CS595—Big Data Technologies

Assignment #2 (Modules 02a & 02b, 10 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.

1. 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

2. 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 “Backboard” portal in the “Free Books and Chapters” section
   - Apache Hadoop 2.8 Map Reduce

3. Only if you have a PC:
   - Go to: https://git-for-windows.github.io/
   - Download and install the software
   - Execute the “Git Bash” shell

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

5. You need to create, if needed, and edit a “config” file in the subdirectory “.ssh”.

As a convenience I have created and uploaded the correct “config” file as another attachment for this assignment. All you need to do is download it and place it in the “.ssh” subdirectory of the top level directory you default to when you open a terminal or bash window. Note if it is not there just create it.

But below are instructions if you want to create the “config” file on your own, whatever you prefer.

Use your favorite editor and (create and) edit your “~/.ssh/config” file. For example:

```
vi ~/.ssh/config
```

Enter the following configuration, Note: Spacing and capitalization is important.

```
Host azureSandbox
  Port 22
  User cs595s2018
  HostName 23.96.218.84
  LocalForward 8080 127.0.0.1:8080
  LocalForward 8888 127.0.0.1:8888
  LocalForward 9995 127.0.0.1:9995
  LocalForward 9996 127.0.0.1:9996
  LocalForward 8886 127.0.0.1:8886
  LocalForward 10500 127.0.0.1:10500
  LocalForward 4200 127.0.0.1:4200
  LocalForward 2222 127.0.0.1:2222
```

Save and close the file.

6. Now SSH into the hadoop cluster by using the command below. This will connect automatically using the IP address specified in the config file:

```
ssh azureSandbox
```

You’ll then be asked for a password. Enter:

Unix79127912

You should always to this whenever you work with the Hadoop cluster for all follow on assignments. Just leave this terminal session alone once you have logged on.

7. Now open up another terminal or “bash” window while leaving the previous one open.

Into this new window enter the following:

ssh –p 2222 maria_dev@localhost

You’ll then be asked for a password. Enter:

maria_dev

8. The next part of the assignment requires you to perform some simple operations on a shared HDFS file system.

You will all be logging on to the same account on a temporary Hadoop instance I have set up on Azure for this one purpose. I will provide instructions in the near future about how to set up your own Hadoop sandbox. But, at any rate, please don’t play around with this instance outside of what you are asked to do in the following steps.

9. The Hadoop file system shell command reference is available online at

https://hadoop.apache.org/docs/r2.8.0/hadoop-project-dist/hadoop-common/FileSystemShell.html

or on the “Blackboard” portal in the books section as “Apache Hadoop 2.8 Command Shell Commands”

Note, some of the following questions ask you to take a few screen snapshots. Please submit them in a word document (with your name) indicating the number of the assignment step with which these items are associated.

10. (1 point) Execute the following hdfs command and take a screen snapshot of the names of the files or directories that are listed (also indicating which is a file and which a directory):

hadoop fs –ls /

11. (1 point) Execute a command (you needed to figure out which one) to list the files and subdirectories 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.

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

/user/cs595/<yournamenospaces>

Note: I created the hdfs directory "/user/cs595" for you already.

Write down the command you executed.

13. (2 points) Execute a command that copies a given local file (that I already created for you) to the given hdfs directory :

Local file: /home/maria_dev/cs595doc.txt

HDFS directory: /user/cs595/<yournamenospaces>

Write down the command you executed.

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

    Source hdfs file: /user/cs595/cs595doc2.txt

    Note, I created this file for you already

    Destination HDFS directory: /user/cs595/<yournamenospaces>

    Write down the command you executed.

15. (2 points) This is to get you familiar with the approach to copy files from your MAC or PC to a local account on the Linux machine running the Hadoop Sandbox.
    - Create a plain text file on your local machine (PC or MAC). Name it firstname_lastname.txt. The file should hold just one line and that line should hold your first and last name.
    - Open a new terminal or bash window
    - Copy the local file firstname_lastname.txt to the maria_dev account as follows:
        - scp –P 2222 ./firstname_lastname.txt maria_dev@localhost:/home/maria_dev
        - When prompted for a password, enter

        - maria_dev
    - Note that the “ssh” command used a small p (-p) to specify the port, the scp command uses a capital P (-P) to do so.
- Now, from your hadoop maria_dev account, execute a command that copies the firstname_lastname.txt file to the given hdfs directory :

Local file: /home/maria_dev/firstname_lastname.txt

HDFS directory: /user/cs595/<yournamenospaces>

Write down the command you executed and take a screenshot of all files in the directory “/user/cs595/<yournamenospaces>”