SQL Options in Spark HW

Alirght let's apply what we learned in the lecture to a new dataset!

But first!

Let's start with Spark SQL. But first we need to create a Spark Session!

In [53]:
```
import findspark
findspark.init()
```

In [1]:
```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName("DfApp").getOrCreate()
```

Read in our DataFrame for this Notebook

For this notebook we will be using the Google Play Store csv file attached to this lecture. Let's go ahead and read it in.

About this dataset

Contains a list of Google Play Store Apps and info about the apps like the category, rating, reviews, size, etc.

Source: https://www.kaggle.com/lava18/google-play-store-apps

In [52]:
```
df=spark.read.csv("D:\googleplaystore.csv",header=True, inferSchema=True)
df.show()
```

```
+--------------------+--------------+------+-------+----+-----------+----+-----+--------------+--------------------+--------------+------------------+---------
---+
|                 App|      Category|Rating|Reviews|Size|   Installs|Type|Price|Content Rating|              Genres|  Last Updated|       Current Ver| Android
Ver|
+--------------------+--------------+------+-------+----+-----------+----+-----+--------------+--------------------+--------------+------------------+---------
---+
|Photo Editor & Ca...|ART_AND_DESIGN|   4.1|    159| 19M|    10,000+|Free|    0|      Everyone|        Art & Design|      7-Jan-18|             1.0.0|4.0.3 and
up|
| Coloring book moana|ART_AND_DESIGN|   3.9|    967| 14M|   500,000+|Free|    0|      Everyone|Art & Design;Pret...|     15-Jan-18|             2.0.0|4.0.3 and
up|
|U Launcher Lite –...|ART_AND_DESIGN|   4.7|  87510|8.7M| 5,000,000+|Free|    0|      Everyone|        Art & Design|      1-Aug-18|             1.2.4|4.0.3 and
up|
|Sketch - Draw & P...|ART_AND_DESIGN|   4.5| 215644| 25M|50,000,000+|Free|    0|          Teen|        Art & Design|      8-Jun-18|Varies with device|  4.2 and
up|
|Pixel Draw - Numb...|ART_AND_DESIGN|   4.3|    967|2.8M|   100,000+|Free|    0|      Everyone|Art & Design;Crea...|     20-Jun-18|               1.1|  4.4 and
up|
|Paper flowers ins...|ART_AND_DESIGN|   4.4|    167|5.6M|    50,000+|Free|    0|      Everyone|        Art & Design|     26-Mar-17|                 1|  2.3 and
up|
|Smoke Effect Phot...|ART_AND_DESIGN|   3.8|    178| 19M|    50,000+|Free|    0|      Everyone|        Art & Design|April 26, 2018|               1.1|4.0.3 and
up|
|    Infinite Painter|ART_AND_DESIGN|   4.1|  36815| 29M| 1,000,000+|Free|    0|      Everyone|        Art & Design|     14-Jun-18|          6.1.61.1|  4.2 and
up|
|Garden Coloring Book|ART_AND_DESIGN|   4.4|  13791| 33M| 1,000,000+|Free|    0|      Everyone|        Art & Design|     20-Sep-17|             2.9.2|  3.0 and
up|
|Kids Paint Free -...|ART_AND_DESIGN|   4.7|    121|3.1M|    10,000+|Free|    0|      Everyone|Art & Design;Crea...|      3-Jul-18|               2.8|4.0.3 and
up|
|Text on Photo - F...|ART_AND_DESIGN|   4.4|  13880| 28M| 1,000,000+|Free|    0|      Everyone|        Art & Design|     27-Oct-17|             1.0.4|  4.1 and
up|
|Name Art Photo Ed...|ART_AND_DESIGN|   4.4|   8788| 12M| 1,000,000+|Free|    0|      Everyone|        Art & Design|     31-Jul-18|            1.0.15|  4.0 and
up|
|Tattoo Name On My...|ART_AND_DESIGN|   4.2|  44829| 20M|10,000,000+|Free|    0|          Teen|        Art & Design| April 2, 2018|               3.8|  4.1 and
up|
|Mandala Coloring ...|ART_AND_DESIGN|   4.6|   4326| 21M|   100,000+|Free|    0|      Everyone|        Art & Design|     26-Jun-18|             1.0.4|  4.4 and
up|
|3D Color Pixel by...|ART_AND_DESIGN|   4.4|   1518| 37M|   100,000+|Free|    0|      Everyone|        Art & Design|      3-Aug-18|             1.2.3|  2.3 and
up|
|Learn To Draw Kaw...|ART_AND_DESIGN|   3.2|     55|2.7M|     5,000+|Free|    0|      Everyone|        Art & Design|      6-Jun-18|               NaN|  4.2 and
up|
|Photo Designer - ...|ART_AND_DESIGN|   4.7|   3632|5.5M|   500,000+|Free|    0|      Everyone|        Art & Design|     31-Jul-18|               3.1|  4.1 and
up|
|350 Diy Room Deco...|ART_AND_DESIGN|   4.5|     27| 17M|    10,000+|Free|    0|      Everyone|        Art & Design|      7-Nov-17|                 1|  2.3 and
up|
|FlipaClip - Carto...|ART_AND_DESIGN|   4.3| 194216| 39M| 5,000,000+|Free|    0|      Everyone|        Art & Design|      3-Aug-18|             2.2.5|4.0.3 and
up|
|        ibis Paint X|ART_AND_DESIGN|   4.6| 224399| 31M|10,000,000+|Free|    0|      Everyone|        Art & Design|     30-Jul-18|             5.5.4|  4.1 and
up|
+--------------------+--------------+------+-------+----+-----------+----+-----+--------------+--------------------+--------------+------------------+---------
---+
only showing top 20 rows
```

