Spark Preparation

We check if we are in Google Colab. If this is the case, install all necessary packages.

To run spark in Colab, we need to first install all the dependencies in Colab environment i.e. Apache Spark 3.3.2 with hadoop 3.3, Java 8 and Findspark to locate the spark in the system. The tools installation can be carried out inside the Jupyter Notebook of the Colab. Learn more from A Must-Read Guide on How to Work with PySpark on Google Colab for Data Scientists!

In [5]:
```
!pip install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.5.1.tar.gz (317.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.0/317.0 MB 8.0 MB/s eta 0:00:
0000:0100:01
  Preparing metadata (setup.py) ... done
Collecting py4j==0.10.9.7 (from pyspark)
  Downloading py4j-0.10.9.7-py2.py3-none-any.whl.metadata (1.5 kB)
Downloading py4j-0.10.9.7-py2.py3-none-any.whl (200 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 200.5/200.5 kB 14.7 MB/s eta 0:00:
00
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=f19ccc5d499fc7a8d4e7d257525732c049d6c687a4c5e1045e4f004c1876e
1c1
  Stored in directory: /Users/jirayuwat/Library/Caches/pip/wheels/92/09/11/aa0
1d01a7f005fda8a66ad71d2be7f8aa341bddafb27eee3c7
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.7 pyspark-3.5.1
```

In [2]:
```
try:
  import google.colab
  IN_COLAB = True
except:
  IN_COLAB = False
```

In [3]:
```
if IN_COLAB:
    !apt-get install openjdk-8-jdk-headless -qq > /dev/null
    !wget -q https://dlcdn.apache.org/spark/spark-3.3.2/spark-3.3.2-bin-hadoop3
    !tar xf spark-3.3.2-bin-hadoop3.tgz
    !mv spark-3.3.2-bin-hadoop3 spark
    !pip install -q findspark
    import os
    os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
    os.environ["SPARK_HOME"] = "/content/spark"
```

Start a Local Cluster

In [18]:
```
import pyspark
from pyspark.sql import SparkSession, SQLContext
from pyspark.sql.functions import *

spark = SparkSession.builder.master("local[*]").appName("pyspark_colab").getOr
sc = spark.sparkContext
sqlContext = SQLContext(sparkContext=sc, sparkSession=spark)
```

In [9]:
```
sc
```

Out[9]:

SparkContext

Spark UI

Version	`v3.5.1`
Master	`local[*]`
AppName	`pyspark_colab`

Spark Assignment

Based on the movie review dataset in 'netflix-rotten-tomatoes-metacritic-imdb.csv', answer the below questions.

Note: do not clean or remove missing data

In [22]:
```
df = spark.read.csv("netflix-rotten-tomatoes-metacritic-imdb.csv", header=True
```

In [23]:
```
df.printSchema()
```

```
root
 |-- Title: string (nullable = true)
 |-- Genre: string (nullable = true)
 |-- Tags: string (nullable = true)
 |-- Languages: string (nullable = true)
 |-- Series or Movie: string (nullable = true)
 |-- Hidden Gem Score: double (nullable = true)
 |-- Country Availability: string (nullable = true)
 |-- Runtime: string (nullable = true)
 |-- Director: string (nullable = true)
 |-- Writer: string (nullable = true)
 |-- Actors: string (nullable = true)
 |-- View Rating: string (nullable = true)
 |-- IMDb Score: string (nullable = true)
 |-- Rotten Tomatoes Score: string (nullable = true)
 |-- Metacritic Score: string (nullable = true)
 |-- Awards Received: double (nullable = true)
 |-- Awards Nominated For: double (nullable = true)
 |-- Boxoffice: string (nullable = true)
 |-- Release Date: string (nullable = true)
 |-- Netflix Release Date: string (nullable = true)
 |-- Production House: string (nullable = true)
 |-- Netflix Link: string (nullable = true)
 |-- IMDb Link: string (nullable = true)
 |-- Summary: string (nullable = true)
 |-- IMDb Votes: string (nullable = true)
 |-- Image: string (nullable = true)
 |-- Poster: string (nullable = true)
 |-- TMDb Trailer: string (nullable = true)
 |-- Trailer Site: string (nullable = true)
```

