CS 6301- Big Data Analytics and Management

Fall 2014

Homework # 1

Due Date : October 12, 2013 , 11.59pm

In this homework you will learn how to solve problems using Map Reduce. Please apply Hadoop map-reduce to derive some statistics from IMDB movie data. You can find the dataset in elearning. Copy the data into your hadoop cluster and use it as input data.
You can use the ***put*** or ***copyFromLocal*** HDFS shell command to copy those files into your HDFS directory.

There are 3 datafiles :: **movies.dat, ratings.dat, users.dat**
Please read the "**README_Important**" file to know about the data organization and to know about the **Attribute** of the data. All are very well explained in that README_Important file. In class there will be brief demo/ discussion about that. Please read the questions carefully and use only the data file that you need. Some question may need only users.dat, or some question may need only ratings.dat

After being familiar with the data - you are required to write efficient Hadoop ***Map-Reduce programs in Java*** to find the following information ::

Q1. **Given a input zipcode, find all the user-ids that belongs to that zipcode. You must take the input zipcode in command line**.
[For example, if the input zipcode is 75252 then you need to find all users who lives in 75252]
[You only need users.dat file to get the answer.]

Q2. **Find top 10 average rated movies with descending order of rating** (Use of **chaining** of multiple map-reduce job is a **must** here)
[**Clue** : From the dataset we know that, each user can rate mulitple movies and each movie can be rated by multiple users and this information is found in **ratings.dat** file . So, first we have to find the average rating of a movie and second, we need to find the top 10 average rated movies.

CS 6301- Big Data Analytics and Management

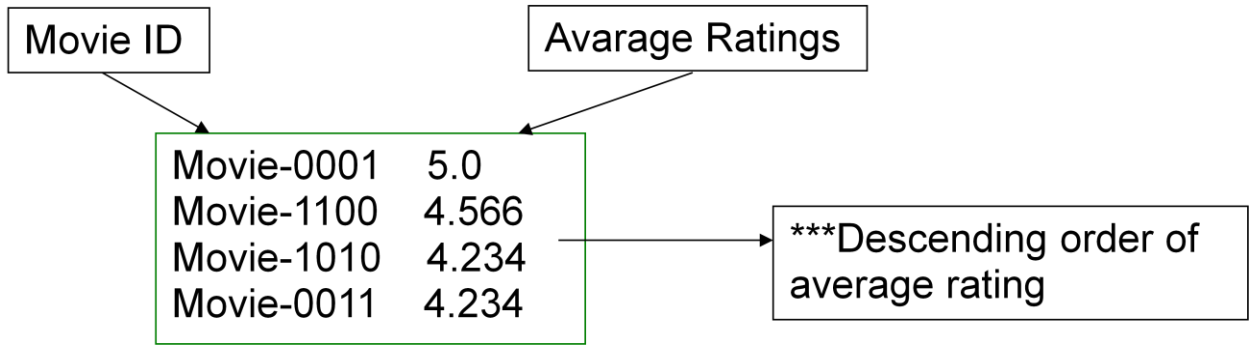
Fall 2014

Homework # 1

Due Date : October 12, 2013 , 11.59pm

So, First map-reduce job will find the average rating of each movies and write into disk. And the second map-reduce job will take the previous output as the input and find the top 10 from them.]
[You only need ratings.dat file to get the answer.]

Sample output of Q2 :



***Map-Reduce Design Pattern – Donald Miner & Adam Shook :: Read this book for **how to get top 10** and do **chaining** multiple map-reduce jobs, the book can be found on **eLearning** : Course Homepage > Big Data Important Resources

Submission ::

You have to upload your submission via e-learning before due date.
Please upload the following to eLearning:

1. Two **jar** files, one for each problem/ One jar file containing all solutions.
2. The Two/one jar-matching **java** files which have the source code.
3. ***A **Readme** text file about how to run your jar file. Give the command to run your jar file.