CSP554—Big Data Technologies

Assignment #2 (Modules 02a & 02b, 20 points)

Due by the start of the next class period

Assignments can be handed in at the beginning of class or uploaded via the Blackboard portal
Note: There may be short quiz questions about readings, assignments or articles (except extra credit) in the class period when they are due.

- Read from (TW)
  - Chapter 2
  - Chapter 4 (optional)
  - Chapter 6 through page 179, no need to become expert in the details, just understand principles and refer back for later reference
  - Chapter 7

- MapReduce is a somewhat challenging topic to approach the first time. So, if you are not satisfied after reading Chapter 4 and 6 above have a look at the following on the "Blackboard" portal in the "Free Books and Chapters" section
  - Apache Hadoop 2.8 Map Reduce

1. Now make sure you have a terminal window capability. All the rest of the assignment, and subsequent assignments will assume you have this capability available.

   Only if you have a PC: Note, some of you might want to use Putty, but please use this instead
   - Go to: https://git-for-windows.github.io/
   - Download and install the software
   - Execute the "Git Bash" shell

   Only if you have a MAC
   - Open Finder.
   - Select Applications. Then chose Utilities.
   - Double click on Terminal.
   - The terminal window will now be open

2.

3. Now we will set up a Hadoop cluster.

   I know this will involve following many detailed steps. But, as a result, you will be prepared to configure a Hadoop cluster in the cloud wherever you need one. And we will be setting up clusters for other assignments. Note, when using the Amazon (AWS) cloud their Hadoop product is called "Elastic Map Reduce" or EMR.
   The general flow of the rest of this assignment will be:
   a. Set up Hadoop cluster
   b. Start up a terminal or bash window
   c. From within the terminal or bash window execute the ssh command to connect to the EMR master node. Don't worry I show you how to work with this command.
   d. Open up another terminal or bash window but do not connect to the EMR cluster
   e. Move some files from your local PC or MAC to an S3 bucket object storage system
   f. Move some files from your local PC or MAC to the Hadoop cluster master node Linux file system using scp (secure copy). Don't worry I show you how to work with this command.
   g. Use the Hadoop Distributed File System (HDFS)
   h. Stop the cluster and release any associated resources (so you don't continue to pay for them)

   To do so you will follow instructions in the file AWS EMR Instructions included in the assignment:

- Set up the Hadoop cluster by following Step 1 through Step 5 in the AWS EMR Instructions document
- Open a terminal or bash window and connect into the Hadoop cluster by following Step 6 in the AWS EMR Instructions document. We will call this window the "Hadoop window".
- Now open up a second terminal or bash window. Don't connect to the Hadoop cluster. We will call this window the "local window".
- We will use the "local window" to execute the secure copy command (scp) to move a file from your local PC to the Hadoop master node.
  a. Download a small text file called "myname.txt" from the assignment file and change the name of the file to some version of your name with no spaces such as "josephrosen.txt". Sometimes, rather than downloading the content of this file, its content just displays in the browser. If so, use your browser's ability to save the content of a page, to save this as yourname.txt. Note, throughout the rest of this document we will still call this file by the name "myname.txt".
  b. Execute the following command (all on one line) to move this file from your local machine to the home directory of your Hadoop master node account. Choose the path and name of your key (.pem) file, the name of your text file, and the DNS name of your Hadoop cluster master node. Append ":/home/hadoop" as the target directory.

scp -i <path-to-file>/emr-key-pair.pem <path-to-file>/josephrosen.txt hadoop@*ec2-###-##-##-###.compute-1.amazonaws.com:/home/hadoop*

You might be asked a security question. If so answer 'yes'.

  c. You can check if this worked by switching to the "Hadoop window" and executing an "ls" command

d. Download another small text file called “myid.txt” from the assignment files and change the name of the file to your student id with no spaces such as “A12345678.txt”. Note throughout the rest of this document we will still call this file by the name “myid.txt”.
e. Create an S3 bucket of whatever name you choose. For purposes of this discussion I will assume it is called mybucket (but of course your bucket must have a name unique across all of AWS)
f. Using the techniques we learned during assignment #1 upload A12345678.txt to the S3 bucket into an object called A12345678.txt (of course use your id). Note throughout the rest of this document we will still call this object by the name “myid.txt”.

