1.Linux 以 apt 管理套件，為了在 Linux 可以安裝最新的套件版本資訊，必須先執行。

sudo apt-get update

2.安裝 SSH

SSH:用來遠端登入其他伺服器使用(sever)，其傳遞資訊維加密過後的資訊。

sudo apt install ssh -y

3.安裝 PDSH

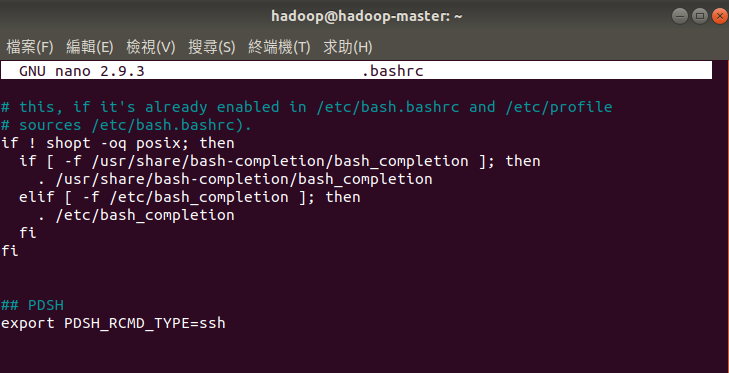
sudo apt install pdsh -y

nano .bashrc

export PDSH\_RCMD\_TYPE=ssh

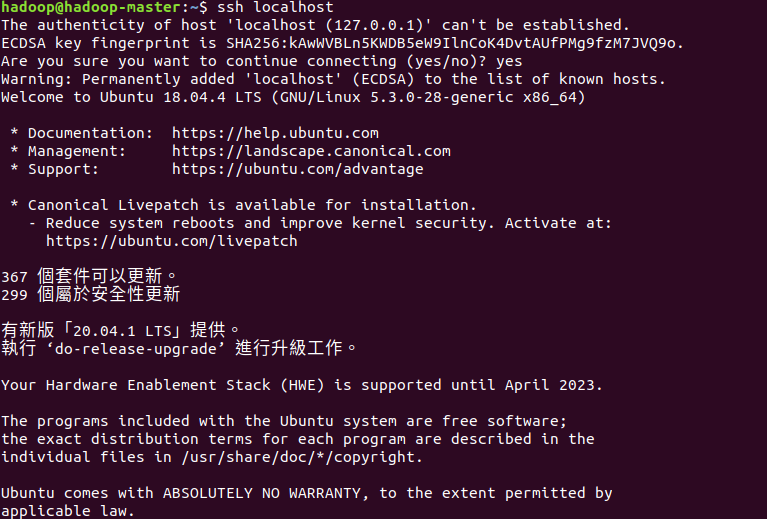
source .bashrc

4. ssh-keygen -t rsa -P "" -f ~/.ssh/id\_rsa



6. 以 SSH 連線到本機

ssh localhost



7. 安裝JAVA

sudo apt install openjdk-8-jdk -y

8. 檢查 JAVA 版本

java -version

9. 下載 Hadoop 3.3.0 binary

Source :未經過 compile 的版本，通常用在 liunx

Binary :已經經過編譯過的版本，需對應相對應的電腦系統。

wget <https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz>

wget <https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz>

10. 解壓縮 Hadoop 3.3.0

sudo tar xzf hadoop-3.3.0.tar.gz

sudo tar xzf hadoop-3.2.1.tar.gz

11.將 Hadoop 移至軟體預設路徑 /usr/local

sudo mv hadoop-3.3.0 /usr/local/hadoop

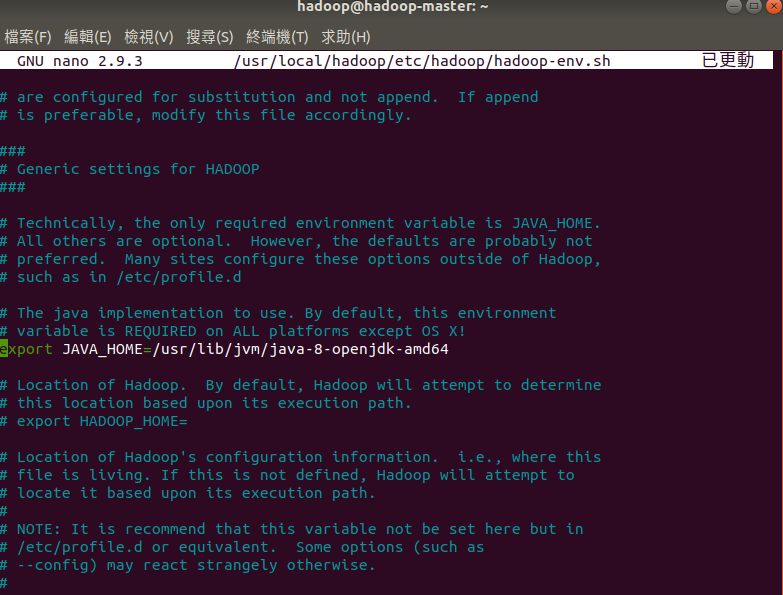
sudo mv hadoop-3.2.1 /usr/local/hadoop

12.更改 Hadoop 設定檔

sudo nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh

在 JAVA\_HOME 中輸入JAVA 的位置

/usr/lib/jvm/java-8-openjdk-amd64



