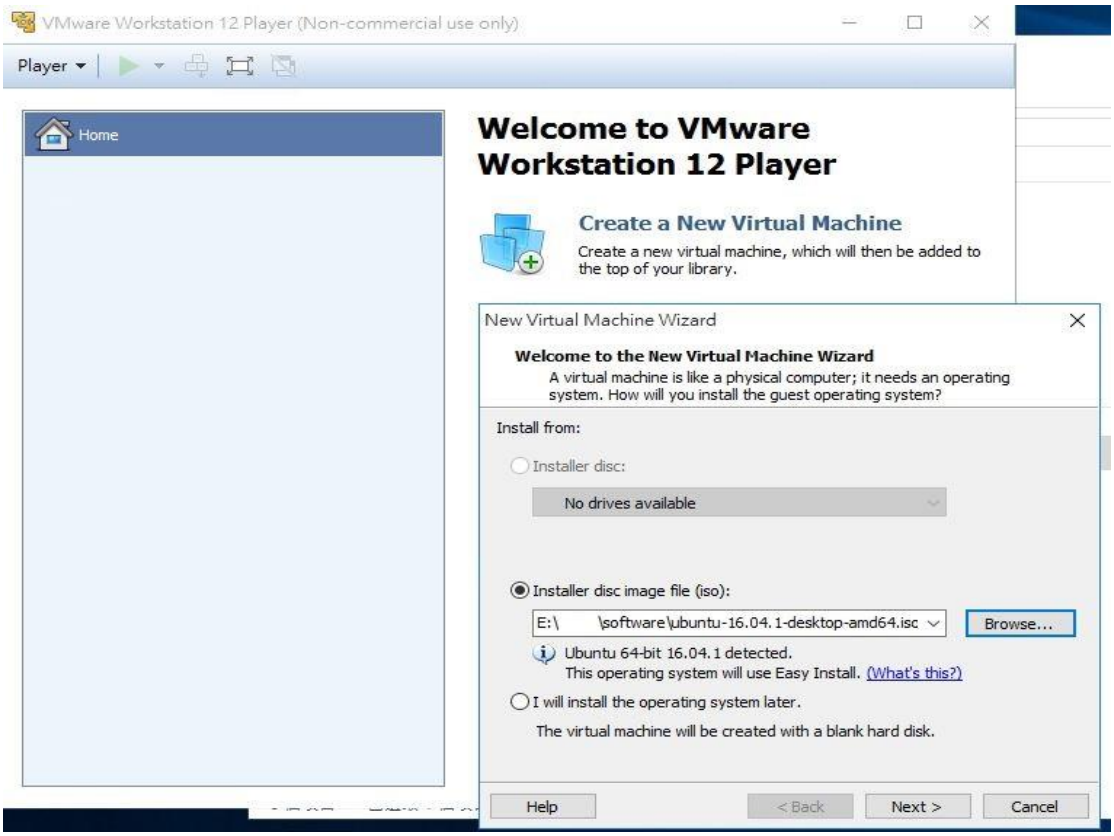
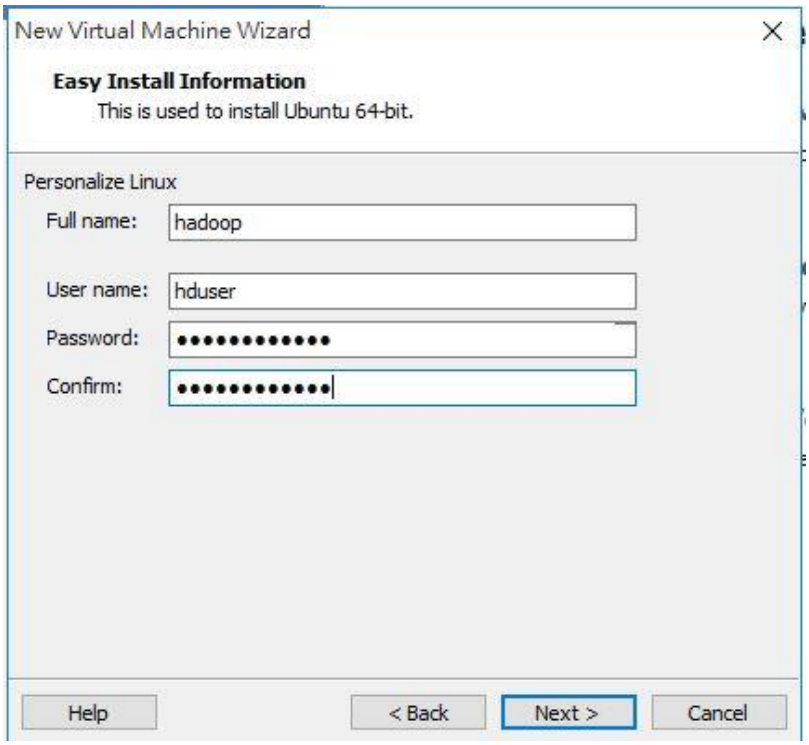
Hadoop Single Node Cluster Setup

