# IST769 Homework Submission Template

## Basic Information

Your Name: Wanyue XiaoYour

SUID: 720633297

Your Email: xwanyue@syr.edu

Date Due: March 4, 2020

Homework #: HW05

## Instructions

For each answer, please include your answer as text, and any screenshot(s) which demonstrate your answer was executed. Most importantly, make sure to include evidence your answer is correct. This will most likely be a screenshot. If you had issues, problems, or had to make assumptions include them in your answer.

## Your Answers:

1.

Upload all the documents in **datasets/text** into a folder called **text** in HDFS:

1/ create the text folder:

**hdfs dfs -mkdir text**

2/ check that if this folder has been created successfully

**hdfs dfs -ls**

**A close up of a keyboard

Description automatically generated**

3/ put the document from datasets to the text folder that we have created in the first step

**hdfs dfs -put datasets/text/\*.txt text/**

4/ list the documents that are under text folder

**hdfs dfs -ls text**

**A close up of a sign

Description automatically generated**

2.

1/ create a **clickstream** folder in HDFS, then create a **logs** and **iplookup** folder inside the clickstream folder

**hdfs dfs -mkdir clickstream**

**hdfs dfs -mkdir clickstream/logs**

**hdfs dfs -mkdir clickstream/iplookup**

A close up of text on a white background

Description automatically generated

2/ Upload all of the files into **clickstream/logs AND clickstream/iplookup** in HDFS.

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

**hdfs dfs -put datasets/clickstream/\*.csv clickstream/iplookup**

3/ Verify the files are there

**hdfs dfs -ls clickstream/logs**

**hdfs dfs -ls clickstream/iplookup**

**A screen shot of a computer

Description automatically generated**

3.

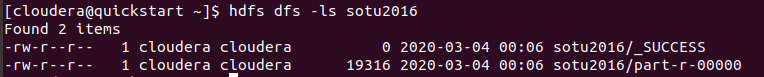
1/ Initiate the java procedure, and use the wordcount function on 2016-state-of-the-union.txt document

**MREX=/usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar**

**yarn jar $MREX wordcount text/2016-state-of-the-union.txt sotu2016/**

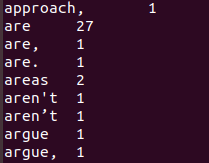
2/ List the documents that are inside the sotu2016 folder

**hdfs dfs -ls sotu2016**



3/ Display the content; And One can find **that ‘are’ appears 27 times**

**hdfs dfs -cat sotu2016/\***



4/ Describe a process which could be done to make the wordcount more useful – **USE grep to catch specific content**

**hdfs dfs -cat sotu2016/\* | grep are**

**A picture containing device, object

Description automatically generated**

4.

1/ List tables available in the database

**sqoop list-tables --connect jdbc:mysql://cloudera/fudgemart\_v3 --usernames=root --password=clouderaA screenshot of a cell phone

Description automatically generated**

2/ Import products and store the results into the **fudgemart-clothing** folder:

**sqoop import --connect jdbc:mysql://cloudera/fudgemart\_v3 --usernames=root --password=cloudera --query"select \* from fudgemart\_products where product\_department = ‘Clothing’ and \$CONDITIONS” --target-dir /user/cloudera/fudgemart-clothing --as-textfile -splite-by product\_id**

**A close up of a sign

Description automatically generated**

3/ extract information from the result documents

**Hadoop fs -cat /user/cloudera/fudgemart-clothing/part-m-00000**

**Hadoop fs -cat /user/cloudera/fudgemart-clothing/part-m-00001**

**Hadoop fs -cat /user/cloudera/fudgemart-clothing/part-m-00002**

**Hadoop fs -cat /user/cloudera/fudgemart-clothing/part-m-00003**

**A picture containing text

Description automatically generated**

5.

