PART 1 – Reading Assignment (Attached)

- Big Data Processing with Hadoop-MapReduce in Cloud Systems.
- Mapreduce is good enough?

PART 2 – Programming Assignment

All hadoop commands are invoked by the bin/hadoop script. Running the hadoop script without any arguments prints the description for all commands.

Usage: hadoop [--config confdir] [--loglevel loglevel] [COMMAND] [GENERIC_OPTIONS] [COMMAND_OPTIONS]

Execute each hadoop command once, and place the screenshots into a word file. If a command cannot be executed for any reason (such as, a distributed environment is needed), you may write the definition of the command, and skip execution.

http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html

PART 3 – Programming Assignment

Copy the attached 'access.log' file into HDFS under /logs directory.

Using the access.log file stored in HDFS, implement MapReduce to find the number of times each IP accessed the website.

PART 4 – Programming Assignment

Download and Copy all the files to a folder in HDFS (http://msis.neu.edu/nyse/)

Write a Java Program to implement PutMerge as discussed in the class to merge the NYSE files in HDFS to find the max price of stock-price-high values for each stock using MapReduce on the single merged-file

PART 5 – Programming Assignment

Write one MapReduce program using each of the classes that extend FileInputFormat<k,v>

(CombineFileInputFormat, FixedLengthInputFormat, KeyValueTextInputFormat, NLineInputFormat, SequenceFileInputFormat, TextInputFormat)

http://hadoop.apache.org/docs/current/api/org/apache/hadoop/mapreduce/lib/input/FileInputFormat.html

You could use any input file of your choice. The size of the input files is not important. The MR programs could simply do counting, or any other analysis you choose.

access.log
MapReduce is Good Enough.pdf
Big Data Processing with Hadoop-MapReduce in Cloud Systems.pdf