Install VMWare tools

sudo add-apt-repository ppa:openjdk-r/ppa

sudo apt-get install openjdk-8-jdk

Sudo apt-get update

sudo apt-get upgrade

sudo apt-get install default-jdk

sudo apt-get install build-essential checkinstall sudo apt-get install libreadline-gplv2-dev libncursesw5-dev libssl-dev libsqlite3-dev tk-dev libgdbm-dev libc6-dev libbz2-dev

sudo update-java-alternatives -s java-1.8.0-openjdk-amd64

sudo apt-get install python-pip

sudo addgroup hadoop

sudo usermod -a -G hadoop oxclo

sudo apt-get install openssh-server

To disable IPv6 on Ubuntu 10.04 LTS, open /etc/sysctl.conf in the editor of your choice and add the following lines to the end of the file:

/etc/sysctl.conf

|  |  |
| --- | --- |
| 1 2 3 4 | # disable ipv6 net.ipv6.conf.all.disable\_ipv6 = 1 net.ipv6.conf.default.disable\_ipv6 = 1 net.ipv6.conf.lo.disable\_ipv6 = 1 |

|  |
| --- |
| cd /usr/local $ sudo tar xzf hadoop-1.0.3.tar.gz $ sudo mv hadoop-1.0.3 hadoop $ sudo chown -R hduser:hadoop hadoop |

sudo mv hadoop-2.7.1/ /usr/local/hadoop

[sudo] password for oxclo:

oxclo@oxclo:~$ sudo chown -R hduser:hadoop /usr/local/hadoop/

export HADOOP\_HOME=/usr/local/hadoop

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

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

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"

**hduser@laptop:~$ sudo mkdir -p /app/hadoop/tmp hduser@laptop:~$ sudo chown hduser:hadoop /app/hadoop/tmp**

**hduser@laptop:~$ vi /usr/local/hadoop/etc/hadoop/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>**

**hduser@laptop:~$ cp /usr/local/hadoop/etc/hadoop/mapred-site.xml.template /usr/local/hadoop/etc/hadoop/mapred-site.xml**

The **mapred-site.xml** file is used to specify which framework is being used for MapReduce.  
We need to enter the following content in between the <configuration></configuration> tag:

**<configuration> <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> </configuration>**

hdfs namenode –format

> ssh-keygen -t dsa -P '' -f ~/.ssh/id\_dsa

cat ~/.ssh/id\_dsa.pub >> ~/.ssh/authorized\_keys

sudo su hduser

cd /usr/local/hadoop/sbin

start-dfs

hadoop fs -mkdir /user/

**hadoop fs -mkdir /user/hduser**

**hadoop fs –put localfile /user/hduser/remotefile**

**hadoop fs –cat /user/hduser/remotefile**

**Copy data over**

**Install Git**

**Need to run sbt/sbt assembly for spark as its slowwwww**

Sbt/sbt -Pyarn -Phadoop-2.x assembly

**Sudo apt-get install r-base**

**echo "deb http://debian.datastax.com/community stable main" | sudo tee -a /etc/apt/sources.list.d/cassandra.sources.list^**

**curl -L http://debian.datastax.com/debian/repo\_key | sudo apt-key add -**

**sudo apt-get install dsc22 cassandra-tools**

**Get Spark 1.5.1 tgz**

**Unzip**

**build/sbt -Pyarn -Phadoop-2.6 assembly**

**Hive?!**

**sudo apt-get install python-dev**

**sudo pip install numpy**