資料量要多大才能稱為
Big Data ?

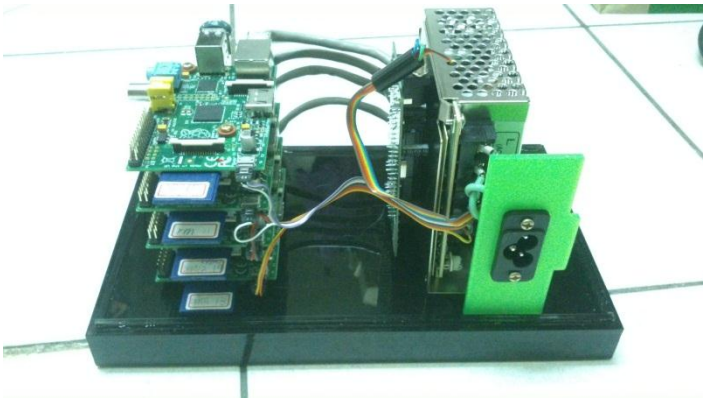
Value	Metric	
1000	**KB**	kilobyte
1000^2	**MB**	megabyte
1000^3	**GB**	gigabyte
1000^4	**TB**	terabyte
1000^5	**PB**	petabyte
1000^6	**EB**	exabyte
1000^7	**ZB**	zettabyte
1000^8	**YB**	yottabyte

Big Data 的資料那裡來？

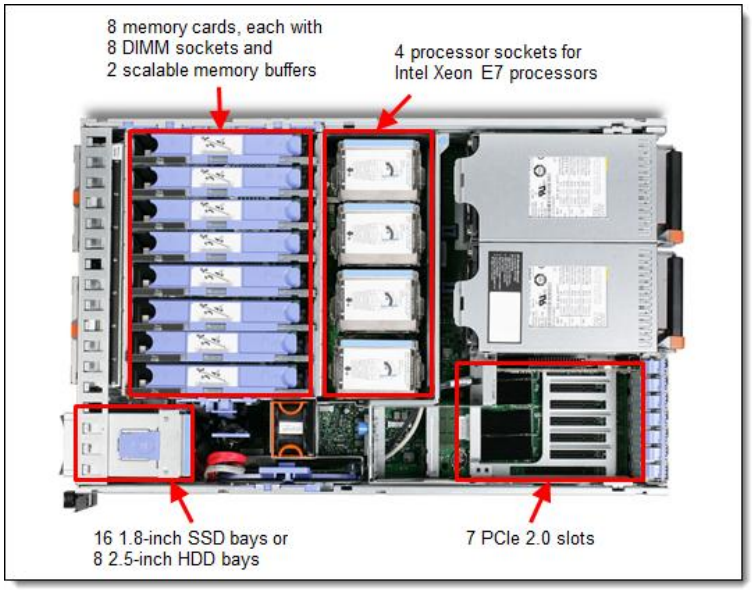
- 企業 IT 資訊系統的網站記錄檔 (Log)
- Open Data (政府資料開放平臺,…)
- 社交網站 (Facebook, Line, ….)
- 物聯網 (監視器, 智慧型穿戴裝置, …..)

罐子裡還有沒有咖啡？

Hadoop 迷你小刀鋒



一部百萬級的 IBM System x3850 X5 可處理 Big Data？

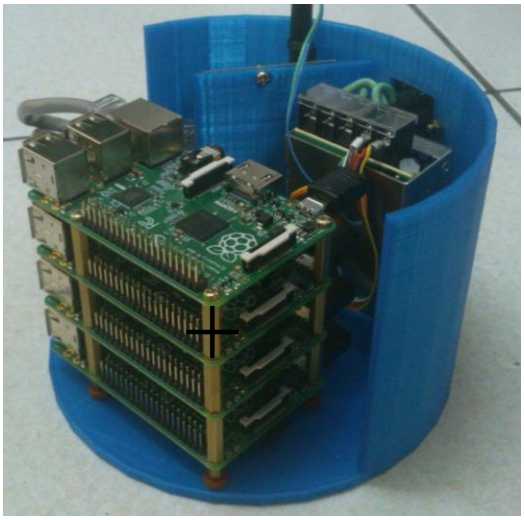


尺寸外型/高度	機架/4U（每機箱）
處理器（最大）	Intel Xeon 高達 2.4 GHz（十核心）/1066 MHz 記憶體存取
處理器數目（標配/最大）	2/4（每節點），（支援選配雙節點）
快取記憶體（最大）	高達 30 MB
記憶體（標配/最大）	**32 GB/2.0 TB 最大 PC3-10600 DDR3L，使用 MAX5 則高達 3.0 TB**
擴充插槽	總計七個 PCIe 半高型
磁碟機櫃（總數/熱抽換式）	8/8 2.5吋串列連接 SCSI（SAS）或 16/16 SAS SSD
內部儲存設備最大容量	每個機箱 4.8 TB（支援 8×73.4 GB、146.8 GB、300 GB、500 GB 及 600 GB SAS 硬碟機、8×160 GB 及 500 GB SATA 硬碟機，或 16×50 GB 及 200 GB 固態硬碟）
網路介面	10 Gbps 乙太網路雙通道光纖通道聚合式網路介面卡，整合式雙十億位元乙太網路（含 TCP-IP 卸載引擎）

要擁有 Big Data 必須花千萬買機器, 百萬買軟體嗎？

Big Data 迷你小刀鋒
硬體架構

樹莓派（英語：**Raspberry Pi**），是一款基於Linux系統的只有一張信用卡大小的單板機電腦。它由英國的樹莓派基金會所開發，目的是以低價硬體及自由軟體刺激在學校的基本的電腦科學教育。



Big Data 迷你小刀鋒
作業系統
Debian Linux + Hadoop (HDFS + YARN)

RASPBIAN
Debian Wheezy

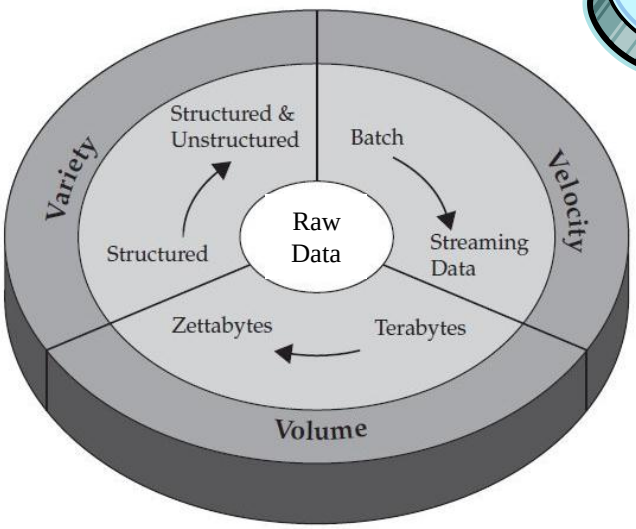
Version:	September 2014
Release date:	2014-09-09
Default login:	pi / raspberry
URL:	raspbian.org
Kernel version:	3.12
Release notes:	Link

