**INFS 774 Big Data Analytics**

**ASSIGNMENT 2**

Due: 6/18/2020 + one week grace period

Please refer to “Access Hadoop VM” posted under “Content” regarding how to access the VM

**TASKS:** In this assignment you need to finish three tasks. In task 1 and task 2, you need to do some data processing with Hadoop - I provide detailed explanations for these two tasks on Page 3. **Whenever you encounter any problem, before you post your question on the discussion forum, please first read “3. Common errors” on page 3 of the document. Common errors.** Quite possibly, you will find your answers there. Task 3 will include a couple of essay questions.

**TASK 1.** You will be working with the Cloudera Hadoop environment. Please complete the Cloudera Homework Labs – Lecture #1 (available on D2l under content -> assignments -> assignment 2). Please submit Screenshots of following steps from tutorial file:

i. Step 2-Uploading Files: After finishing 2.1 – 2.9, please run hadoop fs –ls, submit the results

ii. Step 3-Viewing and manipulating files: Result from step 3

**TASK 2.** You will be working with the Cloudera Hadoop environment. Please complete the Cloudera Homework Labs – Lecture #2 (available on D2l under content -> assignments -> assignment 2). Please submit Screenshots of following steps from tutorial file:

i. Compiling and Submitting a MapReduce Job: step 5

ii. Compiling and Submitting a MapReduce Job: step 7

iii. Compiling and Submitting a MapReduce Job: step 8

iv. Compiling and Submitting a MapReduce Job: step 9 (replicating steps 7 and 8 for pwords)

**TASK3**.

Q1: Using the study materials provided as a starting point, and any others that you may refer to, write in your own words a brief synopsis (200-300 words – roughly half a page, font 12 font, single spacing) of what you understand by the term “big data analytics”. Some ideas for discussion -­- How is the term similar or different than other related terms/disciplines? In building your arguments, consider whether the industry term “big data” is a hype and if so, to what extent. What is novel, unique about this term with regards to the practical and research implications that it presents?

T2. Discuss 2 motivational scenarios (200-300 words – roughly half a page, font 12 font, single spacing) where use of “big data analytics” can play a significant role in drastically changing the current status quo/problem situation. You should use cases other than those mentioned in the book (Harness the Power of Big Data) and video resources. Feel free to look for other resources on the Web and academic databases.

**Task 1 and 2 explanation:**

1. Lab 1:

Lab 1 is mainly about uploading a dataset to the Hadoop machine. Let’s say you have a dataset you want to process using Hadoop, and this dataset is on your windows machine. You first need to transfer the dataset to the Linux machine using ssh or ftp. In our assignment, the dataset “shakespeare” is already in the Linux system, so you don’t need to do file transfer from Windows to Linux. The dataset is now in the Linux file system, but not in the Hadoop HDFS. You still need to upload the file to Hadoop HDFS. How to access HDFS from your Linux system? You usually type “hadoop fs“ or sometimes “hdfs fs”. You use the command “hadoop fs –put sourceOnLinux destinationOnHDFS” to upload a dataset to HDFS. It’s important to remember that the Linux file system and HDFS are two different file systems. They have different sets of commands. You need to be clear about in which file system your target dataset is located and then use the corresponding commands. In order to a MapReduce job, you need to upload the dataset to HDFS.

In lab 1, the target dataset is shakespeare.tar.gz. It is placed under the folder ~/training\_materials/developer/data in the Linux file system. You first process the file in the Linux file system. You need to:

1. Go to the directory ~/training\_materials/developer/data using the command “cd”.
2. Since this file is a zip file, you need to unzip it by typing “tar zxvf shakespeare.tar.gz”. Then you should have a folder called “shakespeare” that includes 5 files.

Next, you upload this “shakespeare” folder to the Hadoop HDFS by inputting “hadoop fs -put shakespeare /user/training/shakespeare”. The Shakespeare folder and its contents will be put into a “remote” HDFS directory named /user/training/shakespeare. Now you can type “hadoop fs -ls shakespeare” to see what’s inside the folder. It turns out that the folder Shakespeare includes five files (glossary, poems, histories, comedies, tragedies) and you want to remove the file glossary by typing “hadoop fs -rm shakespeare/glossary”. Here, you remove the file in HDFS. Actually, you can also remove the file from the shakespeare folder in the Linux system first and then upload the folder to HDFS.

2. Lab 2:

In Lab 1, you upload the shakespeare folder that includes the four files (poems, histories, comedies, tragedies) to HDFS. In lab 2, you want to run a MapReduce job to count the number of occurrences of each word in the folder. Lab 2 is about preparing and running some existing MapReduce code.

The MapReduce code is in the folder ~/workspace/wordcount/src/stubs (stubs is actually a java package, but can be roughly understood as a folder). The folder contains several Java files. You need to first do “ javac -classpath `hadoop classpath` stubs/\*.java” to compile these java files into .class files that contain java bytecode. Then, you type “jar cvf wc.jar stubs/\*.class” to create a jar. A jar file in Java is kind of like a combination of the “zip” and“.exe” file in Windows. It is a zip file that includes the .class files and it is executable. The above is just the common procedure for creating an executable jar file in java. When you run a MapReduce job in java, you always need to create a jar file first.

If you are familiar with java, you should know that to run an executable jar, you need to use the command “java –jar”. In Hadoop, we use “hadoop jar”. You type “hadoop jar wc.jar stubs.WordCount shakespeare wordcounts” to run the mapreduce job. This hadoop jar command says that the JAR file to use is wc.jar, and the main method is in stubs.WordCount (when you do java programming, you always need to have a main method), the input directory is called “shakespere” in the HDFS user root directory (/user/training in our case) and the output directory for storing the results is called wordcounts (the full path should be /user/training/wordcounts). Your java code will then count how many times each word appears in the folder “shakespeare”. The results including the key-value pairs can be found in a file “part-r-0000” in the output folder “wordcount”. You can type “hadoop fs -cat wordcounts/part-r-00000 | less” to view the results

In lab 2, you also need to do “hadoop jar wc.jar stubs.WordCount shakespeare/poems pwords” to count the occurrences of each word in the file “sharepeare/poems”. The output folder is called “pwords”.

3. Common errors:

Please note that in Hadoop, we don’t overwrite files. Hence, if you want to re-run your MapReduce job, you need to first remove the output folder.

Below, I discuss some of the common errors you will probably encounter when you work on this assignment.

1. If you see the error: File does not exist: /user/training/shakespeare/shakespeare when you run the mapreduce code, it basically means that the folder "shakespeare" in your HDFS has somehow been messed up. Obviously, you do not have a file called "shakespeare" under the folder "shakespeare". "hadoop fs -ls shakespeare" should list only 4 text files including comedies, tragedies, histories, and poems. If you see "shakespeare" in the list, it means something went wrong when you upload the dataset - maybe you uploaded the dataset multiple times. In HDFS, you usually do not overwrite an existing dataset. You need to remove the dataset and upload it again. Hence, to deal with the error, you need to remove this shakepeare folder under "/user/training"and then re-upload the data. You can do “hadoop fs –rm –r –f shakespeare” to remove the Shakespeare folder. After you remove the folder,  you can type "hadoop fs -ls" to verify. Now you shouldn't see the folder "shakepeare". Then you can go back to step 2.3 in lecture 1 lab and re-upload the data.
2. Please note that in lab2 step 2. You need type "javac -classpath `hadoop classpath` stubs/\*java". You need to use backquotes (rather than single quotes) to enclose the text "hadoop classpath".