CSOE18 Assignment #2: Hadoop MapReduce for Climate Data Analytics

Due: **0900 hrs IST, 01 December 2023**. Negotiable. Focus must be on learning and discovering new stuff!

Objective: The objective of this project is to familiarize with Hadoop Installation and MapReduce and to apply them in real time data analytics.

General Instructions:
Total Points: **40.** Weightage: **15**
This assignment needs to be done individually and implemented in **Java**.
As always, please feel free to approach me in person in office hours or through Teams, if you have any questions. I'd be happy to help out.

Marking Criteria: Your submission will be marked using the following criteria.

- Clarity of the code with proper indentation and comments.
- Showing good efforts through completed tasks.
- Understanding of fundamental concepts and optimization of the code.
- Showing attention to details through a good quality document.

Your code will be run on Anti-plagiarism software. Copying others contents will be seriously viewed, which will lead to heavy penalties for both the donor and the recipient. I encourage you, not to rush at the last moment.

Task #1 (6 Points)
Consider the following lyrics from the poem "Ungrateful Sorrow" by Indian Literature Nobel Laureate Rabindranath Tagore, which is your input file for this task.

At dawn shey(1) departed
My mind tried to console me –
" Everything is Maya(2)".
Angrily I replied:
"Here's this sewing box on the table,
that flower-pot on the terrace,
this monogrammed hand-fan on the bed---
all these are real."

My mind said: "Yet, think again."
I rejoined: "You better stop.
Look at this storybook,
the hairpin halfway amongst its leaves,
signaling the rest is unread;
if all these things are "Maya",
then why should "shey" be more unreal?"

My mind becomes silent.
A friend arrived and says:
"That which is good is real
it is never non-existent;
entire world preserves and cherishes it its chest
like a precious jewel in a necklace."

I replied in anger: “How do you know?
Is a body not good? Where did that body go?”

Like a small boy in a rage hitting his mother,
I began to strike at everything in this world
that gave me shelter.
And I screamed:” The world is treacherous.”

Suddenly, I was startled.
It seemed like someone admonished me :”You- ungrateful !”

I looked at the crescent moon
hidden behind the tamarisk tree outside my window.
As if the dear departed one is smiling
and playing hide-and-seek with me.

From the depth of darkness punctuated by scattered stars
came a rebuke: “when I let you grasp me you call it an deception,
and yet when I remain concealed,
why do you hold on to your faith in me with such conviction?”

The following two sentences are for your Understanding alone, which is not part of the above input file:
- “Shey” in Bengali can mean either he or she.
- “Maya” meaning Unreal.

The input pairs for the Map phase will be the following:
(0, “At dawn shey(1) departed”) (24, “My mind tried to console me –”) ….

The key is the *byte offset* starting from the beginning of the file. While we won’t need this value in Word Count, it is always passed to the Mapper by the Hadoop framework. The byte offset is a number that can be large if there are many lines in the file.

a. What will the output pairs look like?
b. What will be the types of keys and values of the input and output pairs in the Map phase?

Remember that instead of standard Java data types (String, Int, etc.), Hadoop uses data types from the org.apache.hadoop.io package. You can check Hadoop’s API at the following URL: http://hadoop.apache.org/docs/r2.4.0/api/

For the Reduce phase, some of the output pairs will be the following:
("*Shey*", 2) ("*Maya*", 2) ("*you*", 2)...

c. What will the input pairs look like?
d. What will be the types of keys and values of the input and output pairs in the Reduce phase?
e. Write map () function for Questions a and b.
f. Write reduce () function for Questions c and d.

Task #2 (4 Points)

We will use large (Big!) dataset text files for this task and you can find this dataset in the Assignment-2 folder on MS Teams:

a. Run Word Count on the file Gberg-100M.txt.

Note: First, we have to copy that file to the HDFS.

Output should be as follows:

A 18282
AA 16
AAN 5
AAPRAMI 6
AARE 2
AARON 2
AATELISMIES 1

b. How many Map and Reduce tasks did running Word Count on Gberg-100M.txt produce? Run it again on Gberg-200M.txt and Gberg-500M.txt and write your observations. Additionally, run the following command on the cluster:
$ hdfs getconf -confKey dfs.blocksize

c. What is the link between the input size, the number of Map tasks, and the size of a block on HDFS?

Submission Logistics for Task #1 and #2: Directory name: Task_1 and 2
Contents: A .pdf file containing answers for Task 1, Task 2b & 2c. File Name: Task 1 and 2_Answers
An output file for Task 2a. File Name: Task 2a_WC_Outcome
A single file containing all the functions and classes of word_Count problem. File Name: Your class name.java

Task #3 (20 Points)

National Climatic Data Centre of the U.S. is the world's largest active archive of weather data, which produces CSV (Comma-Separated Values) files with worldwide weather data for each year. Each line of one of these files contains:

- The weather station's code.
- The date, in the ISO-8601 format.
- The type of value stored in that line. All values are integers. TMIN (resp. TMAX) stands for minimum (resp. maximum) temperature. Temperatures are expressed in tenths of degrees Celsius. AWND stands for average wind speed, and PRCP stands for precipitation (rainfall), etc. Several other types of records are used (TOBS, SNOW,...).
- The next field contains the corresponding value (temperature, wind speed, rainfall, etc.)
- All lines contain five more fields that we won't use in this assignment.

