
|helloiamnatalie|  128|
|          nmjnb|  124|
|        AndyyyA|  123|
|     BIGBANG186|  121|
|     mixedvibes|  120|
|       AndyKitt|  115|
+---------------+-----+

39.48111438751221
```

find top 10 famous tracks

```
start_time = time.time()

q3 = listening_df.select('artist','track').groupby('artist','track').agg(count('*').alias('count')).orderBy(desc('count')).limit(10)
q3.show()

end_time = time.time()
execution_time = end_time - start_time
print(execution_time)
```

```
+--------------+----------------+-----+
|        artist|           track|count|
+--------------+----------------+-----+
| Justin Bieber|           Sorry| 3381|
|Arctic Monkeys|Do I Wanna Know?| 2865|
|      Bon Iver|     Skinny Love| 2836|
|          Zayn|      PILLOWTALK| 2701|
|   The Killers|   Mr Brightside| 2690|
|       Rihanna|            Work| 2646|
|      Bastille|         Pompeii| 2606|
|Mumford & Sons| Little Lion Man| 2520|
|Mumford & Sons|        The Cave| 2485|
| Justin Bieber|   Love Yourself| 2481|
+--------------+----------------+-----+

74.8024582862854
```

find top 10 famous tracks of ***Rihanna***

```
start_time = time.time()

q4 = listening_df.select('artist','track').filter(listening_df.artist == 'Rihanna').groupby('artist','track').agg(count('*').alias('cou
q4.show()

end_time = time.time()
execution_time = end_time - start_time
print(execution_time)
```

```
+-------+--------------------+-----+
| artist|               track|count|
+-------+--------------------+-----+
|Rihanna|                Work| 2646|
|Rihanna|Only Girl (In the...| 1749|
|Rihanna|We Found Love (fe...| 1575|
|Rihanna|                 S&M| 1307|
|Rihanna|            Rude Boy| 1303|
|Rihanna|            Diamonds| 1224|
|Rihanna|      Kiss it Better|  945|
|Rihanna| Where Have You Been|  844|
|Rihanna|Cheers (Drink to ...|  697|
|Rihanna|           Needed Me|  679|
+-------+--------------------+-----+

39.12156081199646
```

find top 10 famous albums

```
start_time = time.time()

q5 = listening_df.select('artist','album').groupby('artist','album').agg(count('*').alias('count')).orderBy(desc('count')).limit(10)
q5.show()

end_time = time.time()
execution_time = end_time - start_time
print(execution_time)
```

```
+--------------------+--------------------+-----+
|              artist|               album|count|
+--------------------+--------------------+-----+
|          Kanye West|   The Life Of Pablo|22310|
|              The xx|                  xx|14195|
|      Arctic Monkeys|                  AM|14090|
|               alt-J|     An Awesome Wave|13635|
|      Mumford & Sons|        Sigh No More|13543|
|      Arctic Monkeys|Whatever People S...|12731|
|            Bon Iver|            For Emma|11994|
|              Grimes|          Art Angels|11655|
|Florence + the Ma...|               Lungs|11362|
|               Adele|                  21|11215|
+--------------------+--------------------+-----+

71.71325707435608
```