Hello everyone. Please follow the following instructions for the Assignment – 2.

**Software required:**

1. Putty
2. WinSCP
3. Oracle Virtual Box
4. Cloudera – by default it contains Eclipse and Hadoop packages installed which can be used to program MapReduce programs. (You can find many tutorials for installing Cloudera)

**Basic procedure to follow when executing a MapReduce program in a Hadoop cluster:**

1. The inputs should be transferred to HDFS from the local system
2. The JAR file can reside in the FTP client side(i.e in WinSCP)
3. The output of MapReduce programs will be written on HDFS which can be transferred back to the local system

To understand how MapReduce works, you can see following links along with example

 [https://hadoop.apache.org/docs/r1.2.1/mapred\_tutorial.html](%20https%3A/hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html) - for Hadoop version 1.0

All basic HDFS commands can be found here: <https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-common/FileSystemShell.html>

**Part – 1: Basic text processing using GREP command**

Put following lines into your input file

(a, a1->a2) ^ (c = c2) -> (f, f1->f0) [2, 50%]

(a, a1->a3) ^ (b, ->b1) -> (f, f1->f0) [3, 75%]

(a, a1->a3) ^ (c = c2) -> (f, f1->f0) [1, 80%]

(a, ->a3) ^ (b, b2->b1) -> (f, f1->f0) [3, 50%]

1. See <https://wiki.apache.org/hadoop/Grep> to learn about Grep command
2. Example command:

**hadoop org.apache.hadoop.examples.Grep <input> <output> “.\*a1.\*”**

which matches and returns all lines having “a1”

1. Copy the output to your local system

For Part – 2 and Part – 3, use text from <http://webpages.uncc.edu/aatzache/ITCS6190/Exercises/03_MammalsBook_Text_34848.txt.utf8.txt> for the input

WordCount programs can be found in <https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html>

**Part – 2: Running WordCount v1.0 in single node cluster(cloudera)**

1. When you open Eclipse, you can see a ***training project.*** Locate their external library(Hadoop jar files) paths.
2. Create new JAVA project and import all those jar files.
3. Now copy WordCount v1.0 into new project. Import required jar files if some went missing previously.
4. Convert the project into a jar file.
5. Since it is a single node cluster, you can execute HDFS commands in the terminal itself.
6. Execute and take a screenshot of the output

**Part – 3: Running WordCount v2.0 in dsba cluster**

1. One method is writing the MapReduce code in cloudera itself, zip the project and transfer it to FTP client using the following command

 ***scp filename.zip username@dsba-hadoop.uncc.edu:/users/username***

1. Otherwise you can get Hadoop jar files for Eclipse from the internet and import them in your project
2. Like Part – 2, convert your project into a jar file
3. Use Putty software to run your MapReduce program
4. Get the output and store it in local system

Put all your results in a single folder, zip it and submit.