Install Hadoop on Preconfigured VM cluster on Sheridan Cloud

Original author: Walid Belal, September 2020

Revision by Michael McKee, September 2020

Summary

In this tutorial you will install and configure Hadoop on a three-node Dev cluster of virtual machines on the Sheridan Private Cloud. All the prerequisites are installed on the VMs.

This cluster will have three nodes: **hd-master**, **hd-data-01**, and **hd-data-02**. The node **hadoop1** will be both the namenode and datanode at the same time while **hadoop2** and **hadoop3** are strictly datanodes. These IP addresses are already given, you just need to remember these. There is no need to setup these IP addresses for starting.

Node name	hd-master	hd-data-01	hd-data-02
IP Address (local to cluster)	192.168.56.5	192.168.56.6	192.168.56.7
Function	Name node Data node	Data Node	Data Node

Lab Preparation

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 because it does not have Hadoop configured.

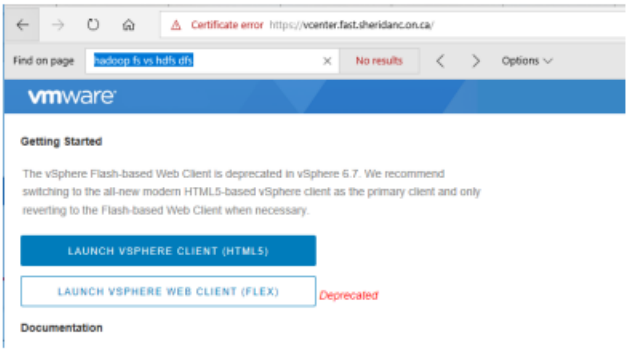
Connect to the Sheridan VPN

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

Note: You will need to remain connected at all time during this exercise.

Find the IP address of the router to access your servers

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



Drill down and click on your router on the Hadoop **Dev** environment. Record the IP address.

Connect to the servers

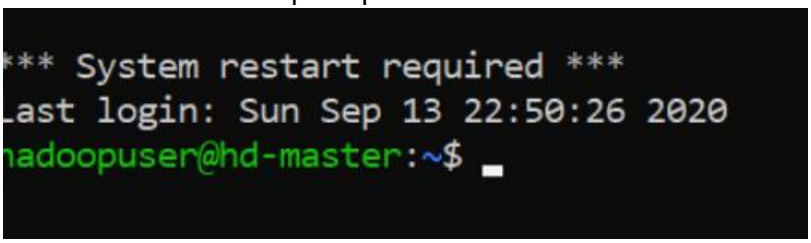
Use ssh from your computer to connect to the router (which will connect you to hd-master). You will be able to connect to the other nodes hd-data-01 (192.168.56.6) and hd-data-02 (192.168.56.7) from hd-master.

ssh hadoopuser@<your router IP address> -p 2221

Note: replace <your router IP address> with the IP address obtained above.

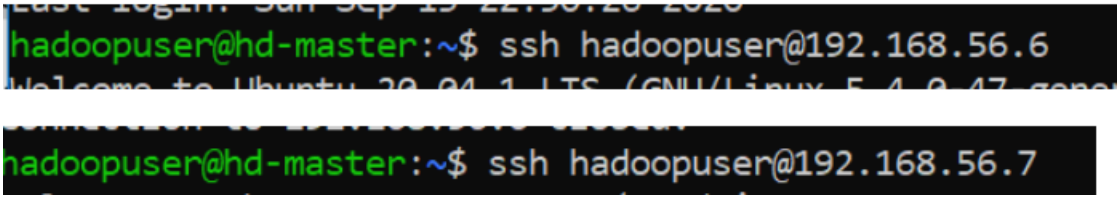
Note: Port 2221 is used to forward you to hd-master. You will be prompted for a password; default is Sher1dan. It is advised to change the password on all your Hadoop nodes.

You should now see a prompt like this:



This means you are now connected to hd-master.

You can then ssh from hd-master to the other servers. You should not be prompted for a password.



Install Hadoop (on all the machines)

Create the below folders on all Hadoop machines one by one.

```
sudo mkdir -p /opt/hadoop/logs
sudo mkdir -p /opt/hdfs/datanode
sudo mkdir -p /opt/hdfs/namenode
sudo mkdir -p /opt/yarn/logs
sudo mkdir -p /opt/yarn/local
sudo mkdir -p /opt/hdfs/tmp
```

```
sudo find /opt -type d -exec chmod -R 775 {} \;
```

```
sudo chown -R hadoopuser /opt/hadoop
sudo chown -R hadoopuser /opt/hdfs
sudo chown -R hadoopuser /opt/yarn
```

Unpack Hadoop on all machines

Hadoop is already downloaded on your virtual machines in the **/home/hadoopuser/resources** directory. You will need to unpack it into the **/opt/hadoop** directory.

```
sudo tar xzf /home/hadoopuser/resources/hadoop-3.3.0.tar.gz --directory=/opt/hadoop --strip 1
```

Configure Hadoop (on all the machines)

Update /etc/profile

Add the following configuration to the profile at the end of the file.

sudo nano /etc/profile

```
export HADOOP_HOME=/opt/hadoop
export PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export HADOOP_CONF_DIR=/opt/hadoop/etc/hadoop
export HDFS_NAMENODE_USER=hadoopuser
export HDFS_DATANODE_USER=hadoopuser
export HDFS_SECONDARYNAMENODE_USER=hadoopuser
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_MAPRED_HOME=/opt/hadoop
export HADOOP_COMMON_HOME=/opt/hadoop
export HADOOP_HDFS_HOME=/opt/hadoop
export PDSH_RCMD_TYPE=ssh
```

Update the environment variables (on all machines)