Download Torrent | Download ZIP

hadoop

開發者	Apache Software Foundation
穩定版本	2.4.0 / 2014年4月7日（6個月前）
開發狀態	活躍
程式語言	Java
作業系統	跨平台
類型	大數據、分佈式系統
許可協議	Apache許可證 2.0
網站	hadoop.apache.org

The Four V's of Big Data



Veracity

Value	Metric	
1000	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	**petabyte**
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	yottabyte

1. 海量 (Volume)
海量資料的特色就在於： 龐大。 企業資料包羅萬端，很容易便達到數兆位元組，甚至千兆位元組之譜。(1)Facebook一天有 30 億筆使用者所產生的 PO 文、按讚、回覆...等資訊，故一個月將近1200億筆的資料，這是關聯式資料庫無法處理的，所以 NoSQL(Not Only SQL)因此誕生。(2)以Youtube 一天的影片上傳量來說，若一個人要全部看完，所需花的時間會來到80年，故窮盡一生若你會長壽的最多看個兩天也很了不起了。

2. 即時性 (Velocity)
海量資料通常具有時效性，一旦串流至企業便須立即使用，方能發揮其最大價值。另外，要注意的是有些網站是24小時流量跟資料不斷湧入，面對這種情況，我們通常可以把他稱為『data stream』，此時 data stream type 的 data mining 將成為一個有趣的議題。因為在此環境下，資料永遠沒有穩態的一天，量隨時再增加，傳統的資料挖掘方式得被迫重新思考其方式。

3. 多樣性 (Variety)
海量資料的範疇不僅止於結構化資料，還包含各類非結構化的資料： 諸如文字、音訊、視訊、點擊串流 (click stream)、日誌檔等等。如何去妥善描繪圖片、影音檔的內容也成為一大議題。資料的運用不在是下了幾行sql語法就可以全部撈出來，因為更多的資訊藏匿在圖片跟影片之中。常見的手法可分為：(1) Meta description mode，在此模式下會將這些影音檔，設定好的描述資料(亦即metadata)來陳述之，之後用一個XML檔來對應一部影片檔，如此才可以從XML去瞭解哪寫圖片想關聯性？哪些影片該如何推薦。(2) Behavior analysis mode，在此模式下會去記錄user behavior的關聯，從使用者行為去歸納，下一個使用者點擊某一內容時，跟它的行為模式最相近的群體其接下來最感興趣的會是什麼，就由大眾的力量來勾勒出群體模式。

4. 真實性 (Veracity)
指的是當資料的來源變得更多元時，這些資料本身的可靠度、品質是否足夠，若資料本身就是有問題的，那分析後的結果也不會是正確的

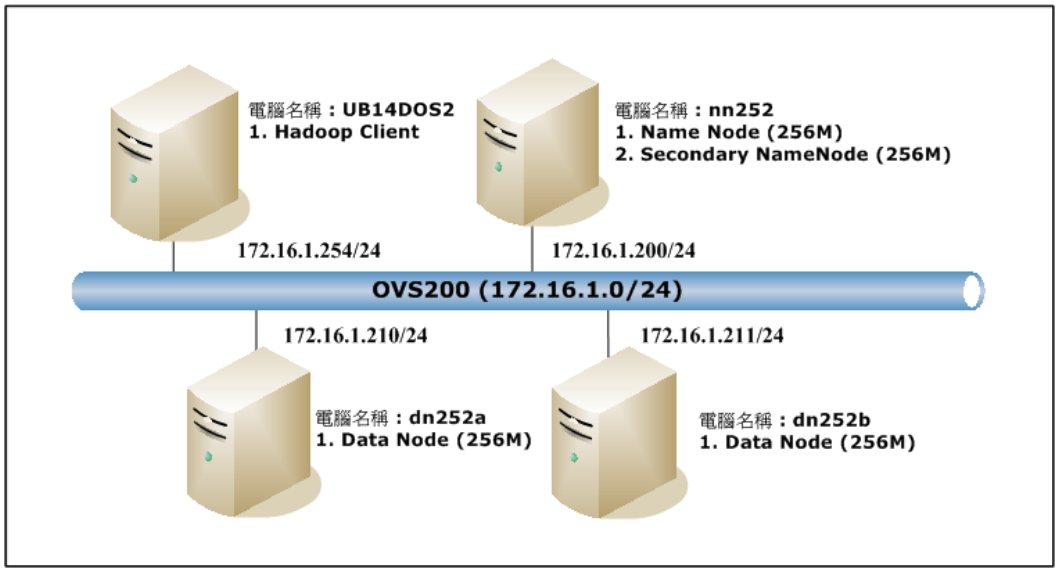
巨量資料（Big Data）要處理的主要三個面向，簡稱為3V。 請問 3V代表以下那三個功能?（請選擇三個答案）
A. 處理時效（Velocity）
B. 資料格式（Variety）
C. 資料量（Volume）
D. 多功能（Versatility）
E. 巨量的（Vast）

答案：

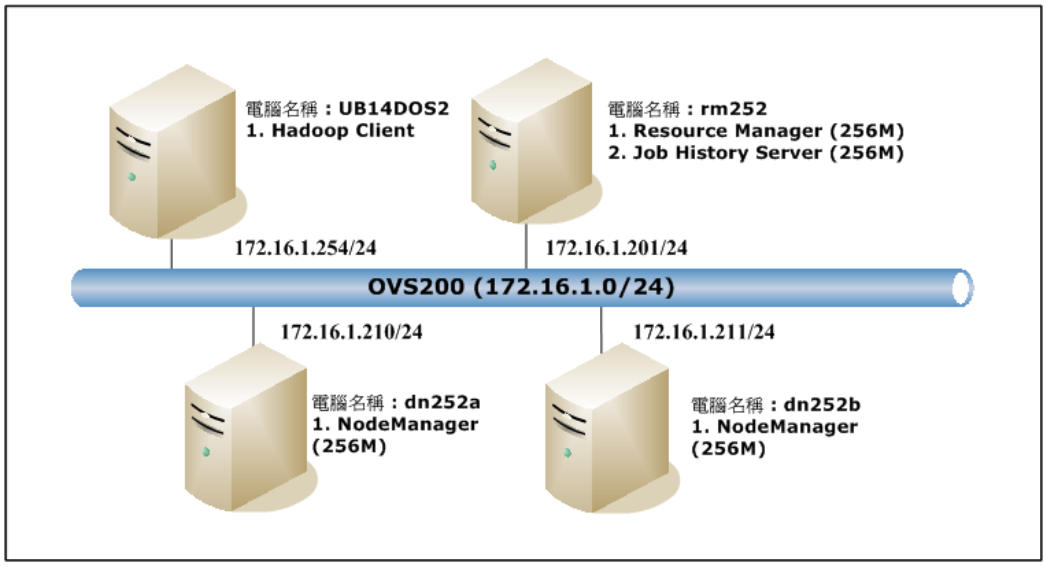
誰能處理 Big Data ?

