Servers

Asif

server1.asif.com [162.242.237.170](mailto:root@162.242.237.170) asifali123

server2.asif.com [162.242.240.134](mailto:root@162.242.240.134) asifali123

server3.asif.com [162.242.240.169](mailto:root@162.242.240.169) asifali123

server4.asif.com [162.242.240.130](mailto:root@162.242.240.130) asifali123

**Multi node**

download java bin file from oracle.com

copy this file from windows to linux in root/ using winscp

change mode of jdk file to 777

# chmod 777 jdkfilename.bin

# ./jdk file name.bin

(doing the above stores java in the path /usr/java/jdk1.7 ... used by hadoop)

java -version

setting JAVA\_HOME variables

(I) in command line just type the following command

# export JAVA\_HOME=/usr/java/jdk1.7/

# export PATH=$PATH:/usr/java/jdk1.7/

or

(II) vi ~/.bash\_profile

export JAVA\_HOME=/usr/lib/jvm/java-1.7.0-openjdk-1.7.0.51.x86\_64/jre

export PATH=$PATH:/usr/lib/jvm/java-1.7.0-openjdk-1.7.0.51.x86\_64

run

# . ~/.bash\_profile

echo $JAVA\_HOME

echo $PATH

**Add all ip address and host name in the file**

**# vi /etc/hosts/**

**162.242.156.85 server196.cloudwick.com**

**23.253.90.24 server197.cloudwick.com**

**2.download cdh3 repository from cloudera**

# wget http://archive.cloudera.com/redhat/6/x86\_64/cdh/cdh3-repository-1.0-1.noarch.rpm

**3.install the repository**

# sudo yum --nogpgcheck localinstall cdh3-repository-1.0-1.noarch.rpm (installs in etc/repos.d)

**4.search hadoop packages**

#yum search hadoop

#yum install hadoop-0.20 (installs in /usr/lib/hadoop/bin)

**5.install system specific deamon packages(namenode)**

for server ([162.242.237.170](mailto:root@162.242.237.170) [server1.asif.com](http://server1.asif.com/))

#yum install hadoop-0.20-namenode

**6.check the installation is done by**

#rpm -ql hadoop-0.20 (displays installed information)

#hadoop version (to check for hadoop version installed)

#man hadoop (view the man page)

**7.create hadoop ,hdfs directory.( /hadoop/hdfs)**

#mkdir hadoop in root path

#mkdir hdfs

**8.in hdfs dir create name and data and namesecondary dir.**

#mkdir name (/hadoop/hdfs/name)

#mkdir data (/hadoop/hdfs/data)

#mkdir namesecondary (/hadoop/hdfs/namesecondary)

**8.1 from root change the ownner of hadoop directory to owner hadoop**

#chown -R hdfs:hadoop hadoop

I) hdfs is user

ii)hadoop is group

III)for hadoop directory

**9.change configuration file properties**

#cd /usr/lib/hadoop/conf/

**10. #vi core-site.xml**

edit file:

(name does not change only the property value changes)

<configuration>

<property>

<name>fs.default.name</name>

<value>hdfs://<162.242.237.170(ip-address)>or<hostname>:8020/</value>

</property>

</configuration>

**11. hdfs-site.xml:**

#vi hdfs-site.xml

configuration>

<property>

<name>dfs.name.dir</name>

<value>/hadoop/hdfs/name</value>

</property>

<property>

<name>dfs.data.dir</name>

<value>/hadoop/hdfs/data</value>

</property>

<property>

<name>fs.checkpoint.dir</name>

<value>/hadoop/hdfs/namesecondary</value>

</property>

<property>

<name>dfs.replication</name>

<value>3</value>

</property>

</configuration>

**12.mapred-site.xml**

# vi mapred-site.xml

<configuration>

<property>

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

<value><162.242.240.134(ip-address)>or<hostname>:8021</value>

</property>

<property>

<name>mapred.tasktracker.map.tasks.maximum</name>

<value>4</value>

</property>

<property>

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

<value>2</value>

</property>

</configuration>

**13. edit master and slave files**

#vi masters

ip address or hostname of name node server

# vi slaves

ip address or hostname of servers running datanode and task tracker.

