Walid Belal
Sept 2020

Summary

In this tutorial you will install and configure Hadoop on a three node cluster of virtual machines. All the prerequisites are installed on the VMs. “Import Hadoop Cluster VMs” tutorial will show you how to import the VMs to your laptop

This cluster will have three nodes hadoop1, hadoop2, and hadoop3. hadoop1 will be 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. No need to setup these IP addresses for starting.

Node name	Hadoop1	Hadoop2	Hadoop3
IP Address	192.168.56.5	192.168.56.6	192.168.56.7
function	Name node Data node	Data Node	Data Node

Before starting this document, make sure all Hadoop clusters are started from VBOx. In order to do that start VBOX and select hadoop1 cluster and click start button in VBOX.

Repeat same step for hadoop2 and 3. Once all clusters started then login to these clusters.
Login : hadoopuser
Password : Sher1dan (Here S is in capital letter)

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/hadoop/mapredhistory/tmp
sudo mkdir –p /opt/hadoop/mapredhistory/done
```

```
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 {} \;
```

unpack Hadoop on all machines

Hadoop is predownloaded on you machines in /home/hadoopuser/resources directory. You will need to unpack inside /opt/Hadoop directory

```
cd /opt/hadoop
```

```
sudo tar xvf /home/hadoopuser/resources/hadoop-3.1.3.tar  --directory=/opt/hadoop  --strip 1
```

(if not gz format)

```
sudo tar xzf  /home/hadoopuser/resources/hadoop-3.1.3.tar.gz  --directory=/opt/hadoop --strip 1
```

*(if gz format – **USE THIS ONE**)*

Configure Hadoop (on all the machines)

Update /etc/profile

Add the following configuration to the profile

```
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

```
sudo 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
```

Restart the server (all machines)

```
sudo shutdown -r now
```

TEST:
Once done try this command

```
hadoop version
```

hadoop’s version will be printed

Following steps are only required for Hadoop1 only;-

Configure the name node

You will apply the below operations only to the name node machine which is called Hadoop1 in this case

Update the hdfs-site.xml to define the nodes

```
sudo 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

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

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://hadoop1:9820/</value>
<description>NameNode URI</description>
</property>
</configuration>
```

Update yarn-site.xml

```
sudo 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

```
sudo 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>hadoop1:10020</value>
<description>Default port is 10020</description>
</property>
<property>
<name>mapreduce.jobhistory.webapp.address</name>
<value> hadoop1:19888</value>
<description>Default port is 19888</description>
</property>
<property>
<name>mapreduce.jobhistory.intermediate-done-dir</name>
<value>/mapredhistory/tmp</value>
<description>Directory where history files are written by MapReduce jobs.</description>
</property>
<property>
<name>mapreduce.jobhistory.done-dir</name>
<value>/mapredhistory/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 the ip addresses of the data nodes to workers file
It would already have entry for localhost, press enter and add following IP addresses to these workers file;-

```
sudo nano $HADOOP_HOME/etc/hadoop/workers
```

```
192.168.56.5
192.168.56.6
192.168.56.7
```

restart the node (only hadoop1 i.e. NameNode)

```
sudo shutdown -r now
```

Perform following steps on Hadoop2 and Hadoop3 clusters

Configure data nodes

Update hdfs-site.xml(both hadoop2 and hadoop3)

```
sudo 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(both hadoop2 and hadoop3)

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

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://hadoop1:9820/</value>
<description>NameNode URI</description>
</property>
</configuration>
```

Update yarn-site.xml(both hadoop2 and hadoop3)

```
sudo 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>
```

restart the node (both hadoop2 and hadoop3)

```
sudo shutdown -r now
```

Start hadoop cluster (from hadoop1 cluster)

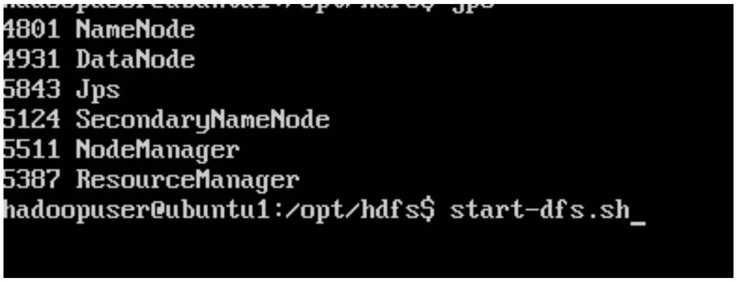
Now everything is configured and you can start the cluster

```
start-all.sh
```

Check the started services on the namenode

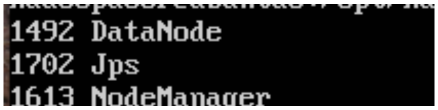
```
jps
```

you should find these services on the namenode



```
4801 NameNode
4931 DataNode
5843 Jps
5124 SecondaryNameNode
5511 NodeManager
5387 ResourceManager
hadoopuser@ubuntu1:/opt/hdfs$ start-dfs.sh_
```

And the following on the datanodes



```
1492 DataNode
1702 Jps
1613 NodeManager
```

if the data node 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

ssh to the cluster

You can always ssh from your local client to the nodes. This will allow you to copy and paste code

Type the following in cmd

ssh hadoopuser@192.168.56.5

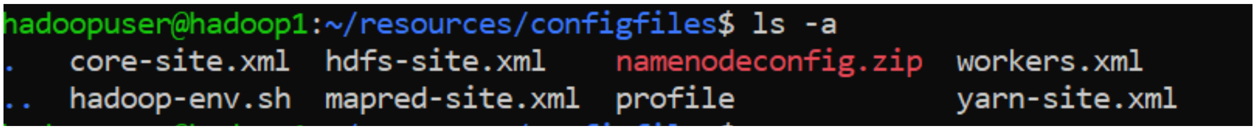
this will ssh you to the name node. The password is Sher1dan

192.168.56.6 and 192.168.56.7 are the ips of the datanodes

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 /home/hadoopuser/resources/configfiles directory

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



```
hadoopuser@hadoop1:~/resources/configfiles$ ls -a
.   core-site.xml  hdfs-site.xml    namenodeconfig.zip  workers.xml
..  hadoop-env.sh  mapred-site.xml  profile             yarn-site.xml
```