1. install Ubuntu
   - 下載 Ubuntu，選擇最新的 64 位元桌面版本

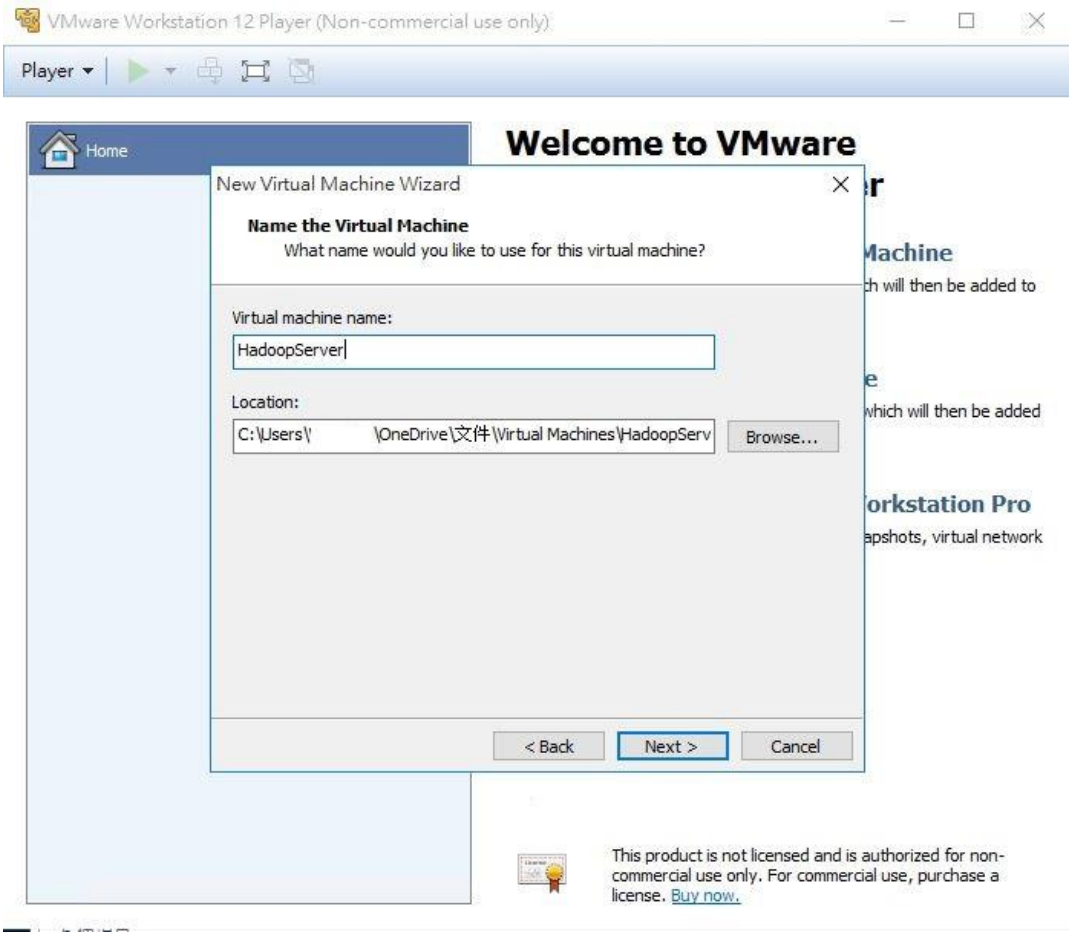
   - create a new VM



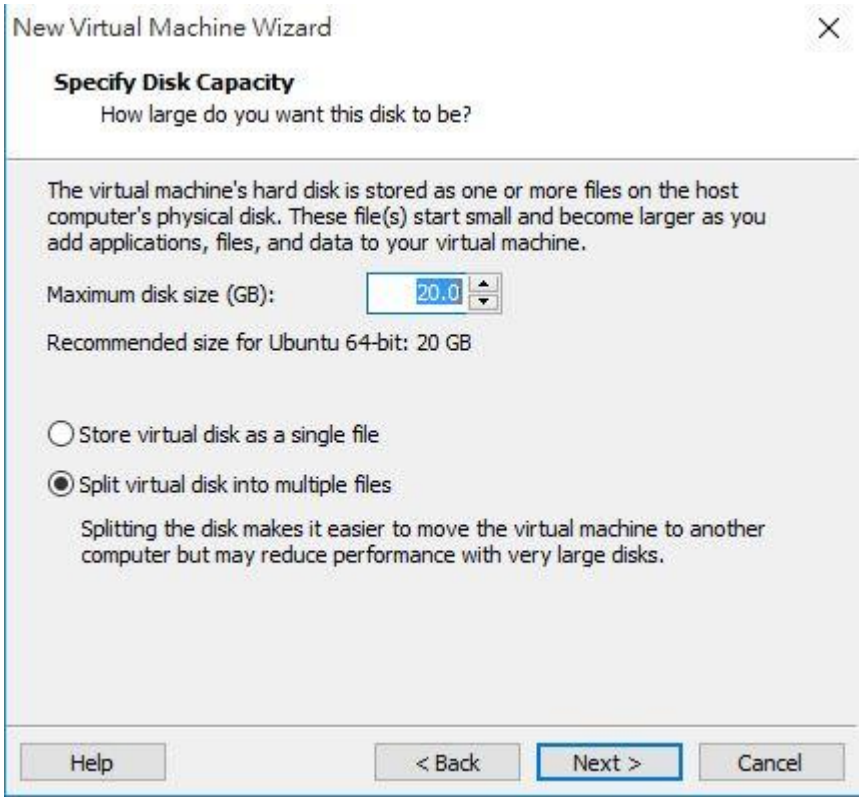
- 輸入管理員帳密➔hduser / hduser

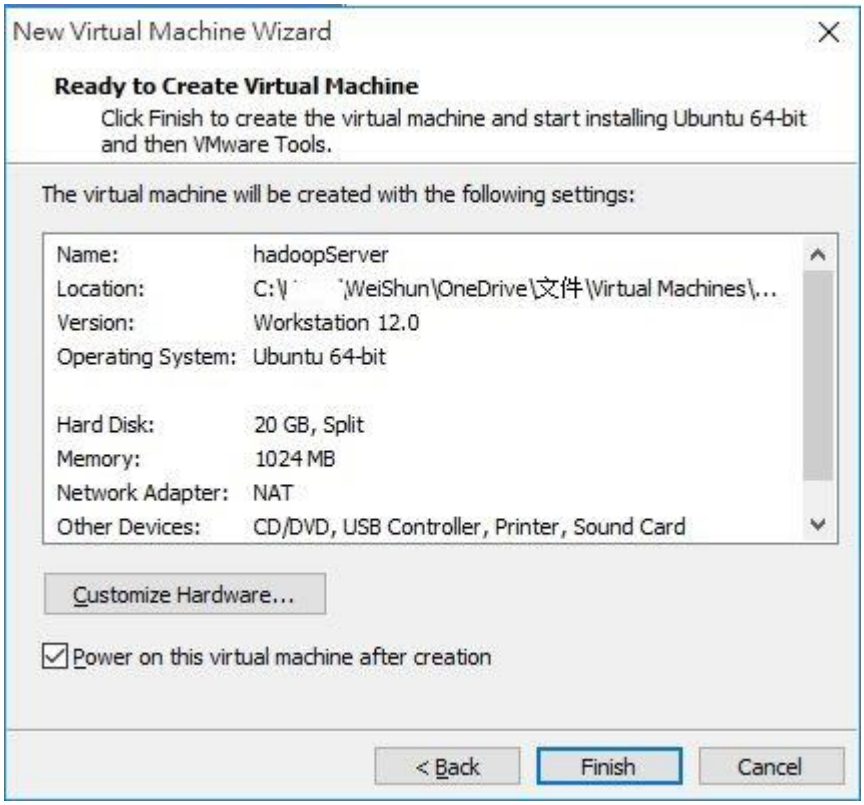


- 輸入伺服器名稱

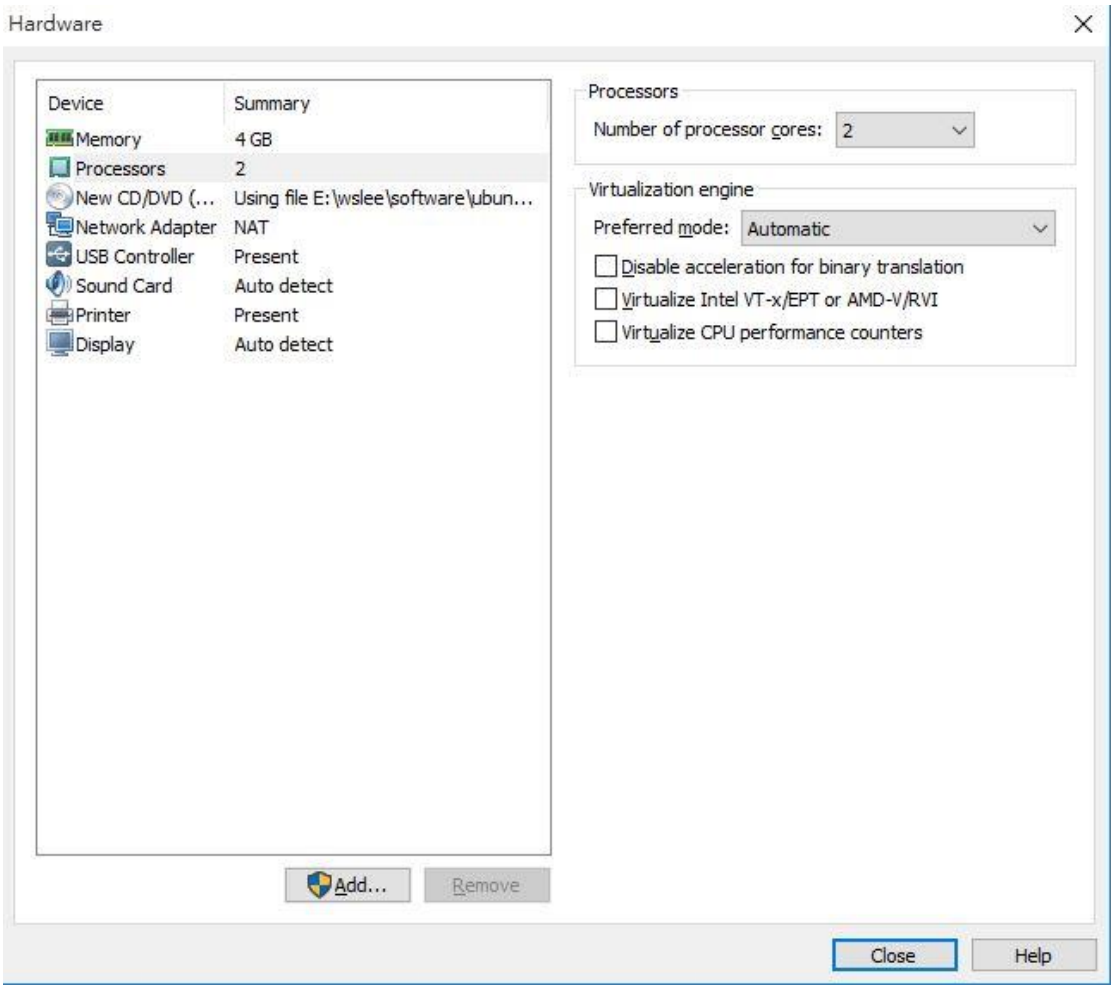


- 一個 VM 分配 20GB

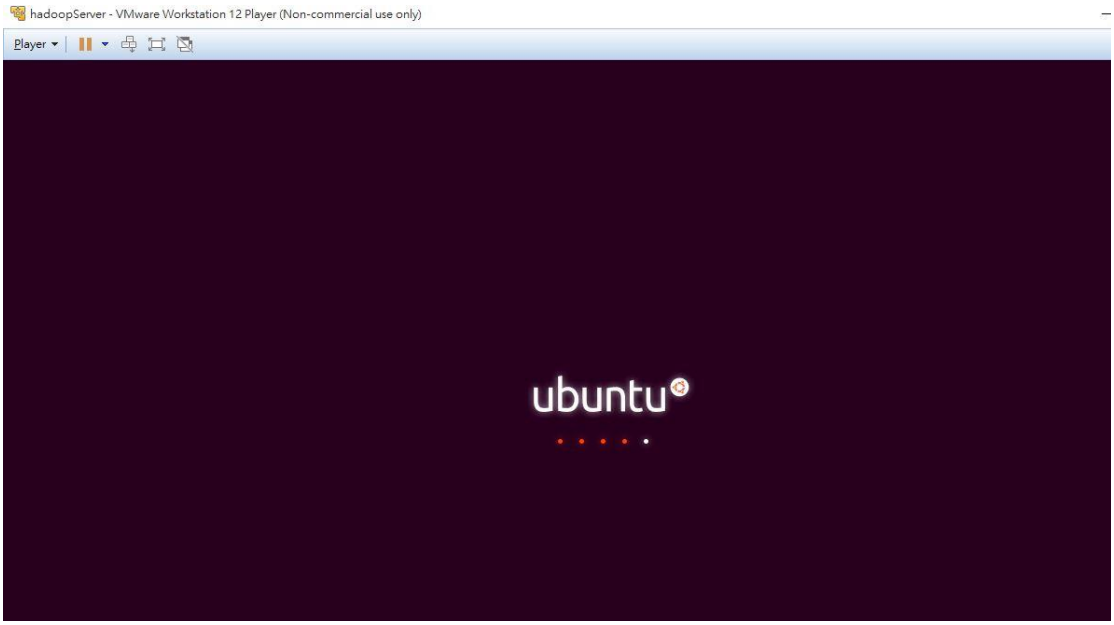

New Virtual Machine Wizard
Specify Disk Capacity
How large do you want this disk to be?
The virtual machine's hard disk is stored as one or more files on the host computer's physical disk. These file(s) start small and become larger as you add applications, files, and data to your virtual machine.
Maximum disk size (GB):
20.0
Recommended size for Ubuntu 64-bit: 20 GB
Store virtual disk as a single file
Split virtual disk into multiple files
