EMORY UNIVERSITY

Managing Big Data

Homework #5 **Due: turned in by Wed 11/01/2017 before class**

___Wenbo Liu______

(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/

The answers for homework assignments should be submitted in a PDF file. When the homework involves script files (e.g. pig, scala, or python scripts), the script files should be submitted in addition to the PDF for purpose of easy-debugging.

Part I: Short Answers (50 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. Define Big Data (10 points)

> In one short paragraph, define what makes an application a big data application. Be sure to cite sources if you include any.

A general way to interpret big data conceptually is that the size of data is becoming extraordinarily huge and new data is being generated at high speed, at the same time different types of data are being analyzed. An application needs to facilitate theses characteristics to become a ‘big data application’: it needs to achieve scalability, meaning it should be capable of processing huge amount of data; it needs to process data fast, sometimes processing data real-time as new data is being generated; it also needs to be able to process various types of data.

2. Advantages and Disadvantages of Hadoop relative to RDBMS (15 points)

> What are the advantages of Hadoop relative to traditional RDBMS? What are the disadvantages of Hadoop? List at least three advantages and at least one disadvantage.

Advantages of Hadoop:

1. Hadoop can easily deal with both structured and unstructured data, while RDBMS is usually used on structured data only.
2. Hadoop is more scalable: you can run a cluster containing two or two thousand nodes depending on your need, RDBMS doesn’t have that agility.

3. Hadoop is fault tolerant: each data block has a back up somewhere within the cluster, and when a node fails there will be no data loss; the system will automatically retrieve copies of data block stored in that failed node from elsewhere, at the same time the failed node will restart itself.

Disadvantages of Hadoop:

1. Hadoop is designed to deal with large amount of data retrieved from a distributed file system, and this type of operation is time consuming. RDBMS, on the other hand, works better when addressing more time-intensive tasks involving smaller datasets.
2. RDMBS has been in the market for quite a while and it is well adapted, while Hadoop has been developed/adapted recently and not a lot of applications are compatible with it.
3. Hadoop system itself is more complex to design compared to traditional RDMBS.

3. Data Locality (10 points)

Describe how the concept of "data locality" contributes to making Hadoop perform well.

Data exist in HDFS as data blocks, and when a Map-Reduce job is executed each mapper will start with retrieving relevant data blocks. To achieve better runtime, Hadoop will assign each mapper to individual nodes such that relevant data blocks are closer to mappers. The most ideal case is that the mapper is being executed on the same node that contains relevant information. The second-most ideal case is that the mapper is being executed on a node located on the same rack as another node that contains relevant data block. This process makes sure data needs to travel less physical distances (between data centers in different locations), as well as relieves data traffic congestion.

4. Understanding MapReduce (15 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.

As the job is created, the system locates data blocks that were split from the text file and assign mapping tasks to nodes that are as close to those data blocks as possible. For each mapper, it reads one line at a time with key=id, value=name&sale amount, and creates a intermediate key-value pair : name-sale amount, and this intermediate pair will be written to local disk. This mapping process is being done by m mappers in a parallel fashion, and each of the M intermediate pairs created by M mappers will be sorted again, aggregated by their intermediate keys(names in this example). Then, the M sorted intermediate pairs will be transferred to R reducers, where pairs with same key will be put to the same reducer. The reducer then returns the total sale of each person, writes the results on HDFS.

Part II. Hands on Linux/HDFS (50 points)

This part of the assignment uses the VM for the first few Hadoop labs. Please include a copy of commands and their step numbers in the PDF file you submit. Please also attach a separate pure-text file that contains all the commands. The latter is for occasional debugging purposes.

1. Linux Commands (30 points; 3 each)

a. Use the Linux command line interface to do the following:

b. Navigate to “$ADIR/exercise/data_mgmt”.

```
cd $ADIR/excercises/data_mgmt
```

c. Find out the size of the text file loyalty_data.txt in that folder.

```
du -h loyalty_data.txt
```

d. Find out the number of lines in the text file loyalty_data.txt.

```
wc -l loyalty_data.txt
```

e. View the content of the file in a controlled manner (that is, you don't want to see the entire file dumped on your screen).

```
head -20 loyalty_data.txt
```

f. Find the lines in the file that contain the world “Cliff”.

```
-grep -n ‘Cliff’ loyalty_data.txt
```

