2. Installation of Hadoop 3.x on Ubuntu

2.1. Java 8 installation

Hadoop requires working java installation. Let us start with steps for installing java 8

b. Add Repository

**sudo add-apt-repository ppa:webupd8team/java**

c. Update the source list

**sudo apt-get update**

d. Install Java 8

**sudo apt-get install oracle-java8-installer**

e. Check if java is correctly installed

**java -version**

2.2. Configure SSH

SSH is used for remote login. SSH is required in Hadoop to manage its nodes, i.e. remote machines and local machine if you want to use Hadoop on it. Let us now see SSH installation of Hadoop 3.x on Ubuntu:

a. Installation of passwordless SSH

**sudo apt-get install ssh**

**sudo apt-get install pdsh**

b. Generate Key Pairs

**ssh-keygen -t rsa -P '' -f ~/.ssh/id\_rsa**

c. Configure passwordless ssh

**cat ~/.ssh/id\_rsa.pub>>~/.ssh/authorized\_keys**

e. Change the permission of file that contains the key

**chmod 0600 ~/.ssh/authorized\_keys**

f. check ssh to the localhost

**ssh localhost**

2.3. Install Hadoop

a. Download Hadoop

<http://redrockdigimark.com/apachemirror/hadoop/common/hadoop-3.0.0-alpha2/hadoop-3.0.0-alpha2.tar.gz>

(Download the latest version of Hadoop hadoop-3.0.0-alpha2.tar.gz)

b. Untar Tarball

**tar -xzf hadoop-3.0.0-alpha2.tar.gz**

2.4. Hadoop Setup Configuration

a. Edit .Bashrc : Open .bashrc

**nano ~/.bashrc**

Go to end to the file and add following lines :

export HADOOP\_PREFIX="/home/saurabh/Downloads/hadoop-3.0.0"

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

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

export HADOOP\_MAPRED\_HOME=${HADOOP\_PREFIX}

export HADOOP\_COMMON\_HOME=${HADOOP\_PREFIX}

export HADOOP\_HDFS\_HOME=${HADOOP\_PREFIX}

export YARN\_HOME=${HADOOP\_PREFIX}

Then run

**source ~/.bashrc**

b. Edit hadoop-env.sh

Edit configuration file hadoop-env.sh (located in HADOOP\_HOME/etc/hadoop) and set JAVA\_HOME:

**SET JAVA\_HOME=/usr/lib/jvm/java-8-oracle/**

c. Edit core-site.xml

Edit configuration file core-site.xml (located in HADOOP\_HOME/etc/hadoop) and add following entries:

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

<property>

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

<value>/home/saurabh/Downloads/hdata</value>

</property>

</configuration>

d. Edit hdfs-site.xml

Edit configuration file hdfs-site.xml (located in HADOOP\_HOME/etc/hadoop) and add following entries:

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

</configuration>

e. Edit mapred-site.xml

Edit configuration file mapred-site.xml (located in HADOOP\_HOME/etc/hadoop) and add following entries:

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

</configuration>

f. Yarn-site.xml

Edit configuration file mapred-site.xml (located in HADOOP\_HOME/etc/hadoop) and add following entries:

<configuration>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

</configuration>

2.5. How to Start the Hadoop services

Let us now see how to start the Hadoop cluster:

The first step to starting up your Hadoop installation is formatting the Hadoop filesystem which is implemented on top of the local filesystem of your “cluster”. This is done as follows:

a. Format the namenode

**bin/hdfs namenode -format**

NOTE: This activity should be done once when you install Hadoop and not for running Hadoop filesystem, else it will delete all your data from HDFS

b. Start HDFS Services

**sbin/start-dfs.sh**

It will give an error at the time of start HDFS services then use:

**echo "ssh" | sudo tee /etc/pdsh/rcmd\_default**

c. Start YARN Services

**sbin/start-yarn.sh**

d. Check how many daemons are running

Let us now see whether expected Hadoop processes are running or not:

**jps**

2961 ResourceManager

2482 DataNode

3077 NodeManager

2366 NameNode

2686 SecondaryNameNode

3199 Jps

2.6. How to Stop the Hadoop services

Let us learn how to stop Hadoop services now:

a. Stop YARN services

**sbin/stop-yarn.sh**

b. Stop HDFS services

**sbin/stop-dfs.sh**

Note:

Browse the web interface for the NameNode; by default, it is available at:

NameNode – <http://localhost:9870/>

Browse the web interface for the ResourceManager; by default, it is available at:

ResourceManager – <http://localhost:8088/>

if it gives ERROR in **sbin/start-dfs.sh Then** :

pdsh -q -w localhost

export PDSH\_RCMD\_TYPE=ssh

nano ~/.bashrc

source ~/.bashrc