```
!pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting pyspark
  Downloading pyspark-3.4.0.tar.gz (310.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 310.8/310.8 MB 4.4 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.9/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-3.4.0-py2.py3-none-any.whl size=311317145 sha256=4fa43b4317c03eca3d473a841f5b9ace0c8bf8a4e
  Stored in directory: /root/.cache/pip/wheels/9f/34/a4/159aa12d0a510d5ff7c8f0220abbea42e5d81ecf588c4fd884
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.4.0
```

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, expr,size, max, length, min, lower, explode, array_contains
spark = SparkSession.builder.getOrCreate()
```

```
spark
```

SparkSession - in-memory

SparkContext

Spark UI

Version
v3.4.0
Master
local[*]
AppName
pyspark-shell

```
df = spark.read.json("movies.json")
```

```
df.head(5)
```

```
[Row(cast=[], genres=[], title='After Dark in Central Park', year=1900),
 Row(cast=[], genres=[], title="Boarding School Girls' Pajama Parade", year=1900),
 Row(cast=[], genres=[], title="Buffalo Bill's Wild West Parad", year=1900),
 Row(cast=[], genres=[], title='Caught', year=1900),
 Row(cast=[], genres=[], title='Clowns Spinning Hats', year=1900)]
```

```
df.printSchema()
```

```
root
 |-- cast: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- genres: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- title: string (nullable = true)
 |-- year: long (nullable = true)
```

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

```
+-------+----------------+------------------+
|summary|           title|              year|
+-------+----------------+------------------+
|  count|           28795|             28795|
|   mean|        Infinity|1959.9489841986456|
| stddev|             NaN| 31.12544556684899|
|    min|   $ aka Dollars|              1900|
|    max|…First Do No Harm|              2018|
+-------+----------------+------------------+
```

Phần trả lời câu hỏi

1. Số lượng phim khác nhau trong tập dữ liệu ?

```
df.distinct().count()
```

```
28789
```

```
df.selectExpr("count(distinct(title)) as Distinct_films").show()
```

```
+--------------+
|Distinct_films|
+--------------+
|         26791|
+--------------+
```

2. Số lượng Phim được phát hành trong năm 2015

```
df.where(expr("year==2015")).distinct().count()
```

```
130
```

3. Năm phát hành phim nhiều nhất là năm nào ?

```
df.groupBy('year').count().orderBy(col('count').desc()).limit(1).show()
```

```
+----+-----+
|year|count|
+----+-----+
|1919|  634|
+----+-----+
```

4. Hãy tìm ra những bộ phim có ít nhất 10 diễn viên và chỉ thuộc 1 thể loại

```
df.where(size(df['cast']) >= 10).where(size(df['genres']) == 1).show()
```

```
+--------------------+-------------+--------------------+----+
|                cast|       genres|               title|year|
+--------------------+-------------+--------------------+----+
|[Walter Huston, (...|      [Drama]|           Dodsworth|1936|
|[Joan Crawford, R...|      [Drama]|  The Gorgeous Hussy|1936|
|[Norma Shearer, (...|      [Drama]|    Romeo and Juliet|1936|
|[John Wayne, Robe...|      [Drama]|The High and the ...|1954|
|[Richard Widmark,...|      [Drama]|          The Cobweb|1955|
|[Henry Fonda, Cha...|      [Drama]|    Advise & Consent|1962|
|[Spencer Tracy, M...|     [Comedy]|It's a Mad, Mad, ...|1963|
|[Lee Marvin, Char...|        [War]|     The Dirty Dozen|1967|
|[Burt Lancaster, ...|   [Disaster]|             Airport|1970|
|[Jimi Hendrix, Th...|[Documentary]|           Woodstock|1970|
|[Steve McQueen, P...|   [Disaster]|The Towering Inferno|1974|
|[Peter Falk, Pete...|     [Comedy]|     Murder by Death|1976|
|[Nick Nolte, Pete...|      [Drama]|  Rich Man, Poor Man|1976|
|[Mel Brooks, Dom ...|     [Comedy]|        Silent Movie|1976|
|[LeVar Burton, Lo...|      [Drama]|               Roots|1977|
|[Tim Matheson, To...|     [Comedy]|National Lampoon'...|1978|
|[Marlon Brando, G...|  [Superhero]|            Superman|1978|
|[Michael Caine, H...|   [Disaster]|           The Swarm|1978|
|[Carol Burnett, L...|     [Comedy]|           A Wedding|1978|
|[John Belushi, Ne...|     [Comedy]|                1941|1979|
+--------------------+-------------+--------------------+----+
only showing top 20 rows
```

5. Hãy chỉ ra bộ phim có tên dài nhất?

```
df_len = df.withColumn("length", length("title"))
maxlen = df_len.agg(max("length")).collect()[0]
df_len.where(col("length")==maxlen['max(length)']).select('title').show()
```

```
+--------------------+
|               title|
+--------------------+
|Cornell-Columbia-...|
+--------------------+
```

6. Hãy chỉ ra những bộ phim có từ "fighting" ?