g. Navigate to folder $ADIR/data

```
cd $ADIR/data
```

h. Create a subfolder called samples

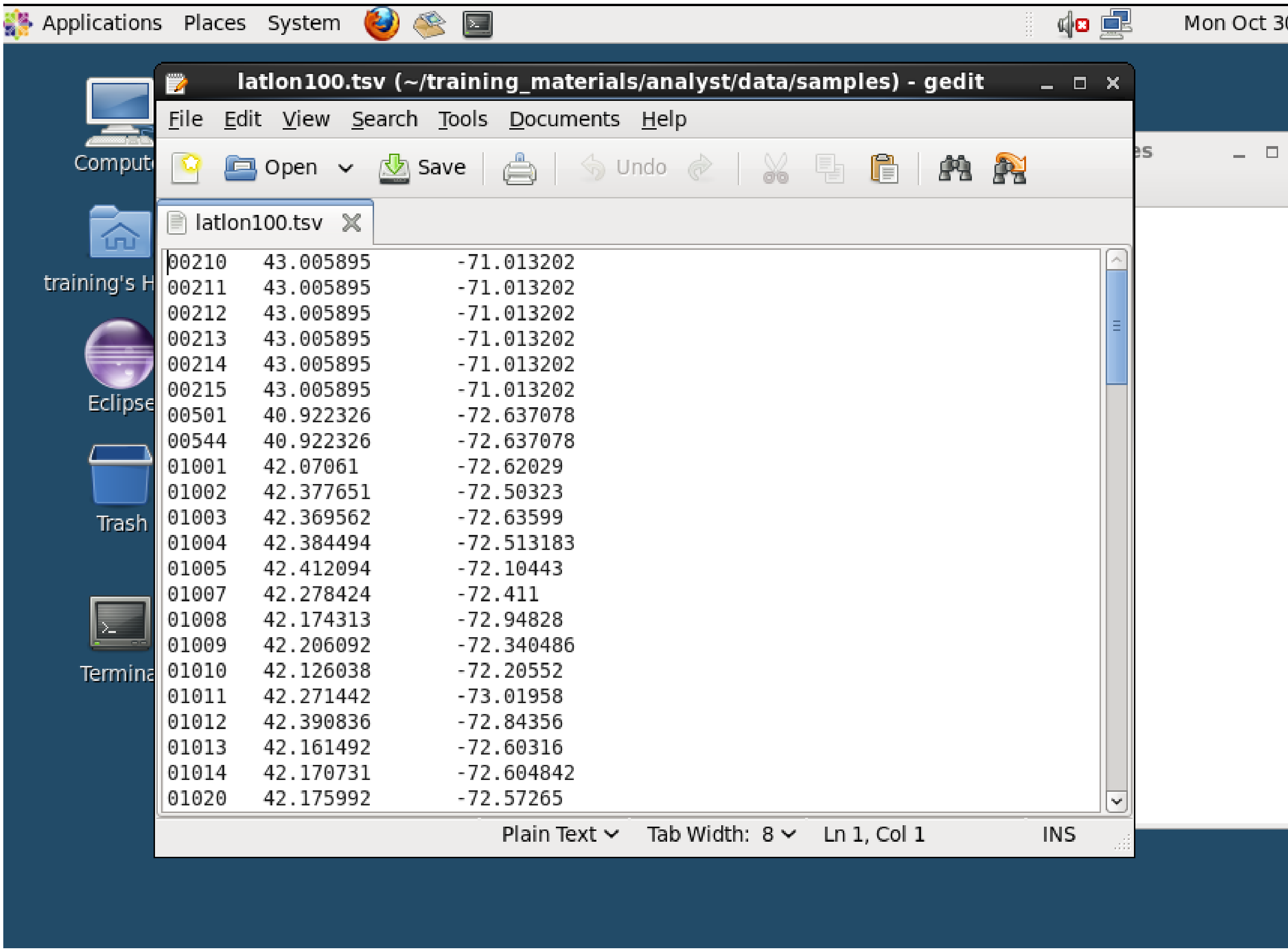
```
mkdir samples
```

i. Note that $ADIR/data/latlon.tsv contains tab delimited latitude and longitude records. Take the first 100 records of latlon.tsv and put them into $ADIR/data/sample/latlon100.tsv.

```
head -n 100 latlon.tsv > samples/latlon100.tsv
```

