**Cloud Computing for Data Analysis**

**Group Activity 01**

**Task 1 - Setup Hadoop Virtual Environment , Simple Commands , Create and Search through a file**

In this activity, you are going to setup a virtual environment for running Hadoop MapReduce code , run simple commands , create a file , search through the file .

**Setup Hadoop Virtual Environment**

View step by step instructions for installing Cloudera in the following links:

**Windows:** https://webpages.uncc.edu/aatzache/ITCS6190/Exercises/GroupActivity01\_Clourera\_Installation\_Windows.docx

**MAC:** https://webpages.uncc.edu/aatzache/ITCS6190/Exercises/GroupActivity01\_Clourera\_Installation\_Mac.docx

For running any Hadoop code, the input should be moved from local system into HDFS. Similarly, the output of your code will be present in HDFS. Files in HDFS can be accessed using **hadoop** commands. In Cloudera, **/user/admin/** is a home directory of all your files and folders.

**Simple Commands**

To list all files and folders in the home directory of HDFS, use:

**hadoop fs -ls /user/admin** (or) **hadoop fs -ls /user/cloudera**

To copy a file from cloudera home folder into HDFS home directoy, use:

**hadoop fs -put /home/cloudera/filename.txt /user/admin/** (or)

**hadoop fs -put /home/cloudera/filename.txt /user/cloudera**

To copy a file from HDFS home directoy into cloudera home folder, use:

**hadoop fs -get /user/admin/filename.txt /home/cloudera/** (or)

**hadoop fs -get /user/cloudera/filename.txt /home/cloudera/**

To open a file in HDFS, use:

**hadoop fs -cat /user/admin/filename.txt** (or) **hadoop fs -cat /user/cloudera/filename.txt**

**fs –** denotes File System to represent that the command is accessing files from HDFS

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

**Create and Search through a file**

1. Open Cloudera
2. Open the Terminal in Cloudera
3. Create a file named **ListOfInputActionRules.txt** using the command

**vi ListOfInputActionRules.txt**

1. Copy and paste following lines in **ListOfInputActionRules.txt**

(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. Make sure the file is present in **/home/cloudera/** folder
2. Again, in the terminal use the command

**hadoop fs -put /home/cloudera/ListOfInputActionRules.txt /user/admin/** (or)

**hadoop fs -put /home/cloudera/ListOfInputActionRules.txt /user/cloudera/**

to copy your file into HDFS

1. See <https://wiki.apache.org/hadoop/Grep> to learn about Grep command
2. Use the command **hadoop org.apache.hadoop.examples.Grep /user/admin/filename.txt /user/admin/Out4Grep01 “.\*a1.\*”** (or)

**/user/cloudera/filename.txt /user/cloudera/Out4Grep01 “.\*a1.\*”**

to return all lines of text which contain the word ‘**a1**’

**(NOTE:** You do not need to create separate output folder for your outputs. Hadoop itself creates the output folder matching the given folder name in the given folder**)**

1. List all files from the output path using

**hadoop fs -ls /user/admin/Out4Grep01** (or)

**hadoop fs -ls /user/cloudera/Out4Grep01**

command

1. Open the file that is named something like **part-r-00000** using

**hadoop fs -cat /user/admin/Out4Grep01/part-r-00000** (or)

**hadoop fs -cat /user/cloudera/Out4Grep01/part-r-00000**

1. You can take a screenshot of this and upload in Canvas

Hadoop outputs will be without any extensions like **part-r-00000**. You can convert this into **.txt** file while copying it into Cloudera home directory using the following command:

**hadoop fs -get /user/admin/ Out4Grep01/part-r-00000 /home/cloudera/outputFile.txt** (or)

**hadoop fs -get /user/cloudera/ Out4Grep01/part-r-00000 /home/cloudera/outputFile.txt**