Non-Graded Assignment

T2.2 Advanced Map Reduce

Assignment: Program is designed to parse user access log files and provide details about user's accessing the system within a given time frame.

Problem Statement:

Given is a User Access log file dataset that has user_id and his login_time into a particular system. The format of the data is:

<user_idDesign> <access_time in system's current milliseconds>

Design a program that will accept start time and end time from the user. The program should tell me the list of all the users and the number of times each user logged in during the given time period.

Example: Consider a sample Log file as follows: U1 123 U2 124 U1 127 U3 140 If the time period provided is between 120 and 130, the result should be: U1 2 U2 1

Sample log files are provided in logfiles directory Logfiles need to be copied to Hadoop's HDFS file system to work upon. Run the program by passing following parameters at run time

a. Root folder location in HDFS file-system where log files are copied

b. Expected folder location in HDFS file-system where results will be stored. Note that the folder should not exist before-hand

c. Start time (in milliseconds) from where, the program should take into account useraccess

d. End time (in milliseconds) till where, the program should take into account user-access

Seismic Analytics:

1. Suppose you have just become the Development Lead for a company which specializes in reading seismic data which measure earthquake magnitudes around the world. There are thousands of such sensors deployed around the world recording earthquake data in log files in the following format:

nc,71920701,1,"Saturday, January 12, 2013 19:43:18 UTC",38.7865,122.7630,1.5,1.10,27,"Northern California"

Each entry consists of lot of details. The items in red are the magnitude of the earthquake and the name of region where the reading was taken, respectively.

Your Director of Software asks you to perform a simple task: for every region where sensors were deployed, find out the highest magnitude of the earthquake recorded.

Example: For the log file:

nc,71920701,1,"Saturday, January 12, 2013 19:43:18 UTC",38.7865,122.7630,1.5,1.10,27,"Northern California"

nc,71920702,1,"Monday, January 14, 2013 19:43:18 UTC",38.7865,122.7630,3.2,1.10,27,"Northern California"

nc,71920703,1,"Saturday, January 12, 2013 19:43:18 UTC",38.7865,122.7630,1.5,1.10,27,"Canada"

nc,71920704,1,"Saturday, January 12, 2013 19:43:18 UTC",38.7865,122.7630,2.5,1.10,27,"New York"

nc,71920705,1,"Sunday, January 13, 2013 19:43:18 UTC",38.7865,122.7630,0.5,1.10,27,"Canada"

The output should be:

Northern California 3.2

Canada 1.5

New York 2.5