We will work on the CSV file for the year 2013, which has been sorted by date first, station second, and value type third, in order to ease its parsing. It can be found in the Assignment-2 folder with name ncdc-2013-sorted.csv. Here is a sample of that file:

```
...
FR000007650,20130102,PRCP,5,,,S,
FR000007650,20130102,TMAX,111,,,S,
FR000007747,20130102,PRCP,3,,,S,
FR000007747,20130102,TMAX,117,,,S,
FR000007747,20130102,TMIN,75,,,S,
FR069029001,20130102,PRCP,84,,,S,
FR069029001,20130102,TMAX,80,,,S,
FS000061996,20130102,PRCP,0,,,S,
FS000061996,20130102,TMAX,206,,,S,
FS000061996,20130102,TMIN,128,,,S,
GG000037279,20130102,TMAX,121,,,S,
GG000037308,20130102,TMAX,50,,,S,
GG000037308,20130102,TMIN,-70,,,S,
GG000037432,20130102,SNWD,180,,,S,
GG000037432,20130102,TMAX,15,,,S,
GG000037432,20130102,TMIN,-105,,,S,
...
```

As you can see, not all stations record all data. For instance, FR069029001 only recorded rainfall and maximum temperature on 01/02/2013. Not all stations provide data for every day of the year either.

NCDC wants to plot the difference between the maximum and the minimum temperature in Central Park for each day in 2013. There is a weather station in Central Park: its code is USW00094728. If we have a look at the TMIN and TMAX records for that weather station, they look like this (USW00094728 provides minimum and maximum temperature data for every day of the year).

```
USW00094728,20130101,TMAX,44,,,X,2400
USW00094728,20130101,TMIN,-33,,,X,2400
USW00094728,20130102,TMAX,6,,,X,2400
USW00094728,20130102,TMIN,-56,,,X,2400
USW00094728,20130103,TMAX,0,,,X,2400
USW00094728,20130103,TMIN,-44,,,X,2400
...
```

a. Let's assume a solution which proposes to use a MapReduce job that does the following:

- The Map task(s) send(s) (,) pairs to the Reducer. Temperatures are converted to degrees Celsius. The output will be: (4.4, -3.3), (0.6, -5.6)...
- For each key/value pair, the Reduce task subtracts the minimum temperature from the maximum temperature, converts it to degrees, and writes the result to a file.

What is wrong with this approach? Can you propose an alternative solution?

Submission Logistics: Directory name: Task_3a
Contents: 1 .pdf files containing your answers with plots. File name: Task3a_Answers
1 Code file and 1 Output file. File name: class_name.java, task3a_op

b. Propose one more solution that the Map phase will be a clean-up phase that discards useless records, and for each day, it will calculate the temperature difference.

i. What will the key and values, output by the Map tasks be? What types will they be?
ii. Can the Mapper produce a key/value pair from a single input? How can we solve this issue?
iii. For each day and weather station, the TMAX record always precedes the TMIN in NCDC's data, and we suppose that no split occurs between a TMAX and a TMIN record. By following this approach, what work will be left to the Reduce task?
iv. Reducer classes extend Hadoop's Reducer class. Read Hadoop's documentation for that class, in particular, what the default behaviour of its reduce() function is. What can you deduce from this? Write the code for both solutions and plot the results.

Submission Logistics: Directory name: Task_3b
Contents: 1 .pdf files containing your answers with plots. File name: Task3b_Answers
1 Code file and 1 Output file. File name: class_name.java, task3b_op

c. Plotting average worldwide temperature variations.

Similarly, you will produce a one-column CSV file. Since not all stations provide all of the data, you will have to be careful that you always subtract the minimum temperature from the maximum temperature of the same station. If a station doesn't provide both the minimum or maximum temperature for that day, it will be ignored. Keep in mind that given the way the file is sorted; the minimum temperature will always follow the maximum temperature for a station on a given day.

i. Write a Reducer that will perform the job, and modify your Mapper accordingly. While your Reducer can use the type FloatWritable for the result, you will use a double when you sum up results to compute the average, in order to make sure to not lose precision when summing up a large number of floating-point values. Write the code for the above job and Plot the results
Note: They should start with 9.857165, 9.882375, 10.542754...

Submission Logistics: Directory name: Task_3c
Contents: A .pdf file with the plots. File name: Task 3c_Answers
One Code file and One Output file. File names: class_name.java, task3c_op

What else?

- Every Task should be implemented in separated code files. Since the dataset and problems are predefined, do not prompt for user input anywhere in the program. When executing each program file for the given problem, specified outputs should be generated automatically (without any user intervention).
- Mention the References in the .pdf files, which you have gone through.

Moodle Submission Logistics:

This is to be completed and submitted to Moodle. By the due date, submit one zip file containing:

- Four directories with appropriate files as mentioned in the submission logistics of each Task.
- A readme file explaining how to compile/run your program without user intervention.
- Name of the folder should be your roll number and assignment number, For example – "**123456789_Assignment_2.zip**".
- Include any other reference materials you have used in this assignment.

END OF ASSIGNMENT