The next part of the assignment requires you to perform some simple operations on a shared HDFS file system and S3. At last!!

- The documentation for the Hadoop 2.8 File System Shell is available on the Blackboard in the Free Books and Chapter section. Note, when this document mentions using the “s3a:” file prefix to access AWS bucket objects, this should be changed to the prefix “s3:” which is what EMR expects.

9. (2 points) Execute the following hdfs command to list the files or directories that are listed (also indicating which is a file and which a directory):

   hadoop fs –ls /

Take a screen snapshot of names of the files or directories that are listed and include it in your assignment submission.

10. (2 points) Execute a command (you needed to figure out which one) to list the files and directories under the hdfs directory listed below:

/user

Write down the command you executed and also take a screen snapshot of names of the files or directories that are listed and include it in your assignment submission.

11. (2 points) Execute a command to create the following HDFS directory:

/user/csp554

Record the command you executed and include it in your assignment submission.

12. (2 points) Execute a command to create the following HDFS directory:

/user/csp554-2

Record the command you executed and include it in your assignment submission.

13. (2 points) Execute a command that copies a given local file to the given hdfs directory :

Source local file: /home/hadoop/myname.txt (where the actual name is your name as described above)

Destination HDFS directory: /user/csp554

Record the command you executed and include it in your assignment submission.

14. (2 points) Copy a file from one hdfs directory to another hdfs directory and write down the command

Source hdfs file: /user/csp554/myname.txt (where the actual name is your name as described above)

Destination HDFS directory: /user/csp554-2

Record the command you executed and include it in your assignment submission.

15. (2 points) Copy the object myid.txt you uploaded to an S3 bucket into the Hadoop master node Linux file system. The actual object includes your student id as above.

Note, Amazon EMR and Hadoop provide a variety of file systems that you can use with EMR. You specify which file system to use with a file system prefix. For example, s3://myawsbucket/path references an Amazon S3 bucket using EMRFS (EMR file system). See:
https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-file-systems.html

The way you do this would be as follows to copy an object from an S3 bucket to the Linux file system of the Hadoop master node.

```
aws s3 cp s3://mybucket/myid.txt   /home/hadoop/myid.txt
```

The above is an AWS CLI (command line interpreter) command. For more information about how to use the CLI to manipulate S3 buckets see: https://docs.aws.amazon.com/cli/latest/reference/s3/index.html

After you executed the above command perform an “ls /home/hadoop” and take a screen snapshot of names of the files or directories that are listed and include it in your assignment submission.

16. (2 points) Copy the same object myid.txt you created in an S3 bucket into HDFS into the directory /users/csp554

```
hadoop fs -cp s3://mybucket/myid.txt    hdfs:///user/csp554-2
```
Note, the three slashes after the “hdfs:”

After you executed the above command, execute another command (you needed to figure out which one) to list the files and directories under the hdfs directory listed below:

/user/csp554-2

Write down the command you executed and also take a screen snapshot of names of the files or directories that are listed and include it in your assignment submission.

17. (2 points) Execute a command to show the contents of the myid.txt file in the hdfs directory /user/csp554-2

    Clue: look up about how to use the “cat” command in the file system shell document.

Write down the command you executed and also take a screen snapshot of the listed content of the file and include it in your assignment submission.

18. (2 points) Execute a command to remove the myid.txt file in the hdfs directory /user/csp554-2

    Clue: look up about how to use the “rm” command in the file system shell document.

Write down the command you executed, then list the content of the /user/csp554-2 HDFS directory and take a screen snapshot of the listed content of the directory and include it in your assignment submission.

19. This might be very important to your wallet ☺.

    Follow Step 7 in the AWS EMR Instructions document to terminate your cluster and delete any buckets you have created. If you forget you may end up paying (a lot) more than you need to.