Splitting the disk makes it easier to move the virtual machine to another computer but may reduce performance with very large disks.
Help
< Back
Next >
Cancel

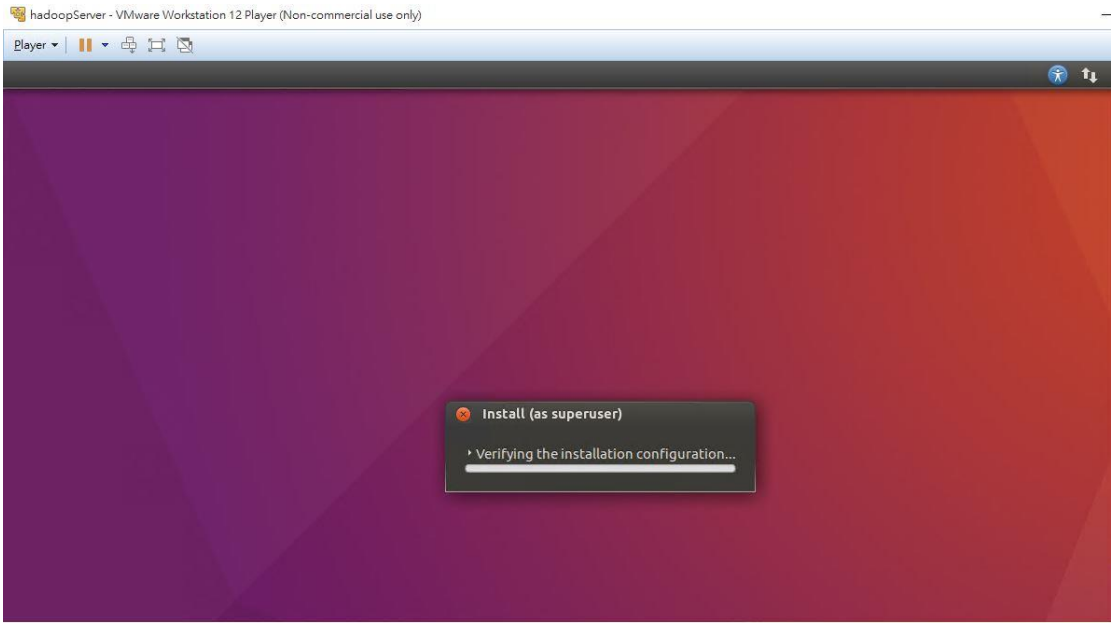

New Virtual Machine Wizard
Ready to Create Virtual Machine
Click Finish to create the virtual machine and start installing Ubuntu 64-bit and then VMware Tools.
The virtual machine will be created with the following settings:
Name: hadoopServer
Location: C:\ \WeiShun\OneDrive\文件\Virtual Machines\...
Version: Workstation 12.0
Operating System: Ubuntu 64-bit
Hard Disk: 20 GB, Split
Memory: 1024 MB
Network Adapter: NAT
Other Devices: CD/DVD, USB Controller, Printer, Sound Card
Customize Hardware...
Power on this virtual machine after creation
< Back
Finish
Cancel