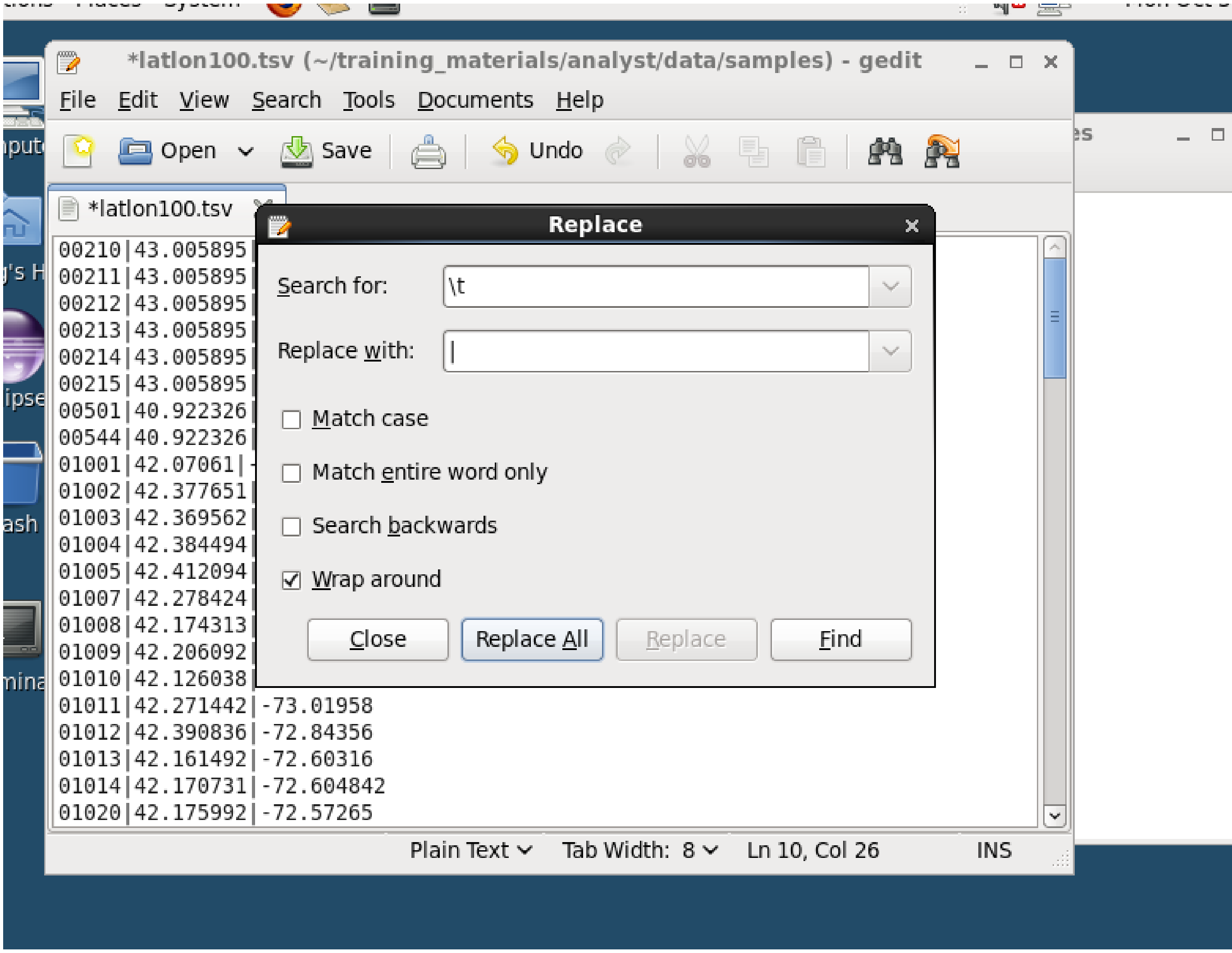
j. Use your favorite text editor to replace tab (\t) with symbol “|” in latlon100.tsv and save the file. (this is the only step where you are allowed to use a GUI tool. Describe how you achieve the replacement)

```
gedit latlon100.tsv
```



```
00210	43.005895	-71.013202
00211	43.005895	-71.013202
00212	43.005895	-71.013202
00213	43.005895	-71.013202
00214	43.005895	-71.013202
00215	43.005895	-71.013202
00501	40.922326	-72.637078
00544	40.922326	-72.637078
01001	42.07061	-72.62029
01002	42.377651	-72.50323
01003	42.369562	-72.63599
01004	42.384494	-72.513183
01005	42.412094	-72.10443
01007	42.278424	-72.411
01008	42.174313	-72.94828
01009	42.206092	-72.340486
01010	42.126038	-72.20552
01011	42.271442	-73.01958
01012	42.390836	-72.84356
01013	42.161492	-72.60316
01014	42.170731	-72.604842
01020	42.175992	-72.57265
```

Then press control-H and do the replacement



*latlon100.tsv (~/training_materials/analyst/data/samples) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
*latlon100.tsv
Replace
Close
Replace All
Replace
Find
Plain Text
Tab Width: 8
Ln 10, Col 26
INS

Search for: \t

Replace with:

☐ Match case

☐ Match entire word only

☐ Search backwards

☑ Wrap around

```
00210|43.005895|
00211|43.005895|
00212|43.005895|
00213|43.005895|
00214|43.005895|
00215|43.005895|
00501|40.922326|
00544|40.922326|
01001|42.07061|-
01002|42.377651|
01003|42.369562|
01004|42.384494|
01005|42.412094|
01007|42.278424|
01008|42.174313|
01009|42.206092|
01010|42.126038|
01011|42.271442|-73.01958
01012|42.390836|-72.84356
01013|42.161492|-72.60316
01014|42.170731|-72.604842
01020|42.175992|-72.57265
```

k. Rename the latlon100.tsv to latlon100.txt.

```
mv latlon100.tsv latlon100.txt
```

2. HDFS Commands (20 points; 5 each)

a. Create a folder latlon in your HDFS home directory.

```
hadoop fs -mkdir latlon
```

b. Put $ADIR/data/latlon.tsv into the newly created folder.

```
hadoop fs -put $ADIR/data/latlon.tsv /user/training/latlon
```

c. List the content of the latlon folder

```
hadoop fs -ls /user/training/latlon
```

d. Remove the folder and the files in it.

```
hadoop fs -rm -r latlon
```