Apache Hadoop 專案網址 : apache.org

Hadoop Distributed File System (HDFS) 分散檔案系統



Yet Another Resource Negotiator (YARN) 分散運算系統



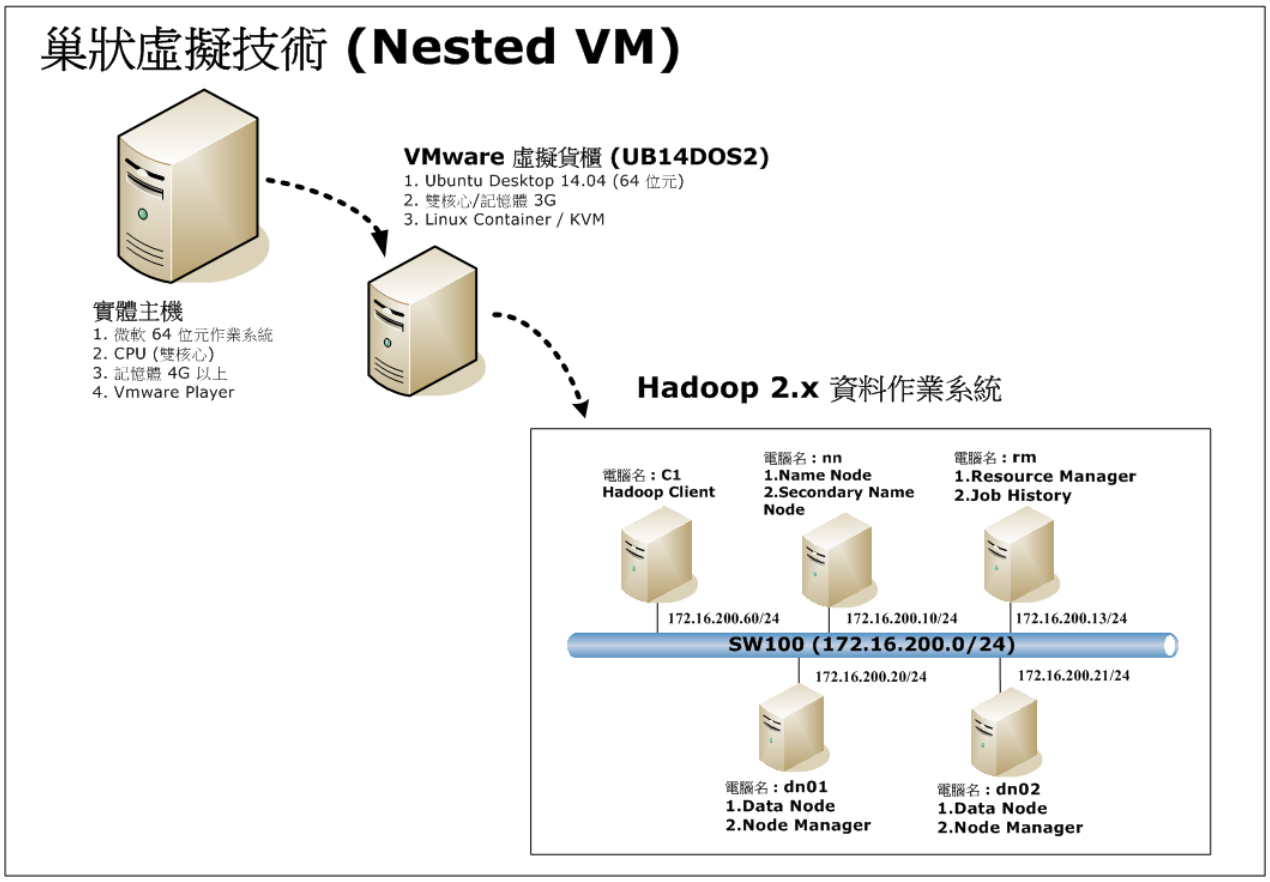
請問以下那一個網址可以下載 Hadoop 套件
A. http://ftp.twaren.net/Unix/Web/apache/hadoop/common/
B. http://ftp.twaren.net/Linux/Web/apache/hadoop/common/
C. http://ftp.twnet.net/Unix/Web/apache/hadoop/getapp/
D. http://ftp.twaren.net/Linux/hadoop/apache/
E. http://ftp.apache.net/Unix/ apache/hadoop/common/

答案：

請問以下那些是Hadoop的主要組成元件 ？(請選擇三個答案)
A. 分散式檔案系統(HDFS)
B. MapReduce 框架
C. 分散運算系統(YARN)
D. RPC 資料序列化系統
E. 分散式資料倉儲 (Hive)