13. sudo nano .bashrc

export PDSH\_RCMD\_TYPE=ssh

#Hadoop Related Options

export HADOOP\_HOME=/usr/local/hadoop

export HADOOP\_INSTALL=$HADOOP\_HOME

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

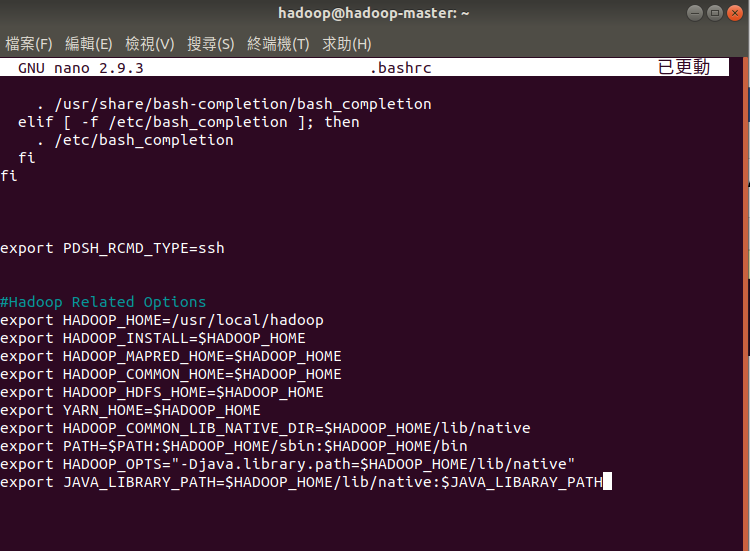
export YARN\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

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

export HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"

export JAVA\_LIBRARY\_PATH=$HADOOP\_HOME/lib/native:$JAVA\_LIBARAY\_PATH



source .bashrc

14. 建立 hadoop 的 namenode 與 datanode

sudo mkdir -p /usr/local/hadoop/hadoop\_data/hdfs/namenode

sudo mkdir -p /usr/local/hadoop/hadoop\_data/hdfs/datanode

15. 新增使用者 hduser

sudo adduser hduser

16.只讓 hduser 只用 Hadoop 的權限

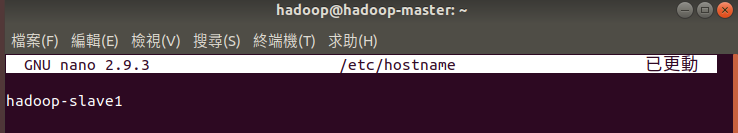
sudo usermod -aG hduser hduser  
sudo chown hduser:root -R /usr/local/hadoop/  
sudo chmod g+rwx -R /usr/local/hadoop/  
sudo adduser hduser sudo

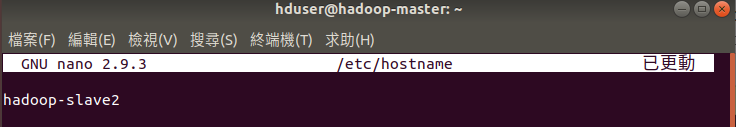
重複 步驟 13~15

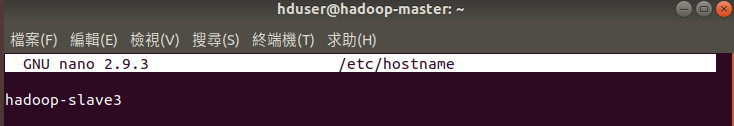
製作 HADOOP CLUSTER

16.針對每個 slave 更改 hostname(主機名稱)

