Homework #4 **Due: turned in by Thu 2/10/2020 end of day**

Carl Xi

(put your name above)

Total grade: _____ out of ___100___ points

General Submission Guidelines

In this and future assignments, there are typically two types of problems: short answers and hands-on exercises. For short answers, if you use others' work as part of your answer, please properly cite your source. If the source involves a URL, the URL should be provided. Please refer to the following example for the bibliography style:

> This phenomenon has been mentioned in several sources include a web page (Kehoe 1992) and a journal paper (Yeh 1996). A recent newspaper article (Greiner 2011) provides further details about this phenomenon.
>
> - Kehoe, Brendan P. "Zen and the Art of the Internet." January 1992, http://freenet.buffalo.edu/~popmusic/zen10.txt
> - Yeh, Michelle. "The 'Cult of Poetry' in Contemporary China." Journal of Asian Studies 55 (1996): 51-80.
> - Greiner, Lynn. "Wrists on fire? Tech gear for what ails you." Globe and Mail (Toronto) January 27, 2011. http://www.theglobeandmail.com/

Part I: Short Answers (40 points)

The answers will be graded along the lines of validity, informativeness, and presentation style. Be sure to include sources if you use any.

1. Understanding MapReduce (40 points)

Suppose you have a big text file that contains order_ID, employee_name, and sale_amount, separated by tabs. You goal is to calculate sum of all sales by employees.

```
0 Alice 3625
1 Bob 5174
2 Alice 893
3 Alice 2139
4 Diana 3581
5 Carlos 1039
6 Bob 4823
7 Alice 5834
8 Carlos 392
9 Diana 1804
...
```

Describe how Hadoop MapReduce carries out such a task, including what steps are involved, their input/out, when do data reading, writing, transferring occur, and when does parallel processing occur.

In a nutshell, Hadoop MapReduce is split into two parts, the 'Map' part and the 'Reduce' part.

First, Hadoop will read the data and pass them to the developer's Mapper code, one entry at a time. Each entry contains a key and a value, which will be employee_name and sale_amount respectively in our case. Next, each mapper will process a single input split from the HDFS (a single HDFS 'block' in most cases) and then write the intermediary data to the user's local disk, which still retains key and value pair from the previous step.

At this stage, the mapping part is basically done. Next, all mapper output gets shuffled and sorted, where all values associated with the same key from the intermediary data are transferred to the same reducer. Each reducer is given one of the unique keys and a list of all values associated with the said key in the order the data is sorted. Lastly, each reducer sums up the value for each key and outputs the result, which is also the key and value pair from the beginning, the only difference is the value in this case will be the sum of each key's sale_amount. The final result is written to HDFS.

Throughout our process, data reading occurs at step 1, data writing occurs at step 5, data transferring occurs at step 3, and parallel processing, where multiple mapper an reducers work in parallel, occurs at steps 2 and 4.

Visually, the following happens:

Data Input:

0	Alice	3625
1	Bob	5174
2	Alice	893
3	Alice	2139
4	Diana	3581
5	Carlos	1039
-	-	-

Gets mapped to:

Mapper Output:

Alice	3625
Bob	5174
Alice	893
Alice	2139
Diana	3581
Carlos	1039
-	-

Gets shuffle and sorted into:

Intermediary Data:

Alice	3625, 893, 2139, 5834, …
Bob	5174, 4823…
Carlos	1039, 392…
Diana	3581,1804…
-	-

Gets Reduced into:

Reducer Output:

Alice	12520
Bob	9997
Carlos	1431
Diana	5385
-	-

Which is basically the same as the final result:

Alice	12520
Bob	9997
Carlos	1431
Diana	5385
-	-

Part II. Hands on Linux/HDFS (60 points)

Please include a copy of commands and their step numbers in the PDF file you submit.

1. HDFS Commands (60 points; 15 each)

a. Create a folder latlon in your HDFS home directory.
b. Put $ADIR/data/latlon.tsv into the newly created folder.
   Note: If you do not already have this directory in your computing environment, please download the corresponding file from Canvas.
c. List the content of the latlon folder
d. Remove the folder and the files in it.

*Creating a folder 'latlon' in my HDFS home directory (I am in home directory by default)

a. `hdfs dfs –mkdir /latlon`

b. `hadoop dfs –put $ADIR/data/latlon.tsv /latlon`

c. `hadoop dfs –ls /latlon`

d. `hadoop fs –rm –r /latlon`