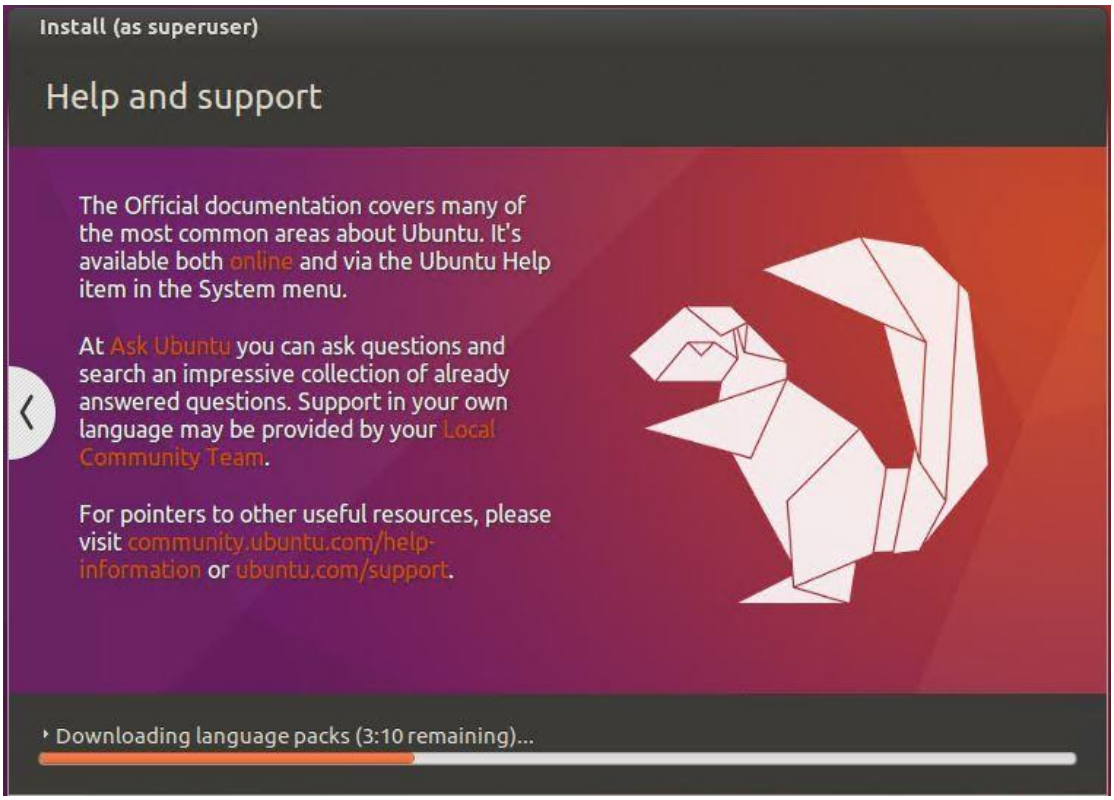
- 選擇 2 processors 和 4GB RAM

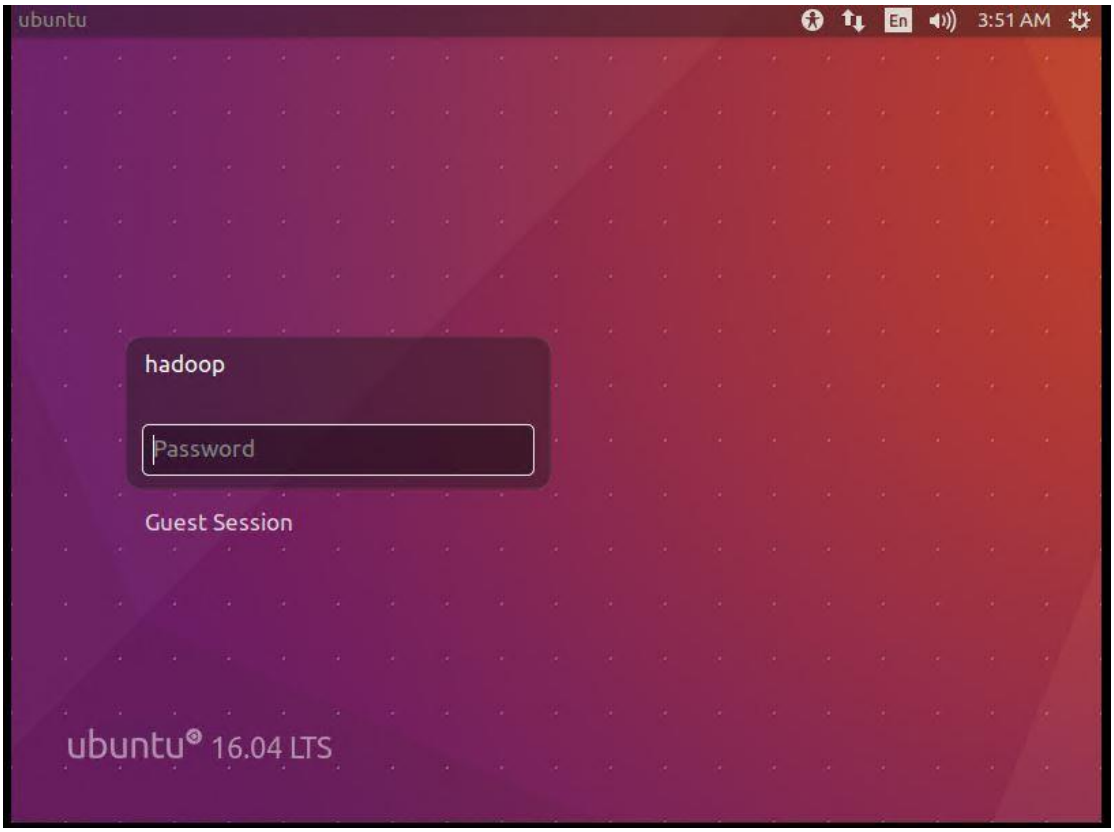


- 安裝中~

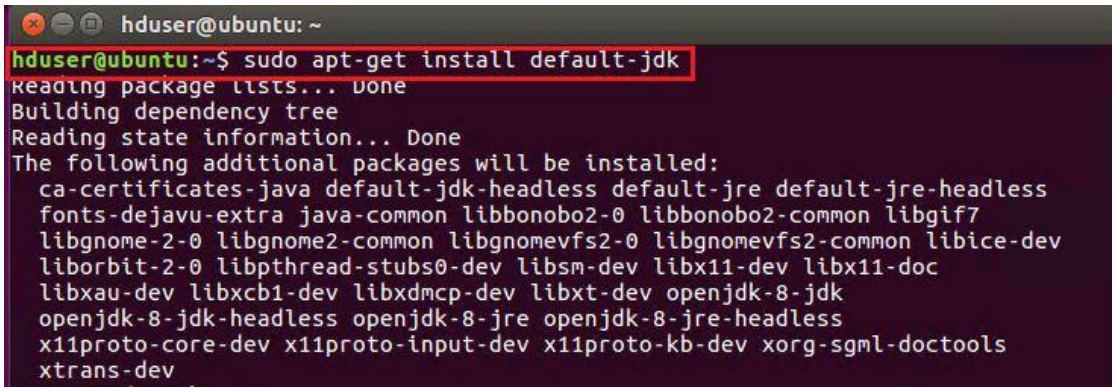



hadoopServer - VMware Workstation 12 Player (Non-commercial use only)
Player
Install (as superuser)
Verifying the installation configuration...

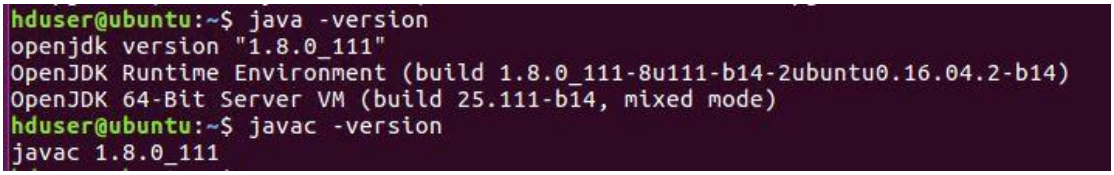

Install (as superuser)
Help and support
The Official documentation covers many of the most common areas about Ubuntu. It's available both online and via the Ubuntu Help item in the System menu.
At Ask Ubuntu you can ask questions and search an impressive collection of already answered questions. Support in your own language may be provided by your Local Community Team.
