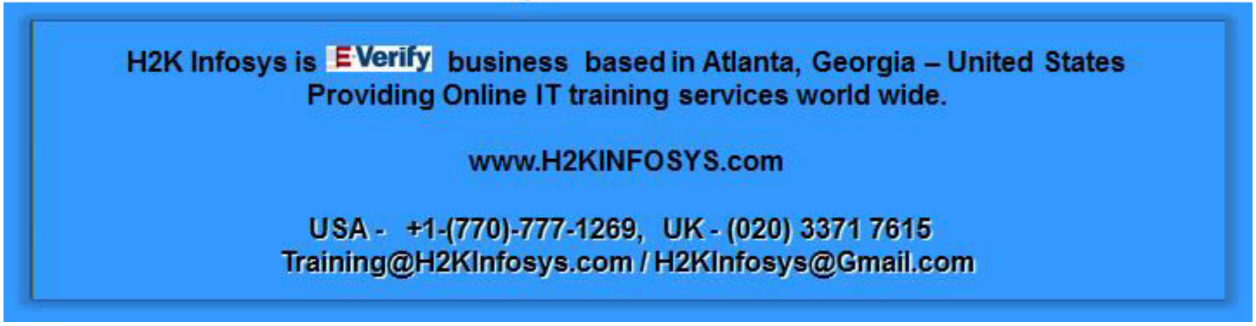
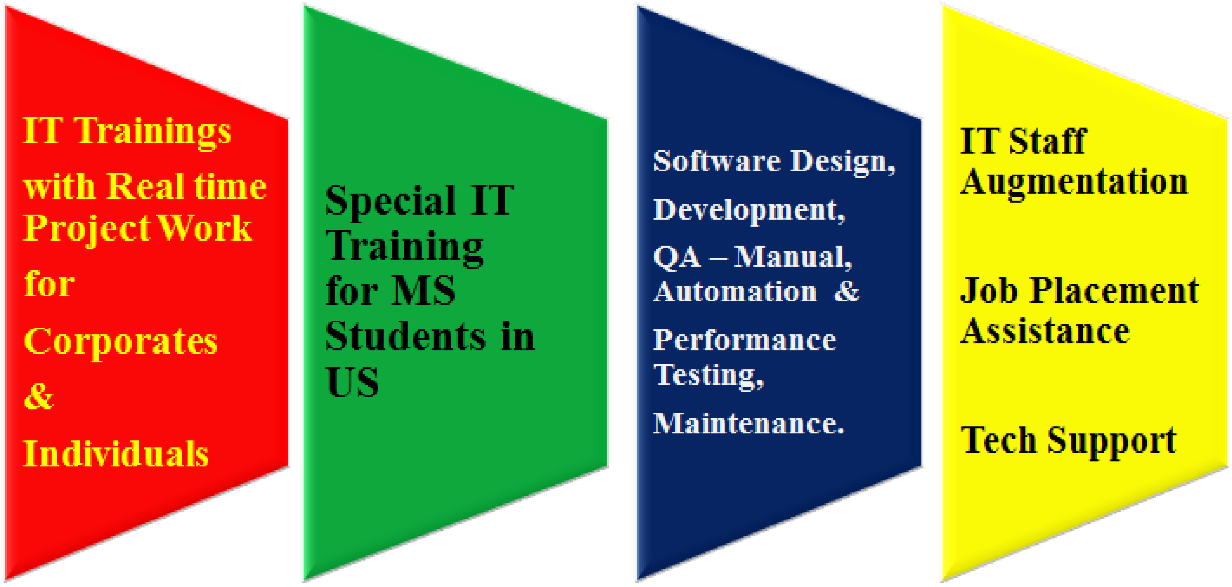


H2K Infosys is E-Verify business based in Atlanta, Georgia – United States
Providing Online IT training services world wide.

www.H2KINFOSYS.com

USA - +1-(770)-777-1269, UK - (020) 3371 7615
Training@H2KInfosys.com / H2KInfosys@Gmail.com

H2K INFOSYS PROVIDES WORLD CLASS SERVICES IN



IT Trainings with Real time Project Work for Corporates & Individuals

Special IT Training for MS Students in US

Software Design, Development, QA – Manual, Automation & Performance Testing, Maintenance.

