Problem Statement:

CSV file contains following columns:

color,director_name,num_critic_for_reviews,duration,director_facebook_likes,actor_3_facebook_likes,actor_2_name,actor_1_facebook_likes,gross,genres,actor_1_name,movie_title,num_voted_users,cast_total_facebook_likes,actor_3_name,facenumber_in_poster,plot_keywords,movie_imdb_link,num_user_for_reviews,language,country,content_rating,budget,title_year,actor_2_facebook_likes,imdb_score,aspect_ratio,movie_facebook_likes

Copy data to local system:

Wget
https://raw.githubusercontent.com/ruchisahu/dataAnalysis_cloud9/master/IMDB/imdb-5000-movie-dataset/movie_metadata.csv
Transfer this data to new folder IMDB(from downloads to home directory)

Copy data into HDFS:

Created new directory IMDB in tmp folder of hdfs:

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /tmp/IMDB/
```

Transfer IMDB local folder to IMDB hdfs:

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/IMDB /tmp/IMDB
```

Let's check our folder:

```
[cloudera@quickstart ~]$ hdfs dfs -ls /tmp/IMDB/
```

Files:

Start command line interface:

Beeline as the command-line interface

```
[cloudera@quickstart ~]beeline -u jdbc:hive2://
```

Show database:

```
jdbc:hive2://> show databases;
OK
+-----------+--+
| tab_name  |
+-----------+--+
+-----------+--+
.....................
```

Creating the Table in Hive:

```
jdbc:>>>
CREATE TABLE imdb(
 color string,
 director_name string,
 num_critic_for_reviews string,
 duration string,
 director_facebook_likes string,
 actor_3_facebook_likes bigint,
 Actor_2_name string,
Actor_1facebooklikes string,
 Gross bigint,
 genres int,
 actor_1_name string,
movie_title int,
 Num_voted_users  int,
Cast_total_facebook_likes int,
Actor_3_name string,
Facenumber_in_poster string,
Plot_keywords string,
Movie_imdb_link string,
Num_user_for_reviews int,
Language string,
Country string,
Content_rating string,
Budget int,
Title_year int,
Actor_2_facebook_likes int,
Imdb_score int,
Aspect_ratio int,
Movie_facebook_likes int)
ROW FORMAT DELIMITED
  FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');
```

Loading IMDB data into hive table:

```
0: jdbc:hive2:///> LOAD DATA INPATH '/tmp/IMDB/IMDB' OVERWRITE INTO TABLE imdb;
Loading data to table default.imdb
Table default.imdb stats: [numFiles=1, numRows=0, totalSize=1489644, rawDataSize=0]
OK
```

Retrieving the Data:

To retrieve the data, the select command is used.

```
Select DISTINCT
director_name from imdb;
.....
Select COUNT(DISTINCT director_name )from imdb;

8/04/26 16:18:45 [HiveServer2-Background-Pool: Thread-56]: WARN mapreduce.Counters: Group
org.apache.hadoop.mapred.Task$Counter is deprecated. Use
org.apache.hadoop.mapreduce.TaskCounter instead
2018-04-26 16:18:45,824 Stage-1 map = 0%,  reduce = 0%
2018-04-26 16:18:52,147 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.33 sec
2018-04-26 16:18:58,471 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.38 sec
MapReduce Total cumulative CPU time: 4 seconds 380 msec
Ended Job = job_1524770188000_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.38 sec   HDFS Read: 1501458 HDFS Write: 5
SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 380 msec
OK
+-------+--+
|  _c0  |
+-------+--+
| 2399  |
+-------+--+
1 row selected (22.041 seconds)
.....
SELECT actor_3_facebook_likes
 FROM imdb
WHERE Imdb_score>7;

04                |
| 250             |
| 670             |
| 6               |
| 21              |
| 51              |
| 93              |
| 0               |
| 131             |
| 178             |
| 27              |
| 0               |
```

| 27 |
| NULL |
| 56 |
| 50 |
| 117 |
| 476 |
| 130 |
+--+
383 rows selected (0.786 seconds)
.....

Difference between avg,min and max Budget

select avg(Budget) from imdb;;

Total MapReduce CPU Time Spent: 3 seconds 790 msec
OK
+--+
| _c0 |
+--+
| 3.527481639331849E7 |
+--+
1 row selected (22.644 seconds)

.....

select min(Budget) from imdb;
Total MapReduce CPU Time Spent: 3 seconds 420 msec
OK
+--+
| _c0 |
+--+
| 218 |
+--+

.....

select max(Budget) from imdb;

Total MapReduce CPU Time Spent: 3 seconds 440 msec
OK
+--+
| _c0 |
+--+
| 2127519898 |
+--+
.....

```
select Actor_2_name  from imdb  where Budget>4.5 limit 5;
OK
+-------------------+--+
|   actor_2_name    |
+-------------------+--+
| Joel David Moore  |
| Orlando Bloom     |
| Rory Kinnear      |
| Christian Bale    |
| Samantha Morton   |
+-------------------+--+
.....................................
0: jdbc:hive2:///> select Actor_2_name  from imdb  where Budget>4.5 and Imdb_score>8 limit 5;
OK
18/04/27 15:37:28 [main]: WARN lazy.LazyStruct: Extra bytes detected at the end of the row! Ignoring
similar problems.
+-----------------+--+
|  actor_2_name   |
+-----------------+--+
| Heath Ledger  |
| Jeffrey DeMunn  |
| Al Pacino         |
| Marlon Brando   |
| T.J. Storm        |
+-----------------+--+
```