Spark Movielens Analysis – Movie Preference by Profession

**Skills Required:** Apache Spark SQL

**Description:**

This case study is used to analyze Movie lens Data set which has details about users, ratings, movies, genre using Spark SQL.

**Input Data Set:**

Sample Data:

**Movies (u.item)**

MovieID::Movie Name::Genre

1|Toy Story (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)|0|0|0|1|1|1|0|0|0|0|0|0|0|0|0|0|0|0|0

2|GoldenEye (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?GoldenEye%20(1995)|0|1|1|0|0|0|0|0|0|0|0|0|0|0|0|0|1|0|0

3|Four Rooms (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Four%20Rooms%20(1995)|0|0|0|0|0|0|0|0|0|0|0|0|0|0|0|0|1|0|0

**Ratings (u.data)**

userID MovieID Rating UnixTimeStamp

196 242 3 881250949

186 302 3 891717742

22 377 1 878887116

244 51 2 880606923

**Users (u.user)**

UserID::Gender::Age::Occupation::Zip-code

1|24|M|technician|85711

2|53|F|other|94043

3|23|M|writer|32067

4|24|M|technician|43537

**Genre (u.genre)**

unknown|0

Action|1

Adventure|2

Animation|3

Children's|4

Comedy|5

Crime|6

Documentary|7

Drama|8

Fantasy|9 Spark Movielens Analysis – Movie Preference by Profession

Film-Noir|10

Horror|11

Musical|12

Mystery|13

Romance|14

Sci-Fi|15

Thriller|16

War|17

Western|18

Source Data: http://files.grouplens.org/datasets/movielens/ml-100k.zip

**Requirement**

Use the movielens dataset as input and perform the following analysis:

Find the preference in movie genre based on profession:

**Output required:**

Profession, Genre, wa\_rating

Wa\_rating = (no ratings > 3) / total ratings

Output should be sorted by Profession and Genre in ascending order and wa\_rating in descending order

The above analysis should be performed using Spark SQL Data Frame API.