1/ Load **datasets/tweets/tweets.psv** into the HDFS folder **tweets**

**hdfs dfs -put** **datasets/tweets/tweets.psv tweets/**



2/ Create a database twitter and create a table called tweets inside the database twitter

**create database twitter;**

**use twitter;**

**create tables tweets (ID varchar(50) Primary Key, timestamp varchar(50), datetime varchar(50), username varchar(50) , tweet\_text varchar(100));**

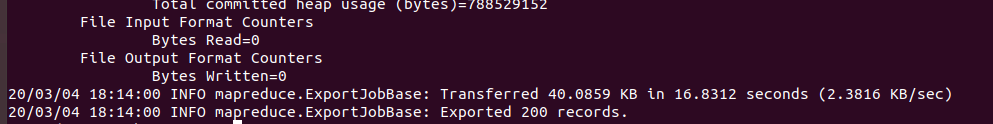
**describe tweets;**

**A screenshot of a cell phone

Description automatically generated**

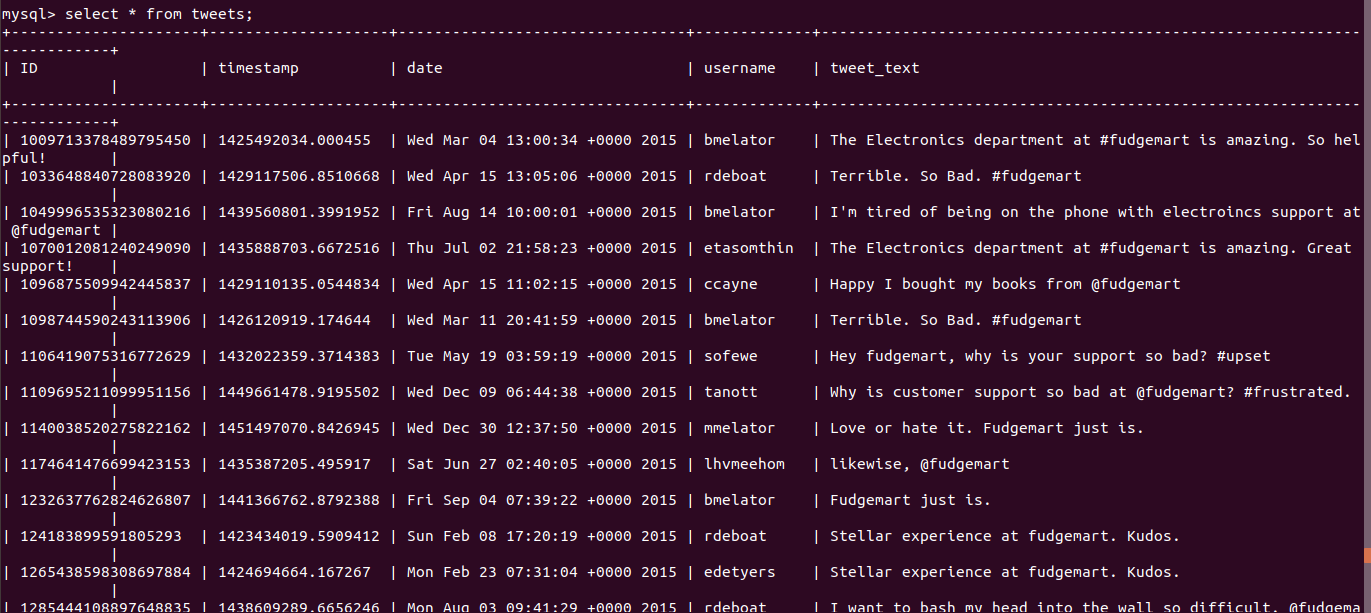
3/ Export the data from HDFS into the MySQL table.

**Sqoop export –connect jdbc://cloudera/twitter –username=root –password=cloudera –table tweets –export-dir tweets/ --input-fields-terminated-by ‘|’**

****

4/ Select all the records from the tweets table

**SELECT \* FROM tweets;**

****

## Student Reflection:

To achieve the highest grade on the assignment you must be as descriptive and personal as possible with your reflection. Ask yourself the following questions.

1. What new information have I learned from completing this assignment?

Mapreduce, Sqoop etc. noSQL databases

1. What were my challenges / roadblocks I encountered on the way to completing it?

Operation of Linux System

1. To be better prepared to attempt this assignment I should \_\_\_\_\_\_\_ ?

Revew

1. If I could make this assignment better, I would \_\_\_\_\_?

Google

1. Rate your comfort level with completing this assignment:

1.

1 ==> I can do this on my own and explain how to do it.

2 ==> I can do this on my own without any help.

3 ==> I can do this with help or guidance from others. If you choose this level, please list those who helped you.

4 ==> I don't understand this at all yet and need extra help. If you choose this, please try to articulate that which you do not understand.