For pointers to other useful resources, please visit community.ubuntu.com/help-information or ubuntu.com/support.
Downloading language packs (3:10 remaining)...


ubuntu
En
3:51 AM
hadoop
Password
Guest Session
ubuntu® 16.04 LTS

2. install Java JDK
   - 開啟一個 terminal，輸入 sudo apt-get install default-jdk，讓 apt-get 去抓最新的 JDK

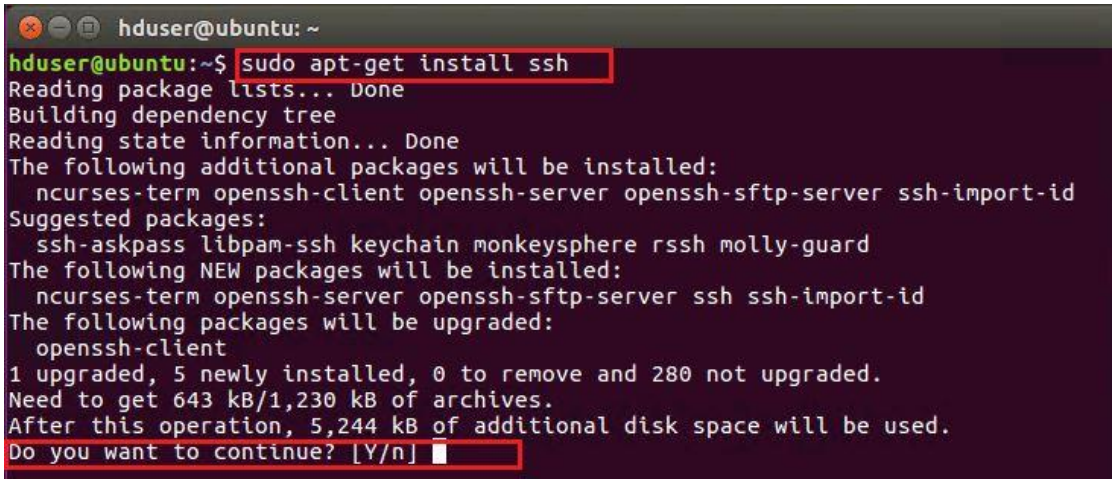


   - 安裝完畢，可試著輸入 java -version 和 javac -version 檢查

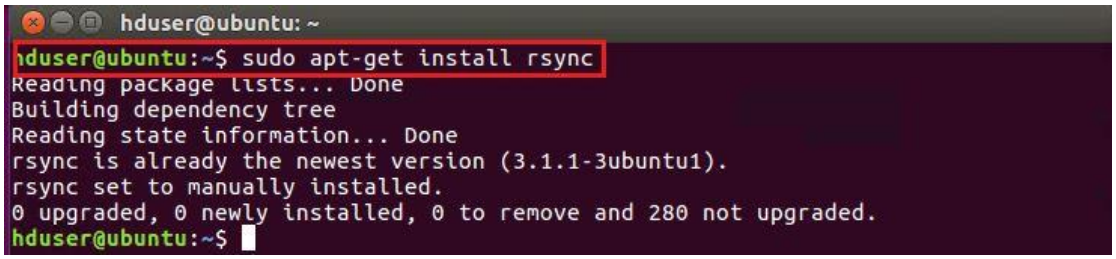


   - 輸入 sudo update-alternatives --display java 取得 JDK 的安裝路徑

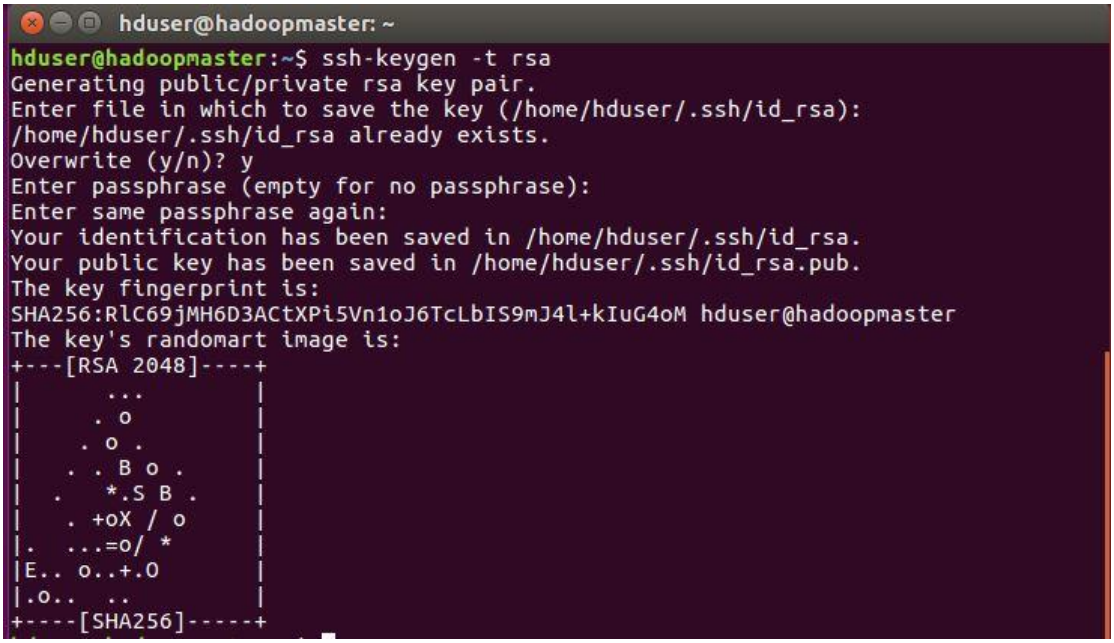
3. Install SSH
   - 之後連線到其他伺服器時可重複輸入密碼
   - 輸入 sudo apt-get install ssh 安裝 SSH