In [24]:
```
df.show(5)
```

```
+--------------------+--------------------+--------------------+--------------
-+--------------+---------------+--------------------+------------+---------
------+--------------------+--------------------+----------+----------+------
--------------+---------------+---------------+--------------------+--------
--+------------+--------------------+--------------------+--------------------
+--------------------+--------------------+----------+--------------------+---
----------------+--------------------+------------+
|               Title|               Genre|                Tags|       Language
s|Series or Movie|Hidden Gem Score|Country Availability|     Runtime|       Di
rector|              Writer|              Actors|View Rating|IMDb Score|Rotten
Tomatoes Score|Metacritic Score|Awards Received|Awards Nominated For| Boxoffic
e|Release Date|Netflix Release Date|    Production House|        Netflix Link|
IMDb Link|             Summary|IMDb Votes|               Image|              P
oster|        TMDb Trailer|Trailer Site|
+--------------------+--------------------+--------------------+--------------
-+--------------+---------------+--------------------+------------+---------
------+--------------------+--------------------+----------+----------+------
--------------+---------------+---------------+--------------------+--------
--+------------+--------------------+--------------------+--------------------
+--------------------+--------------------+----------+--------------------+---
----------------+--------------------+------------+
|   Lets Fight Ghost|Crime, Drama, Fan...|Comedy Programmes...|Swedish, Spanis
h|         Series|             4.3|            Thailand|< 30 minutes|Tomas Alf
redson|John Ajvide Lindq...|Kåre Hedebrant, P...|          R|       7.9|
98.0|            82.0|           74.0|               57.0|$2,122,065| 12 Dec
2008|          2021-03-04|Canal+, Sandrew M...|https://www.netfl...|https://ww
w.imdb....|A med student wit...|  205926.0|https://occ-0-470...|https://m.medi
a-a...|                NULL|        NULL|
|HOW TO BUILD A GIRL|              Comedy|Dramas,Comedies,F...|         Englis
h|          Movie|             7.0|              Canada|    1-2 hour|  Coky Gi
edroyc|       Caitlin Moran|Paddy Considine, ...|          R|       5.8|
79.0|            69.0|            1.0|                NULL|   $70,632| 08 May
2020|          2021-03-04|Film 4, Monumenta...|https://www.netfl...|https://ww
w.imdb....|When nerdy Johann...|    2838.0|https://occ-0-108...|https://m.medi
a-a...|https://www.youtu...|     YouTube|
|         Centigrade|     Drama, Thriller|           Thrillers|         Englis
h|          Movie|             6.4|              Canada|    1-2 hour|  Brendan
Walsh|Brendan Walsh, Da...|Genesis Rodriguez...|    Unrated|       4.3|
NULL|            46.0|           NULL|                NULL|   $16,263| 28 Aug
2020|          2021-03-04|                NULL|https://www.netfl...|https://ww
w.imdb....|Trapped in a froz...|    1720.0|https://occ-0-108...|https://m.medi
a-a...|https://www.youtu...|     YouTube|
|              ANNE+|              Drama|TV Dramas,Romanti...|         Turkis
h|         Series|             7.7| Belgium,Netherlands|< 30 minutes|
NULL|                NULL|Vahide Perçin, Go...|       NULL|       6.5|
NULL|            NULL|            1.0|                NULL|      NULL| 01 Oct
2016|          2021-03-04|                NULL|https://www.netfl...|https://ww
w.imdb....|Upon moving into ...|    1147.0|https://occ-0-148...|https://m.medi
a-a...|                NULL|        NULL|
|              Moxie|Animation, Short,...|Social Issue Dram...|         Englis
h|          Movie|             8.1|Lithuania,Poland,...|    1-2 hour|  Stephen
Irwin|                NULL|        Ragga Gudrun|       NULL|       6.3|
NULL|            NULL|           NULL|                 4.0|      NULL| 22 Sep
2011|          2021-03-04|                NULL|https://www.netfl...|https://ww
w.imdb....|Inspired by her m...|      63.0|https://occ-0-403...|https://m.medi
a-a...|                NULL|        NULL|
+--------------------+--------------------+--------------------+--------------
-+--------------+---------------+--------------------+------------+---------
------+--------------------+--------------------+----------+----------+------
--------------+---------------+---------------+--------------------+--------
```

```
--+-----------+-------------------+-------------------+-------------------
+-------------------+-------------------+----------+-------------------+---
----------------+-------------------+-----------+
only showing top 5 rows
```

What is the maximum and average of the overall hidden gem score?

In [21]:
```
df.select(avg('Hidden Gem Score'), max('Hidden Gem Score')).show()
```

```
+---------------------+---------------------+
|avg(Hidden Gem Score)|max(Hidden Gem Score)|
+---------------------+---------------------+
|    5.937551386501234|                  9.8|
+---------------------+---------------------+
```

How many movies that are available in Korea?

In [32]:
```
# Korean in Languages column
df.filter(df['Languages'].contains('Korean')).count()
```

Out[32]:
```
735
```

Which director has the highest average hidden gem score?

In [37]:
```
df.groupBy('Director').agg({'Hidden Gem Score': 'avg'}).orderBy('avg(Hidden Ge
```

```
+-----------+---------------------+
|   Director|avg(Hidden Gem Score)|
+-----------+---------------------+
|Dorin Marcu|                  9.8|
+-----------+---------------------+
only showing top 1 row
```

How many genres are there in the dataset?

In [46]:
```
# convert df to do map reduce
rdd = df.rdd

# map
rdd = rdd.map(lambda x: x['Genre'].split(', ') if x['Genre'] else []).flatMap(

# reduce
rdd = rdd.reduceByKey(lambda x, y: x + y)

len(rdd.collect())
```

Out[46]:
```
28
```

In [ ]: