|  |  |
| --- | --- |
|  | **Managing Big Data** |

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

\_Xi (Athena) Li\_

(put your name above)

Total grade: \_\_\_\_\_\_\_ out of \_\_\_100\_\_\_ points

# **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 big data application processes datasets whose size exceeds the typical reach of an RDBMS to capture, store, manage, and analyze that data. In other words, big data application can process dataset in many forms with extensible data storage and integration and volume of terabyte to petabyte. It can also analyze streaming data to enable decisions within fraction of a second, and measure the reliability and predictability of inherently imprecise data types.

### 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:
  + Lower cost
  + Less time
  + Greater flexibility
  + Near-linear scalability
* Disadvantages:
  + Slow query speed
  + No transaction support

### 3. Data Locality (10 points)

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

* The major drawback of Hadoop was cross-switch network traffic due to the huge volume of data. To overcome this drawback, Data Locality came into the picture. Data locality refers to the ability to move the computation close to where the actual data resides on the node, instead of moving large data to computation. This minimizes network congestion and increases the overall throughput of the system (Singh 2017).

### 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.

* The Mapper reads data in the form of key/value pairs (a file offset key (employee\_name), and a value with line content (sale\_amount))
* After the Map phase is over, all intermediate values for a given intermediate key are combined together into a list. we can break this down into map task, where the mapper task goes through the data and returns the sum of all sales
  + For instance, result produced from one mapper task would look like:
    - (Alice 12491) (Bob 9997) (Diana 5385) (Carlos 1431)
* This list is given to a Reducer
* There may be a single Reducer, or multiple Reducers
* All values associated with a particular intermediate key (employee name)are guaranteed to go to the same Reducer
* The intermediate keys, and their value lists, are passed to the Reducer in sorted key order
* This step is known as the ‘shuffle and sort’
* The Reducer outputs zero or more final key/value pairs
  + For instance: (Alice 12491) (Bob 9997) (Diana 5385) (Carlos 1431)
* These are written to HDFS

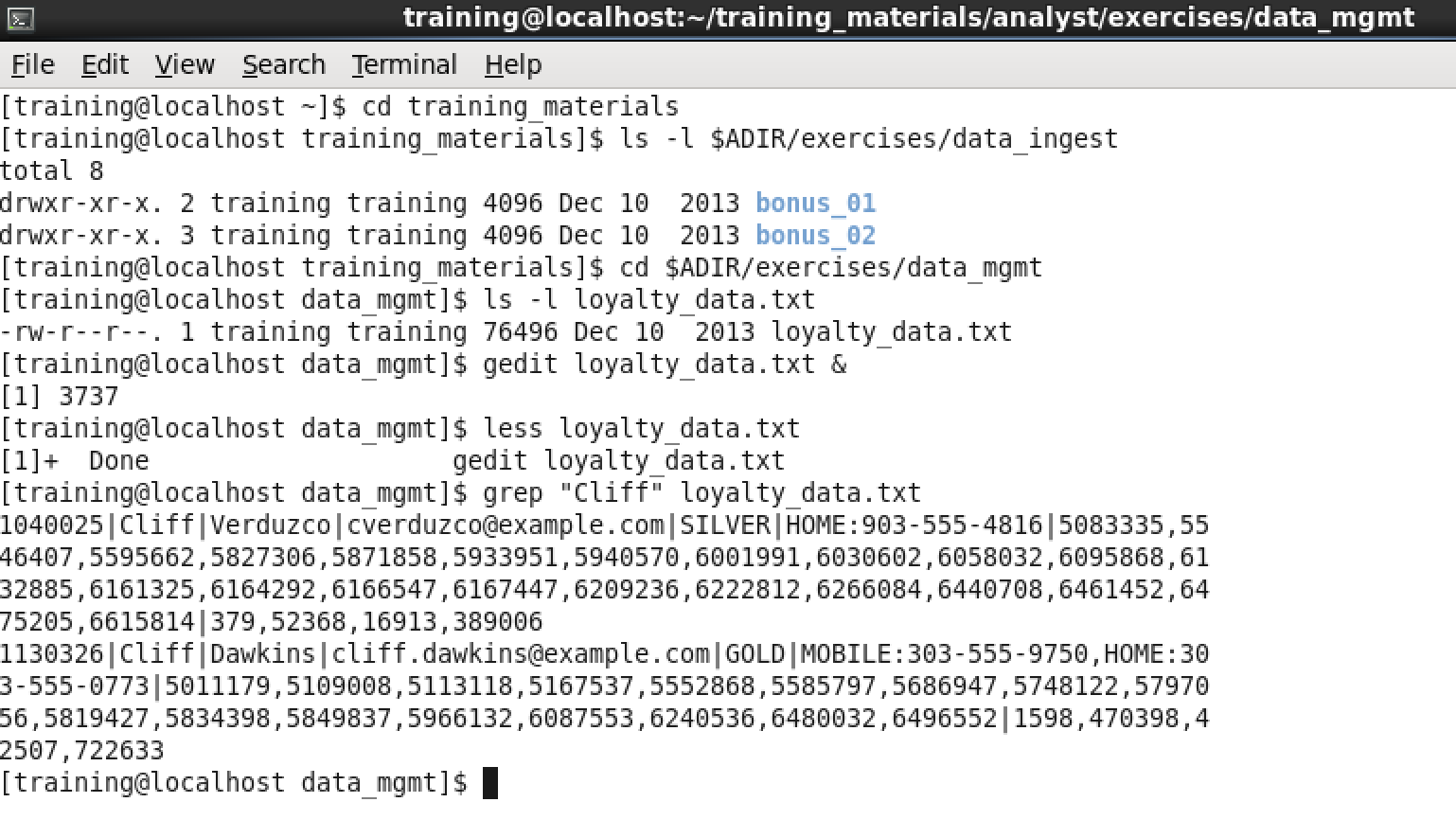
Work Cited

Singh, M. (2017, August 9). What does the term "data locality" mean in Hadoop? Retrieved October 30, 2017, from https://www.quora.com/What-does-the-term-data-locality-mean-in-Hadoop