IT Staff Augmentation

Job Placement Assistance

Tech Support

Hadoop Installation Guide

Software's that needs to be downloaded:

VMware player: as per your machine configuration (preferably vmware-player-6.0.1 for 64 bit machine)

https://2ra5-downloads.phpnuke.org/en/c59571/vmware-player#.VCH-86jrZMt

Ubuntu - (12.04 if it's a 64 bit machine (or) go to Ubuntu download site of select which suits your machine) (Note: Link is not shared as they keep on organizing the archives of Ubuntu)

Hadoop – hadoop-1.0.3 version from apache mirrors

http://archive.apache.org/dist/hadoop/core/

STEP 1: Installing the JDK and JRE

Run the following commands sequentially

Clean up the historical open jdk:

```
    sudo apt-get purge openjdk*
```

To install jdk and jre

```
    sudo apt-get install python-software-properties
    sudo add-apt-repository ppa:webupd8team/java
    sudo apt-get update
    sudo apt-get install oracle-java6-installer
```

Check Installation

`java –version` - Should print the Java version

For ex: java version "1.6.0_45"

```
        Java(TM) SE Runtime Environment (build 1.6.0_45-b06)
        Java HotSpot(TM) 64-Bit Server VM (build 20.45-b01,
mixed mode)
```

STEP 2 : Adding a dedicated system user

It is better to create a separate Hadoop user, as it will be helpful to isolate Hadoop installations from other services running on the same machine

Create the group first

```
    sudo addgroup hadoop
```

Create the user of your choice

```
    sudo adduser --ingroup hadoop hduser (sample username)
```

In case if you don’t want to create a new user, you can simply skip this step.

STEP 3 : Configuring SSH

SSH is required to perform cluster-wide operations. So, Hadoop requires SSH access to manage its nodes, i.e. remote machines plus your local machine if you want to use Hadoop on it.

For our single-node setup of Hadoop, we therefore need to configure SSH access to `localhost` for the `hduser` user we created in the previous step.

Generate the SSH key for the user hduser

Change the User

```
su – hduser
```

Generate the RSA key pair with an empty password. If password is given then every time Hadoop tries to connect to a node you need to enter the password, so it's better to create without any password. To achieve this we are generating a public key pair which will be shared across the cluster.

```
ssh-keygen -t rsa -P ""
```

when this command is run it promotes for the file to store the key, give a path as ex: /home/lib/.ssh/id_rsa (some valid path)

Enable SSH access

You have to enable SSH access to your local machine with this newly created key.

```
cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
```

This commands adds the public key generated to the authorized keys list file.

Connecting to local machine

The final step is to test the SSH setup by connecting to your local machine with the hduser user. The step is also needed to save your local machines host key fingerprint to the hduser user's known_hosts file.

```
ssh localhost
```

Some common Problems: **when I ran this command I got an error like "connect to host localhost port 22 connection refused"**

Solution 1 : **I removed and reinstalled the ssh on my system and it worked fine. The commands to remove and reinstall are as below respectively**

```
sudo apt-get remove openssh-client openssh-server
sudo apt-get install openssh-client openssh-server
```

Solution 2 : **Check for the installation of both ssh client and ssh server, run the below two commands one after the other and they should display a path to directory**

```
which ssh
which sshd
```

Mostly the second command will not display a directory path, that means ssh server is not installed, to install run the below command

```
    sudo apt-get install openssh-server
```

STEP 4 : Disabling IPv6

____We need to disable IPv6 to avoid Hadoop getting bounded to IPv6 instead of IPv4.

First check if IPv6 is enabled or disabled in the system, command is

```
    cat /proc/sys/net/ipv6/conf/all/disable_ipv6
```

A return value of 0 means IPv6 is enabled, a value of 1 means disabled.

In case enabled, please disable it by adding below line conf/hadoop-env.sh after STEP 5, use the below commands

```
    sudo echo export HADOOP_OPTS=-Djava.net.preferIPv4Stack=true >> hadoop-env.sh
```

STEP 5 : Download Hadoop distribution

This step would require us create new directories, to create a directory the newly created user(hduser) should have sudo permissions, to add the user to sudo user run the below command

```
    sudo adduser hduser sudo
```

Now create a directory “hadoop-inst” in usr

```
    cd /usr
    sudo mkdir hadoop-inst
```

sudo has to be used while executing any command outside the users home directory

Copy the dowloaded tar file of hadoop distribution to the /usr/hadoop-inst directory

```
cd /usr/hadoop-inst
```

Move the hadoop distribution in the Downloads folder to /usr/hadoop-inst directory

```
sudo mv /<path in which hadoop tar file is downloaded>/hadoop-1.0.3.tar.gz
/usr/hadoop-inst
```

Extract the tar file, you should be in usr/hadoop-inst directory

```
sudo tar xzf hadoop-1.0.3.tar.gz
```

Give full permission with the current user to the directory extracted

```
sudo chown -R hduser:hadoop hadoop-1.0.3
```