答案：

如何實作 Hadoop 資料作業系統？

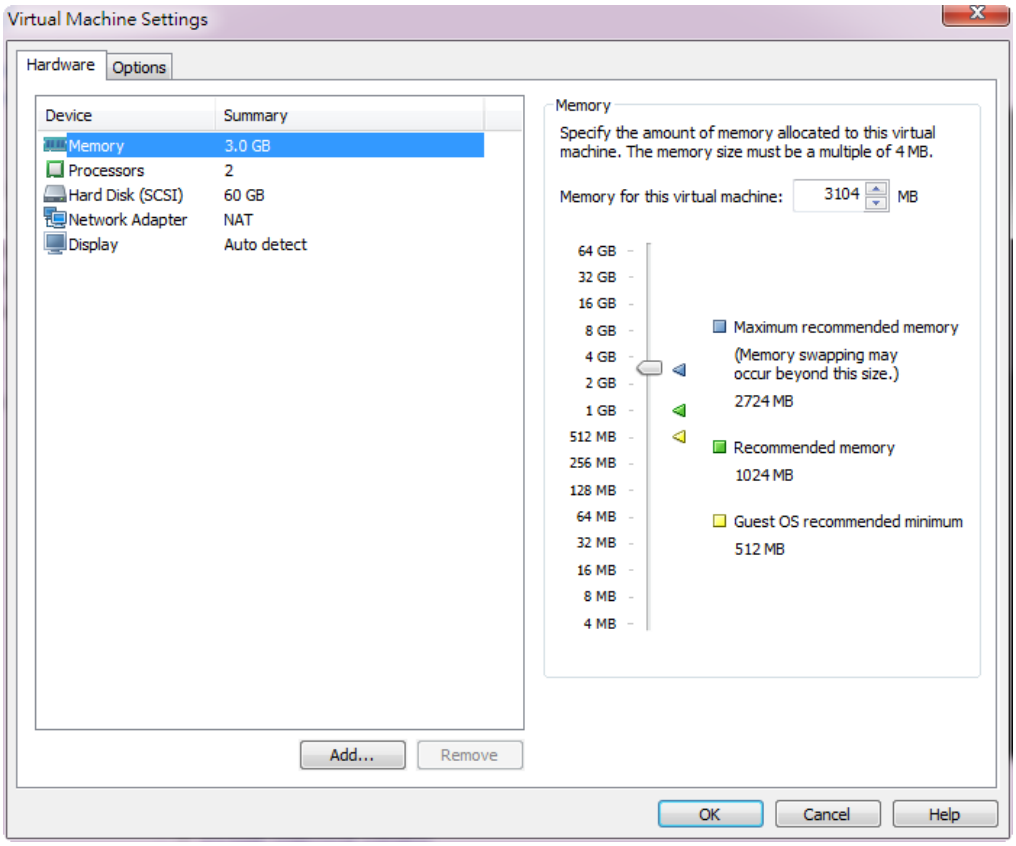


安裝 Vmware 虛擬貨櫃主機

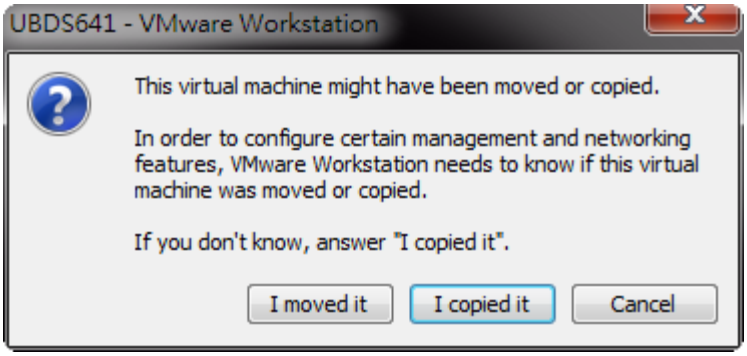
1. 將 UB14DOS2v1.zip 解壓縮至 C 磁碟機的根目錄 （\）

2. 使用 VMware Player 開啟 UB14DOS2 虛擬貨櫃主機 （UB1404.vmx）

3. 設定虛擬記憶體



4. 第一次啟動 UB14DOS2 虛擬貨櫃主機
務必點選 "I copied it" 按鈕，產生新的虛擬貨櫃主機



請問當您執行以下何者指令時，會出現以下畫面

```
eth1      Link encap:Ethernet  HWaddr 00:16:3e:53:a7:bc
          inet6 addr: fe80::216:3eff:fe53:a7bc/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:245 errors:0 dropped:0 overruns:0 frame:0
          TX packets:218 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:37273 (37.2 KB)  TX bytes:28598 (28.5 KB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:16436  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:0 (0.0 B)  TX bytes:0 (0.0 B)
```

A. ifconfig -a

B. ipconfig -a

C. ifconfig -t

D. ipconfig -t

E. ifconfig lxc

答案：

管理員看見以下結果:

```
tcp   0   0 0.0.0.0:50370    0.0.0.0:*  LISTEN  1001  8236   2471/java
tcp   0   0 0.0.0.0:50310    0.0.0.0:*  LISTEN  1001  9898   2428/java
tcp   0   0 0.0.0.0:48159    0.0.0.0:*  LISTEN  1001  8896   2428/java
tcp   0   0 0.0.0.0:53321    0.0.0.0:*  LISTEN  1001  9828   2457/java
tcp   0   0 127.0.0.1:54310 0.0.0.0:*  LISTEN  1001  8843   2471/java
tcp   0   0 127.0.0.1:54311 0.0.0.0:*  LISTEN  1001  8230   2457/java
tcp   0   0 0.0.0.0:59305    0.0.0.0:*  LISTEN  1001  88141  2471/java
tcp   0   0 0.0.0.0:50360    0.0.0.0:*  LISTEN  1001  9857   4005/java
tcp   0   0 0.0.0.0:49300    0.0.0.0:*  LISTEN  1001  9837   4785/java
tcp   0   0 0.0.0.0:50030    0.0.0.0:*  LISTEN  1001  8773   2857/java
MicroSys@Globalit:~$，
```

請問以下那一個指令完成此工作?

A. netstat

B. ping

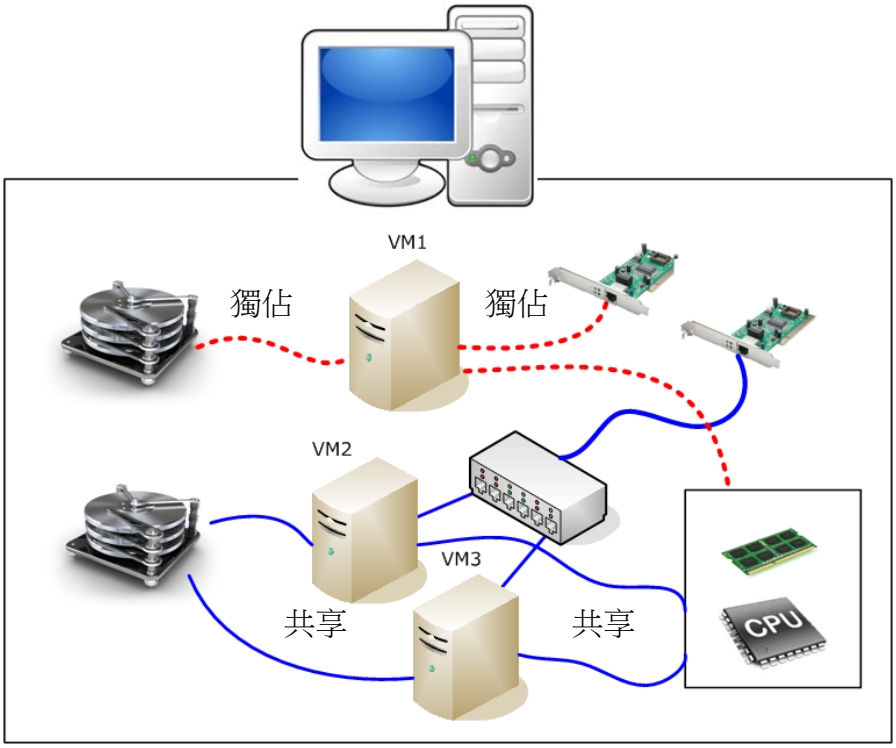
C. tracer

D. ipconfig

E. ifconfig

答案：

Linux Container (LXC) 高效能虛擬技術



1. 安裝 LXC

student@UB14DOS2:~/dos$ **sudo apt-get install lxc**

2. 產生 LXC 虛擬電腦 (Ubuntu Server 14.04)

student@UB14DOS2:~/dos$ **sudo lxc-create -t ubuntu -n myUS14**
Checking cache download in /var/cache/lxc/trusty/rootfs-amd64 ...
Copy /var/cache/lxc/trusty/rootfs-amd64 to /var/lib/lxc/myUS14/rootfs ...
Copying rootfs to /var/lib/lxc/myUS14/rootfs ...