sudo nano /etc/hostname

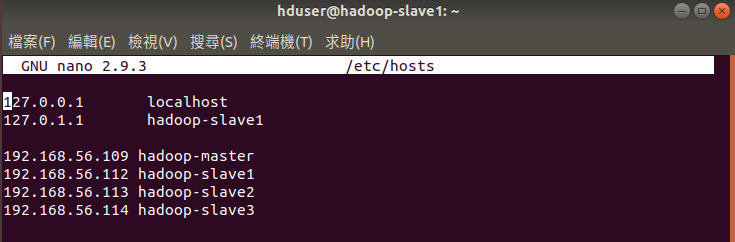






17. 更改 /etc/hosts， 該檔案能將主機名稱對應到相對的IP位置。

sudo nano /etc/hosts

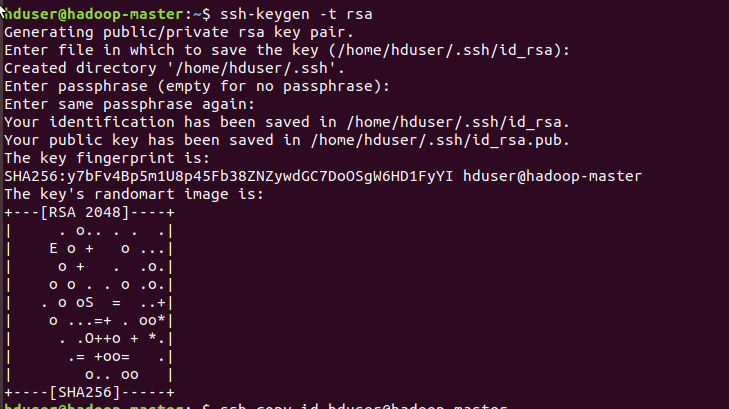


18. 重新開機讓更改的 hostname 生效

sudo reboot

19.產生 ssh 金鑰

ssh-keygen -t rsa



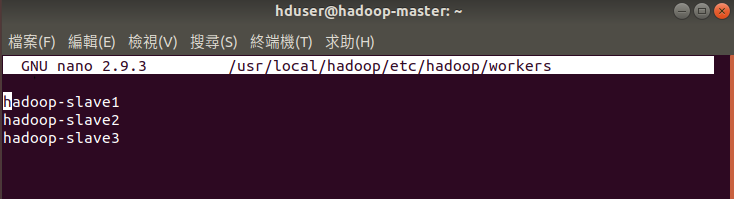
20. 執行 ssh-copy-id 參數為要登入的主機，就可以將 public key 傳送到該主機，並幫忙設定好免密碼登入的信任主機。

ssh-copy-id hduser@hadoop-master  
ssh-copy-id hduser@hadoop-slave1  
ssh-copy-id hduser@hadoop-slave2

ssh-copy-id hduser@hadoop-slave3

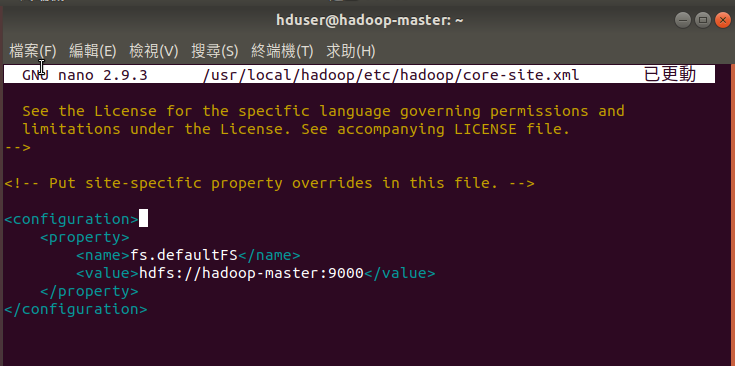
21. 告訴 Hadoop 那些主機是 slave

sudo nano /usr/local/hadoop/etc/hadoop/workers



22. 設定 core-site.xml 檔案

sudo nano /usr/local/hadoop/etc/hadoop/core-site.xml



<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://hadoop-master:9000</value>

</property>

</configuration>

23. 設定 yarn-site.xml 檔案

sudo nano /usr/local/hadoop/etc/hadoop/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>