**5.1.install system specific deamon packages(jobtracker and secondary name node)**

for server ([162.242.240.134](mailto:root@162.242.237.170) [server2.asif.com](http://server1.asif.com/))

#yum install hadoop-0.20-jobtracker

yum install hadoop-0.20-secondarynamenode

**7.create hadoop ,hdfs directory.( /hadoop/hdfs)**

#mkdir hadoop in root path

#mkdir hdfs

**8.in hdfs dir create name and data , namesecondary dir.**

#mkdir name (/hadoop/hdfs/name)

#mkdir data (/hadoop/hdfs/data)

#mkdir namesecondary (/hadoop/hdfs/namesecondary)

**8.1 from root change the ownner of hadoop directory to owner hadoop**

#chown -R hdfs:hadoop hadoop

I) hdfs is user

ii)hadoop is group

III)for hadoop directory

**9.change configuration file properties**

#cd /usr/lib/hadoop/conf/

**10. #vi core-site.xml**

edit file:

(name does not change only the property value changes)

<configuration>

<property>

<name>fs.default.name</name>

<value>hdfs://<ip-address>or<hostname>:8020/</value>

</property>

</configuration>

**11. hdfs-site.xml:**

#vi hdfs-site.xml

configuration>

<property>

<name>dfs.name.dir</name>

<value>/hadoop/hdfs/name</value>

</property>

<property>

<name>dfs.data.dir</name>

<value>/hadoop/hdfs/data</value>

</property>

<property>

<name>fs.checkpoint.dir</name>

<value>/hadoop/hdfs/namesecondary</value>

</property>

<property>

<name>dfs.replication</name>

<value>3</value>

</property>

</configuration>

**12.mapred-site.xml**

# vi mapred-site.xml

<configuration>

<property>

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

<value><ip-address>or<hostname>:8021</value>

</property>

<property>

<name>mapred.tasktracker.map.tasks.maximum</name>

<value>4</value>

</property>

<property>

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

<value>2</value>

</property>

</configuration>

**13. edit master and slave files**

#vi masters

ip address or hostname of jobtracker server

# vi slaves

ip address or hostname of servers running datanode and task tracker

**5.install system specific deamon packages(datanode and task tracker)**

for server ([162.242.240.169](mailto:root@162.242.237.170) [server3.asif.com](http://server1.asif.com/))

for server ([162.242.240.130](mailto:root@162.242.237.170) [server4.asif.com](http://server1.asif.com/))

#yum install hadoop-0.20-datanode

#yum install hadoop-0.20-tasktracker

**6.check the installation is done by**

#rpm -qa | hadoop-0.20 (displays installed information)

#hadoop version (to check for hadoop version installed)

#man hadoop (view the man page)

**7.create hadoop ,hdfs directory.( /hadoop/hdfs)**

#mkdir hadoop in root path

#mkdir hdfs

**8.in hdfs dir create name and data and namesecondary dir.**

#mkdir name (/hadoop/hdfs/name)

#mkdir data (/hadoop/hdfs/data)

#mkdir namesecondary (/hadoop/hdfs/namesecondary)

**8.1 from root change the ownner of hadoop directory to owner hadoop**

#chown -R hdfs:hadoop hadoop

I) hdfs is user

ii)hadoop is group

III)for hadoop directory

**9.change configuration file properties**

#cd /usr/lib/hadoop/conf/

**10. #vi core-site.xml**

edit file:

(name does not change only the property value changes)

<configuration>

<property>

<name>fs.default.name</name>

<value>hdfs://<162.242.237.170(ip-address)>or<hostname>:8020/</value>

</property>

</configuration>

**11. hdfs-site.xml:**

#vi hdfs-site.xml

configuration>

<property>

<name>dfs.name.dir</name>

<value>/hadoop/hdfs/name</value>

</property>

<property>

<name>dfs.data.dir</name>

<value>/hadoop/hdfs/data</value>

</property>

<property>

<name>fs.checkpoint.dir</name>

<value>/hadoop/hdfs/namesecondary</value>

</property>

<property>

<name>dfs.replication</name>

<value>3</value>

</property>

</configuration>

**12.mapred-site.xml**

# vi mapred-site.xml

<configuration>

<property>

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

<value><162.242.240.134(ip-address)>or<hostname>:8021</value>

</property>

<property>

<name>mapred.tasktracker.map.tasks.maximum</name>

<value>4</value>

</property>

<property>

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

<value>2</value>

</property>

</configuration>

**13. edit master and slave files**

#vi masters

does not have any

# vi slaves

ip address or hostname of servers running datanode and task tracker.