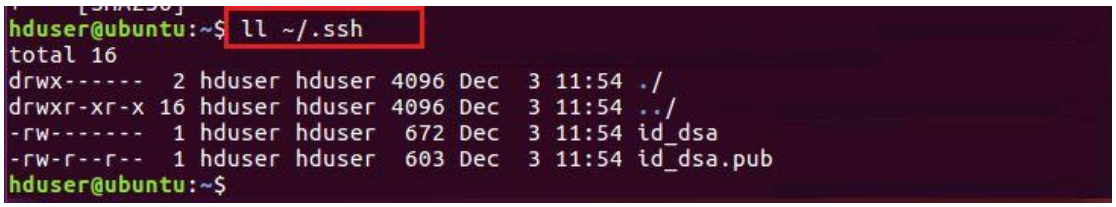
   - 安裝 rsync，輸入 sudo apt-get install rsync



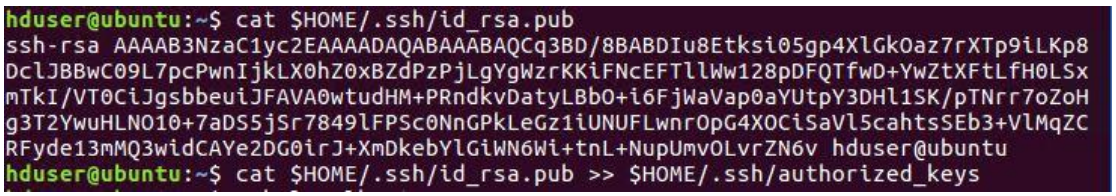
- 輸入 ssh-keygen -t rsa 產生 public/private key pair



- 查看產生的 SSH key，輸入 ll ~/.ssh



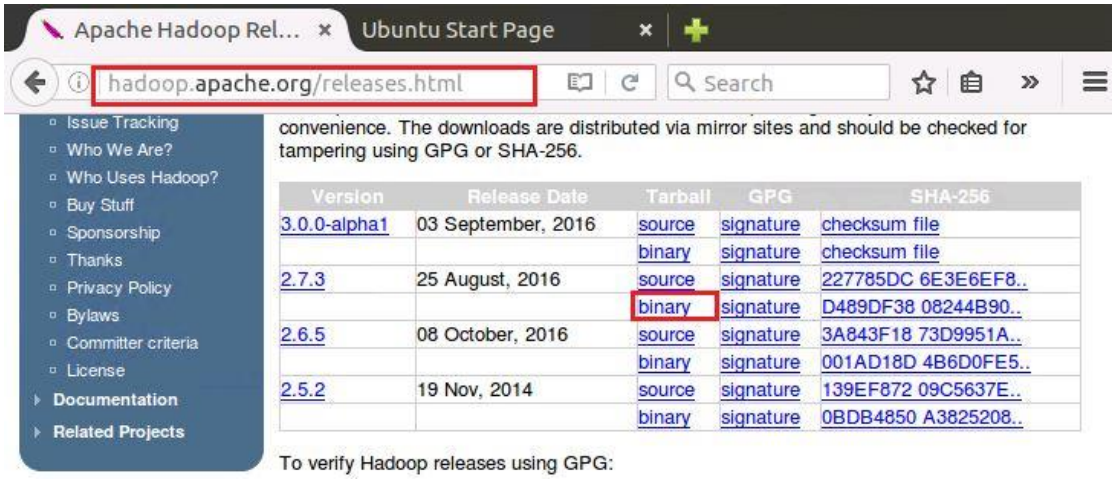
- 輸入 cat $HOME/.ssh/id_rsa.pub >> $HOME /.ssh/authorized_keys 設定 key



- 啟動 SSH ssh localhost

```
hduser@ubuntu:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:wN0apnZoV21KYAfdRRk/+r0WU9BVJKCjH91g/oYvk04.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 16.04.1 LTS (GNU/Linux 4.4.0-31-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

236 packages can be updated.
82 updates are security updates.


The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.
```

4. Install Hadoop
   - 連線至 Hadoop (網址)



   - 下載最新版 Hadoop-2.7.3.tar.gz，輸入 wget http://apache.stu.edu.tw/hadoop/common/hadoop-2.7.3.tar.gz

```
hduser@ubuntu: ~
hduser@ubuntu:~$ wget http://apache.stu.edu.tw/hadoop/common/hadoop-2.7.3/hadoop
-2.7.3.tar.gz
--2016-12-03 15:22:50--  http://apache.stu.edu.tw/hadoop/common/hadoop-2.7.3/had
oop-2.7.3.tar.gz
Resolving apache.stu.edu.tw (apache.stu.edu.tw)... 120.119.118.1, 2001:e10:c41:e
eee::1
Connecting to apache.stu.edu.tw (apache.stu.edu.tw)|120.119.118.1|:80... connect
ed.
HTTP request sent, awaiting response... 200 OK
Length: 214092195 (204M) [application/x-gzip]
Saving to: 'hadoop-2.7.3.tar.gz'

hadoop-2.7.3.tar.gz    0%[                    ] 967.11K   263KB/s    eta 14m 3s
```

- 解壓縮 Hadoop-2.7.3.tar.gz 在 Terminal 中輸 sudo tar -zxvf hadoop-2.7.3.tar.gz

```
hduser@ubuntu:~$ sudo tar -zxvf hadoop-2.7.3.tar.gz
```

- 預設安裝路徑為 /usr/local ，輸入 sudo mv hadoop-2.7.3 /usr/local/hadoop
- 查看 hadoop 安裝目錄，輸入 ll /usr/local/hadoop

```
hduser@ubuntu: ~
hduser@ubuntu:~$ sudo mv hadoop-2.7.3 /usr/local/hadoop
hduser@ubuntu:~$ ll /usr/local/hadoop
total 140
drwxr-xr-x  9 root root  4096 Aug 18 09:49 ./
drwxr-xr-x 11 root root  4096 Dec  3 15:36 ../
drwxr-xr-x  2 root root  4096 Aug 18 09:49 bin/
drwxr-xr-x  3 root root  4096 Aug 18 09:49 etc/
drwxr-xr-x  2 root root  4096 Aug 18 09:49 include/
drwxr-xr-x  3 root root  4096 Aug 18 09:49 lib/
drwxr-xr-x  2 root root  4096 Aug 18 09:49 libexec/
-rw-r--r--  1 root root 84854 Aug 18 09:49 LICENSE.txt
-rw-r--r--  1 root root 14978 Aug 18 09:49 NOTICE.txt
-rw-r--r--  1 root root  1366 Aug 18 09:49 README.txt
drwxr-xr-x  2 root root  4096 Aug 18 09:49 sbin/
drwxr-xr-x  4 root root  4096 Aug 18 09:49 share/
hduser@ubuntu:~$
```