In the above command `hduser:hadoop` **is** `userName:groupName`

STEP 6 : Update $HOME/.bashrc

open the .bashrc file(which can been in home-view hidden files) in the text editor and add the following at the end of the file

```
# Set Hadoop-related environment variables
export HADOOP_HOME=/usr/hadoop-inst/hadoop-1.0.3

# Set JAVA_HOME (we will also configure JAVA_HOME directly for Hadoop
later on)
export JAVA_HOME=/usr/lib/jvm/java-6-oracle

# Some convenient aliases and functions for running Hadoop-related
commands
unalias fs &> /dev/null
alias fs="hadoop fs"


unalias hls &> /dev/null
alias hls="fs -ls"

# If you have LZO compression enabled in your Hadoop cluster and
# compress job outputs with LZOP (not covered in this tutorial):
# Conveniently inspect an LZOP compressed file from the command
# line; run via:
#
# $ lzohead /hdfs/path/to/lzop/compressed/file.lzo
#
# Requires installed 'lzop' command.
#
lzohead () {
    hadoop fs -cat $1 | lzop -dc | head -1000 | less
}

# Add Hadoop bin/ directory to PATH
export PATH=$PATH:$HADOOP_HOME/bin
```

If Above is not working properly, we can run the individual command like below in the terminal

```
export HADOOP_HOME=/usr/hadoop-inst/hadoop-1.0.3
export JAVA_HOME=/usr/lib/jvm/java-6-oracle
export PATH=$PATH:$HADOOP_HOME/bin
```

Configuration (will be available in /usr/hadoop-inst/hadoop-1.0.3/CONF folder)

STEP 7 : hadoop-env.sh

Change the JAVA_HOME to `/usr/lib/jvm/java-6-oracle` **by opening the file in the text editor**

STEP 8 : conf/*-site.xml

You can leave the settings below “as is” with the exception of the hadoop.tmp.dir parameter – this parameter you must change to a directory of your choice. We will use the directory /app/hadoop/tmp in this tutorial. Hadoop’s default configurations use hadoop.tmp.dir as the base temporary directory both for the local file system and HDFS, so don’t be surprised if you see Hadoop creating the specified directory automatically on HDFS at some later point.

Now we create the directory and set the required ownerships and permissions:

```
sudo mkdir -p /app/hadoop/tmp
sudo chown hduser:hadoop /app/hadoop/tmp
sudo chmod 750 /app/hadoop/tmp
```

Add the following snippets between the <configuration> ... </configuration> tags in the respective configuration XML file.

In file conf/core-site.xml:

```
<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>
```

In file conf/mapred-site.xml:

```
<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>
```

In file conf/hdfs-site.xml:

```
<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>
```

STEP 9 : Format Name node

Format the Name Node by using the below commands

```
     /usr/hadoop-inst/hadoop-1.0.3/bin/hadoop namenode -format
```

(or)

```
          hadoop namenode -format
```

NAME NODE SHOULD NOT BE FORMATTED (TWICE) IT SHOULD BE DONE ONLY ONCE

STEP 10 : Start the pseudo cluster

Start the cluster by using the command

```
bin/start-all.sh
```

The output would be as below

```
hduser@ubuntu:/usr/hadoop-inst/hadoop-1.0.3$ bin/start-all.sh

Warning: $HADOOP_HOME is deprecated.



starting namenode, logging to /usr/hadoop-inst/hadoop-1.0.3/libexec/../logs/hadoop-hduser-namenode-ubuntu.out

localhost: starting datanode, logging to /usr/hadoop-inst/hadoop-1.0.3/libexec/../logs/hadoop-hduser-datanode-ubuntu.out

localhost: starting secondarynamenode, logging to /usr/hadoop-inst/hadoop-1.0.3/libexec/../logs/hadoop-hduser-secondarynamenode-ubuntu.out

starting jobtracker, logging to /usr/hadoop-inst/hadoop-1.0.3/libexec/../logs/hadoop-hduser-jobtracker-ubuntu.out

localhost: starting tasktracker, logging to /usr/hadoop-inst/hadoop-1.0.3/libexec/../logs/hadoop-hduser-tasktracker-ubuntu.out

hduser@ubuntu:/usr/hadoop-inst/hadoop-1.0.3$
```

After this type jps in the terminal and enter

```
hduser@ubuntu:/usr/hadoop-inst/hadoop-1.0.3$ jps

12662 DataNode

13352 Jps

12958 JobTracker

12430 NameNode

12874 SecondaryNameNode

13166 TaskTracker

hduser@ubuntu:/usr/hadoop-inst/hadoop-1.0.3$
```

