Movie Data Analytics Project

Dataset

https://edureka.wistia.com/medias/7qd5lgmko4

Dataset Description

Column1: Movie ID
Column2: Movie name
Column3: Year of release
Column4: Rating of the movie
Column5: Movie duration in seconds

Problem Statement

A. Find the number of movies released between 1950 and 1960.
B. Find the number of movies having rating more than 4.
C. Find the number of movies with duration more than 2 hours (7200 second).
D. Find the list of years and number of movies released each year.
E. Find the total number of movies in the dataset.

Solution (using Hive):

Step 1: Create a database using following command.

```
create database project;
```

Step 2: Use your created database.

```
use project;
```

Step 3: Create a table using following command.

```
CREATE TABLE moviedata (movieid INT, name STRING, yearofrelease  INT,rating FLOAT,duration
INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

Step 4: Load data to the table created using following command.

```
 LOAD  DATA LOCAL INPATH 'moviedata.txt' OVERWRITE INTO TABLE moviedata;
```

Solution for problem (A)---Find the number of movies released between 1950 and 1960.

```
Select count(*) from (select distinct name from moviedata where yearofrelease between 1950 AND 1960)a;
```

Output:

```
Total MapReduce CPU Time Spent: 4 seconds 760 msec
OK
545
Time taken: 30.112 seconds, Fetched: 1 row(s)
```

Solution for problem (B)--Find the number of movies having rating more than 4.

```
select  count(distinct name) from moviedata where rating>4.0;
```

Output:

```
Total MapReduce CPU Time Spent: 8 seconds 70 msec
OK
841
Time taken: 29.037 seconds, Fetched: 1 row(s)
```

Solution for problem (C)-- Find the number of movies with duration more than 2 hours (7200 second).

```
select count(distinct name) from moviedata where duration>7200;
```

Output:

```
Total MapReduce CPU Time Spent: 4 seconds 900 msec
OK
641
Time taken: 14.998 seconds, Fetched: 1 row(s)
```

Solution for problem (D)-- Find the list of years and number of movies released each year.

```
select yearofrelease,count(distinct name) from moviedata group by yearofrelease;
```

Output:(due to limited space I have taken first 10 results)

```
OK
1913 3
1914 20
1915 1
1916 1
1918 1
1919 3
1920 6
1921 2
1922 2
1923 4
```

Solution for problem (E)-- Find the total number of movies in the dataset.

```
Select count(distinct name) from moviedata;
```

Output:

```
Total MapReduce CPU Time Spent: 4 seconds 470 msec
OK
49143
Time taken: 15.096 seconds, Fetched: 1 row(s)
```