# Lab Exercise 11: Hadoop, HDFS, MapReduce and Pig

## Step 1

### Part 3

Hadoop variables:  
HADOOP\_DATANODE\_OPTS=-Dcom.sun.management.jmxremote  
HADOOP\_LOG\_DIR=/usr/local/hadoop-0.20.1/logs  
HADOOP\_JOBTRACKER\_OPTS=-Dcom.sun.management.jmxremote  
HADOOP\_HOME=/usr/local/hadoop-0.20.1  
HADOOP\_HEAPSIZE=64  
HADOOP\_VERSION=0.20.1  
HADOOP\_SECONDARYNAMENODE\_OPTS=-Dcom.sun.management.jmxremote  
HADOOP\_BALANCER\_OPTS=-Dcom.sun.management.jmxremote  
HADOOP\_NAMENODE\_OPTS=-Dcom.sun.management.jmxremote

### Part 5

Directory of /user/gpadmin [on Hadoop]

Found 3 items  
drwxr-xr-x - gpadmin supergroup 0 2012-03-05 10:18 /user/gpadmin/input  
-rw-r--r-- 1 gpadmin supergroup 354 2011-10-11 09:14 /user/gpadmin/kmeans.sql  
drwxr-xr-x - gpadmin supergroup 0 2012-04-11 20:07 /user/gpadmin/output

## Step 5

### Part 2

It took 0.055 seconds [real] to count the words in greeneggsandham.txt.

### Part 5

When running MRwordcount.sh, the Hadoop logs are shown. Map 0% Reduce 0% went to Map 100% Reduce 100%. Shows that the output file was placed in output/d4710.

### Part 6

It took 30.475s [real] to run the script.

### Part 7

Yes it was slower than the UNIX command. This is perhaps the initialization of Hadoop system.

### Question 1a

The results of wf.sh and MRwordcount.sh were consistent, however one run locally was sorted, where the one run on Hadoop was not sorted.

### Question 1b

The word “I” was repeated the most with a total of 84 times in both cases.

### Question 1c

Dr. Seuss is successful in using just 50 words.

## Step 6

### Part 2

When running Hadoop web JobTracker, the name of the job in the console is the same as the name of the job in the website. But much more statistical information about how many bytes read and graphs about mapping and reducing.

### Part 4

The TaskTracker was empty because there were no tasks running when I pulled it up. But after attempting to run it again, I notice that the job number is used at a suffix to the word attempt.

## Step 7

### Part 4a

There were 3 jobs runs on the back end with job ids: job\_201501301142\_0011, job\_201501301142\_0012, and job\_201501301142\_0013.

### Part 4b

There were 192 records read by the input.

### Part 4c

The term frequency of Sam was 19, of Anywhere was 8, and of Not was 82.

### Part 4d

The words with the maximum term frequency was “I” with 84 occurrences. The terms with the minimum term frequency was “if” and “mot” with 1 occurrence each.