**SOFTWARE DOWNLOAD LINKS:**

Hadoop- <http://archive.cloudera.com/cdh5/cdh/5/hadoop-2.5.0-cdh5.3.2.tar.gz>

Hive - <http://archive.cloudera.com/cdh5/cdh/5/hive-0.13.1-cdh5.3.2.tar.gz>

Flume - <http://archive.cloudera.com/cdh5/cdh/5/flume-ng-1.5.0-cdh5.3.2.tar.gz>

**Why Linux?**

* Hadoop is built on the top of Linux.
* GNU/Linux is supported as a development and production platform.
* Hadoop has been demonstrated on GNU/Linux clusters with 2000 nodes.

**Why Java?**

* Java is a pre-requisite to Hadoop.
* Java should be installed in both master and slave nodes.
* Make sure to check java installation by executing the command

$java -version

This document describes how to install and configure a single-node Hadoop cluster. Once the installation is done you can perform Hadoop Distributed File System (HDFS) and Hadoop Map-Reduce operations.

Pre-requisite: Ubuntu and java installation

**Step 1: Configure SSH**

**Install Open SSH Server-Client**

$sudo apt-get install openssh-server openssh-client

**Generate Key Pairs**

$ssh-keygen -t rsa -P ""

**Configure password-less SSH**

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

**Check by SSH to localhost**

$ssh localhost

**Step2: Install Hadoop**

**Untar Tar ball**

$tar xzf hadoop-2.5.0-cdh5.3.2.tar.gz

Note: Hadoop can be downloaded from the link mentioned at the top.

**Step3: Edit Configuration**

**Edit .bashrc:**

Edit .bashrc file located in user’s home directory and add following parameters:

export HADOOP\_PREFIX="/home/previn/hadoop-2.5.0-cdh5.3.2"

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}

Note: After above step restart the terminal, so that all the environment variables will come into effect

**Edit hadoop-env.sh:**

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

export JAVA\_HOME=/<path-to-the-root-of-your-Java-installation>

(eg: /usr/lib/jvm/java-8-oracle/)

**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/previn/hdata</value>

</property>

</configuration>

Note: /home/previn/hdata is a sample location; please specify a location where you have Read Write privileges

**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>

**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>

**Edit 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>

**Step4: Start the cluster**

**Format the name node:**

$bin/hdfs namenode -format

NOTE: This activity should be done once when you install hadoop, else it will delete all your data from HDFS

**Start HDFS Services:**

$sbin/start-dfs.sh

**Start YARN Services:**

$sbin/start-yarn.sh

**Check whether services have been started**

$jps

Make sure you are seeing all 5 daemons running

NameNode, DataNode, ResourceManager, NodeManager, SecondaryNameNode

**Step5: Stop the cluster**

**Stop HDFS Services:**

$sbin/stop-dfs.sh

**Stop YARN Services:**

$sbin/stop-yarn.sh