Stopping the pseudo cluster is as simple as running this job

```
bin/stop-all.sh
```

```
hduser@ubuntu:/usr/hadoop-inst/hadoop-1.0.3$ bin/stop-all.sh

Warning: $HADOOP_HOME is deprecated.



stopping jobtracker

localhost: stopping tasktracker

stopping namenode

localhost: stopping datanode

localhost: stopping secondarynamenode

hduser@ubuntu:/usr/hadoop-inst/hadoop-1.0.3$
```

Once everything is installed properly, install eclipse on your machine

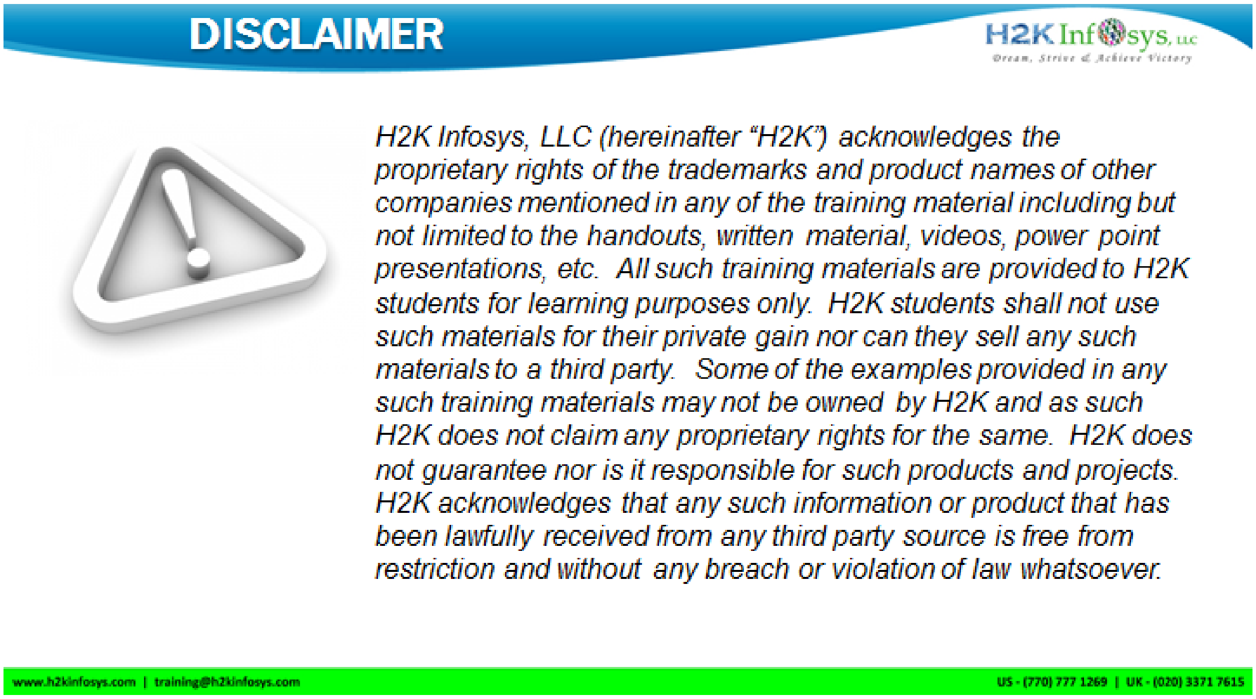
Download Eclipse from the download section of the official website (http://www.eclipse.org/downloads/). Remember to choose the correct package for your architecture (32bit or 64 bit).

The package will have the name:

eclipse-standard-kepler-SR1-linux-gtk-x86_32.tar.gz

or

eclipse-standard-kepler-SR1-linux-gtk-x86_64.tar.gz



DISCLAIMER

H2K Infosys, LLC (hereinafter "H2K") acknowledges the proprietary rights of the trademarks and product names of other companies mentioned in any of the training material including but not limited to the handouts, written material, videos, power point presentations, etc. All such training materials are provided to H2K students for learning purposes only. H2K students shall not use such materials for their private gain nor can they sell any such materials to a third party. Some of the examples provided in any such training materials may not be owned by H2K and as such H2K does not claim any proprietary rights for the same. H2K does not guarantee nor is it responsible for such products and projects. H2K acknowledges that any such information or product that has been lawfully received from any third party source is free from restriction and without any breach or violation of law whatsoever.

www.h2kinfosys.com | training@h2kinfosys.com US - (770) 777 1269 | UK - (020) 3371 7615