format name node:

# sudo –u hdfs /usr/lib/hadoop/bin/hadoop namenode –format

**to stop service tables (file wall)**

# service iptables stop

**start name node service**:

# etc/init.d/hadoop-0.20-namenode start

**to check services are running**

netstat -plten

in web : http://namenodeserver.com:50070

**start secondarynamenode**

# etc/init.d/hadoop-0.20-secondarynamenode start

**start datanode**

# etc/init.d/hadoop-0.20-datanode start

**start jobtracker**

# etc/init.d/hadoop-0.20-jobtracker start

**if permission denied error occurs when checked in log file then create the directory tmp**

sudo -u hdfs hadoop fs -mkdir hdfs://server1.asif.com/tmp/

or

sudo -u hdfs hadoop fs -mkdir /tmp

**change ownership**

sudo -u hdfs hadoop fs -chown -R mapred:supergroup /tmp

or

in mapred-site-xml include the following property

**<property>**

**<name>****mapreduce.jobtracker.system.dir</name>**

**<value>/tmp/mapred/system</value>**

**</property>**

**start tasktracker**

# etc/init.d/hadoop-0.20-tasktracker start

**After cluster started running to execute a sample mapreduce program of wordcount**

**1) create a file in local file system in linux in root directory(/.)**

vi abc.txt

--> **create a directory in hdfs**

# sudo -u hdfs hadoop fs -mkdir /user/asif

**2) move abc.txt to hdfs**

#sudo -u hdfs hadoop fs -put <abc.txt path> <hdfs path(/user/asif)>

**examp:** **sudo -u hdfs hadoop fs -put examp.txt /user/asif**

**To run example wordcount program**

#sudo -u hdfs hadoop jar /usr/lib/hadoop/hadoop.examples.jar wordcount <input path of abc.txt file in hdfs> <output path in hdfs>

examp: sudo -u hdfs hadoop jar /usr/lib/hadoop/hadoop-examples.jar wordcount /user/varun /user/asif2

DISTCP command ( to copy data from one hdfs cluster to another)

**1) it should be run from the destination cluster only**

**--> if both the clusters are of same hadoop version then**

sudo -u hdfs hadoop distcp hdfs://namenode ip address/path of file hdfs://namenode ipaddress/destination path

or

sudo -u hdfs hadoop distcp -i hdfs://namenode ip address/path of file hdfs://namenode ipaddress/destination path

**if destination cluster is of higher hadoop version and source cluster is of lower hadoop version then**

sudo -u hdfs hadoop distcp -i hftp://namenode ipaddress:50070/path of file hdfs://namenode ipaddress:8020/destination path

**cominission (adding datanode to cluster)**

-->> in namenode server, create an include file ie (# vi include ) in the path you give in **dfs.hosts.include** in hdfs-site-xml, this include file should contain ip address of the datanode you want to add.

I)set the core-site, hdfs-site, mapred-site-xml files to the one which you want to add this data node.(conf files of new data node should be same as the name node conf xml file properties)

2)in hdfs-site-xml include the following property

**<property>**

**<name>dfs.hosts.include</name>**

**<value>/usr/lib/hadoop/conf/include</value>**

**</property>**

3) in mapred-site-xml include the following property

**<property>**

**<name>mapred.hosts.include</name>**

**<value>/usr/lib/hadoop/conf/include</value>**

**</property>**

4)include ip address of namenode and jobtracker( as the task tracker should also be comissioned) in /etc/hosts file in the new datanode.

-->> start the datanode service.

5)refresh the name node with the following command

**sudo -u hdfs hadoop dfsadmin -refreshNodes**

**sudo -u mapred hadoop mradmin –refreshNodes**

check the UI of namenode and jobtracker....it shows the updated no of datanodes.

**Decommissioning of data node :**

-->> in namenode server, create an exclude file ie (# vi exclude ) in the path you give in **dfs.hosts.exclude** in hdfs-site-xml and mapred-site-xml, the exclude file should contain ip address of the datanode you want to remove.

