CS6350:Big Data Analytics and Management

Spring 2015

DUE DATE: Feb 17,2015

TA: Gbadebo Ayoade
gga110020@utdallas.edu

Homework 1

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" file to know about the data organization and to know about the Attribute of the data. All are very well explained in that README 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 movies.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 **list all male user id whose age is less or equal to 7 .**

Using the users.dat file, list all the male userid filtering by age. This demonstrates the use of mapreduce to filter data.

Q2 **Find the count of female and males users in each age group**
The age distribution is given below (same as in read me file)

* 7: "Under 18"
* 24: "18-24"
* 31: "25-34"
* 41: "35-44"
* 51: "45-55"
* 56: "55-61"

* 62: "62+"

Use the **users.dat** file.

A sample output is given below

//Age Gender and Count
7 M 200
24 F 120

where age is 7, gender is male and count is 200.

Q3 **List all movie title where genre is “fantasy”**

The **genre input** must be taken from **command line.**

Use the **movies.dat** file

NB:
To run your jobs use the following synthax
hadoop jar name_of_jar_file Classname <input dir> <output dir> [<extra input paramter>(may be optional due to question e.g genre input)]

Submission ::
You have to upload your submission via e-learning before due date.
Please upload the following to eLearning:
1. Three jar files, one for each problem/ One jar file containing all solutions.
2. 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.