CS 6301- Big Data Analytics and Management

Spring 2014

Homework/Assignment# 1

Due Date : TBD

In this homework-1 you will learn how to solve problems using Map Reduce. Please apply Hadoop map-reduce to derive some statistics from IMDB movie data.
Currently, the dataset are currently available at :: **cs6360.utdallas.edu/tmp/IMDB_DataSet**
To get the data please login into **cs6360.utdallas.edu** using putty, login name and password is your netID and Password. After login you can find the dataset here :: **/tmp/IMDB_DataSet**

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.

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

Q1. Find all the user ids who has rated at least n movies. (n=30 or 40) [use only ratings.dat as input file]

Q2. Given some movie titles - find all the genres of the movies**. (Taking the movie titles as command line input is a must here)** [use only movies.dat as input file]

Q3. Find the top 10 zipcodes based on the avarage age of users belong to that zipcode, in the descending order of the avarage age. Top 10 means the youngest 10 avarage age of users of that zipcode**. (Use of Chaining of two map-reduce job is a must here.)** [use only users.dat as input file]

Please DO NOT copy the input files again in the HDFS as they are already copied into the HDFS here : **/Spring2014_HW-1/input_HW-1/**
Please use the above directory as the input. AND ::
Please use **/tmp/<your net-id>/output** directory to generate your output.

Detail discussion/demo about the questions will be presented in the class and sample output data of each of 3 questions will be given as example soon, so that you can understand which output format is needed.
Submission ::
Probably, you have to come and show us your submission on some date OR, you might be given similar kind of problem on spot to solve with enough time given . We will take decision about that later. You have to upload your submission **before due date** and on the demo day, you will download your submission and run on the spot. Do not upload any fake/false code – which will be highly penalized.
Please upload the following to eLearning:

1. Three jar files, one for each problem/ One jar file containing all solutions.
2. The Three/One jar-matching java files which have the source code.
3. A ReadMe file explaining the commands to run your program.