Your Name: STEPHEN OMONDI  
Your SUID: 946934043  
Your Email: SOOMONDI@SYR.EDU  
Date Due: 2019-11-02  
Homework #: LABWORK 5

# Exercises

Complete each of the following exercises. If you are unsure how to accomplish the task, please consult the coursework videos where there are explanations and demos. For each of the following write down the commands required to complete the task.

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 –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

hdfs dfs –mkdir clickstream /logs

hdfs dfs –mkdir clickstream /iplookup

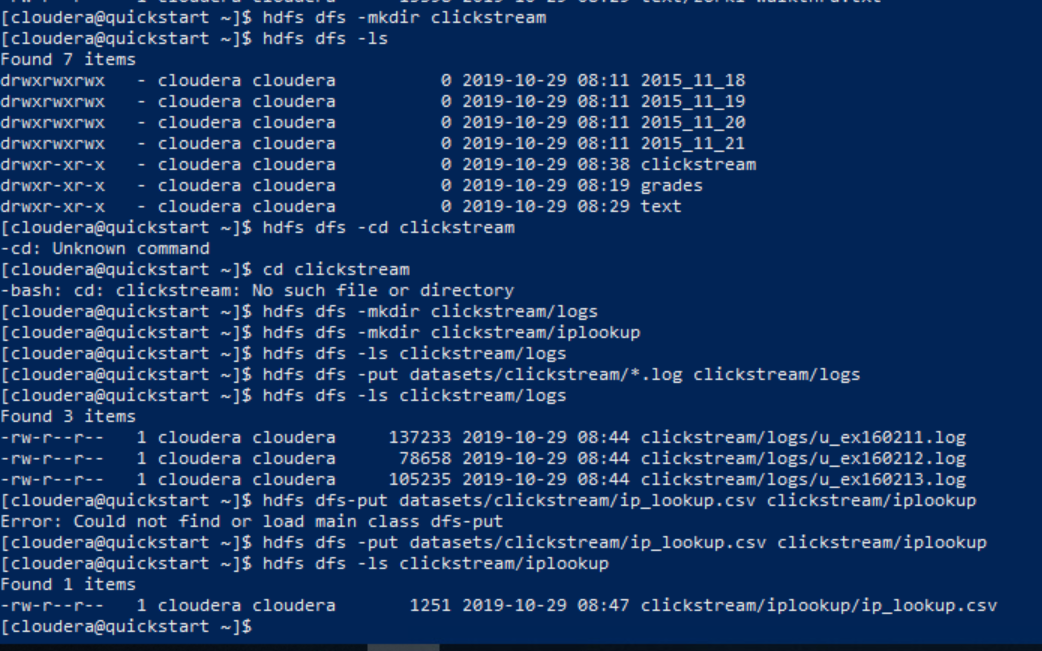
hdfs dfs –ls clickstream

hdfs dfs –put datasets/clickstream/\*.log clickstream/logs

hdfs dfs –ls clickstream/logs

hdfs dfs –put datasets/clickstream/ip\_lookup.csv clickstream/iplookup

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 wordcount text/2016-state-of-the-union.txt sotu2016