3. 啟動及登入 myUS14 虛擬電腦

student@UB14DOS2:~/dos$ **sudo lxc-start -n myUS14**
Ubuntu 14.04 LTS myUS14 console

myUS14 login: <4>init: setvtrgb main process (415) terminated with status 1
<4>init: plymouth-upstart-bridge main process ended, respawning
<ENTER>
Ubuntu 14.04 LTS myUS14 console

myUS14 login: **ubuntu**
Password:

4. 關閉 myUS14 虛擬電腦

ubuntu@myUS14:~$ **sudo shutdown -h now**

5. 列出所有 **LXC** 虛擬電腦

```
student@UB14DOS2:~/dos$  sudo lxc-ls -f
NAME       STATE       IPV4    IPV6    AUTOSTART
----------------------------------------
dn01       STOPPED     -       -       NO
dn02       STOPPED     -       -       NO
myCT65     STOPPED     -       -       NO
myUS14     STOPPED     -       -       NO
nn         STOPPED     -       -       NO
rm         STOPPED     -       -       NO
```

6. LXC 虛擬電腦的系統更新及安裝所需套件

```
student@UB14DOS2:~/dos$  sudo lxc-start -n myUS14
ubuntu@myUS14:~$  sudo apt-get update
ubuntu@myUS14:~$  sudo apt-get upgrade
ubuntu@myUS14:~$  sudo apt-get install nano tree wget curl
[sudo] password for ubuntu: ubuntu
```

7. 建立 **LXC** 虛擬電腦的自訂命令

```
ubuntu@myUS14:~$  nano .bashrc
                  ::
alias ping='ping -c 4'
alias bye='sudo shutdown -h now'
```

8. 測試自定命令

```
ubuntu@myUS14:~$  exit                          #  登出
ubuntu@myUS14:~$  ping www.hinet.net
ubuntu@myUS14:~$  bye
[sudo] password for ubuntu: ubuntu
```

9. 以 **Console** 模式 登入 **LXC** 虛擬電腦

```
student@UB14DOS2:~/dos$  sudo lxc-start -n myUS14 -d
student@UB14DOS2:~/dos$  sudo lxc-console -n myUS14
```

按 **Ctrl+a q** 離開 **Console** 模式

```
student@UB14DOS2:~/dos$  sudo lxc-ls -f
NAME       STATE       IPV4              IPV6    AUTOSTART
----------------------------------------------
dn01       STOPPED     -                 -       NO
dn02       STOPPED     -                 -       NO
myCT65     STOPPED     -                 -       NO
myUS14     RUNNING     172.16.1.187      -       NO
nn         STOPPED     -                 -       NO
rm         STOPPED     -                 -       NO
student@UB14DOS2:~/dos$  sudo lxc-stop  -n  myUS14
```

10. 複製 **LXC** 虛擬電腦

```
student@UB14DOS2:~/dos$  sudo lxc-clone  myUS14  myClone
Created container myClone as copy of myUS14
```

11. 刪除 **LXC** 虛擬電腦

```
student@UB14DOS2:~/dos$  sudo lxc-destroy -n myClone
```

請問以下何者 指令可以建立 Linux Container 虛擬電腦? 設定內容: . 作業為 ubuntu . 名稱為 myUS1204
A. sudo lxc-open -t ubuntu -n myUS1204
B. sudo lxc-create -t ubuntu -n myUS1204
C. sudo lxc-run -t ubuntu -n myUS1204
D. sudo lxc-reboot -t ubuntu -n myUS1204

答案:

請問以下何者 指令可以啟動 Linux Container 虛擬電腦? 設定內容: . 作業為 ubuntu . 名稱為 myUS1204
A. sudo lxc-reboot -n myUS1204
B. sudo lxc-start -n myUS1204
C. sudo lxc-run -n myUS1204
D. sudo lxc-power_on -n myUS1204

答案:

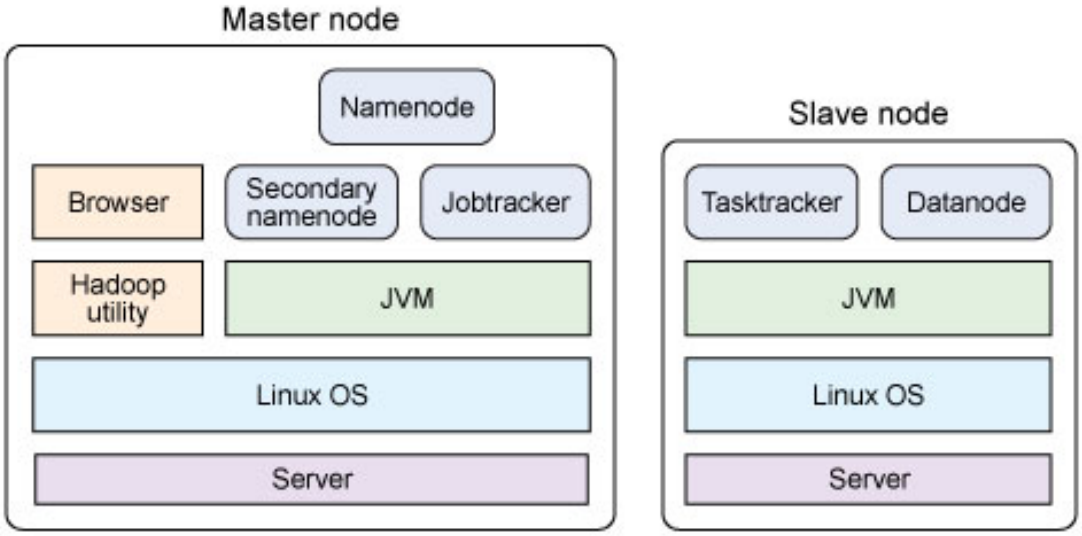
當你執行 brctl show 後:出現以下畫面

請問以下何者描述為正確 :

A. lxcbr0 具備 brigde 功能
B. lxcbr0 預設提供 DHCP 及 NAT 功能
C. lxcbr0 具備 Routing Protocol 功能
D. lxcbr0 具備 tracer IP 功能

答案:

Hadoop 分散運算系統 –建置 Hadoop 模板運算主機



1. 建立 **nn252** 運算主機

student@UB14DOS2:~/dos$ **sudo lxc-clone myUS14 nn252**

2. 啟動並登入 **nn252**

student@UB14DOS2:~/dos**$ sudo lxc-start -n nn252**

3. 設定 **/etc/hosts**

ubuntu@nn252:~$ **sudo nano /etc/hosts**
127.0.0.1 localhost
172.16.1.200 nn252
172.16.1.201 rm252
172.16.1.210 dn252a
172.16.1.211 dn252b

[註] 確認 nn252 這名稱不可指定 127.0.0.1 這 Loop Back 位址, 如指定會造成 DataNode 無法連接到 NameNode

4. 設定固定 **IP** 位址

ubuntu@nn252:~$ **sudo nano /etc/network/interfaces**

::

auto eth0
iface eth0 inet **static**
address 172.16.1.200
netmask 255.255.255.0
gateway 172.16.1.1
dns-nameservers 168.95.1.1

5. 重新啟動 **nn252** 運算主機

ubuntu@nn252:~$ **sudo reboot**

5. 產生 **ssh** 自動登入憑證

ubuntu@nn252:~$ **ssh-keygen -t dsa -P '' -f ~/.ssh/id_dsa**

ubuntu@nn252:~$ **ls .ssh**

id_dsa id_dsa.pub

ubuntu@nn252:~$ **cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys**

6. 第一次執行 **SSH** 本機登入

ubuntu@nn252:~$ **ssh nn252**

The authenticity of host 'nn252 (172.16.1.200)' can't be established.

ECDSA key fingerprint is 06:c9:82:35:c1:9e:1a:5b:b8:dd:da:91:eb:d5:e1:36.

Are you sure you want to continue connecting (yes/no)? **yes**

Warning: Permanently added 'nn252,172.16.1.200' (ECDSA) to the list of known hosts.