<property>

<name>yarn.resourcemanager.hostname</name>

<value>hadoop-master</value>

</property>

<property>

<name>yarn.acl.enable</name>

<value>0</value>

</property>

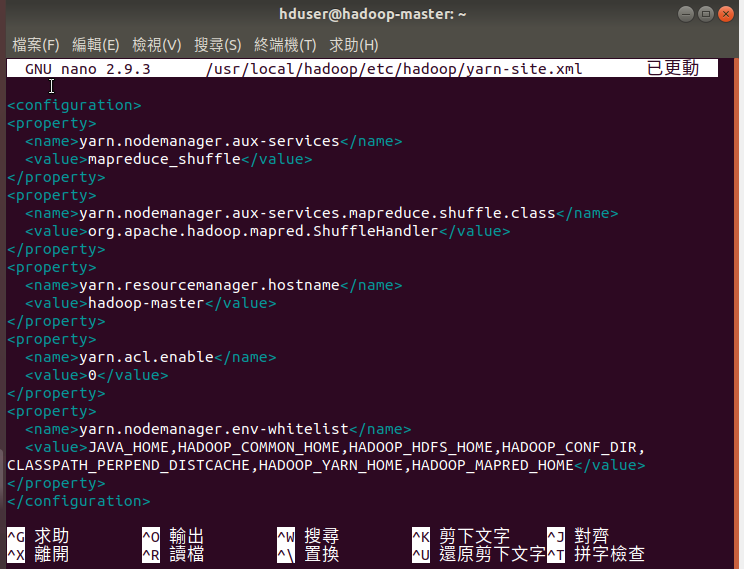
<property>

<name>yarn.nodemanager.env-whitelist</name>

<value>JAVA\_HOME,HADOOP\_COMMON\_HOME,HADOOP\_HDFS\_HOME,HADOOP\_CONF\_DIR,CLASSPATH\_PERPEND\_DISTCACHE,HADOOP\_YARN\_HOME,HADOOP\_MAPRED\_HOME</value>

</property>

</configuration>



24. 設定 hdfs-site.xml 檔案

其中 /usr/local/hadoop/hadoop\_data/namenode與 /usr/local/hadoop/hadoop\_data/datanode 為放namenode 與 datanode 的路徑，需要創立否則hadoop datanode啟動會失敗

sudo nano /usr/local/hadoop/etc/hadoop/hdfs-site.xml

<configuration>  
<property>  
<name>dfs.namenode.name.dir</name><value>/usr/local/hadoop/hadoop\_data/hdfs/namenode</value>  
</property>  
<property>  
<name>dfs.datanode.data.dir</name><value>/usr/local/hadoop/hadoop\_data/hdfs/datanode</value>  
</property>  
<property>  
<name>dfs.replication</name>  
<value>3</value>  
</property>

<property>

<name>dfs.webhdfs.enabled</name>

<value>true</value>

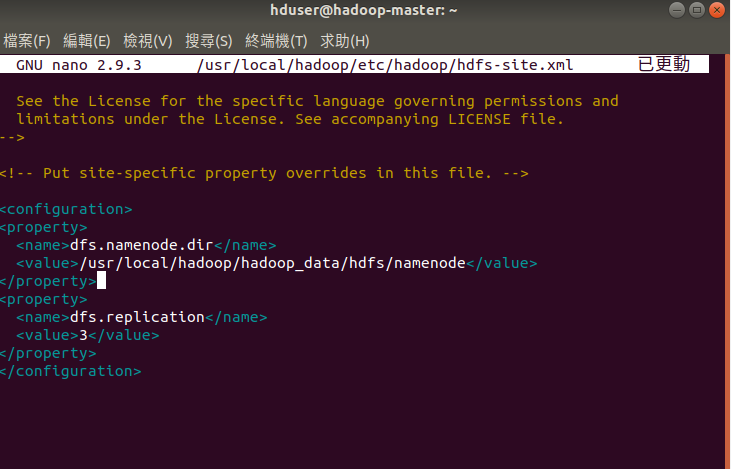
</property>

<property>

<name>dfs.namenode.http-address</name>

<value>hadoop-master:50070</value>

</property>  
</configuration>



25. 設定 mapred-site.xml 用於設定監控Map 與 Reduce 程式的 JobTracker 工作分配以及 TaskTracker 工作執行狀況。

sudo nano /usr/local/hadoop/etc/hadoop/mapred-site.xml

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

</configuration>



26. 將 /usr/local/hadoop/etc/hadoop/ 目錄下的檔案傳輸到 slave 的目錄下。同步 slave與 master 的設定檔。

scp /usr/local/hadoop/etc/hadoop/\* hadoop-slave1:/usr/local/hadoop/etc/hadoop/

scp /usr/local/hadoop/etc/hadoop/\* hadoop-slave2:/usr/local/hadoop/etc/hadoop/

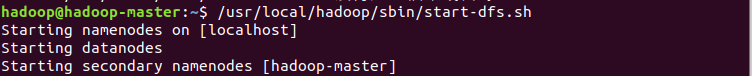
scp /usr/local/hadoop/etc/hadoop/\* hadoop-slave3:/usr/local/hadoop/etc/hadoop/