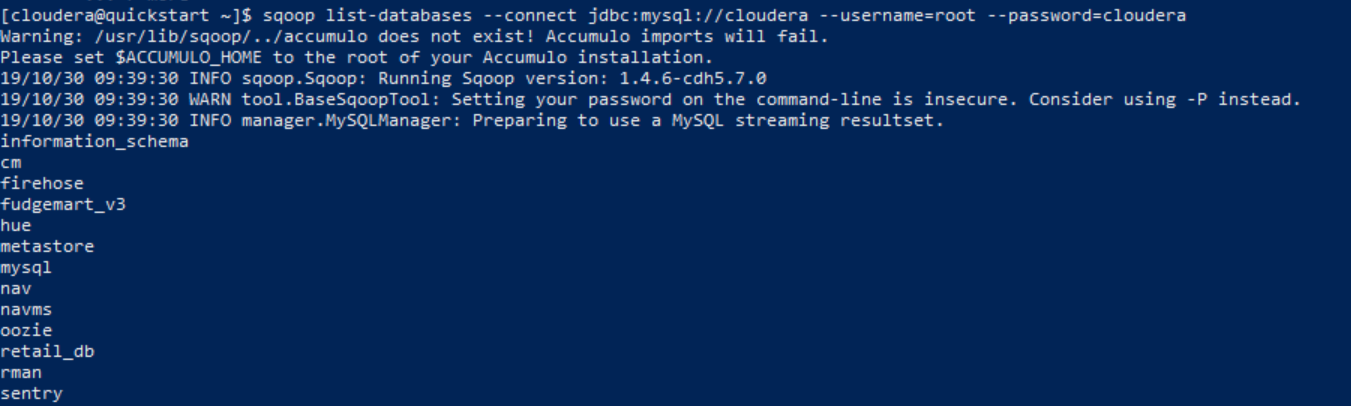
hdfs dfs –cat sotu2016/\*

The word “**are”** appears **27** times in the 2016 State of the Union address. A more meaningful usage of wordcount would be to distinguish between root words and punctuations, such that the following three words are seen as one, not three: “are”, “are.”, “are,”

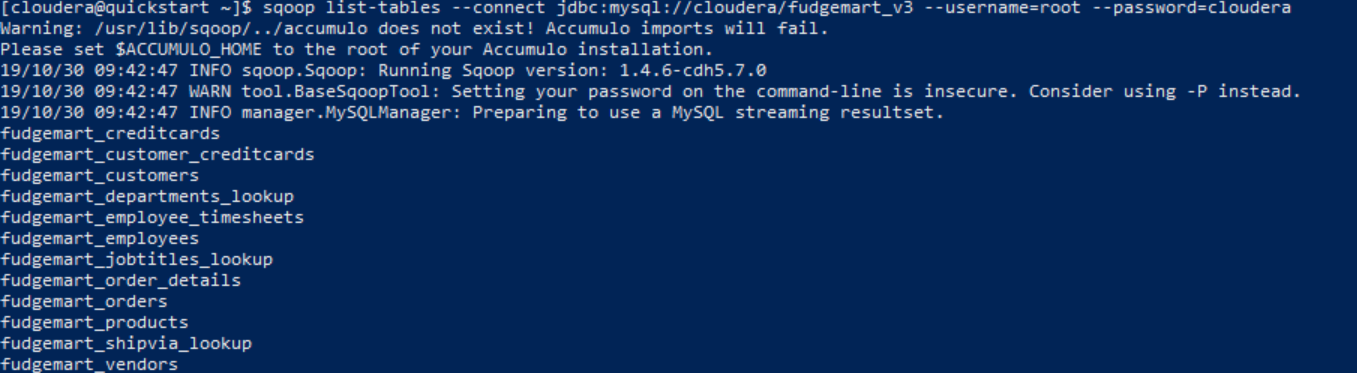
1. Type the following command to import the **fudgemart\_v3** database into the local **mysql** instance on the Hadoop client:  
   mysql -u root -p < ~/datasets/fudgemart/mysql.sql   
   The password is **cloudera**. Write down the commands you used to complete these tasks:  
   Use the **sqoop** utility 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**

**Using SQOOP client to connect to MySQL, show databases, and query tables in the database**

sqoop list-databases –-connect jdbc:mysql://cloudera –-username=root –password=cloudera

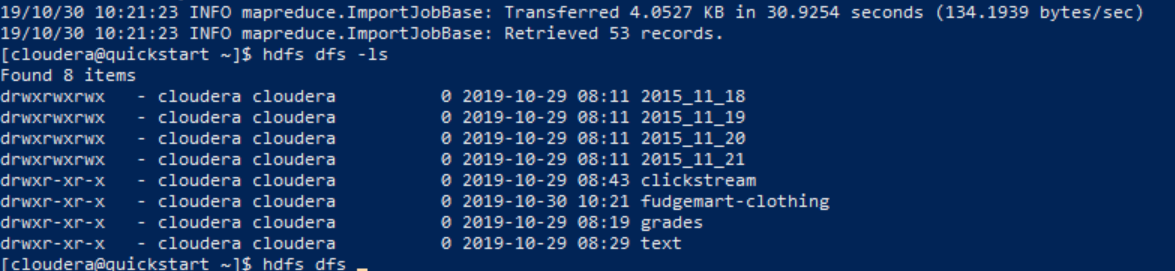


sqoop list-tables –-connect jdbc:mysql://cloudera/fudgemart\_v3 –-username=root –password=cloudera