- 設定環境變數，編輯 .bashrc 輸入 sudo gedit ~/.bashrc ，在檔案的最後加入下列字串(注意 JAVA_HOME 為本機的路徑)

```
hduser@ubuntu:~$ sudo gedit ~/.bashrc
[sudo] password for hduser:
```

```
hduser@ubuntu: ~
hduser@ubuntu:~$ sudo gedit ~/.bashrc
Open    *.bashrc    Save
fi

#Hadoop Variables
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib"
export JAVA_LIBRARY_PATH=$HADOOP_HOME/lib/native:$JAVA_LIBRARY_PATH
#Hadoop Variables
```

```
#Hadoop Variables
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib"
export JAVA_LIBRARY_PATH=$HADOOP_HOME/lib/native:$JAVA_LIBRARY_PATH
#Hadoop Variables
```

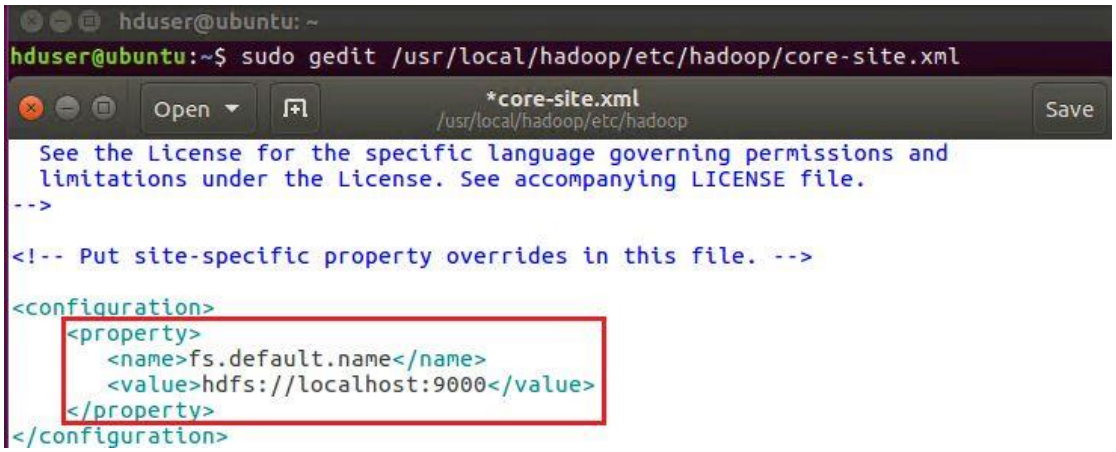
- 讓 .bashrc 生效，輸入 source ~/.bashrc

- 接下來編輯 Hadoop-env.sh，輸入 sudo gedit /usr/local/hadoop/etc/hadoop/hadoop-env.sh
- 在檔案中輸入下列字串(注意 JAVA_HOME 為本機的路徑)

```
export JAVA_HOME=<YOUR JAVA HOME PATH>
```

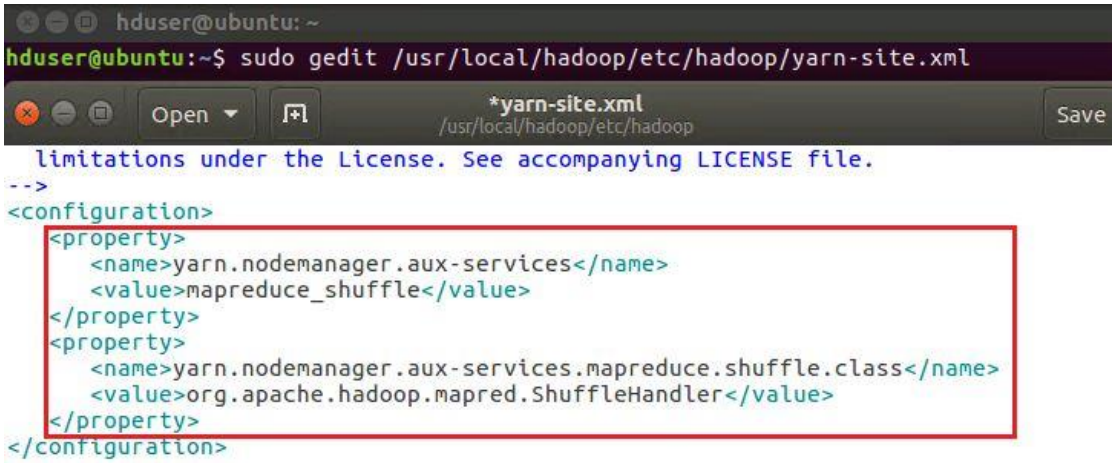
- 編輯 core-site.xml，輸入 sudo gedit /usr/local/hadoop/etc/hadoop/core-site.xml
- 在 xml 檔中，<configuration> tag 間輸入下列字串

```
<property>
   <name>fs.default.name</name>
   <value>hdfs://localhost:9000</value>
</property>
```



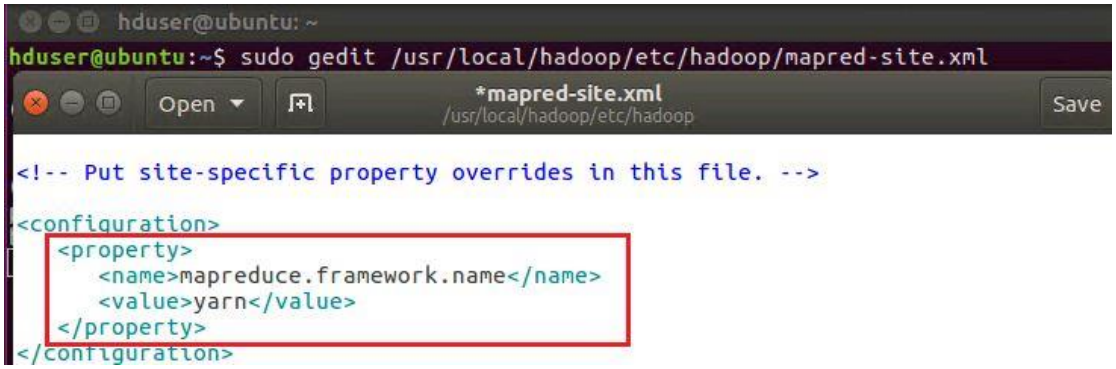
- 編輯 yarn-site.xml，輸入 sudo gedit /usr/local/hadoop/etc/hadoop/yarn-site.xml
- 在 xml 檔中，<configuration> tag 間輸入下列字串

```
<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>
```



