Hadoop Linux - Pseudo Distribution

Download Ubuntu 16: <https://www.ubuntu.com/download/desktop>

Installing Hadoop Step by Step:

<http://www.bogotobogo.com/Hadoop/BigData_hadoop_Install_on_ubuntu_16_04_single_node_cluster.php>

**Note**: Above link is missing YARN configuration

Refer following link for YARN configuration

<https://www.ibm.com/developerworks/community/blogs/d9a07ec3-11e2-467d-b758-6861c4cb1d44/entry/How_to_install_Hadoop_2_7_0_in_ubuntu_16_04?lang=en>

**Commands as following:**

1. Install Java

*sudo apt-get update*

*sudo apt-get install default-jdk  
 java -version*

1. Install ssh if doesn’t exist

*sudo apt-get install ssh*

*which ssh*

1. Generate rsa key pair (answer to the key name , only Enter)

*ssh-keygen -t rsa -P ""*

1. Copy rsa Public key in authorized key file

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

1. Test if your ssh is working fine

*ssh localhost*

1. Download Hadoop binaries and extract it

*wget http://apache.mirror.rafal.ca/hadoop/common/hadoop-2.6.5/hadoop-2.6.5.tar.gz*

*tar -xvzf hadoop-2.6.5.tar.gz*

1. Move it to User Local folder (You can keep it anywhere)

*sudo mv hadoop-2.6.5 /usr/local/hadoop*

1. Change the ownership of hadoop folder if need be (ID:GROUP can be something like Ali:Home or whatever)

*sudo chown -R <ID:GROUP> /usr/local/Hadoop*

1. Install vim to be able to edit with vi

*Sudo apt-get install vim*

1. Modify .bashrc file and add following variables in it

*#HADOOP VARIABLES START*

*export JAVA\_HOME=/usr/lib/jvm/java-8-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/native"*

*export HADOOP\_CONF\_DIR=$HADOOP\_INSTALL/etc/hadoop*

*#HADOOP VARIABLES END*

1. Apply .bashrc file

*source ~/.bashrc*

1. Create directory structure for hadoop

*sudo mkdir -p /app/hadoop/tmp*

*sudo chown <USER:GROUP> /app/hadoop/tmp*

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

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

*sudo chown -R <USER:GROUP> /usr/local/hadoop\_store*

1. Modify certain files specific to hadoop configurations (All files are at location **/usr/local/hadoop/etc/hadoop/**):
   1. hadoop-env.sh

*export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64*

* 1. core-site.xml

*<configuration>*

*<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>*

*</configuration>*

* 1. hdfs-site.xml

*<configuration>*

*<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>*

*</configuration>*

* 1. yarn-site.xml

*<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>*

1. Format namenode

*hadoop namenode -format*

1. Start all services

*cd /usr/local/hadoop/sbin*

*start-dfs.sh*

*start-yarn.sh*

1. Check if all services are up and running

*jps*

Hadoop Windows - Pseudo Distribution

1. Extract hadoop-2.6.2 file on your local system where you want to install Hadoop
2. Copy java JDK in C: drive if already not there
3. Go to location **hadoop-2.6.2/etc/hadoop** 
   1. Modify core-site.xml file to update hadoop.tmp.dir property

Create temp directory somewhere and use that as value to mentioned property

* 1. Modify Hadoop-env.cmd to update JAVA\_HOME

Should be the location of your JDK in C: directory

* 1. Hdfs-site.xml:

Create directories /hadoop-2.6.2/data/namenode and /hadoop-2.6.2/data/datanode

Update dfs.namenode.name.dir and dfs.datanode.name.dir properties accordingly

* 1. yarn-site.xml

Update yarn.nodemanager.log-dirs property

1. set HADOOP\_HOME environment variable to bin directory within hadoop-2.6.2 folder
2. open windows command prompt cmd and execute following commands
   1. hdfs namenode –format
   2. change directory to hadoop-2.6.2/sbin
   3. start-dfs.cmd
   4. start-yarn.cmd

# Cloudera Quick Start VM:

<https://www.cloudera.com/downloads/quickstart_vms/5-12.html>

Select VMWare as platform