28. 將 namenode 格式化

hdfs namenode -format

29. 啟動 hdfs

/usr/local/hadoop/sbin/start-dfs.sh



30. 啟動 yarn

/usr/local/hadoop/sbin/start-yarn.sh



31.查看目前所有 JAVA執行的 process。

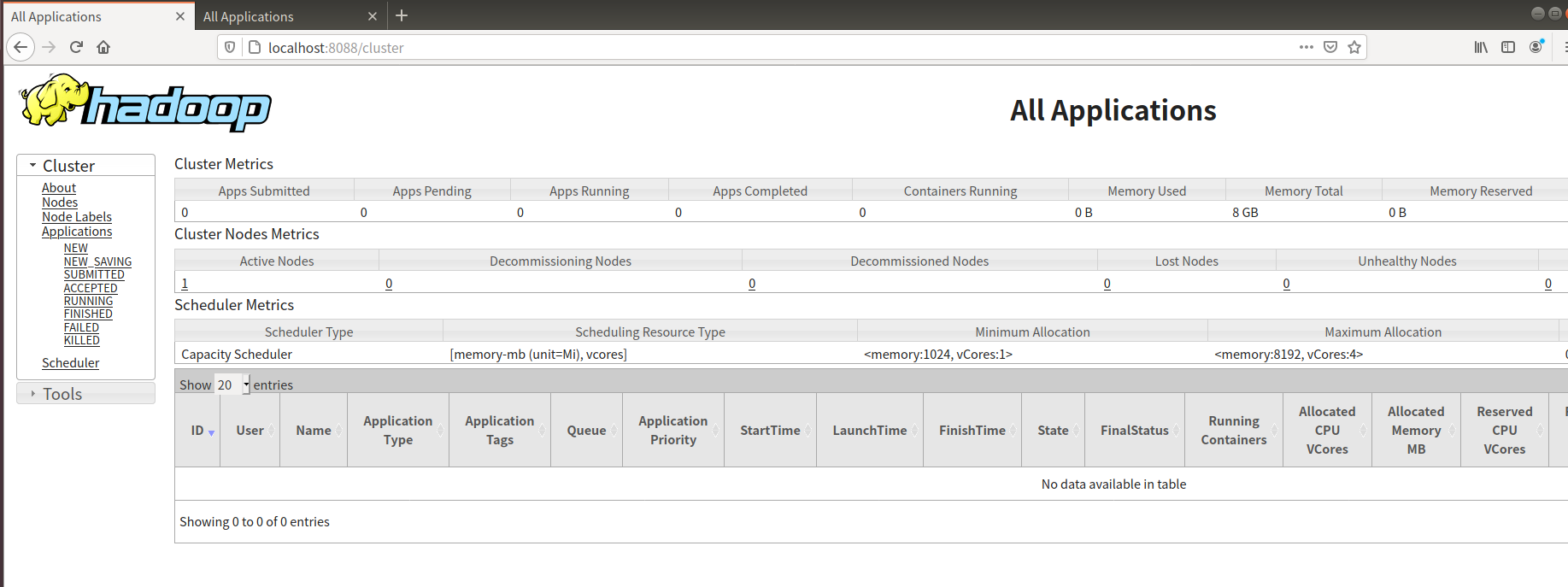
jps



32. 進入 Hadoop ResourceManager Web 介面網址

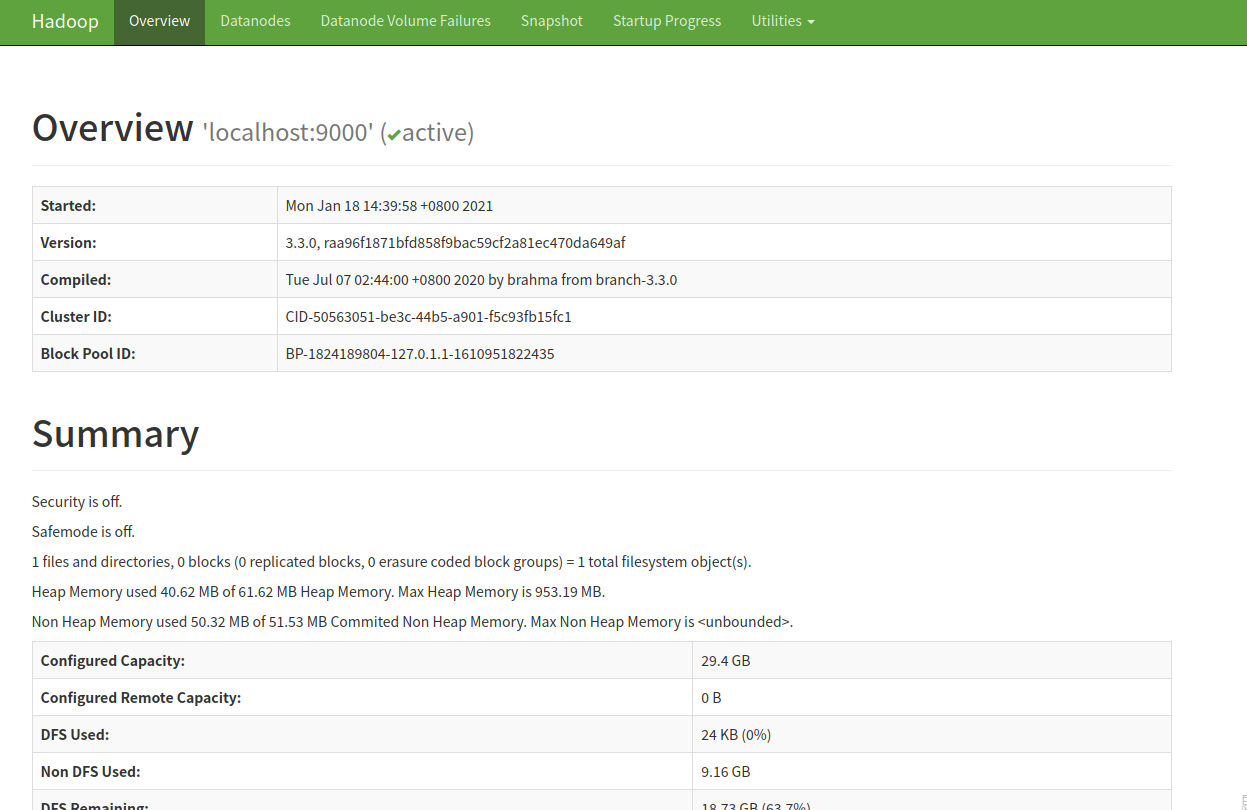
在瀏覽器輸入

http://localhost:8088/



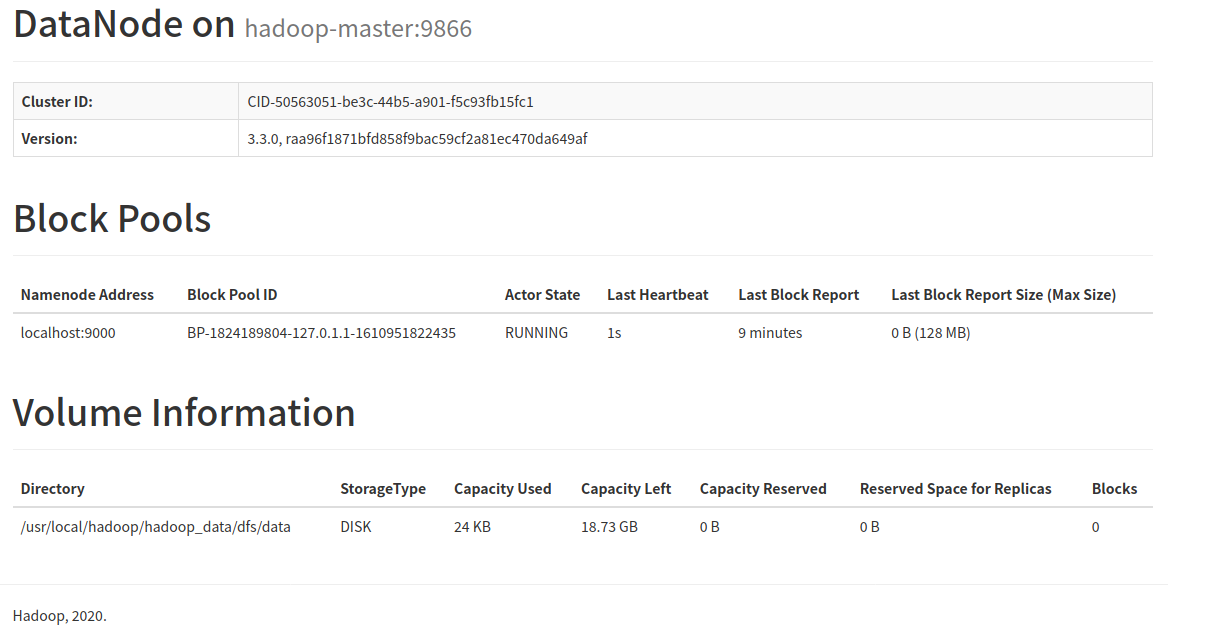
33. 進入 NameNode HDFS Web介面

http://localhost:9870/



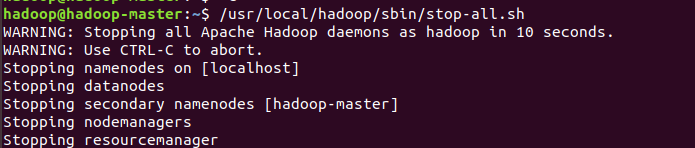
34. 進入 DataNode HDFS Web介面

http://localhost:9864/



35. 停止 HDFS

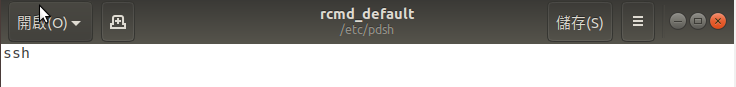
$HADOOP\_HOME/sbin/stop-all.sh



遇到

## hadoop-slave1: rcmd: socket: Permission denied





SPARK 安裝

1. Spark 由 scala 寫的，先安裝 scala

sudo apt-get install scala -y

2. 下載spark需對應相對的 Hadoop 版本

wget [**https://downloads.apache.org/spark/spark-3.0.1/spark-3.0.1-bin-hadoop3.2.tgz**](https://downloads.apache.org/spark/spark-3.0.1/spark-3.0.1-bin-hadoop3.2.tgz)

