Big Data (ECE 595-004/007), Fall 2018
Hands-on 3: Flume, Sqoop, and Hive
Due date: November 21, 2018, 11:59 PM

1. (**10 pts**) Software installation:
   A. (2pts) MySQL
      Execute command: **`sudo apt-get install mysql-server mysql-common mysql-client`**
   B. (2.5 pts) Flume
      a. Download Apache Flume from http://www.apache.org/dyn/closer.lua/flume/1.8.0/apache-flume-1.8.0-bin.tar.gz
      b. Extract the setup file in the downloaded folder using command:
         `tar xvfz apache-flume-1.8.0-bin.tar.gz`
      c. Move the extracted folder to /usr/local/flume using command
         `sudo mv apache-flume-1.8.0-bin /usr/local/flume`
      d. Change the ownership of /usr/local/flume using command
         `sudo chown –R bigdata:bigdata /usr/local/flume`
         Note: here **bigdata** is user name. In your case it may be different. You can get the user name by executing command **`whoami`**
      e. Update bashrc file in your home directory using command:
         `gedit ~/.bashrc`
         Add following lines in .bashrc file at the end:
         `export FLUME_HOME=/usr/local/flume`
         `export PATH=$PATH:$FLUME_HOME/bin/`
         Close .bashrc file and execute **`source .bashrc`** command
      f. Start Hadoop service using **`start-all.sh`** command and execute **`flume-ng --help`**. If you see any output, it means Flume has installed in your system.
   C. (2.5 pts) Sqoop
      a. Download Sqoop from http://ftp.wayne.edu/apache/sqoop/1.4.7/sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
      b. Extract the setup file in the downloaded folder using command:
         `tar xvfz sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz`
      c. Move the extracted folder to /usr/local/sqoop using command
         `sudo mv sqoop-1.4.7.bin__hadoop-2.6.0 /usr/local/sqoop`
      d. Change the ownership of /usr/local/sqoop using command
         `sudo chown –R bigdata:bigdata /usr/local/sqoop`
      e. Update bashrc file in your home directory using command:
         `gedit ~/.bashrc`
         Add following lines in .bashrc file at the end:
         `export SQOOP_HOME=/usr/local/sqoop`

`export PATH=$PATH:$SQOOP_HOME/bin/`

Close .bashrc file and execute **`source .bashrc`** command

f. Download MySQL connector from https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-java-5.1.47.tar.gz

g. Extract it using command
`tar xvzf mysql-connector-java-5.1.47.tar.gz`

h. Copy mysql-connector-java-5.1.47-bin.jar file from the extracted folder and paste it into /usr/local/sqoop/lib folder. You can do it from command or GUI.

i. Execute **`sqoop help`**. If you see any output, it means Sqoop has installed in your system.

D. (3 pts) Hive

a. Download Hive from http://apache.osuosl.org/hive/stable-2/apache-hive-2.3.4-bin.tar.gz

b. Extract the setup file in the downloaded folder using command:
`tar xvfz apache-hive-2.3.4-bin.tar.gz`

c. Move the extracted folder to /usr/local/hive using command
`sudo mv apache-hive-2.3.4-bin /usr/local/hive`

d. Change the ownership of /usr/local/hive using command
`sudo chown –R bigdata:bigdata /usr/local/hive`

e. Update bashrc file in your home directory using command:
`gedit ~/.bashrc`

Add following lines in .bashrc file at the end:

```
export HIVE_HOME=/usr/local/hive
export PATH=$PATH:$HIVE_HOME/bin/
export HIVE_CONF_DIR=$HIVE_HOME/conf
```

Close .bashrc file and execute **`source .bashrc`** command

f. Open hive-env.sh file using **`gedit /usr/local/hive/conf/hive-env.sh`** command

Add **HADOOP_HOME=/usr/local/hadoop** in the file

g. Create hive-site.xml inside /usr/local/hive/conf folder using command **`touch hive-site.xml`**

Open it using **`gedit /usr/local/hive/conf/hive-site.xml`**

Add following lines in the file

```
<configuration>
<property>
        <name>system:java.io.tmpdir</name>
        <value>/tmp/hive/java</value>
</property>
<property>
        <name>system:user.name</name>
        <value>${user.name}</value>
</property>
```

```
<property>
        <name>javax.jdo.option.ConnectionURL</name>

<value>jdbc:derby:;databaseName=metastore_db;create=true</value>
</property>
</configuration>
```

h. Copy hive-default.xml.template to hive-default.xml inside hive/conf folder

i. Execute **`schematool -initSchema -dbType derby`**

j. Start Hadoop service using **`start-all.sh`** and then execute **`hive`** command. If you see any output, it means Hive has installed in your system.

2. (**20 pts**)

A. (6 pts) Create a Flume agent to copy data files **flume.txt** and **sqoop.txt** (uploaded in Blackboard) from **/tmp/flume** directory to **/user/flume** directory in HDFS cluster. You have to copy each file in **/tmp/flume** folder one by one and Flume should copy them in HDFS. Finally verify that files have been copied into the cluster. Show the files content in HDFS cluster. You have to write and save the agent file in **/usr/local/flume/conf** folder.

B. (8 pts) Create a database **bigdata** in MySQL. Create a table **student** table inside the database. The table has attributes as follows: **name** (varchar), **PUID** (fixed chars, say 5 character digit) as primary key, and **major** (2-4 chars, e.g. ECE, CS, MCE). Name and major should not be null. Insert a few entries into the table. Finally import the table into HDFS and show the content of the table in HDFS cluster.

C. (6 pts) Upload **students.txt** (uploaded in Blackboard) file into HDFS cluster. Find the average score for each subject by creating a Hive table for the file and execute query on it.