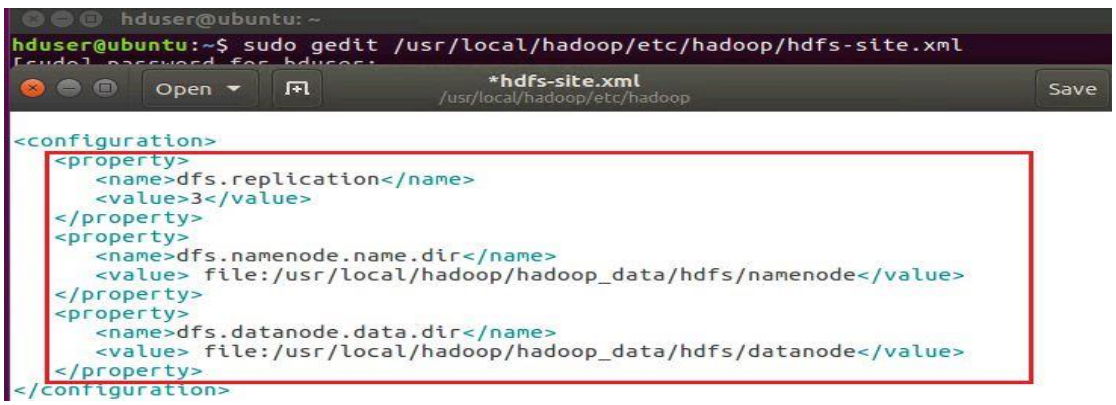
- 編輯 mapred-site.xml，從樣本複製後再編輯，輸入
  i. sudo cp /usr/local/hadoop/etc/hadoop/mapred-site.xml.template /usr/local/hadoop/etc/hadoop/mapred-site.xml
  ii. sudo gedit /usr/local/hadoop/etc/hadoop/mapred-site.xml

```
<property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
</property>
```



- 編輯 hdfs-site.xml，輸入 sudo gedit /usr/local/hadoop/etc/hadoop/hdfs-site.xml
- 在 xml 檔中，<configuration> tag 間輸入下列字串

```
<property>
    <name>dfs.replication</name>
    <value>3</value>
</property>
<property>
    <name>dfs.namenode.name.dir</name>
    <value> file:/usr/local/hadoop/hadoop_data/hdfs/namenode</value>
</property>
<property>
    <name>dfs.datanode.data.dir</name>
    <value> file:/usr/local/hadoop/hadoop_data/hdfs/datanode</value>
 </property>
```



- 建立 namenode，datanode 資料夾目錄，並將擁有者改為 hduser
  i. sudo mkdir -p /usr/local/hadoop/hadoop_data/hdfs/namenode
  ii. sudo mkdir -p /usr/local/hadoop/hadoop_data/hdfs/datanode
  iii. sudo chown hduser:hduser -R /usr/local/hadoop

```
hduser@ubuntu: ~
hduser@ubuntu:~$ sudo mkdir -p /usr/local/hadoop/hadoop_data/hdfs/namenode
hduser@ubuntu:~$ sudo mkdir -p /usr/local/hadoop/hadoop_data/hdfs/datanode
hduser@ubuntu:~$ sudo chown hduser:hduser -R /usr/local/hadoop/
hduser@ubuntu:~$
```

- 格式化 Hadoop file system，輸入 Hadoop namenode -format

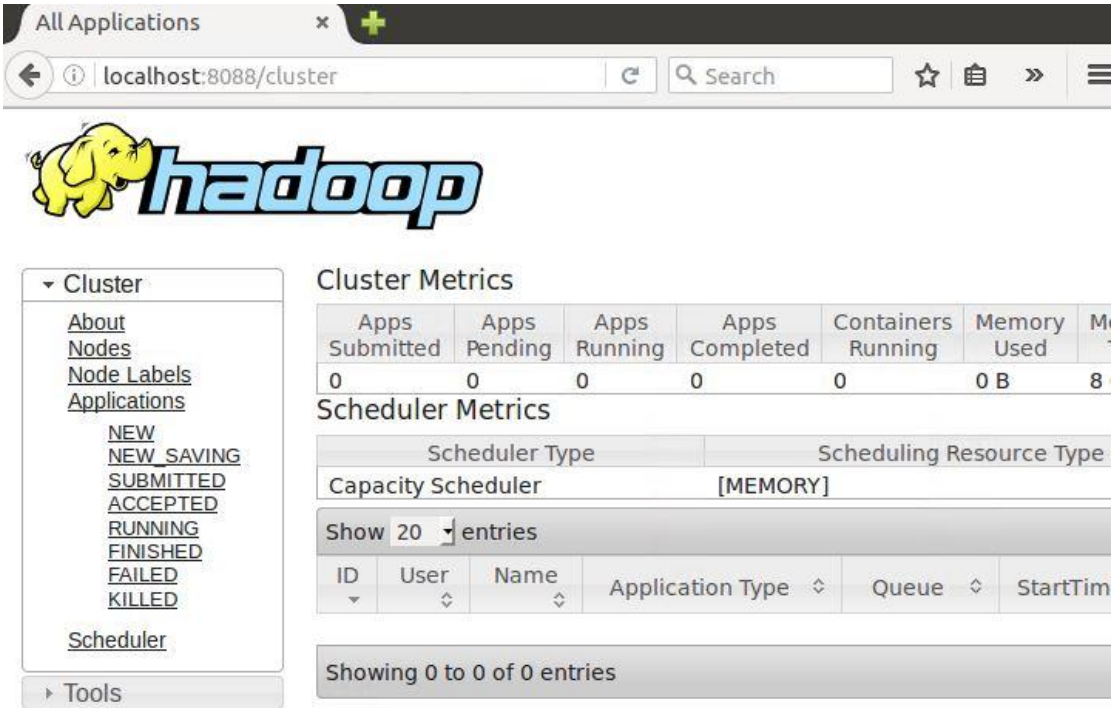
```
hduser@ubuntu: ~
hduser@ubuntu:~$ hadoop namenode -format
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

16/12/03 17:10:41 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = ubuntu/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.7.3
STARTUP_MSG:   classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/
hadoop/common/lib/hadoop-auth-2.7.3.jar:/usr/local/hadoop/share/hadoop/common/li
b/jsch-0.1.42.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-logging-1.1.
```

```
53 bytes saved in 0 seconds.
16/12/03 17:10:47 INFO namenode.NNStorageRetentionManager: Going to retain 1 ima
ges with txid >= 0
16/12/03 17:10:47 INFO util.ExitUtil: Exiting with status 0
16/12/03 17:10:47 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at ubuntu/127.0.1.1
************************************************************/
```

- 輸入 start-all.sh 啟動 Hadoop

```
hduser@ubuntu: ~
hduser@ubuntu:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
hduser@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser-na
menode-ubuntu.out
hduser@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser-da
tanode-ubuntu.out
Starting secondary namenodes [0.0.0.0]
hduser@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hd
user-secondarynamenode-ubuntu.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resource
manager-ubuntu.out
hduser@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduser-n
odemanager-ubuntu.out
hduser@ubuntu:~$
```

- 若有啟動成功，http://localhost:8088 可看到下圖畫面



- http://localhost:50070 檢視 datanode 或 file system