**Week 5 Assignment 1 - Using HDFS**

*Please click on the link above to submit this week's assignment.*

*NOTE: Submit either the Paper or the Hands-on assignment.*

NOTE:  Refer to this site for HDFS commands : *https*://*hadoop*.apache.*org*/docs/r2.4.1/*hadoop*-project-dist/*hadoop*-common/*FileSystemShell*.*html*

*You will either write this paper or do the hands-on assignment*

*=================================   Paper ====================================================*

By using information from the lecture notes and online sources, answer the following questions. Where required, please provide the complete command line with proper spacing and syntax.

1. Briefly explain the following terms:  HDFS, MapReduce, YARN, Hadoop Distribution
2. Read about AWS EMR and provide a page or two on EMR benefits and components
3. What is the role of an HDFS NameNode? What would happen if the NameNode is lost permanently?
4. What is the role of a secondary NameNode?
5. What is the role of an HDFS DataNode? What would happen if a DataNode is lost permanently?
6. HDFS Replication
   1. What is the default HDFS replication factor?
   2. Can you make the replication factor to be 1? If so, when and why would you do that?
   3. Are there any risks in making replication factor = 1?
   4. Is there a downside to increasing the replication factor to be more than 3?
   5. What is hdfs rebalancing?
7. Provide the complete HDFS command syntax to do the following:  ( Please refer to *https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-common/FileSystemShell.html* for the hdfs commands)
   1. List the contents of the root directory in HDFS
   2. Create a directory in HDFS ( ex. /tmp/test)
   3. Upload a file named "abc.txt" to the /tmp/test folder in HDFS.
   4. Download the file named "abc.txt" from the /tmp/test folder in HDFS
   5. View the contents of a file in HDFS ( ex. /tmp/test/abc.txt)
   6. Remove the file name "abc.txt" from the /tmp/test folder in HDFS
8. Using the following lines of text describe how the various MapReduce phases are involved in counting the number of occurrences of the word "sentence"  . You must show how the word counting progresses through all the MapReduce phases (Split, Shuffle, Map, Reduce etc..) with  diagrams and description.

This is test sentence number one.

This is test sentence number 2.

This is test sentence number three.

This is sentence no 4.

sentence 5.

**Requirements for the written assignments:**

* Assignment file must have a .doc or .docx extension; screen shots should be in .jpg, .gif, or .pdf
* Paper length must be at least  7 pages long
* Points for this assignment = 95

*==================================================*

*For this hands-on assignment, choose any one of the following options  .*

***Hands-on Option 1:  Level of difficulty = Medium***

1. Download the Hadoop VirtualBox image from one  of the following : (these are pre-installed with Hadoop tools)
   * Bitnami.com  (https://bitnami.com/stack/hadoop) OR
   * Cloudera.com

Note: Bitnami, Cloudera distributions need more memory and CPUs on the host PC/Mac

NOTE:  You MUST give the VM enough memory to run the Mapreduce  Word Count example. Otherwise, the program will be stuck for a very long time.  Recommendation is 8 GB.

1. Import the image into your VirtualBox environment
2. Start the VM and login
3. Execute the HDFS commands to do the following:
   * 1. Create an HDFS folder
     2. Copy a file into the HDFS folder
     3. List contents of the HDFS folder
     4. Download a file from HDFS
     5. View the contents of a file in HDFS
     6. Remove a file from HDFS
4. Create a text file containing some sentences. Upload this file into HDFS and then Execute the sample MapReduce WordCount program to provide a count of all words from the text file
5. Display the contents of the file produced by the MapReduce operation. Take a screen shot of the output.
6. Shutdown the VM when finished

**Hands-on option 2:  Using Azure HDInsight or AWS EMR -  Level of difficulty = easy**

1. Login to Azure and create a small HDInsight Hadoop cluster, OR, login to AWS and create a 2 node (use M1.medium) EMR cluster
   1. **NOTE: AWS EMR is not part of the free tier. You will be charged for the number of hours the cluster is running.**
   2. **NOTE:  If you want to use Azure, sign up for Azure's 30-day free tier. We will be using Azure for Weeks 6, 7 and 8 also.**
2. Connect to the cluster using ssh
3. Execute the commands to do the following :
   * 1. Create an HDFS folder
     2. Copy a file into the HDFS folder
     3. List contents of the HDFS folder
     4. Download a file from HDFS
     5. View the contents of a file in HDFS
     6. Remove a file from HDFS
4. Create a text file containing some sentences. Upload this file into HDFS and then Execute the sample MapReduce WordCount program to provide a count of all words from the text file
5. Display the contents of the file produced by the MapReduce operation. Take a screen shot of the output.
6. Shutdown the cluster when finished

**Hands-on option 3:  Install Hadoop -  Level of difficulty = high**

* NOTE: This option is to be done in a Linux VM.

1. Download and install Apache Hadoop binaries in a Linux VM in VirtualBox or any other hypervisor
2. Follow the instructions to configure and format HDFS
3. Start the services
4. Execute the HDFS commands :
   1. Create an HDFS folder
   2. Copy a file into HDFS
   3. List contents of an HDFS folder
   4. Download a file from HDFS
   5. View the contents of a file in HDFS
   6. Remove a file from HDFS
5. Create a text file containing some sentences. Upload this file into HDFS and then Execute the sample MapReduce WordCount program to provide a count of all words from the text file
6. Display the contents of the file produced by the MapReduce operation. Take a screen shot of the output.
7. Shutdown the VM when finished

**Requirements for the assignments:**

* Submit write up and screen shots for all activities .
* Assignment file must have a .doc or .docx extension; screen shots should be in .jpg, .gif, or .pdf

Points for this assignment = 95