3. 解壓縮 spark

tar xvf spark-3.0.1-bin-hadoop3.2.tgz

4. 將spark 移至 /usr/local/spark

sudo mv spark-3.0.1-bin-hadoop3.2 /usr/local/spark

5. 新增 spark 相關變數至 ~/.bashrc

#export YARN\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop 會有 warning 故註解調

sudo nano ~/.bashrc

export SPARK\_HOME=/usr/local/spark

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

export HDFS\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop

export HADOOP\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop

#export YARN\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop

source .bashrc

6. 修改 spark-env.sh

cd /usr/local/spark/conf  
複製 spark-env.sh.template 為 spark-env.sh

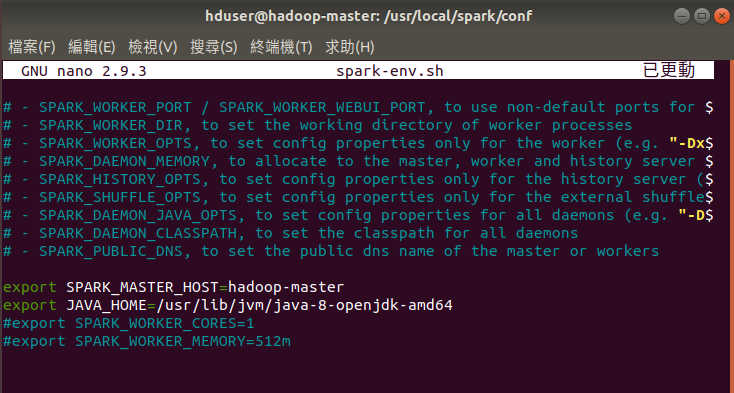
cp spark-env.sh.template spark-env.sh

修改 spark-env.sh

sudo nano spark-env.sh

export SPARK\_MASTER\_HOST=hadoop-master

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

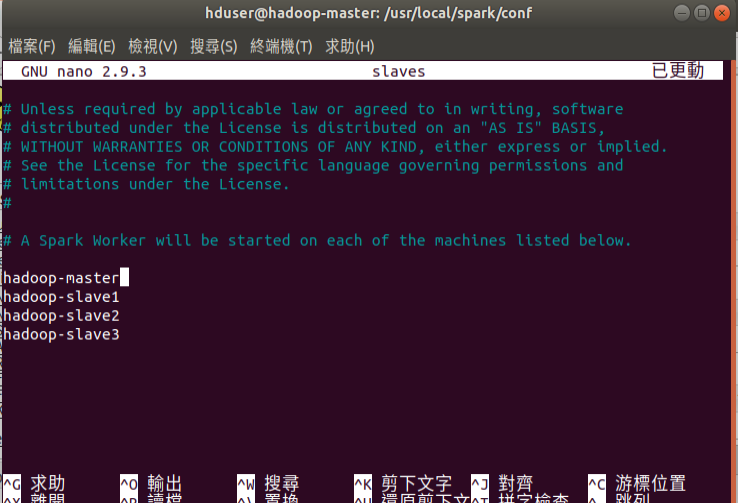
8. 修改 slaves 檔

複製 slaves.tempalte 檔為 slaves 檔

cp slaves.template slaves

新增 slaves 到檔案中 (此處 Master 也加入運算)

sudo nano slaves



11.將spark 複製到各個 slave(此處以 slave1 作為示範)

