Homework 3 - Hadoop and MapReduce

This homework consists of three parts: Questions about Hadoop file systems HDFS, questions about programming MapReduce Tasks, and a programming exercise.
Please send in one file containing the answers to the question, as well as either a java file or two python files for the programming exercise (see details below). For the programming exercise please also send a README.txt file that contains the instructions how to compile (if necessary) and how to execute the program.
Submit until Sunday the 13th April 4pm.

1) Write the corresponding HDFS commands to perform the following tasks. Each of these tasks must be accomplished with a single HDFS command. Hint: type hadoop fs -help for the list of commands available. Note the difference between the local (Linux) file directory and HDFS directory. To double-check your answers, you're encouraged to test the commands (on your own HDFS directory) to make sure they work correctly.

. (a) Moving a file named abc.txt located in the local directory path /user/cse491 to the following path in HDFS /user/hduser/data/

. (b) Displaying the entire content of the file data.txt located at the following path in HDFS /user/hduser/data/

. (c) Displaying only the last 1kB of the file data.txt located at the following path in HDFS /user/hduser/data/.

. (d) Copying a file named part-r-00000 from the HDFS directory /user/ cse491/ output/ to the local directory /user/cse491

. (e) Deleting all the contents (including subdirectories) of the following HDFS directory /user/cse491/output.

. (f) Listing all the filenames and subdirectories stored under the directory named / user/yourusername/output.

2) For each problem and data set described below, state how would you setup the (key,value) pairs as inputs and outputs for the mappers and reducers. If the mapper must perform a filtering step before outputting its key-value pairs, make sure you specify the type of filtering needed. Also, explain the operation performed by the reduce function given its input. Assume your Hadoop program uses TextInputFormat as its input format (where each record corresponds to a line of the input file). Since the inputs for the mappers are the same (byte offset, content of the line), you only need to specify the mappers' outputs.

Example:

Data set: Collections of text documents.

Problem: Count the frequency of words that appear at least 100 times in the documents.

Answer:

Mapper output: key is a word, value is 1 (no filtering needed).

Reducer input: key is a word, value is list of 1's.

Reduce function: sums up the 1's for each key (word).

Reducer output: key is a word, value is frequency of the word (filter words whose frequencies are below 100).

(a) Data set: Student transcript database. Each record consists the following information: student ID, course code, semester enrolled, number of credits, and GPA for the course. For example:

A123456074 CSE260 FS2013 3 3.5

GPA is a numeric-valued attribute ranging from 0.0 to 4.0. Problem: Compute the average GPA for each student. Ignore all the incomplete grades (GPA = NULL) when computing average GPA. Answer:

Mapper output:

Reducer input:

Reduce function:

Reducer output:

(b) Data set: Twitter follower graph. Each record corresponds to a 2-tuple (follower,followee). For example, the record

john123 mary456
means john123 is a follower of the tweets posted by mary456. Problem: Find all pairs of users who have reciprocal relation. For example, if john123 and mary456 has a reciprocal relation, then john123 is a follower of mary456, and vice-versa.
Answer:

Mapper output:

Reducer input:

Reduce function:

Reducer output:

(c) Data set: Online review data. Each line in the data file has 3 columns (user id, product id, rating), where ratings are integer-valued ranging from 1 (bad) to 5 (good). Problem: For each user, list all the products they liked (i.e., prod- ucts that have ratings higher than 3).

Answer:

Mapper output:

Reducer input:

Reduce function:

Reducer output:

(d) Data set: Maximum and minimum daily temperature readings for weather stations from around the world. Each line in the data files has 3 columns (station id, date, max temperature, min temperature). Problem: Find the station id and date of anomalous temperature readings. A temperature reading is anomalous if its min temperature exceeds max temperature for the given day.
Answer:
Mapper output:

Reducer input:

Reduce function:

Reducer output:

(e) Data set: Online dating data. Each line in the data file has information about an attribute and the list of users who possesses such attribute. For example, the line:

classical music 1412 2313 2553 3901 8101 12005
indicates users 1412, 2313, 2553, 3901, 8101, and 12005 enjoy listening to classical music.
Problem: Compute the jaccard similarity between all pairs of users. Answer:

Mapper output:

Reducer input:

Reduce function:

Reducer output:

3) On the cloudera image you find a folder called datasets that has a zip file in it called:
median_income_by_zip_code_census2000.zip

Unzip that file by double cicking it in the file browser. It will unpack several files, but the one we are interested in is the .csv file. It contains a couple of comma separated columns:
ID, ZIP, “ZIP detail”,income

We want to use Hadoop and MapReduce to analyze this file. You can do this in Java or python streaming API. Please hand in your results and your code (.java file, or the mapper and reducer .py file)

For each interval of 10.000 zip codes we want to know the minimum and maximum salary in the table for all those zip code ranges. Observe that there are zip codes like 999HH, please ignore all zip codes that are not integer numbers.