Welcome to Ubuntu 14.04.1 LTS (GNU/Linux 3.13.0-36-generic x86_64)

* Documentation: https://help.ubuntu.com

Last login: Sat Oct 4 23:23:32 2014

ubuntu@nn252:~$ **exit**

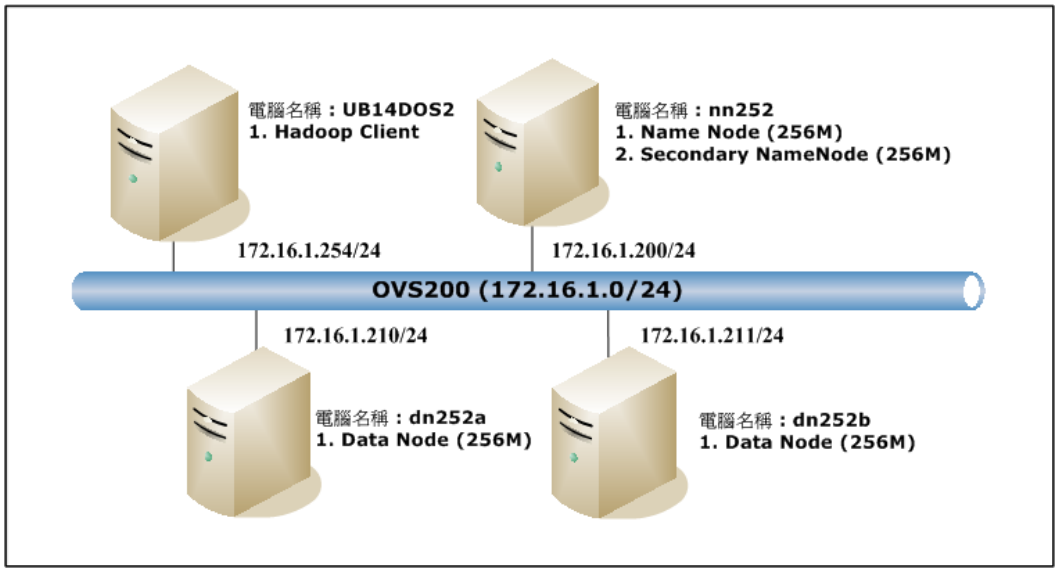
ubuntu@nn252:~$ **ls .ssh**

authorized_keys id_dsa id_dsa.pub known_hosts

7. 關閉 **nn252** 運算主機

ubuntu@nn252:~$ **bye**

建置 Hadoop 分散運算系統 –SSH 網路管理系統



1. 建立 rm252, dn252a 及 dn252b 運算主機

student@UB14DOS2:~/dos$ **sudo lxc-clone nn252 rm252**

student@UB14DOS2:~/dos$ **sudo lxc-clone nn252 dn252a**

student@UB14DOS2:~/dos$ **sudo lxc-clone nn252 dn252b**

2. 設定 rm252, dn252a 及 dn252b 運算主機的 IP 位址

student@UB14DOS2:~/dos$ **sudo lxc-start -n rm252 -d**

student@UB14DOS2:~/dos$ **sudo lxc-console -n rm252**

ubuntu@rm252:~$ **sudo nano /etc/network/interfaces**

::

```
auto eth0
iface eth0 inet static
  address 172.16.1.201
  netmask 255.255.255.0
  gateway 172.16.1.1
  dns-nameservers 168.95.1.1
```

3. 關閉 rm252 運算主機

ubuntu@rm252:~$ **bye**

[重要] 請重複步驟 2 , 3 設定 dn252a 及 dn252b 運算主機的 IP 位址

3. 撰寫 start252.sh 程式, 啟動 nn252, rm252, dn252a 及 dn252b 運算主機

```
student@UB14DOS2:~/dos$ nano start252.sh
#!/bin/bash
sudo lxc-start -n nn252 -d
sudo lxc-start -n rm252 -d
sudo lxc-start -n dn252a -d
sudo lxc-start -n dn252b -d

student@UB14DOS2:~/dos$ chmod +x start252.sh
student@UB14DOS2:~/dos$ ./start252.sh
student@UB14DOS2:~/dos$ sudo lxc-ls -f
NAME    STATE    IPV4          IPV6  AUTOSTART
----------------------------------------------
dn01    STOPPED  -             -     NO
dn02    STOPPED  -             -     NO
dn252a  RUNNING  172.16.1.210  -     NO
dn252b  RUNNING  172.16.1.211  -     NO
myCT65  STOPPED  -             -     NO
myUS14  STOPPED  -             -     NO
nn      STOPPED  -             -     NO
nn252   RUNNING  172.16.1.200  -     NO
rm      STOPPED  -             -     NO
rm252   RUNNING  172.16.1.201  -     NO
```

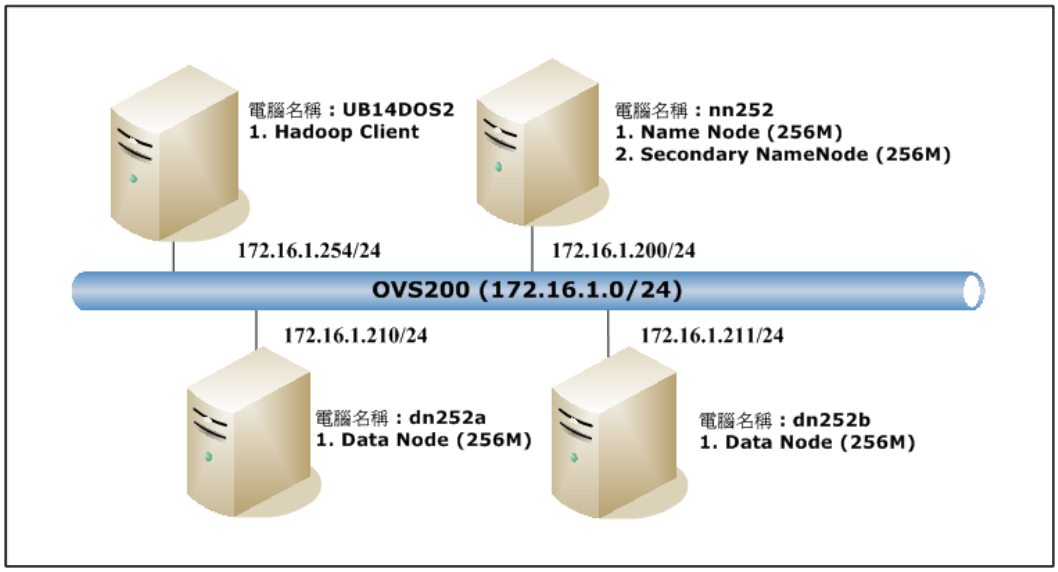
4. 設定從 nn252 使用 ssh 自動登入 rm252, dn252a 及 dn252b 運算主機

```
student@UB14DOS2:~/dos$ sudo lxc-console -n nn252
ubuntu@nn252:~$ ssh rm252
The authenticity of host 'rm252 (172.16.1.201)' can't be established.
ECDSA key fingerprint is 06:c9:82:35:c1:9e:1a:5b:b8:dd:da:91:eb:d5:e1:36.
Are you sure you want to continue connecting (yes/no)? yes
                                        ::
ubuntu@rm252:~$ exit

ubuntu@nn252:~$ ssh dn252a
The authenticity of host 'dn252a (172.16.1.210)' can't be established.
ECDSA key fingerprint is 06:c9:82:35:c1:9e:1a:5b:b8:dd:da:91:eb:d5:e1:36.
Are you sure you want to continue connecting (yes/no)? yes
                                        ::
ubuntu@dn252a:~$ exit

ubuntu@nn252:~$ ssh dn252b
The authenticity of host 'dn252b (172.16.1.211)' can't be established.
ECDSA key fingerprint is 06:c9:82:35:c1:9e:1a:5b:b8:dd:da:91:eb:d5:e1:36.
Are you sure you want to continue connecting (yes/no)? yes
                                        ::
ubuntu@dn252b:~$ exit
```

