Walid Belal
Sept 2020
Revisions by Michael McKee September 2020

Summary

In this exercise, you will install Sqoop on your Sheridan private cloud **Dev** environment Hadoop cluster.

Pre-requisites

You must have completed the Hadoop installation on your Sheridan cloud Dev environment. If you have not done that, please see the document **Week 2 - Install Hadoop on preconfigured VM cluster on Sheridan Cloud**.

You must also be able to connect to your **Dev** Hadoop cluster in the Sheridan private cloud. See the document **Software Required for the Course** and the video on Slate: Connect to Your Hadoop Cluster on Sheridan Cloud.
Note: Your **Dev** cluster is specified since it has Hadoop already installed and configured, and it is ready for installation of Sqoop.

Connect to the VPN

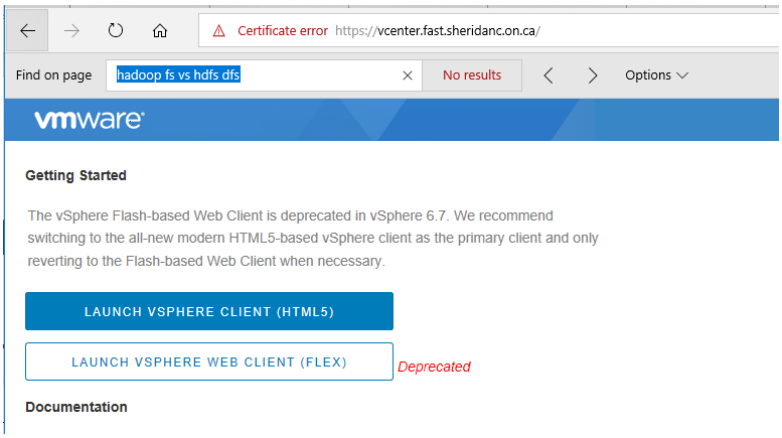
Open Cisco Any Connect Security Mobile Client

Type *vpn.sheridancollege.ca* and connect

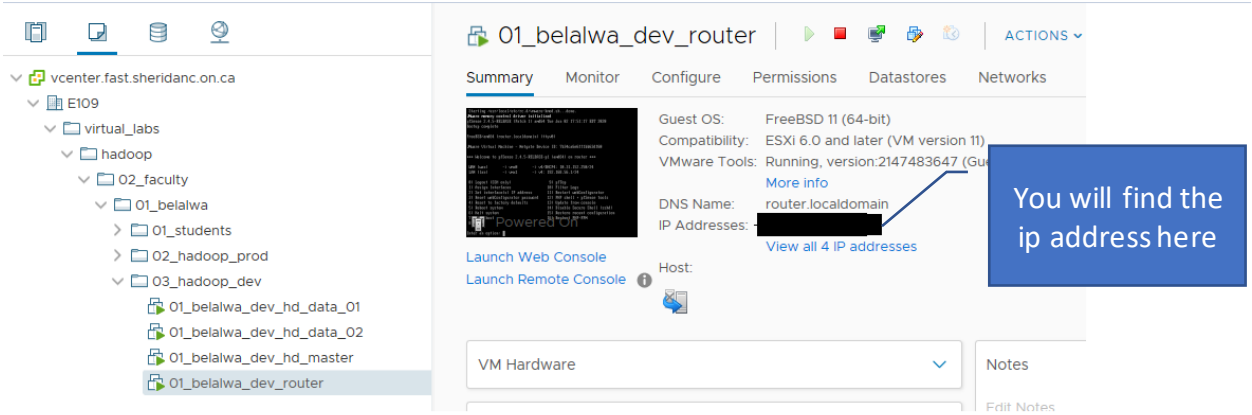
You will need to use your Sheridan credentials to connect NOTE: You need to be connected all the time to the VPN

Find the IP address of the router that you will use to access the servers

Open https://vcenter.fast.sheridanc.on.ca and launch the HTML5 client



Drill down the router on the Hadoop **Dev** environment, and you will find the IP address of the server on the summary page



Make note of the IP address. Write it down.

1. Lab preparation

Connect to the Sheridan VPN

Open Cisco AnyConnect Secure Mobility Client.
Specify the VPN *vpn.sheridancollege.ca* and click **Connect**.
Provide your Sheridan credentials and click OK.

2. Connect and Start the Hadoop cluster

Connect to Hadoop cluster's namenode from your laptop using ssh

You should have already connected to the Sheridan VPN at this point. Please see above if you have not.

Open **cmd**

Type *ssh hadoopuser@<insert your IP address> -p 2221*

Note: replace *<insert your IP address>* above with the IP address of your Sheridan cloud router.

You will be prompted to enter password. The default password is **Sher1dan**

```
C:\Users\belalwa>ssh hadoopuser@10.31.[redacted] -p 2221
hadoopuser@10.31.[redacted]'s password:
```

Now you should get a prompt that looks like this

```
*** System restart required ***
Last login: Sun Sep 13 22:50:26 2020
hadoopuser@hd-master:~$
```

If you get it then you are connected to hd-master.

Start Hadoop

Type *jps* to check if Hadoop is started. If all the following services are started, proceed to the next step.

```
hadoopuser@hd-master:~$ jps
33170 NameNode
33781 ResourceManager
33942 NodeManager
42232 Jps
33340 DataNode
33566 SecondaryNameNode
hadoopuser@hd-master:~$
```

If you do not see the above, start Hadoop by performing the following steps:

In your ssh connection type *start-all.sh*

Type *jps* again and confirm that the processes started without error.