Add the following configuration to **hadoop-env.sh**

nano /opt/hadoop/etc/hadoop/hadoop-env.sh

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/opt/hadoop
export HADOOP_CONF_DIR=/opt/hadoop/etc/hadoop
export HADOOP_LOG_DIR=/opt/hadoop/logs
```

Logout of all nodes (all machines)

logout

Log back in again

TEST:
Once done execute this command:

hadoop version

Hadoop's version will be printed

Troubleshooting tip:
If you will get an error that Hadoop is not a command, then most likely there is a typo in the PATH variable of the profile. There may be a return after you copy and paste. The PATH variable should not be fragmented with spaces.

Following steps are required for hd-master only:

Configure the name node

You will apply the below operations only to the **name node** machine which is called **hd-master** in this case

Update the hdfs-site.xml to define the nodes

nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml

```
<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
      <value>file:///opt/hdfs/namenode</value>
      <description>NameNode directory for namespace and transaction logs storage.</description>
  </property>

  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:///opt/hdfs/datanode</value>
    <description>DataNode directory</description>
  </property>

  <property>
    <name>dfs.replication</name>
    <value>3</value>
  </property>

  <property>
    <name>dfs.permissions</name>
    <value>false</value>
  </property>

  <property>
    <name>dfs.datanode.use.datanode.hostname</name>
    <value>false</value>
  </property>

  <property>
    <name>dfs.namenode.datanode.registration.ip-hostname-check</name>
    <value>false</value>
  </property>

</configuration>
```

Update core-site.xml on hd-master

nano $HADOOP_HOME/etc/hadoop/core-site.xml

```
<configuration>
 <property>
  <name>fs.defaultFS</name>
  <value>hdfs://hd-master:9820/</value>
  <description>NameNode URI</description>
 </property>
</configuration>
```

Update yarn-site.xml on hd-master

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

```
<configuration>

 <property>
  <name>yarn.nodemanager.local-dirs</name>
  <value>file:///opt/yarn/local</value>
 </property>

 <property>
  <name>yarn.nodemanager.log-dirs</name>
  <value>file:///opt/yarn/logs</value>
 </property>

</configuration>
```

Update mapreduce config file on hd-master

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

```
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
<description>MapReduce framework name</description>
</property>
<property>
<name>mapreduce.jobhistory.address</name>
<value>hd-master:10020</value>
<description>Default port is 10020</description>
</property>
<property>
<name>mapreduce.jobhistory.webapp.address</name>
<value> hd-master:19888</value>
<description>Default port is 19888</description>
</property>
<property>
<name>mapreduce.jobhistory.intermediate-done-dir</name>
<value>/mr-history/tmp</value>
<description>Directory where history files are written by MapReduce jobs.</description>
</property>
<property>
<name>mapreduce.jobhistory.done-dir</name>
<value>/mr-history/done</value>
<description>Directory where history files are managed by the MR JobHistory Server.</description>
</property>
</configuration>
```

Format the name node

hdfs namenode –format

Add IP addresses to workers file.

Add the IP addresses of the data nodes to the **workers** file.
Replace any existing entries (e.g. **localhost) with the** following IP addresses for the **workers** file:

nano $HADOOP_HOME/etc/hadoop/workers

```
192.168.56.5
192.168.56.6
192.168.56.7
```

Perform following steps on hd-data-01 and hd-data-02 VMs

Configure data nodes

ssh to every data node (i.e. ssh hadoopuser@192.168.56.6 and ssh hadoopuser@192.168.56.6) from hd-master. Perform the following steps.

Update hdfs-site.xml (on both hd-data-01 and hd-data-02)

nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml

```
<configuration>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:///opt/hdfs/datanode</value>
<description>DataNode directory</description>
</property>
</configuration>
```

Update core-site.xml (on both hd-data-01 and hd-data-02)

nano $HADOOP_HOME/etc/hadoop/core-site.xml

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://hd-master:9820/</value>
<description>NameNode URI</description>
</property>
</configuration>
```

Update yarn-site.xml (on both hd-data-01 and hd-data-02)

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

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
<description>Yarn Node Manager Aux Service</description>
</property>
</configuration>
```

Start Hadoop cluster (from hd-master)

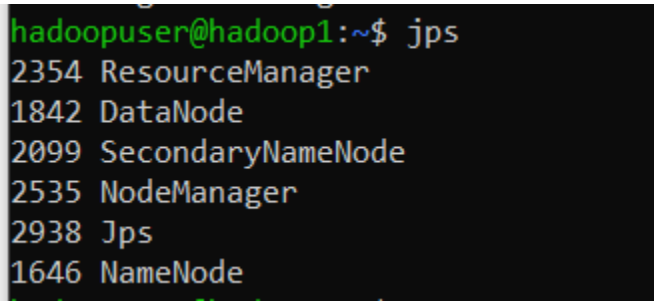
Now everything is configured, and you can start the cluster

start-all.sh

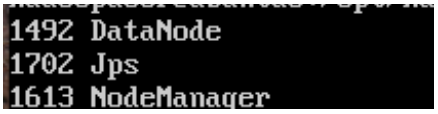
Check the started services on the namenode

jps

You should find these services on the namenode



And the following on the datanodes



If the DataNode service is not started, then most likely you need to recreate datanode folder

Test the cluster

Create a folder on Hadoop

Type the following on the namenode

hadoop fs -mkdir /test

Check if the directory was created

hadoop fs -ls /

Tips

Pasting a copied line may add an unwanted carriage return if what is being copied spans more than one line. Make sure any such extra carriage returns are removed.

Copy config files (this step is only required if you messed up with configuration files; otherwise this is not required)

Every node comes preloaded with the correct config files in the following directory:
/home/hadoopuser/resources/configfiles

You can use linux cp command to replace the original files with these preconfigured files.