Hadoop分散運算系統 – 安裝 Java 及 Hadoop 套件



1. 設定共用目錄 (/opt)

請務必先關閉所有 LXC 虛擬主機 (nn252, rm252, dn252a, dn252b)

student@UB14DOS2:~/dos$ **sudo nano /var/lib/lxc/nn252/fstab**
/opt /var/lib/lxc/nn252/rootfs/opt none bind 0 0

student@UB14DOS2:~/dos$ **sudo nano /var/lib/lxc/rm252/fstab**
/opt /var/lib/lxc/rm252/rootfs/opt none bind 0 0

student@UB14DOS2:~/dos$ **sudo nano /var/lib/lxc/dn252a/fstab**
/opt /var/lib/lxc/dn252a/rootfs/opt none bind 0 0

student@UB14DOS2:~/dos$ **sudo nano /var/lib/lxc/dn252b/fstab**
/opt /var/lib/lxc/dn252b/rootfs/opt none bind 0 0

2. 下載 JDK 及 Hadoop 套件

jdk 下載網址：**http://www.oracle.com/technetwork/java/javase/downloads/index.html**

Hadoop 下載網址：**http://ftp.twaren.net/Unix/Web/apache/hadoop/common/**

3. 請將下載的套件解壓縮至 /opt 目錄, 命令如下：
student@UB14DOS2:~/dos$ **sudo tar xvfz jdk-7u71-linux-x64.tar.gz -C /opt**
student@UB14DOS2:~/dos$ **sudo tar xvfz hadoop-2.5.2.tar.gz -C /opt**

4. 設定環境變數

student@UB14DOS2:~$ **sudo nano /var/lib/lxc/nn252/rootfs/home/ubuntu/.bashrc**

::

export JAVA_HOME**=/opt/jdk1.7.0_71/**

export HADOOP_HOME**=/opt/hadoop-2.5.2**

export PATH=$PATH:$JAVA_HOME/bin

export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

student@UB14DOS2:~$ **sudo nano /var/lib/lxc/rm252/rootfs/home/ubuntu/.bashrc**

::

export JAVA_HOME**=/opt/jdk1.7.0_71/**

export HADOOP_HOME**=/opt/hadoop-2.5.2**

export PATH=$PATH:$JAVA_HOME/bin

export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

student@UB14DOS2:~$ **sudo nano /var/lib/lxc/dn252a/rootfs/home/ubuntu/.bashrc**

::

export JAVA_HOME**=/opt/jdk1.7.0_71/**

export HADOOP_HOME**=/opt/hadoop-2.5.2**

export PATH=$PATH:$JAVA_HOME/bin

export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

student@UB14DOS2:~$ **sudo nano /var/lib/lxc/dn252b/rootfs/home/ubuntu/.bashrc**

::

export JAVA_HOME**=/opt/jdk1.7.0_71/**

export HADOOP_HOME**=/opt/hadoop-2.5.2**

export PATH=$PATH:$JAVA_HOME/bin

export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

5. 執行 **java** 及 **hadoop** 命令

5-1 啟動 **Hadoop** 分散運算系統的所有運算主機

student@UB14DOS2:~$ **./start252.sh**

student@UB14DOS2:~$ **sudo lxc-console -n nn252**

5-2 顯示 **jdk** 版本代號

ubuntu@nn252:~$ **java -version**
java version "1.7.0_71"
Java(TM) SE Runtime Environment (build 1.7.0_71-b14)
Java HotSpot(TM) 64-Bit Server VM (build 24.71-b01, mixed mode)

5.3 顯示 **hadoop** 版本代號

ubuntu@nn252:~$ **hadoop version**

Hadoop 2.5.2

::

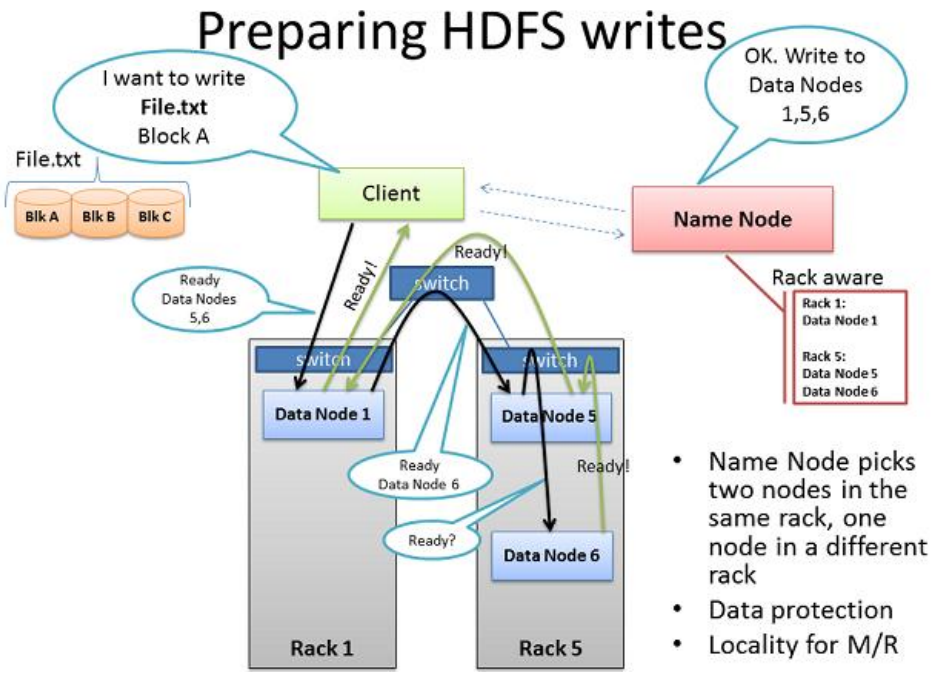
機櫃 (Rack)

