Assignment 3

20 November 2019

- ⇒ Movie information is in the file "movies.csv" and is in the following format:
  - ₪ MovieID::Title::Genres
- ⇒ All ratings are contained in the file "ratings.csv" and are in the following format:
  - ₪ UserID::MovieID::Rating::Timestamp
    - δ UserIDs range between 1 and 6040
    - δ MovieIDs range between 1 and 3952
    - δ Ratings are made on a 5-star scale (whole-star ratings only)
    - δ Timestamp is represented in seconds since the epoch as returned by time(2)
    - δ Each user has at least 20 ratings
- ⇒ User information is in the file "users.csv" and is in the following format:
  - ₪ UserID::Gender::Age::Occupation::Zip-code
    - δ Gender is denoted by a "M" for male and "F" for female
    - δ Age is chosen from the following ranges:
      - ϖ 1: "Under 18"
      - ϖ 18: "18-24"
      - ϖ 25: "25-34"
      - ϖ 35: "35-44"
      - ϖ 45: "45-49"
      - ϖ 50: "50-55"
      - ϖ 56: "56+"
- ⇒ Occupation information is in “occupation.csv”

1. Create a database called moviedetails and use it
2. Create tables to hold the movie and ratings data in Hive.
3. Display 20 records of movies and ratings.
4. Find out numbers of non-adults as per Indian standard, who has rated movies
5. Find the age of the most rated user with counts of rating
6. Find the count of the ratings based on age
7. Find the movie with the maximum number of ratings.
8. Find the movie with the lowest rating.
9. Find the movie with the maximum people from age 45 and 50 rating it.
10. Create a database called userdetails and use it.
11. Create tables to hold user and occupation data in Hive.
12. Find out occupation of all the users
13. Find out the no of users with same occupation and having age more than 25 along with occupation details
14. Find the occupation of all female users.
15. Find the number of female and male working as doctor/health care.