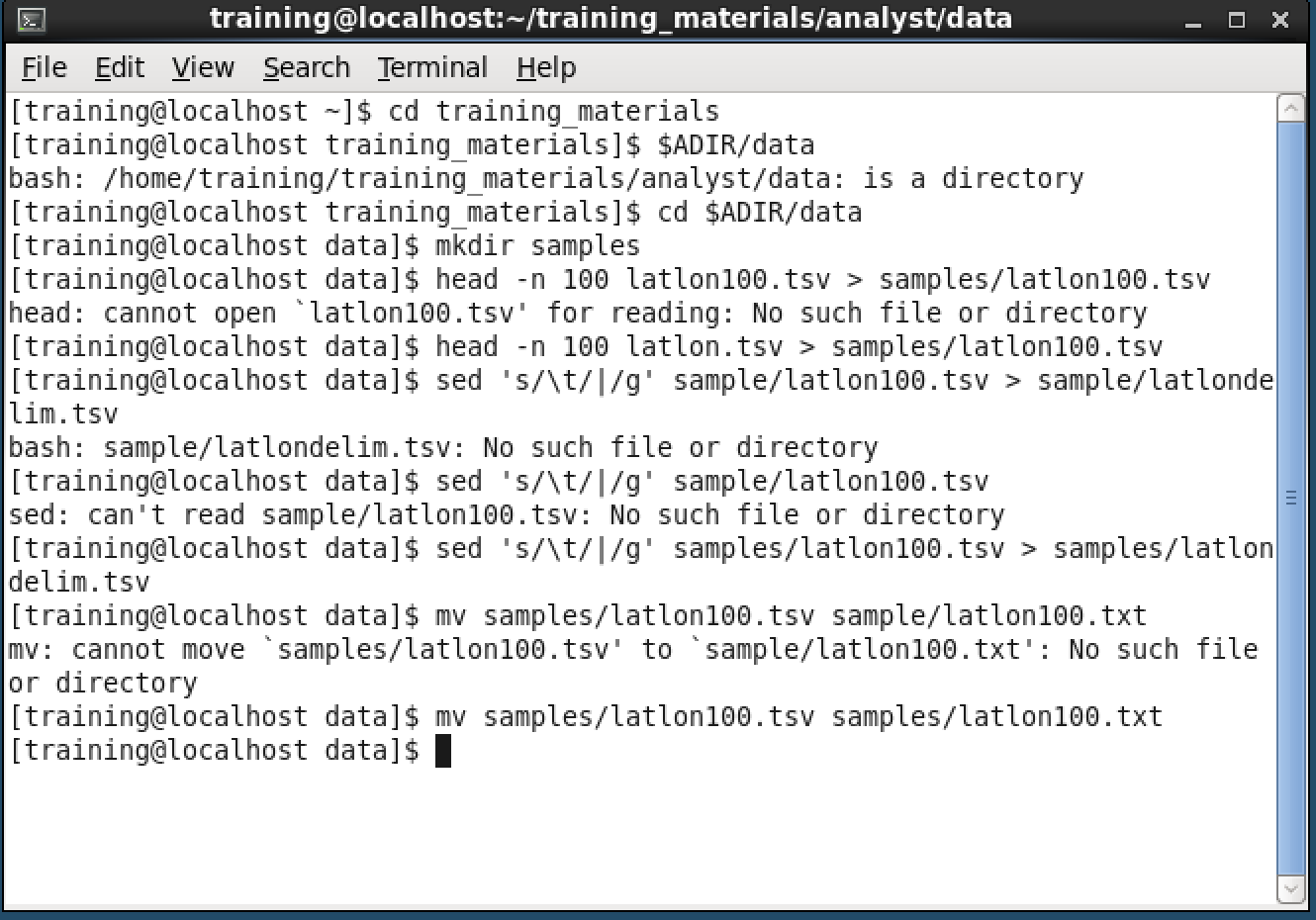
# **Part II. Hands on Linux/HDFS (50 points)**

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. Use the Linux command line interface to do the following:
2. Navigate to “$ADIR/exercise/data\_mgmt”.
   1. cd $ADIR/exercise/data\_mgmt
3. Find out the size of the text file loyalty\_data.txt in that folder.
   1. ls –l loyalty\_data.txt
4. Find out the number of lines in the text file loyalty\_data.txt.
   1. gedit loyalty\_data.txt &
5. 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).
   1. less loyalty\_data.txt
6. Find the lines in the file that contain the world “Cliff”.
   1. grep “Cliff” loyalty\_data.txt



1. Navigate to folder $ADIR/data
   1. cd $ADIR/data
2. Create a subfolder called samples
   1. mkdir samples
3. 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.
   1. head -n 100 latlon.tsv > sample/latlon100.tsv
4. 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)
   1. sed 's/\t/|/g' sample/latlon100.tsv > sample/latlon100delim.tsv
5. Rename the latlon100.tsv to latlon100.txt.
   1. mv sample/latlon100.tsv sample/latlon100.txt



### 2. HDFS Commands (20 points; 5 each)

1. Create a folder latlon in your HDFS home directory.
   1. hadoop fs -mkdir -p /user/training/latlon
2. Put $ADIR/data/latlon.tsv into the newly created folder.
   1. hadoop fs -put $ADIR/data/latlon.tsv /user/training/latlon
3. List the content of the latlon folder
   1. hadoop fs -ls /user/training/latlon
4. Remove the folder and the files in it.
   1. hadoop fs -rm /user/training/latlon/latlon.tsv

