Project

MovieLens data sets were collected by the GroupLens Research Project at the University of Minnesota. It represent users' reviews of movies.

This data set consists of:

- 100,000 ratings (1-5) from 943 users on 1682 movies
- Each user rating for at least 20 movies
- Simple demographic info for the users (age, gender, occupation, zip)

u.data: The full u data set, 100000 ratings by 943 users on 1682 items

- Each user has rated at least 20 movies.
- Users and items are numbered, consecutively from 1.
- The data is randomly ordered.
- This is a tab separated list of user id | item id | rating | timestamp.
- The time stamps are unix seconds since 1/1/1970 UTC.

u.user: Demographic information about the users

- This is a tab separated list of user id | age | gender | occupation | zip code.
- The user IDs are the ones used in the u.data data set.

PS**:** Stores u.data and u.user in HDFS by running the HDFS command

Perform the following steps using Spark SQL:

1. Create a u_data table in Hive using Spark SQL.
2. See the field descriptions of the u_data table.
3. Load data into the u_data table from a HDFS text file.
4. Show all the data in the newly-created u_data table.
5. Show the number of items reviewed by each user in the newly-created u_data table.
6. Show the number of users that reviewed each item in the newly created u_data table.
7. Create a u_user table in Hive.
8. See the field descriptions of the u_user table.
9. Load data into the u_user table from a HDFS text file.
10. Show all the data in the newly-created user table.
11. Count the number of data in the u_user table.
12. Count the number of users in the u_user table, gender-wise.
13. Join the u_data table and u_user tables based on user IDs.