First things first

Let's check out the first few lines of the dataframe to see what we are working with

In [51]:
```
df.columns
```

Out[51]:
```
['App',
 'Category',
 'Rating',
 'Reviews',
 'Size',
 'Installs',
 'Type',
 'Price',
 'Content Rating',
 'Genres',
 'Last Updated',
 'Current Ver',
 'Android Ver']
```

As well as the schema to make sure all the column types were correctly infered

In [10]:
```
df.printSchema()
```

```
root
 |-- App: string (nullable = true)
 |-- Category: string (nullable = true)
 |-- Rating: string (nullable = true)
 |-- Reviews: string (nullable = true)
 |-- Size: string (nullable = true)
 |-- Installs: string (nullable = true)
 |-- Type: string (nullable = true)
 |-- Price: string (nullable = true)
 |-- Content Rating: string (nullable = true)
 |-- Genres: string (nullable = true)
 |-- Last Updated: string (nullable = true)
 |-- Current Ver: string (nullable = true)
 |-- Android Ver: string (nullable = true)
```

Looks like we need to edit some of the datatypes. We need to update Rating, Reviews and Price as integer (float for Rating) values for now, since the Size and Installs variables will need a bit more cleaning. Since we haven't been over this yet, I'm going to provide the code for you here so you can get a quick look at how it used (and how often we need it!).

make sure to change the df name to whatever you named your df

In [6]:
```
from pyspark.sql.types import IntegerType, FloatType
newdf = df.withColumn("Rating", df["Rating"].cast(FloatType())) \
            .withColumn("Reviews", df["Reviews"].cast(IntegerType())) \
            .withColumn("Price", df["Price"].cast(IntegerType()))
print(newdf.printSchema())
newdf.limit(5).toPandas()
```

```
root
 |-- App: string (nullable = true)
 |-- Category: string (nullable = true)
 |-- Rating: float (nullable = true)
 |-- Reviews: integer (nullable = true)
 |-- Size: string (nullable = true)
 |-- Installs: string (nullable = true)
 |-- Type: string (nullable = true)
 |-- Price: integer (nullable = true)
 |-- Content Rating: string (nullable = true)
 |-- Genres: string (nullable = true)
 |-- Last Updated: string (nullable = true)
 |-- Current Ver: string (nullable = true)
 |-- Android Ver: string (nullable = true)

None
```

Out[6]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	7-Jan-18	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	15-Jan-18	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	1-Aug-18	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	8-Jun-18	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	20-Jun-18	1.1	4.4 and up

Looks like that worked! Great! Let's dig in.

