Setting up Hadoop made Easy

# Word of Motivation

Hadoop installation is the most complex step when you start out to learn Hadoop, especially when you are new to Linux as well. I have given it my best to make it as simple as possible. At some point of time it may test you, please be patient and just check the new error on the internet or just check with me incase, if you still find trouble. I try to update the document with the new error and better ways to do the same thing.

Just in case, if you fail to get the right advice on internet and are stuck for long (2 days or more), please contact me. I would help you out.

# Version details

Following are the details of components used, all license free:

1. Apache Hadoop 2.x.x
2. Ubuntu (32 bits) 14.04 (running on virtual Machine)
3. Windows 8. (The same thing can be done on mac, i.e., install a virtual machine on mac and follow the below procedure). Any windows machine would do well.

# Basic Idea in a Nutshell

Following are the steps that would be taken in a nutshell:

1. Install virtual machine on windows or OS.
2. Install Ubuntu on the virtual machine.
3. Download and untar Hadoop package on Ubuntu.
4. Download and install Java on Ubuntu. (Hadoop is written completely in Java).
5. Tell Ubuntu where the Hadoop installation has been done.
6. Tell Hadoop where Java installation has been done.
7. Update the following for your setup:
   1. Core-site.xml -> to set default Schema and authority.
   2. Hdfs-site.xml -> to set def.replication to 1 rather than the default three, otherwise all the blocks would always be alarmed with under replication.
   3. Mapred-site.xml -> To let know of host and port pair where the Jobtrackers runs at.
8. Format the name node and you are ready.

# Installing Ubuntu

## Step 1: Download and Install Virtual layer

You can install either VMBox or VMWare. Both have free version. Free version of VMware is more user friendly but VMBox has more functional capabilities than VMWare.

So if you are new, please go for VMWare and skip the first two videos of the course as they deal with VMBox.

## Step 2: Download and Install Ubuntu on Virtual layer

Download 32 bit Ubuntu OS ISO file and point the ISO to boot up the virtual Machine.

# Installing Java

The following commands downloads and installs jdk in /usr/lib/jvm and jre can be found inside jdk folder at path something like /usr/lib/jvm/java\*/jre

Type/Copy/Paste: sudo apt-get install openjdk-7-jdk

Type/Copy/Paste: java -version

java version "1.7.0\_25"

OpenJDK Runtime Environment (IcedTea 2.3.12) (7u25-2.3.12-4ubuntu3)

OpenJDK 64-Bit Server VM (build 23.7-b01, mixed mode)

# SSH and create a new username

Install OpenSSH

Type/Copy/Paste: sudo apt-get install openssh-server

Create account: hduser and Group: hadoop

Type/Copy/Paste: sudo addgroup hadoop

Type/Copy/Paste: sudo adduser --ingroup hadoop hduser

Type/Copy/Paste: sudo adduser hduser sudo

After user is created, re-login into ubuntu using hduser

Setup SSH Certificate

Type/Copy/Paste: ssh-keygen -t rsa -P ''

...

Your identification has been saved in /home/hduser/.ssh/id\_rsa.

Your public key has been saved in /home/hduser/.ssh/id\_rsa.pub.

...

Type/Copy/Paste: cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys

Type/Copy/Paste: ssh localhost

Welcome to Ubuntu 14.04 LTS (GNU/Linux 3.13.0­24­

generic i686)

\* Documentation: https://help.ubuntu.com/

206 packages can be updated.

33 updates are security updates.

Last login: Fri Jun 13 18:26:50 2014 from localhost

# Download Untar Hadoop 2.2.0

Download hadoop.

Open the browser and search Download Hadoop. You should be able to find link to Apache Hadoop as the very first.

Untar hadoop and change the ownership

Type/Copy/Paste: cd downloads

Type/Copy/Paste: sudo tar vxzf hadoop-2.2.0.tar.gz -C /usr/local

Type/Copy/Paste: cd /usr/local

Type/Copy/Paste: sudo mv hadoop-2.2.0 hadoop

Type/Copy/Paste: sudo chown -R hduser:hadoop hadoop

# Setup Hadoop Environment Variables

## hduser@a-VirtualBox:~$ gedit ~/.bashrc

## Edit bashrc

Paste the following at the end of file.

#Hadoop variables

export JAVA\_HOME=/usr/lib/jvm/java-7-openjdk-i386/

export JAVA\_HOME=/usr/lib/jvm/jdk/

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

###end of paste

## hadoop-env.sh

Change JAVA\_HOME variable

export JAVA\_HOME=/usr/lib/jvm/jdk/

## Refresh and check hadoop script is accessible.

Restart the terminal to refresh the changes.

Type/Copy/Paste: hadoop version

At this point, hadoop is installed.

# Configure Hadoop

## core-site.xml

Place the following property between the <configuration> tags

<property>

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

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

</property>

## yarn-site.xml

Place the following property between the <configuration> tags

<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>

## mapred-site.xml

1. Copy mapred-site.xml.template mapred-site.xml
2. Paste the following between the <configuration> tags

<property>

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

<value>yarn</value>

</property>

hdfs-site.xml

Paste following between <configuration> tag

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

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

<value>file:/home/hduser/mydata/hdfs/namenode</value>

</property>

<property>

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

<value>file:/home/hduser/mydata/hdfs/datanode</value>

</property>

## Format Namenode

hduser@ubuntu40:~$ hdfs namenode -format

# Start Hadoop Service

Type/Copy/Paste: start-dfs.sh

....

Type/Copy/Paste: start-yarn.sh

....

Type/Copy/Paste: jps

If everything is successful, you should see following services running

2583 DataNode

2970 ResourceManager

3461 Jps

3177 NodeManager

2361 NameNode

2840 SecondaryNameNode

Taste

<http://localhost:50070/>

# Run an example

Type/Copy/Paste: hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.2.0.jar pi 2 5

# Stopping services

Type/Copy/Paste: stop-dfs.sh

....

Type/Copy/Paste: stop-yarn.sh

....