3. Download and extract Sqoop

You will install sqoop in the /opt directory

cd /opt
sudo wget https://downloads.apache.org/sqoop/1.4.7/sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz

Now extract the tar file

sudo tar xvf /opt*/sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz*

Rename the extracted directory to sqoop

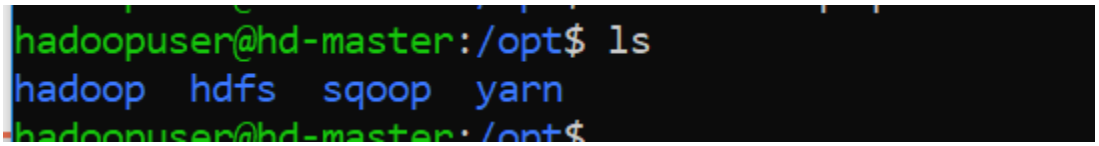
sudo mv sqoop-1.4.7.bin__hadoop-2.6.0sqoop

Remove the tar file

sudo rm sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz

ls

you should see the sqoop directory



4. Add Sqoop path environment variable to profile

sudo nano /etc/profile

Modify the file to add a new environment variable SQOOP_HOME and modify your PATH.
Note: Do not just cut and paste the content into the file.

SQOOP_HOME=/opt/sqoop
Add $SQOOP_HOME/bin to PATH variable

It should look like this:

```
## NEW VARIABLES FOR HADOOP GO BELOW THIS LINE ##
export HADOOP_HOME=/opt/hadoop
export SQOOP_HOME=/opt/sqoop
export PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$SQOOP_HOME/bin
export HADOOP_CONF_DIR=/opt/hadoop/etc/hadoop
```

Logout and login again

logout

Login and try

sqoop version

You should see the below

```
hadoopuser@hd-master:~$ sqoop version
Warning: /opt/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /opt/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /opt/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /opt/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
2020-09-28 23:50:38,833 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Sqoop 1.4.7
git commit id 2328971411f57f0cb683dfb79d19d4d19d185dd8
Compiled by maugli on Thu Dec 21 15:59:58 STD 2017
```

5. Configure Sqoop

Copy the Sqoop environment file from a template

sudo cp $SQOOP_HOME/conf/sqoop-env-template.sh $SQOOP_HOME/conf/sqoop-env.sh

sudo nano $SQOOP_HOME/conf/sqoop-env.sh

Add the variables below to the end of the file.

```
export HADOOP_COMMON_HOME=/opt/hadoop
export HADOOP_MAPRED_HOME=/opt/hadoop
```

6. Make changes to Hadoop config files

Add the below properties to yarn-site.xml between the <configuration> and </configuration> tags.

sudo nano $HADOOP_HOME/etc/hadoop/yarn-site.xml

```
<property>
  <name>yarn.nodemanager.vmem-check-enabled</name>
  <value>false</value>
</property>
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property>
```

Add the properties below to mapred-site.xml between the <configuration> and </configuration> tags.

sudo nano $HADOOP_HOME/etc/hadoop/mapred-site.xml

```
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>

<name>yarn.app.mapreduce.am.env</name>
<value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
</property>
<property>
<name>mapreduce.map.env</name>
<value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>
</property>
<property>
<name>mapreduce.reduce.env</name>
<value>HADOOP_MAPRED_HOME=${HADOOP_HOME}</value>

</property>
```

Copy the MySql-Java connector jar file to /opt/sqoop/lib folder.
You would find jar file in mysql-connector-java-8.0.20 folder.

sudo cp /usr/share/java/mysql-connector-java.jar /opt/sqoop/lib

You will need to copy **commons-lang-2.6.jar** file to the /opt/sqoop/lib directory.
Download the jar file as provided on Slate.
Using WinSCP, copy the file to the folder **upload-example**.
Then from your Dev hd-master, copy the file to /opt/sqoop/lib.

sudo cp upload-example/commons-lang-2.6.jar /opt/sqoop/lib

7. Connect to MySQL to create a table

MySQL is already installed on the hd-master node.
In the following steps you will connect to the MySQL as root user and you will create a user that will be used to work with Sqoop.

mysql -u root -p

You will be prompted for password. The password is Sher1dan@

```
hadoopuser@hd-master:~$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 14
Server version: 8.0.21-0ubuntu0.20.04.4 (Ubuntu)

Copyright (c) 2000, 2020, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>
```

At the mysql> prompt, type the following to create the database, use the database, create a table Employee and insert into it some values:

CREATE dbtest;

USE dbtest;

CREATE TABLE Employee
(Eid VARCHAR(3) PRIMARY KEY,
Ename VARCHAR(30),
Eaddress VARCHAR(50));

After that completes, type the following at the mysql> prompt:

INSERT INTO Employee VALUES(101,'tr','333 akad rd');
INSERT INTO Employee VALUES(103,'tt','344 akad rd');
COMMIT;

exit

Now test Sqoop

sqoop list-databases --connect jdbc:mysql://localhost/ --username root --password Sher1dan@

sqoop import --connect jdbc:mysql://localhost/dbtest --username root --password Sher1dan@ --table Employee --m 1 --target-dir /sqoop-r2

You can verify the data was written to HDFS.
With the previous exercises you have completed, you have everything you need to examine the newly written data.