2)in hdfs-site-xml include the following property

**<property>**

**<name>dfs.hosts.exclude</name>**

**<value>/usr/lib/hadoop/conf/exclude</value>**

**</property>**

3) in mapred-site-xml include the following property

**<property>**

**<name>mapred.hosts.exclude</name>**

**<value>/usr/lib/hadoop/conf/exclude</value>**

**</property>**

4)refresh the name node with the following command

**sudo -u hdfs hadoop dfsadmin -refreshNodes**

**sudo -u mapred hadoop mradmin –refreshNodes**

check the UI of namenode and jobtracker ....it shows the updated no of datanodes.

upgrading from cdh3 to cdh4:

first put the name node into safe mode:

#sudo -u hdfs hadoop dfsadmin -safemode enter

Perform a saveNamespace operation:

#sudo -u hdfs hadoop dfsadmin -saveNamespace

1. This will result in a new fsimage being written out with no edit log entries.
2. With the NameNode still in safe mode, shut down all services as instructed below.

**Make sure the Hadoop services are shut down across your entire cluster by**

for x in /etc/init.d/hadoop-\* ; do sudo $x stop ; done

**Check each host to make sure that there are no processes running as the hdfs or mapred users from root:**

ps -aef | grep java

**Back up the HDFS Metadata**

cd /hadoop/hdfs/name

# tar -cvf /root/nn\_backup\_data.tar .

**this will store a copy of name directory into /root/nn\_backup\_data.tar**

## Copy the Hadoop Configuration files

## cd ~

## mkdir conf

cp /etc/hadoop/conf/\* ~/conf/

cp /etc/hadoop/conf.empty/log4j.properties ~

**remove hadoop -0.20**

#sudo yum remove hadoop-0.20

**remove cdh3 repo**

#sudo yum remove cdh3-repository-1.0-1.noarch.rpm

download cdh4 repo

#wget http://archive.cloudera.com/cdh4/one-click-install/redhat/6/x86\_64/cloudera-cdh-4-0.x86\_64.rpm

**remove cdh3-repository-1.0-1.noarch**

#sudo yum remove cdh3-repository-1.0-1.noarch

**install cdh4 repo**

# sudo yum --nogpgcheck localinstall cloudera-cdh-4-0.x86\_64.rpm

**install demons/ services**

namenode

sudo yum install hadoop-hdfs-namenode

**copy conf files and log4j files back to their directores**

cp -rf conf/\* /etc/hadoop/conf/

cp -rf log4j.properties /etc/hadoop/conf/

cp -rf log4j.properties /etc/hadoop/conf.empty/

sudo service hadoop-hdfs-namenode upgrade

jobtracker

sudo yum install hadoop-0.20-mapreduce-jobtracker

secondary name node

sudo yum install hadoop-hdfs-secondarynamenode

data and task tracker

sudo yum install hadoop-0.20-mapreduce-tasktracker hadoop-hdfs-datanode

**from namenode server ( as these files are not present in jobtracket datanodes and task tracker)**

scp -r /etc/hadoop/conf/\* root@jt.server2.com:/etc/hadoop/conf/

scp -r /etc/hadoop/conf/\* root@dn1.server3.com:/etc/hadoop/conf/

scp -r /etc/hadoop/conf/\* root@dn2.server4.com:/etc/hadoop/conf/

Start up the DataNodes:

sudo service hadoop-hdfs-datanode start

Wait for NameNode to exit safe mode, and then start the Secondary NameNode

1. To check that the NameNode has exited safe mode, look for messages in the log file, or the NameNode's web interface, that say "...no longer in safe mode."
2. To start the Secondary NameNode , enter the following command on the Secondary NameNode host:

**sudo service hadoop-hdfs-secondarynamenode start**

**create tmp directory in hdfs (if not already present )**

sudo -u hdfs hadoop fs -mkdir /tmp

sudo -u hdfs hadoop fs -chmod -R 1777 /tmp

**start the tasktracker**

sudo service hadoop-0.20-mapreduce-tasktracker start

**start the job tracker**

sudo service hadoop-0.20-mapreduce-jobtracker start

**finalize the upgrade**

sudo -u hdfs hdfs dfsadmin -finalizeUpgrade

**restart the data nodes**

sudo service hadoop-hdfs-datanode restart

**upgrade is done!!**