FRAGMA DATA

CODING PROBLEM

The problem mentioned below revolves around movies dataset. The dataset contains 4 files which are follows,

File Name	Description / Schema
movies.dat	MovieID – Title – Genres
ratings.dat	UserID – MovieID – Rating – Timestamp
users.dat	UserID – Gender – Age – Occupation – ZipCode
README	Additonal information / explanation about the above three files

The dataset can be downloaded from the link : http://grouplens.org/datasets/movielens/1m/

Please submit your code in

Hadoop Map Reduce OR JAVA OR Spark scala ONLY

[DO NOT use HIVE QUERIES or SPARK SQL] to solve below questions:

1. Top ten most viewed movies with their movies Name (Ascending or Descending order)

2. Top twenty rated movies (Condition: The movie should be rated/viewed by at least 40 users)

3. We wish to know how have the genres ranked by Average Rating, for each profession and age group. The age groups to be considered are: 18-35, 36-50 and 50+.

You need to formulate results in following table:

Occupation	Age Group	Genre Ranking by Avg. Rating				
		Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Programmer	18-35	Action	Suspense	Thriller	Romance	Horror
Programmer	36-50	Action	Suspense	Thriller	Romance	Horror
Programmer	50+	Action	Suspense	Thriller	Romance	Horror
Farmer	18-35	Action	Suspense	Thriller	Romance	Horror
Farmer	36-50	Action	Suspense	Thriller	Romance	Horror

Note that values populated in following table are just representative, and will change with the actual data.

The table should be output as a single CSV file, rows sorted by Occupation followed by Age Group.

Instructions:

1) Please make your program readable and well structured. Showcase your object-oriented skills and/or functional programming skills.

2) Your solution should be scalable to larger data sets.

3) We are not particular about formatting of output. It should just be readable.

4) You would get 3 Days to revert with the solution.

5) You can send us the ZIP file with the source code.