Movie Data Case Study

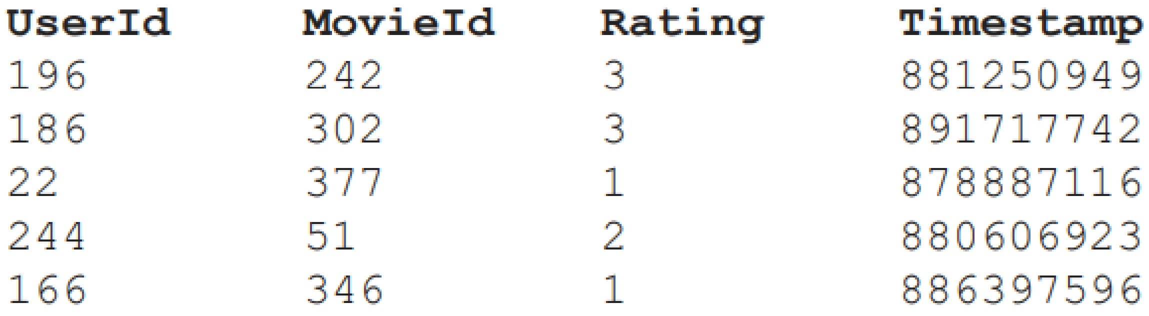
Business Scenario:

MovieLens data sets were collected by the GroupLens Research Project at the University of Minnesota. The data sets contain reviews of movies by different users. This data set consists of 100,000 ratings given by 943 users on 1682 movies. Each user has rated at least 20 movies.

Data Sets and Data Dictionary:

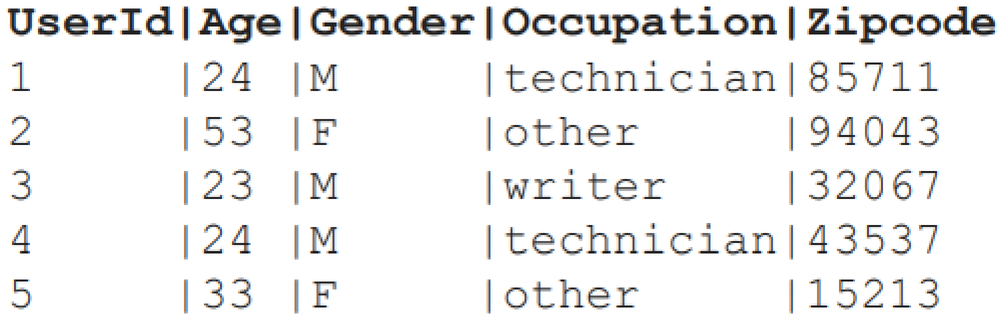
The data is given in delimited files. The data dictionaries for each of the file are given below:

u.data contains 100,000 records in tab limited format as show below. The data file given does not have the header row



UserId	MovieId	Rating	Timestamp
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
166	346	1	886397596

u.user contains details of 1682 users with columns of each row delimited by | character. The given file does not contain the header row.

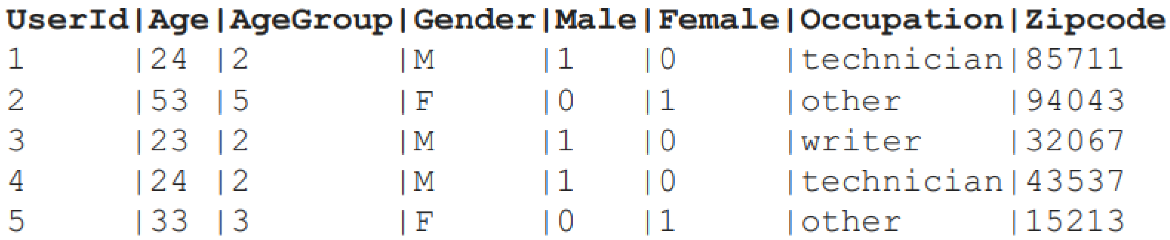


```
UserId|Age|Gender|Occupation|Zipcode
1     |24 |M     |technician|85711
2     |53 |F     |other     |94043
3     |23 |M     |writer    |32067
4     |24 |M     |technician|43537
5     |33 |F     |other     |15213
```

Problem Statement:

Use PySpark and load the data into data frames. Write code to perform the following tasks.

1. Give gender-wise breakup of the users
2. Give the top 5 movies which are reviewed maximum number of times
3. Give the top 5 users who reviewed maximum number of movies
4. List the top 10 movies which received highest number of 5 star ratings
5. List the top 10 users who gave highest number of 5 star ratings
6. Prepare a dataframe from user data that has the records transformed as below.
7. Save the dataframe as a Hive table in your database.



```
UserId|Age|AgeGroup|Gender|Male|Female|Occupation|Zipcode
1     |24 |2       |M     |1   |0     |technician|85711
2     |53 |5       |F     |0   |1     |other     |94043
3     |23 |2       |M     |1   |0     |writer    |32067
4     |24 |2       |M     |1   |0     |technician|43537
5     |33 |3       |F     |0   |1     |other     |15213
```

The transformations to be done are as below:

1. Add age group and populate it like age 24 = 2, age 53 = 5, age 30 =3 etc as shown above
2. Adding a dummy variable for the column Gender and filling in 1 or 0 accordingly