```
df.filter(lower(df.title).contains("fighting")).show()
```

```
+--------------------+---------------+--------------------+----+
|                cast|         genres|               title|year|
+--------------------+---------------+--------------------+----+
|[Bessie Love, Ann...|[Comedy, Drama]|  A Fighting Colleen|1919|
|[Blanche Sweet, R...|      [Western]|     Fighting Cressy|1919|
|[Harry T. Morey, ...|        [Drama]|    Fighting Destiny|1919|
|[Tom Mix, Teddy S...|      [Western]|   Fighting for Gold|1919|
|[Jack Perrin, Hoo...|      [Western]|  The Fighting Heart|1919|
|[Art Acord, Mildr...|      [Western]|   The Fighting Line|1919|
|[William Duncan, ...|       [Action]|  The Fighting Guide|1922|
|[Tom Mix, Patsy R...|      [Western]| The Fighting Streak|1922|
|[Richard Barthelm...|   [Historical]|  The Fighting Blade|1923|
|[Ernest Torrence,...|       [Comedy]| The Fighting Coward|1924|
|[Jack Hoxie, Hele...|      [Western]|       Fighting Fury|1924|
|[Pat O'Malley, Ma...|        [Drama]|The Fighting Adve...|1924|
|[Fred Thomson, Ha...|      [Western]|    The Fighting Sap|1924|
|[Richard Talmadge...|       [Action]|  The Fighting Demon|1925|
|[Billy Sullivan, ...|       [Sports]|       Fighting Fate|1925|
|[George O'Brien, ...|        [Drama]|  The Fighting Heart|1925|
|[Bob Reeves, Lew ...|      [Western]|       Fighting Luck|1925|
|[Bill Cody, Jean ...|      [Western]|  The Fighting Smile|1925|
|[William Haines, ...|        [Drama]| Fighting the Flames|1925|
|[William Fairbank...|       [Action]|      Fighting Youth|1925|
+--------------------+---------------+--------------------+----+
only showing top 20 rows
```

7. Hãy chỉ ra các thể loại có trong bộ dữ liệu ?

```
df.select(explode("genres").alias("genres")).distinct().show()
```

```
+-------------+
|       genres|
+-------------+
|        Crime|
|      Romance|
|     Thriller|
|      Slasher|
|Found Footage|
|    Adventure|
|         Teen|
| Martial Arts|
|       Sports|
|        Drama|
|          War|
|  Documentary|
|       Family|
|      Fantasy|
|       Silent|
|     Disaster|
|        Legal|
|      Mystery|
| Supernatural|
|     Suspense|
+-------------+
only showing top 20 rows
```

8. Hãy chỉ ra những bộ phim có diễn viên **Harrison Ford** tham gia

```
temp = df.filter(array_contains(df.cast,"Harrison Ford"))
temp.select("title").show()
```

```
+--------------------+
|               title|
+--------------------+
|Experimental Marr...|
| Happiness a la Mode|
|Romance and Arabella|
|      The Third Kiss|
|The Veiled Adventure|
|          Who Cares?|
|You Never Saw Suc...|
| The Wonderful Thing|
|      Find the Woman|
| The Primitive Lover|
|     Smilin' Through|
|     When Love Comes|
| Little Old New York|
|     Three Miles Out|
|           The Wheel|
|       Almost a Lady|
| Hell's Four Hundred|
|   The Nervous Wreck|
|  Up in Mabel's Room|
|         Golf Widows|
+--------------------+
only showing top 20 rows
```

9. Hãy chỉ ra những bộ phim có diễn viên tham gia có tên chứa từ "Lewis" ?

```
temp = df.withColumn("name_cast",explode(df.cast))
temp.filter(lower(temp.name_cast).contains("lewis")).select("title").show()
```

```
+--------------------+
|               title|
+--------------------+
|       The Butterfly|
|The Exploits of E...|
| Mein Lieber Katrina|
|      Going Straight|
|Gretchen the Gree...|
|     A Sister of Six|
| The Bride's Silence|
|Nine-Tenths of th...|
|The Faith of the ...|
|         The Hoodlum|
|Jacques of the Si...|
|The Last of His P...|
|        Man's Desire|
|   Yvonne from Paris|
|Nine-Tenths of th...|
|                 813|
|    Huckleberry Finn|
|             Salvage|
|The Five Dollar Baby|
|    A Fool There Was|
+--------------------+
only showing top 20 rows
```

10. Top 5 diễn viên tham gia nhiều bộ phim nhất ?

```
temp = df.withColumn("actor", explode(df.cast))
temp.groupby("actor").count().sort('count', ascending = False).show(5)
```

```
+----------------+-----+
|           actor|count|
+----------------+-----+
|    Harold Lloyd|  190|
|     Hoot Gibson|  142|
|      John Wayne|  136|
|Charles Starrett|  116|
|    Bebe Daniels|  103|
+----------------+-----+
only showing top 5 rows
```