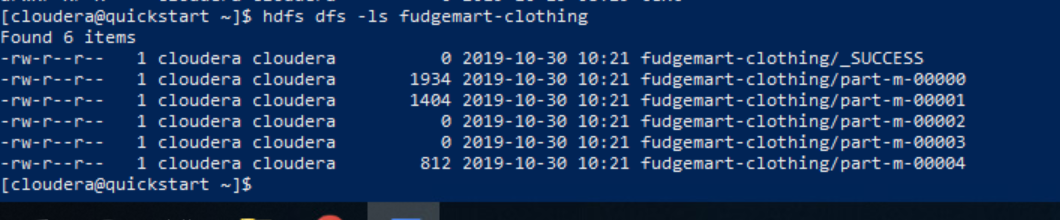


sqoop import –-connect jdbc:mysql://cloudera/fudgemart\_v3 –-username=root –password=cloudera \

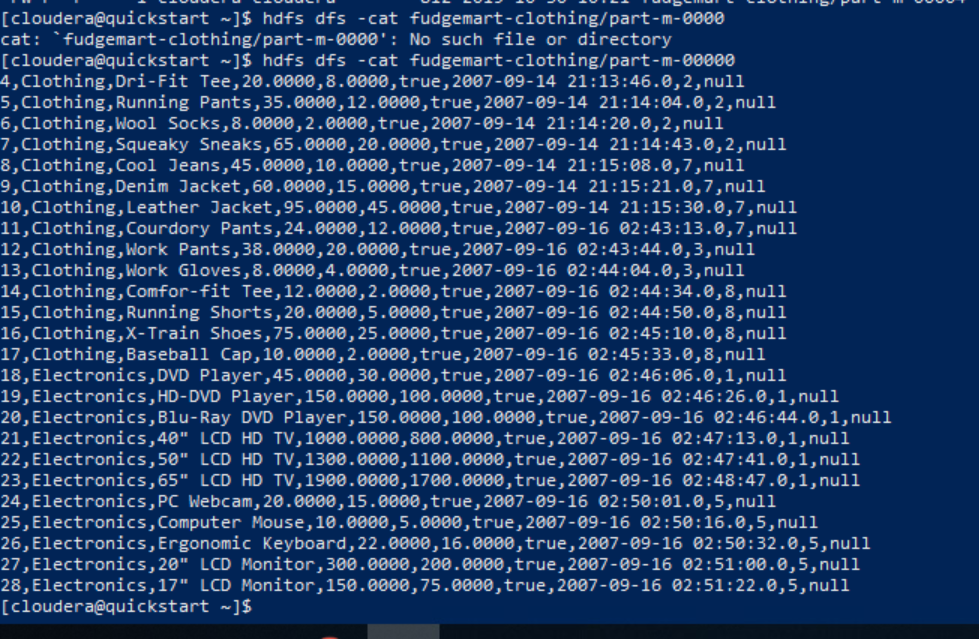
--query “SELECT \* FROM fudgemart\_products WHERE \$CONDITIONS order by product\_category” --target-dir --as-textfile --split-by product\_category



hdfs dfs –ls fudgemart-clothing



hdfs dfs –cat fudgemart-clothing/part-m-00000



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.

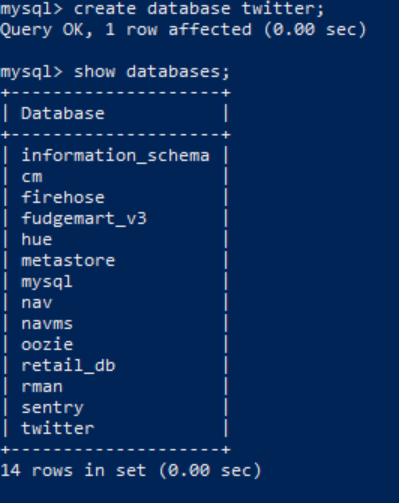
hdfs dfs –put datasets/tweets/tweets.psv tweets

mysql –uroot – pcloud;

create database twitter;

use twitter;

create table tweets ()



use twitter;

create table tweets(

id int NOT NULL AUTO\_INCREMENT,

time\_stamp timestamp NOT NULL,

date\_time datetime NOT NULL,

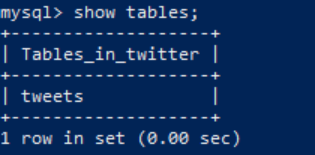
username varchar (255) NOT NULL,

tweet\_text varchar(1000) NOT NULL,

PRIMARY KEY(id)

);

show tables;



sqoop export –-connect jdbc:mysql://cloudera/twitter –username=root password=cloudera –-tweets –-export-dir /user/cloudera/tweets --input-fields-terminated-by “\t”