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

Adil Ashish Kumar

(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/

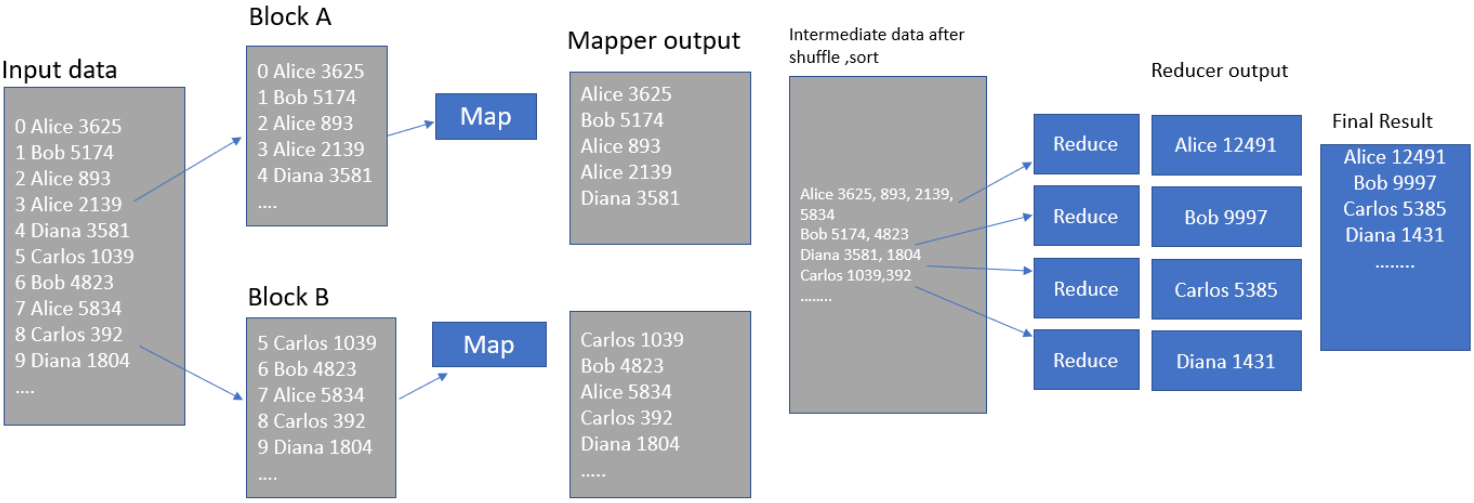
Part I: Short Answers (40 points)

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.

Input data is split into blocks as it is stored (data reading). Each block is a representation of data node. Map operation is then applied on each block so that each key value pair is mapped to a list of key value pairs. Intermediate mapper output is then written to disk. After the map operation the shuffle and sort is carried out so that the input to every reducer is sorted by key. The reduce operation is then done by taking key, list of values as input. The output of reducer is written to HDFS. The parallel processing occurs when input data is split into blocks.



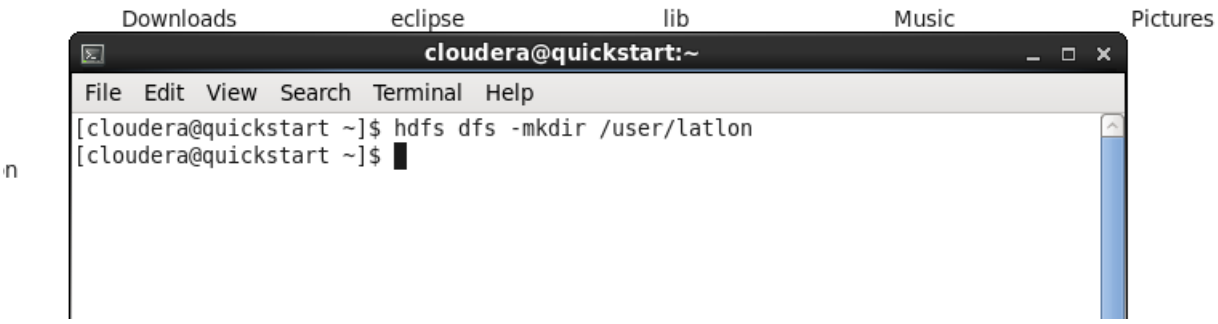
Part II. Hands on Linux/HDFS (60 points)

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

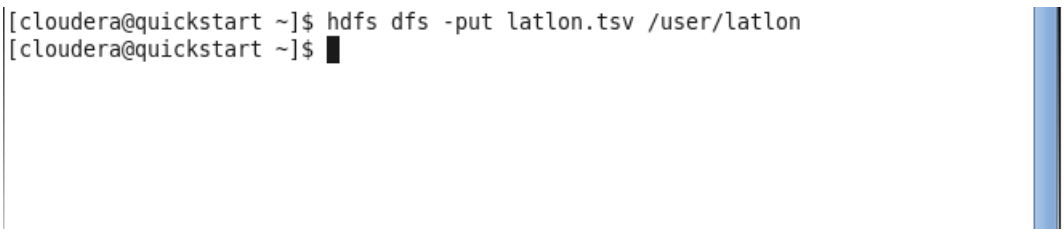
1. HDFS Commands (60 points; 15 each)

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.

a



b



c

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/latlon
Found 1 items
-rw-r--r--   1 cloudera supergroup    1127316 2020-02-09 17:26 /user/latlon/latl
on.tsv
[cloudera@quickstart ~]$
```

d

```
[cloudera@quickstart ~]$ hdfs dfs -rm /user/latlon/latlon.tsv
Deleted /user/latlon/latlon.tsv
```

```
[cloudera@quickstart ~]$ hdfs dfs -rmdir /user/latlon
[cloudera@quickstart ~]$ hdfs dfs -ls /user
Found 9 items
drwxr-xr-x   - cloudera cloudera            0 2017-10-23 10:28 /user/cloudera
drwxr-xr-x   - mapred   hadoop              0 2017-10-23 10:29 /user/history
drwxrwxrwx   - hive     supergroup          0 2017-10-23 10:31 /user/hive
drwxrwxrwx   - hue      supergroup          0 2017-10-23 10:30 /user/hue
drwxrwxrwx   - jenkins  supergroup          0 2017-10-23 10:30 /user/jenkins
drwxrwxrwx   - oozie    supergroup          0 2017-10-23 10:30 /user/oozie
drwxrwxrwx   - root     supergroup          0 2017-10-23 10:30 /user/root
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 10:31 /user/spark
drwxr-xr-x   - cloudera supergroup          0 2020-01-27 13:09 /user/training
[cloudera@quickstart ~]$
```