1. Create Tempview

Go ahead and create a tempview of the dataframe so we can work with it in spark sql.

In [10]:
```
df.createOrReplaceTempView("tempview")
newdf.createOrReplaceTempView("tempview1")
```

2. Select all apps with ratings above 4.1

Use your tempview to select all apps with ratings above 4.1

In [54]:
```
res=spark.sql("select * from tempview where Rating>4.1")
res.toPandas()
```

Out[54]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	1-Aug-18	1.2.4	4.0.3 and up
1	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	8-Jun-18	Varies with device	4.2 and up
2	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	20-Jun-18	1.1	4.4 and up
3	Paper flowers instructions	ART_AND_DESIGN	4.4	167	5.6M	50,000+	Free	0	Everyone	Art & Design	26-Mar-17	1	2.3 and up
4	Garden Coloring Book	ART_AND_DESIGN	4.4	13791	33M	1,000,000+	Free	0	Everyone	Art & Design	20-Sep-17	2.9.2	3.0 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
7560	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	25-Jul-17	1.48	4.1 and up
7561	Fr. Mike Schmitz Audio Teachings	FAMILY	5	4	3.6M	100+	Free	0	Everyone	Education	6-Jul-18	1	4.1 and up
7562	Parkinson Exercices FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Everyone	Medical	20-Jan-17	1	2.2 and up
7563	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	19-Jan-15	Varies with device	Varies with device
7564	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	0	Everyone	Lifestyle	25-Jul-18	Varies with device	Varies with device

7565 rows × 13 columns

3. Now pass your results to an object

(ie create a spark dataframe)

Select just the App and Rating column where the Category is in the Comic category and the Rating is above 4.5.

In [84]:
```
res=spark.sql("select App,Rating from tempview1 where Category=='COMICS' and Rating>4.5")
res.toPandas()
```

Out[84]:

	App	Rating
0	Manga Master - Best manga & comic reader	4.6
1	GANMA! - All original stories free of charge f...	4.7
2	Röhrich Werner Soundboard	4.7
3	Unicorn Pokez - Color By Number	4.8
4	Manga - read Thai translation	4.6
5	Dragon Ball Wallpaper - Ringtones	4.7
6	Children's cartoons (Mithu-Mina-Raju)	4.6
7	【Ranobbe complete free】 Novelba - Free app tha...	NaN
8	Faustop Sounds	4.7
9	Hojiboy Tojiboyev Life Hacks	5.0
10	Best Wallpapers Backgrounds(100,000+ 4K HD)	4.7
11	Laftel - Watching and Announcing Snooping, Str...	4.6
12	WebComics	4.8
13	Superheroes, Marvel, DC, Comics, TV, Movies News	5.0
14	Pepsi Cards DC	NaN

4. Which category has the most cumulative reviews

Only select the one category with the most reivews.

Note: will require adding all the review together for each category

In [61]:
```
res=spark.sql("select category,sum(reviews) from tempview1 group by category")
res.show(1)
```

```
+--------+------------+
|category|sum(reviews)|
+--------+------------+
|  EVENTS|      161018|
+--------+------------+
only showing top 1 row
```

5. Which App has the most reviews?

Display ONLY the top result

Include only the App column and the Reviews column.

In [79]:
```
res=spark.sql("select App,Reviews from tempview1 order by Reviews desc")
res.show(10)
```

```
+--------------------+--------+
|                 App| Reviews|
+--------------------+--------+
|            Facebook|78158306|
|            Facebook|78128208|
|  WhatsApp Messenger|69119316|
|  WhatsApp Messenger|69119316|
|  WhatsApp Messenger|69109672|
|           Instagram|66577446|
|           Instagram|66577313|
|           Instagram|66577313|
|           Instagram|66509917|
|Messenger – Text ...|56646578|
+--------------------+--------+
only showing top 10 rows
```

5. Select all apps that contain the word 'dating' anywhere in the title

Note: we did not cover this in the lecture. You'll have to use your SQL knowledge :) Google it if you need to.

