**REPORT – Assignment 6**

In this project, we had to find the movie recommendations for every user along with their scores. First I installed Mahout in my local machine ran in standalone mode. For this, we used ‘GroupLens Movie’ dataset which provides rating of movies.

This archive contains:

1) **u.data:** contains several tuples(user\_id, movie\_id, rating, timestamp)

2) **u.user:** contains several tuples(user\_id, age, gender, occupation, zip\_code)

3) **u.item:** contains several tuples(movie\_id, title, release\_date, video\_release\_data, imdb\_url, cat\_unknown, cat\_action, cat\_adventure, cat\_animation, cat\_children, cat\_comedy, cat\_crime, cat\_documentary, cat\_drama, cat\_fantasy, cat\_film\_noir, cat\_horror, cat\_musical, cat\_mystery, cat\_romance, cat\_sci\_fi, cat\_thriller, cat\_war, cat\_western)

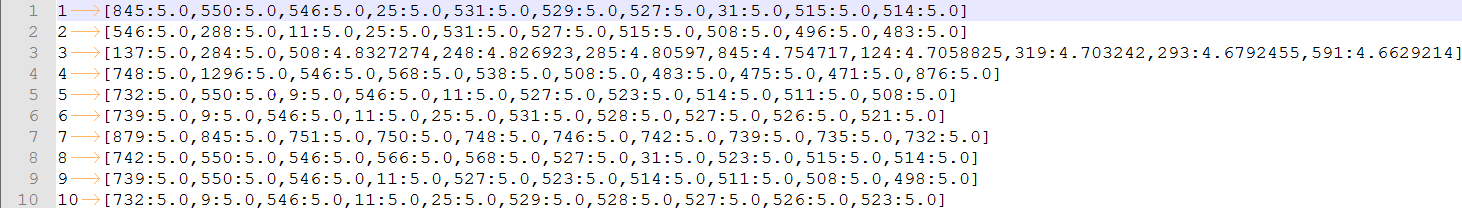
This data set contains 943 users, 1,682 movies and 100,000 ratings.

To run, I used the modified command:-

**-core-0.7-job.jar org.apache.mahout.cf.taste.hadoop.item.RecommenderJob --input mahout/u.data --output mahout --usersFile mahout/u.user --itemsFile u.item --booleanData true --similarityClassname SIMILARITY\_COOCCURRENCE**

Mahout computes the recommendations by running several Hadoop mapreduce jobs.

After 30-50 minutes, the jobs are finished and each user will have the 10 movies that she might mostly like based on the co-occurrence of each movie in users’ reviews



The key is the userId of the user and the value is the ‘movie ID : Similarity Rating’.

1. **How long does it take in Standalone mode?**

In standalone mode, it took 1min and 36 seconds to run all the Map reduce programs.

1. **How long does it take in HDFS pseudo-distributed?**

In distributed mode, it took 7min and 50sec to run all the Map Reduce programs.

1. **What if you had 100 VMs on separate physical hardware?**

In fully distributed mode, it would have been much faster since, the hardware is not shared. It would have been almost 100 times faster.

It was interesting to note that standalone mode was faster than HDFS pseudo distributed mode. According to my understanding, it was due to the networking overhead caused due to communication between the daemons and JVMs. If the datasets were larger, pseudo distributed mode would have performed much better.