IST769 HDFS, MapReduce and YARN

Your Name: Tajudeen Abdulazeez   
Your SUID: 69687-7373-0   
Your Email: toabdula@syr.edu   
Date Due:   
Homework #: 5

1. Upload all the documents in **datasets/text** into a folder called **text** in HDFS. What HDFS command must you run to verify the files are there after they are uploaded? Your answer should consist of the command you typed to complete the task.

* Hdfs dfs -mkdir text
* Hdfs dfs -put datasets/text/\*.txt text/
* Hdfs dfs -ls text

1. In this part you will upload the **clickstream** dataset to HDFS. Specifically, create a **clickstream** folder in HDFS, then create a **logs** and **iplookup** folder inside the clickstream folder. Upload all of the **\*.log** files from the **datasets/clickstream** local folder into **clickstream/logs** in HDFS. Upload the **ip\_lookup.csv** file from the same folder into **clickstream/iplookup** on HDFS. Verify the files are there. Your answer should consist of the commands you typed to complete the task.

* Hdfs dfs -mkdir clickstream/logs
* Hdfs dfs -mkdir clickstream/iplookup
* Hdfs dfs -ls clickstream
* Hdfs dfs -put datasets/clickstream/\*.csv clickstream/iplookup/
* Hdfs dfs -put datasets/clickstream/\*.log clickstream/logs/
* Hdfs dfs -ls clickstream/logs
* Hdfs dfs -ls clickstream/iplookup

1. Use the MapReduce examples:  
   export MREX=/usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar   
   to perform a wordcount on the 2016 State of the Union address, saving the output to the HDFS folder **sotu2016**. Write down the commands to complete the task. How many times does the word **are** appear in the 2016 State of the Union address? Describe a process which could be done to make the wordcount more useful?

* export MREX=/usr/lib/Hadoop-mapreduce/Hadoop-mapreduce-examples.jar
* yarn jar $MREX
* hdfs dfs -ls text
* yarn jar $MREX wordcount /user/cloudera/text/2016-state-of-the-union.txt /user/cloudera/sotu2016
* hdfs dfs -ls sotu2016
* hdfs dfs -cat sotu2016/part-r-00000
* hdfs dfs -get sotu2016/part-r-00000 sotu2016.csv
* Get the wordcount file into a local drive from the hdfs for further processing.
* Plots the histogram of the wordcounts [sotu2016.csv] to get the word distribution/words with the most occurrence.

1. Type the following command to import the **fudgemart\_v3** database into the local **mysql** instance on the Hadoop client:  
   mysql -u root -p < ~/datasets/fudgemart/mysql.sql   
   The password is **cloudera**. Write down the commands you used to complete these tasks:  
   Use the **sqoop** utility to verify there are tables in the database by listing them from the **fudgemart\_v3** database. Next write a sqoop command to import Fudgemart products in the ‘Clothing’ department into a HDFS folder **/user/cloudera/fudgemart-clothing**

* Mysql -u root -p < ~/datasets/fudgemart/mysql.sql
* Sqoop help
* Sqoop list-databases –connect jdbc:mysql://cloudera –username=root –password=cloudera
* Sqoop list-tables –connect jdbc:mysql://cloudera/fudgemart\_v3 –username=root –password=cloudera
* Sqoop import –connect jdbc:mysql://cloudera/fudgemart\_v3 –username=root –password=cloudera –target-dir /user/cloudera/fudgemart-clothing –query “ select \* from fudgemart\_products where product\_department=’clothing’ AND \$CONDITIONS” –as-textfile –split-by product\_name
* Hdfs dfs -ls fudgemart-clothing
* Hdfs dfs -cat fudgemart-clothing/part-m-00001

1. Let’s import HDFS data into MySQL. Record each command you type as your solution:
   1. Load **datasets/tweets/tweets.psv** into the HDFS folder **tweets**
   2. Login to MySQL: mysql -u root -p The password is **cloudera**. Create a database **twitter**
   3. Create a table called **tweets** inside the database **twitter** the table should have columns for id, timestamp, date time, username, and tweet\_text.
   4. Export the data from HDFS into the MySQL table.  
      **TIPS:** If your SQOOP job fails it is likely due to the table constraints such as selecting a data type too small for the imported data. Try to insert a row in the table using a sample from the HDFS data. This will help you to ensure your chosen data types will work.

* Hdfs dfs -mkdir tweets
* Hdfs dfs -ls
* Hdfs dfs -put datasets/tweets/tweets.psv tweets/
* Hdfs dfs -ls tweets
* Mysql -u root -p
* Create database twitter;
* Show databases;
* Use twitter;
* Create table tweets (tweet\_id bigint, tweet\_timestamp double, tweet\_date datetime, username text, tweet\_text text)
* Sqoop export –connect jdbc:mysql://cloudera/twitter –username=root –password=cloudera –table tweets –export-dir /user/cloudera/tweets –input-field-terminated “|”