In [76]:
```
res=spark.sql("select App from tempview1 where App like '%dating%' ")
res.toPandas()
```

Out[76]:

	App
0	Meet, chat & date. Free dating app - Chocolate...
1	Friend Find: free chat + flirt dating app
2	Spine- The dating app
3	Princess Closet : Otome games free dating sim
4	happn – Local dating app

6. Use SQL Transformer to display how many free apps there are in this list

In [70]:
```
from pyspark.ml.feature import SQLTransformer
res=SQLTransformer(statement="select App,Type from __THIS__ where Type='Free'")
print("Free Apps in this list :",res.transform(newdf).count())
print(res.transform(newdf).show())
```

```
Free Apps in this list : 10037
+--------------------+----+
|                 App|Type|
+--------------------+----+
|Photo Editor & Ca...|Free|
| Coloring book moana|Free|
|U Launcher Lite –...|Free|
|Sketch - Draw & P...|Free|
|Pixel Draw - Numb...|Free|
|Paper flowers ins...|Free|
|Smoke Effect Phot...|Free|
|    Infinite Painter|Free|
|Garden Coloring Book|Free|
|Kids Paint Free -...|Free|
|Text on Photo - F...|Free|
|Name Art Photo Ed...|Free|
|Tattoo Name On My...|Free|
|Mandala Coloring ...|Free|
|3D Color Pixel by...|Free|
|Learn To Draw Kaw...|Free|
|Photo Designer - ...|Free|
|350 Diy Room Deco...|Free|
|FlipaClip - Carto...|Free|
|        ibis Paint X|Free|
+--------------------+----+
only showing top 20 rows

None
```

7. What is the most popular Genre?

Which genre appears most often in the dataframe. Show only the top result.

In [78]:
```
res=spark.sql("select App,Reviews,Genres from tempview1 order by Genres desc")
res.show(10)
```

```
+--------------------+-------+------+
|                 App|Reviews|Genres|
+--------------------+-------+------+
|Korean Dungeon: K...|    703|  Word|
|    Puzzle for CS:GO|    247|  Word|
|Draw N Guess Mult...|  29505|  Word|
|Word Crossy - A c...| 240416|  Word|
|Words With Friend...|1704112|  Word|
|         Word Search| 295305|  Word|
|Draw Something Cl...|1092106|  Word|
|          Wordscapes| 230710|  Word|
|Guess the Class ?...|     57|  Word|
|         Word Search| 295576|  Word|
+--------------------+-------+------+
only showing top 10 rows
```

8. Select all the apps in the 'Tools' genre that have more than 100 reviews

In [25]:
```
res=spark.sql("select App,Reviews,Genres from tempview1 where Genres='Tools' and Reviews>100")
res.show()
```

```
+--------------------+--------+------+
|                 App| Reviews|Genres|
+--------------------+--------+------+
|   Moto File Manager|   38655| Tools|
|              Google| 8033493| Tools|
|    Google Translate| 5745093| Tools|
|        Moto Display|   18239| Tools|
|      Motorola Alert|   24199| Tools|
|     Motorola Assist|   37333| Tools|
|Cache Cleaner-DU ...|12759663| Tools|
|  Moto Suggestions ™|     308| Tools|
|          Moto Voice|   33216| Tools|
|          Calculator|   40770| Tools|
|         Device Help|   28860| Tools|
|     Account Manager|   76604| Tools|
|             myMetro|   26189| Tools|
|        File Manager|  739329| Tools|
|           My Telcel|   45838| Tools|
|Calculator - free...|   25592| Tools|
| ASUS Sound Recorder|   34126| Tools|
|  iWnn IME for Nexus|    2394| Tools|
|Samsung Max - Dat...|  330468| Tools|
|          ZenUI Help|  136874| Tools|
+--------------------+--------+------+
only showing top 20 rows
```

That's all folks! Great job!