CS6350:Big Data Analytics and Management

Spring 2015

DUE DATE: March 6, 11:59pm
TA: Gbadebo Ayoade
gga110020@utdallas.edu

Homework 2

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 (**Use the same data as in homework 1**)
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: Find top 5 average movies rated by female users and print out the titles and the average rating given to the movie by female users.

This question involves filtering , joining data from multiple files and job chaining.
You should **use reduce side join** for this question.

Note: First of all get all movies rated by female users, then find the average rating given to each movie by female users.

You will need all the three data files for this question.

Please check out the lecture on hadoop programming given in class for implementation sample. **Hadooplecture2withsamplecode.zip. You can find the document at the url**

below from elearning (Requires elearning username and password.)
https://elearning.utdallas.edu/bbcswebdav/pid-727586-dt-content-rid-6316587_1/xid-6316587_1

e.g
Given ratings.dat as (**Note this is just an example data format for simplicity**.)

userid	movied	ratings
1	30	4
1	40	3
2	20	3
2	30	4
3	30	2

users.dat as

userid	gender
1	M
2	F
3	F

we can see from this data that user 2 and 3 are Females.

We can see movie id 30 is rated by users 2 and 3, therefore the average rating given to the movieid with id 30 by female users is
4+2/2 = 3.
since user 2 and 3 are females and they rated the movie with id 30 ratings 4 and 2 respectively.

Note: we are ignoring the rating given by the male user with id 1 even though the user rated the movie with id 30 also.

You will then join the results obtained above to the movies.dat file to get the title.

Your final result can be in the following format

"title of movies"	"avg rating by females"
Toy Story	3.5

To run your jobs use the following syntax
hadoop jar name_of_jar_file Classname <input dir> <output dir> [<extra input paramter>

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. 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.

Q2.

Given the id of a movie, find all userids,gender and age of users who rated the movie 4 or greater.

You will input the movie id from command line.

For this question **use map side join** to implement join in hadoop.

To implement map side join, you will be loading users.dat in the hadoop **distributed cache**.

Please, check out the second lecture on hadoop programming given in class for implementation sample. **Hadooplecture2withsamplecode.zip. You can find the document at the url below on elearing ((Requires elearning username and password.)**

https://elearning.utdallas.edu/bbcswebdav/pid-727586-dt-content-rid-6316587_1/xid-6316587_1

Use the users.dat and ratings.dat