**Installing HADOOP**

1. The first step is to make sure that all our servers are running Oracle Java Version 7.

To install Oracle Java, you can download Oracle JDK from Oracle’s website. The

Oracle Java installation file that I have downloaded is jdk-7u51-linux-x64.rpm.

After downloading the file, perform the following instructions as the root user:

**$ chmod +x jdk-7u51-linux-x64.rpm**

**$ rpm -ivh jdk-7u51-linux-x64.rpm**

**$ export JAVA\_HOME="/usr/java/jdk1.7.0\_51/jre/"**

**$ export PATH=$JAVA\_HOME/bin:$PATH**

The preceding instructions will install Oracle Java 7, set the JAVA\_HOME environment

variable, and add the bin folder for the Java runtime to PATH.

2. Create a user, for example, hduser, and set a password for the user, using the

following commands:

**$ useradd hduser**

**$ passwd hduser**

3. Provide the user with sudo privileges by editing the sudoers file placed at /etc and

adding the following line:

**hduser ALL=(ALL) ALL**

4. Download the CDH5 RPM for CentOS (here, I am using CentOS 6) using the

following link:

http://archive.cloudera.com/cdh5/one-click-install/redhat/6/x86\_64/cloudera-cdh-5-

0.x86\_64.rpm

5. Log in as hduser and install the downloaded RPM using the following command:

**$ sudo yum --nogpgcheck localinstall cloudera-cdh-5-0.x86\_64.rpm**

6. Perform steps 1, 2, 3, 4, and 5 on all servers that will be part of the cluster.

7. Execute the following command from the user hduser on node1.hcluster to install

the namenode daemon:

**$ sudo yum install hadoop-hdfs-namenode**

**Yum list | grep namenode**

8. Execute the following command from the user hduser on node1.hcluster to install

the jobtracker daemon:

**$ sudo yum install hadoop-0.20-mapreduce-jobtracker**

9. Execute the following command from the user hduser on node1.hcluster to install

the Hadoop client:

**$ sudo yum install hadoop-client**

10. After executing the preceding steps, you will find the file core-site.xml under

/etc/hadoop/conf/. Initially, this file is empty. Edit the file and update it as follows:

<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://node1.hcluster:8020</value>

<description>

Defines the name of the filesystem.

</description>

</property>

</configuration>

For a complete listing of configurable properties for the core-site.xml file, refer to

http://hadoop.apache.org/docs/r2.3.0/hadoop-project-dist/hadoop-common/coredefault.

xml.

11. Similarly, the file hdfs-site.xml under /etc/hadoop/conf/ will be empty. Update it

as follows:

<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>

<name>dfs.namenode.servicerpc-address</name>

<value>node1.hcluster:8022</value>

<description>

This is the RPC address for the namenode. This address is used by

services like the datanodes to connect to the namenode.

</description>

</property>

<property>

<name>dfs.namenode.http-address</name>

<value>node1.hcluster:50070</value>

<description>

This is the HTTP address for the namenode web user interface.

</description>

</property>

<property>

<name>dfs.replication</name>

<value>3</value>

<description>

This property defines the replication factor of the data blocks

in HDFS.

</description>

</property>

<property>

<name>dfs.blocksize</name>

<value>134217728</value>

<description>

This property defines block size for files in HDFS (bytes).

</description>

</property>

</configuration>

For a complete list of configurable properties for the hdfs-site.xml file, refer to

http://hadoop.apache.org/docs/r2.3.0/hadoop-project-dist/hadoop-hdfs/hdfsdefault.

xml.

12. Create the mapred-site.xml file in /etc/hadoop/conf if is not already present, and

update it as follows:

<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>

<name>mapred.job.tracker</name>

<value>node1.hcluster:8021</value>

<description>

This property defines the address at which the jobtracker service

runs.

</description>

</property>

<property>

<name>mapred.job.tracker.http.address</name>

<value>0.0.0.0:50030</value>

<description>

This property defines HTTP address for jobtracker web user

interface.

</description>

</property>

<property>

<name>mapred.reduce.tasks</name>

<value>4</value>

<description>

This property defines the number of reduce tasks that can run on

the cluster.

</description>

</property>

</configuration>

For a complete list of configurable properties for the mapred-site.xml file, refer to

http://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduceclient-core/mapred-default.xml.

13. Update the slaves file placed at /etc/hadoop/conf as follows:

node2.hcluster

node3.hcluster

node4.hcluster

14. Execute the following command as hduser on node1.hcluster to format the

namenode daemon:

**$ sudo -u hdfs hdfs namenode -format**

15. Execute the following command from hduser on node2.hcluster, node3.hcluster,

and node4.hcluster to install tasktracker and datanode:

**$ sudo yum install hadoop-0.20-mapreduce-tasktracker hadoop-hdfsdatanode**

Copy the core-site.xml, hdfs-site.xml, and mapred-site.xml files from

node1.hcluster to node2.hcluster, node3.hcluster, and node4.hcluster in

/etc/hadoop/conf/.

16. Execute the following command from hduser on each node to start HDFS:

**$ for x in 'cd /etc/init.d ; ls hadoop-hdfs-\*' ; do sudo service $x**

**start ; done**

The preceding command will start the namenode daemon on node1.hcluster and the

datanode daemon on node2.hcluster, node3.hcluster, and node4.hcluster.

17. Execute the following command from hduser on node1.hcluster:

**$ sudo -u hdfs hdfs dfs -mkdir /tmp**

**$ sudo -u hdfs hdfs dfs -chmod -R 1777 /tmp**

**$ sudo -u hdfs hdfs dfs -mkdir -p /var/lib/hadoophdfs/cache/mapred/mapred/staging**

**$ sudo -u hdfs hdfs dfs -chmod 1777 /var/lib/hadoophdfs/cache/mapred/mapred/staging**

**$ sudo -u hdfs hadoop fs -chown -R mapred /var/lib/hadoophdfs/cache/mapred**

**$ sudo -u hdfs hadoop fs -mkdir -p /tmp/mapred/system**

18. Execute the following command from hduser on node1.hcluster to start the

jobtracker daemon:

**$ sudo service hadoop-0.20-mapreduce-jobtracker start**

19. Execute the following command from hduser on node2.hcluster, node3.hcluster,

and node4.hcluster to start the tasktracker daemon:

**$ sudo service hadoop-0.20-mapreduce-tasktracker start**

Your four-node Hadoop cluster should now be up and running. You can test the cluster by

visiting the URLs for the namenode UI and the jobtracker UI using a browser. The URL

for the namenode UI is http://node1.hcluster:50070. The URL for the jobtracker UI is

http://node1.hcluster:50030.

To modify the properties of the HDFS and MapReduce, edit the configuration files present

under the location /etc/hadoop/conf.

**Stopping Hadoop services**

Execute the following command from the user hduser to stop the namenode daemon on

node1.hcluster and the datanode daemon on node2.hcluster, node3.hcluster, and

node4.hcluster:

**$ for x in 'cd /etc/init.d ; ls hadoop-hdfs-\*' ; do sudo service $x stop ;**

**done**

Execute the following command from the user hduser to stop the jobtracker daemon on

node1.hcluster:

**$ sudo service hadoop-0.20-mapreduce-jobtracker stop**

Execute the following command from the user hduser to stop the tasktracker daemon on

node2.hcluster, node3.hcluster, and node4.hcluster:

**$ sudo service hadoop-0.20-mapreduce-tasktracker**