ssh hadoop-slave1

cd /usr/local/

sudo mkdir spark

sudo chown hduser:root -R /usr/local/spark/

sudo chmod g+rwx -R /usr/local/spark/

exit

回到master

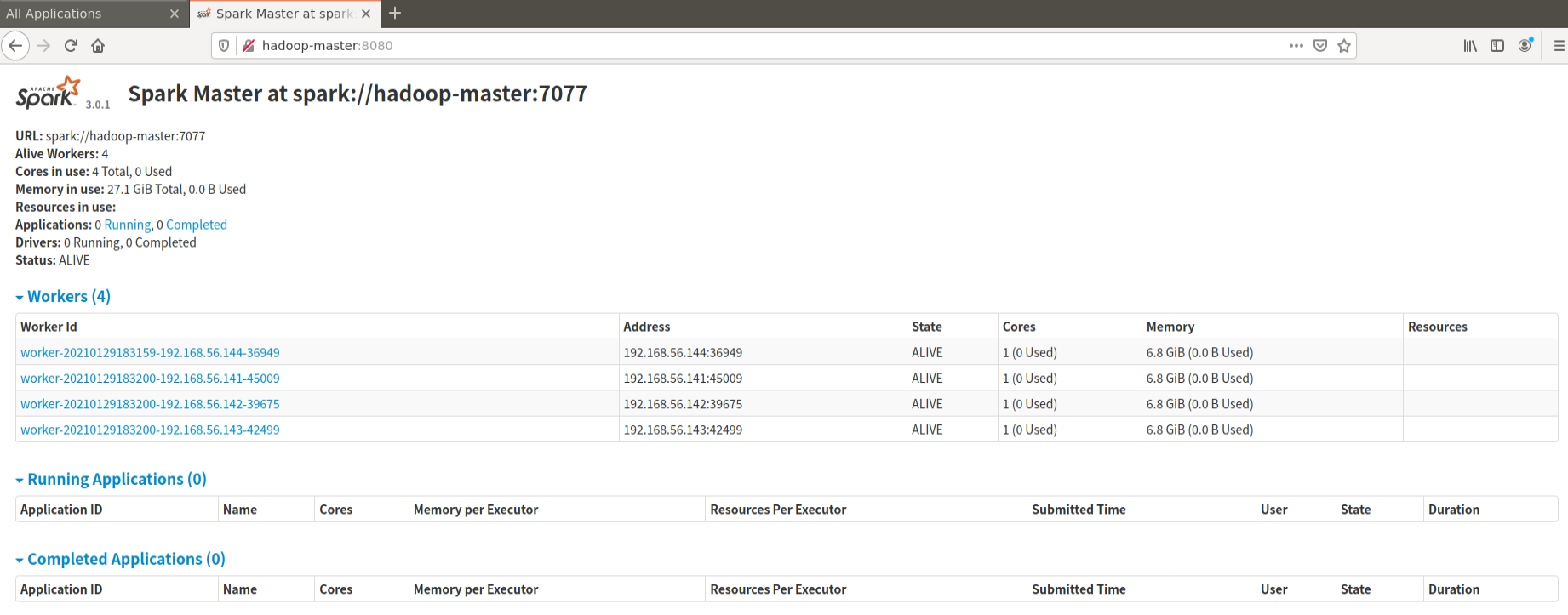
scp -r /usr/local/spark/ hadoop-slave1:/usr/local/

12.回到 master 啟動 spark

cd /usr/local/spark  
./sbin/start-all.sh

./sbin/stop-all.sh

13.打開網頁瀏覽器輸入 master-hadoop:8080 檢查是否成功



安裝 Anaconda ( 每台都要安裝 master、slaves)

1. 上網下載 Anaconda

wget <https://repo.anaconda.com/archive/Anaconda3-2020.11-Linux-x86_64.sh>

1. 安裝 Anaconda™

bash Anaconda3-2020.11-Linux-x86\_64.sh -b

1. 在 ~/.bashrc 增加 Anaconda 及 pyspark路徑

sudo nano ~/.bashrc

# Anaconda related options

export PATH=/home/hduser/anaconda3/bin:$PATH

export ANACONDA\_PATH=/home/hduser/anaconda3

export PYSPARK\_DRIVER\_PYTHON=$ANACONDA\_PATH/bin/ipython

export PYSPARK\_PYTHON=$ANACONDA\_PATH/bin/python

sour ce ~/.bashrc

# 啟動 pyspark 以 jupyter notebook 呈現

PYSPARK\_DRIVER\_PYTHON=ipython PYSPARK\_DRIVER\_PYTHON\_OPTS="notebook" HADOOP\_CONF\_DIR=/usr/local/hadoop/etc/hadoop MASTER=yarn-client pyspark

cd /usr/local/spark

./sbin/start-all.sh

./sbin/stop-all.sh

hdfs namenode -format

sudo rm -r /usr/local/hadoop/hadoop\_data/hdfs/namenode

sudo rm -r /usr/local/hadoop/hadoop\_data/hdfs/datanode

sudo mkdir -p /usr/local/hadoop/hadoop\_data/hdfs/namenode

sudo mkdir -p /usr/local/hadoop/hadoop\_data/hdfs/datanode

sudo chown hduser:root -R /usr/local/hadoop/  
sudo chmod g+rwx -R /usr/local/hadoop/