NAME- KUNAL JAIN

ROLL- 2203022

CLASS-LY IS1

AIM: Installation of Single Node Hadoop Cluster on Ubuntu

THEORY: Hadoop is an open source framework based on Java that manages the storage and processing of large amounts of data for applications. Hadoop uses distributed storage and parallel processing to handle big data and analytics jobs, breaking workloads down into smaller workloads that can be run at the same time.

1.Installation of Hadoop:

Hadoop software can be installed in three modes of operation:

Stand Alone Mode: Hadoop is a distributed software and is designed to run on a commodity of machines. However, we can install it on a single node in stand-alone mode .In this mode, Hadoop software runs as a single monolithic java process. This mode is extremely useful for debugging purpose. You can first test run your Map-Reduce application in this mode on small data, before actually executing it on cluster with big data.

2.Pseudo Distributed Mode: In this mode also, Hadoop software is installed on a Single Node .Various daemons of Hadoop will run on the same machine as separate java processes.

Hence all the daemons namely Name Node, Data Node, Secondary Name Node, Job Tracker, Task Tracker run on single machine.

Hadoop Installation: Ubuntu Operating System in stand-alone mode

1. Install Java

sudo apt-get update

sudo apt-get install default-jdk

java -version

2. Disable IPv6

sudo apt-get install vim

sudo vim /etc/sysctl.conf

# disable ipv6

net.ipv6.conf.all.disable\_ipv6 = 1

net.ipv6.conf.default.disable\_ipv6 = 1

net.ipv6.conf.lo.disable\_ipv6 = 1

cat /proc/sys/net/ipv6/conf/all/disable\_ipv6 ... (should return zero)

3. Adding a dedicated Hadoop User

sudo addgroup hadoop

sudo adduser --ingroup hadoop hduser

4. Install SSH

sudo apt-get install ssh

5. Give hduser Sudo Permission

sudo adduser hduser sudo

6. Setup SSH Certificates

su hduser

ssh-keygen -t rsa -P ""

cat $HOME/.ssh/id\_rsa.pub >> $HOME/.ssh/authorized\_keys

ssh localhost

7. Install Hadoop

su hduser

wget http://mirrors.sonic.net/apache/hadoop/common/hadoop-2.6.0/hadoop-2.6.0.tar.gz

tar xvzf hadoop-2.6.0.tar.gz

cd hadoop-2.6.0

sudo mkdir /usr/local/hadoop

sudo mv \* /usr/local/hadoop

8. Set up the Configuration files

a. vim ~/.bashrc

#HADOOP VARIABLES START

export JAVA\_HOME=/usr/lib/jvm/java-7-openjdk-amd64

export HADOOP\_INSTALL=/usr/local/hadoop

export PATH=$PATH:$HADOOP\_INSTALL/bin

export PATH=$PATH:$HADOOP\_INSTALL/sbin

export HADOOP\_MAPRED\_HOME=$HADOOP\_INSTALL

export HADOOP\_COMMON\_HOME=$HADOOP\_INSTALL

export HADOOP\_HDFS\_HOME=$HADOOP\_INSTALL

export YARN\_HOME=$HADOOP\_INSTALL

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_INSTALL/lib/native

export HADOOP\_OPTS="-Djava.library.path=$HADOOP\_INSTALL/lib"

#HADOOP VARIABLES END

b. vim /usr/local/hadoop/etc/hadoop/hadoop-env.sh

export JAVA\_HOME=/usr/lib/jvm/java-7-openjdk-amd64

c. sudo mkdir -p /app/hadoop/tmp

sudo chown hduser:hadoop /app/hadoop/tmp

vim /usr/local/hadoop/etc/hadoop/core-site.xml

<property>

<name>hadoop.tmp.dir</name>

<value>/app/hadoop/tmp</value>

<description>A base for other temporary directories.</description>

</property>

<property>

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

<value>hdfs://localhost:54310</value>

<description>The name of the default file system. A URI whose

scheme and authority determine the FileSystem implementation. The

uri's scheme determines the config property (fs.SCHEME.impl) naming

the FileSystem implementation class. The uri's authority is used to

determine the host, port, etc. for a filesystem.</description>

</property>

d. cp /usr/local/hadoop/etc/hadoop/mapred-site.xml.template /usr/local/hadoop/etc/hadoop/mapred-site.xml

vim /usr/local/hadoop/etc/hadoop/mapred-site.xml

<property>

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

<value>localhost:54311</value>

<description>The host and port that the MapReduce job tracker runs

at. If "local", then jobs are run in-process as a single map

and reduce task.

</description>

</property>

e. sudo mkdir -p /usr/local/hadoop\_store/hdfs/namenode

sudo mkdir -p /usr/local/hadoop\_store/hdfs/datanode

sudo chown -R hduser:hadoop /usr/local/hadoop\_store

vim /usr/local/hadoop/etc/hadoop/hdfs-site.xml

<property>

<name>dfs.replication</name>

<value>1</value>

<description>Default block replication.

The actual number of replications can be specified when the file is created.

The default is used if replication is not specified in create time.

</description>

</property>

<property>

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

<value>file:/usr/local/hadoop\_store/hdfs/namenode</value>

</property>

<property>

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

<value>file:/usr/local/hadoop\_store/hdfs/datanode</value>

</property>

9. Format Hadoop filesystem

hadoop namenode -format

10. Starting Hadoop

su hduser

sudo chown -R hduser:hadoop /usr/local/hadoop/

cd /usr/local/hadoop/sbin

start-all.sh

11. Testing if it is working

Jps

netstat -plten | grep java

http://localhost:50070/

12. Stopping Hadoop

stop-all.sh