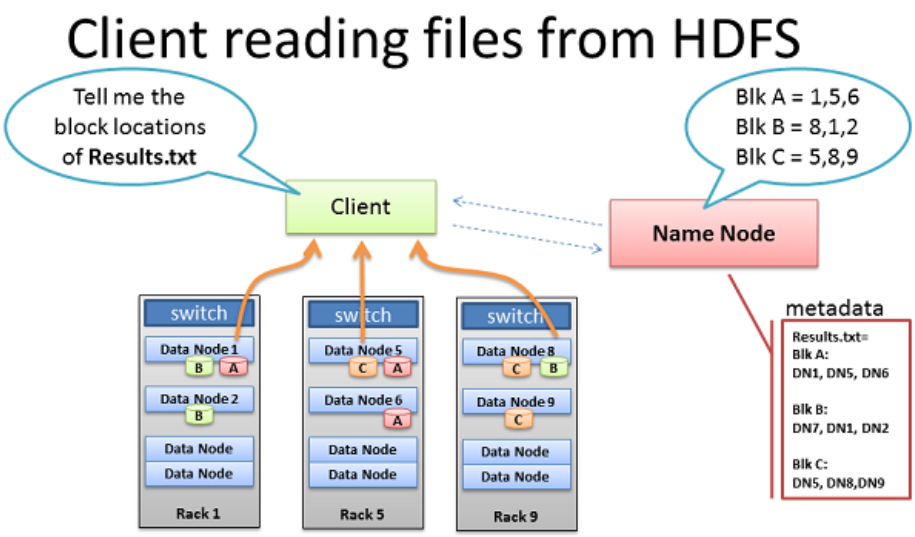
42U Rack

DL160 Gen8 Server

iSeries Rack Filled with DL360p Gen8 Servers

HDFS 分散檔案系統運作原理





你公司的 Hadoop cluster 內的 HDF 儲存具有以下內容: -25 nodes -總共 100 TB 儲存空間（每一個 node 有 4TB） 請問，在預試的情況下，有多少空間可以讓你儲存資料 ？
A. 大約 100TB
B. 大約 25TB
C. 大約 10TB
D. 大約 33TB
E. 大約 400TB

答案:

當管理員要建置 Hadoop cluster 的 slave node 時， 請問要將 slave node 格式化成以下何種檔案系統（filesystem）？
A. 任何 Linux 檔案系統（filesystem)均可。
B. HDFS 的檔案系統（filesystem)。
C. Ext3 or Ext4 檔案系統（filesystem)。
D. 不需事先進行格式化，在 HDFS 儲存時，就直自動進行資料格式化

答案:

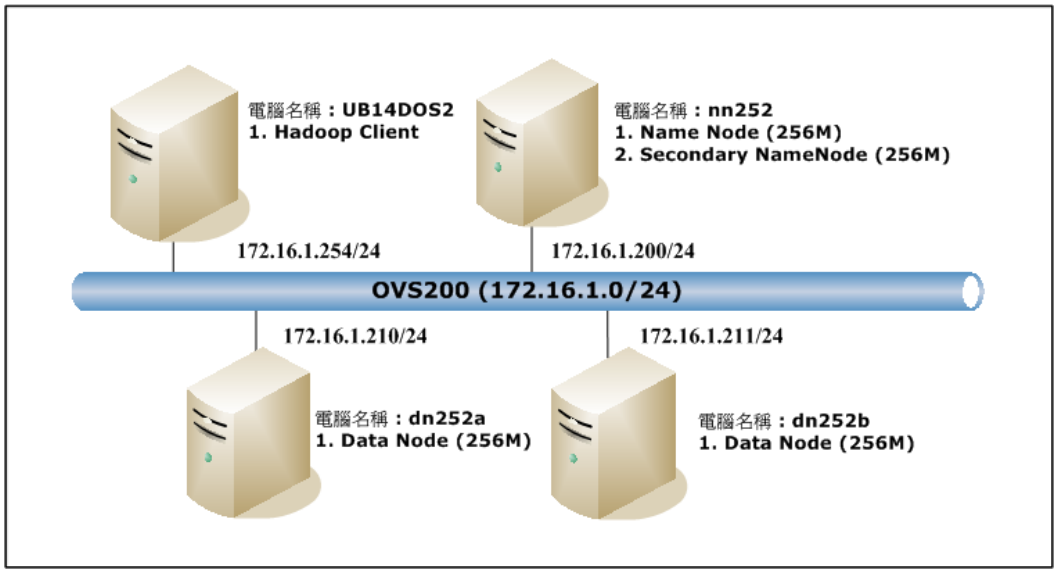
當管理員要建置 HDFS 的高可用性（High Availability;HA)時，兩個 Namenode 的名稱為 MicroSys01 及 MicroSys02 請問當你執行以下指令時， hdfs haadmin -failover MicroSys01 MicroSys02 會出現以下何者結果 ？
A. MicroSys02 會被設定為 Namenode，並且被設定成 active Namenode。
B. MicroSys01 會被設定為 Namenode，並且被設定成 active Namenode。
C. MicroSys01 會被設定為 停止模式，並且被設定成 active Namenode。
D. MicroSys01 會被設定為 停止模式，MicroSys02 被設定成 active Namenode。
E. MicroSys02 會被設定為 停止模式，MicroSys01 被設定成 active Namenode。

答案:

當使用者要儲存 310MB 檔案到 HDFS 中，HDFS 內的 block.siez 設定為 128 MB，其它都是 Hadoop 預設值? 請問以下描述那些為正確 ?(請選擇 4 個)
A. 當儲存此檔案到 HDFS 內，此檔案最終佔 1200MB 儲存空間。
B. 依儲存順序而言的第三個 block 為 64MB。
C. 依儲存順序而言的第三個 block 為 54MB。
D. 依儲存順序而言的前兩個 block 為 128MB。
E. 每個 block 都被重複儲存 3 次。
F. 當儲存時，此檔案先被分成 三個 block。
G. 每個 block 都被重複儲存 9 次。
H. 依儲存順序而言的前三個 block 為 128MB。

答案:

HDFS 分散檔案系統 – 設定與管理



1. 設定 core-site.xml

student@UB14DOS2:~/dos$ **sudo nano /opt/hadoop-2.5.2/etc/hadoop/core-site.xml**

::

```
<configuration>
  <property>
     <name>fs.default.name</name>
     <value>hdfs://nn252:8020</value>
  </property>
</configuration>
```

2. 設定 hdfs-site.xml

student@UB14DOS2:~/dos$ **sudo nano /opt/hadoop-2.5.2/etc/hadoop/hdfs-site.xml**

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:/home/ubuntu/nn</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:/home/ubuntu/dn</value>
  </property>
  <property>
    <name>fs.permissions.umask-mode</name>
    <value>002</value>
  </property>
```

```
  <!-- Secondary NameNode -->
  <property>
    <name>dfs.http.address</name>
    <value>nn252:50070</value>
  </property>
  <property>
    <name>dfs.secondary.http.address</name>
    <value>nn252:50090</value>
  </property>
  <property>
    <name>dfs.checkpoint.dir</name>
    <value>file:/home/ubuntu/sn</value>
  </property>
  <property>
    <name>dfs.checkpoint.edits.dir</name>
    <value>file:/home/ubuntu/sn</value>
  </property>

  <property>
    <name>dfs.webhdfs.enabled</name>
    <value>true</value>
  </property>
</configuration>
```

3. 設定 hadoop-env.sh

student@UB14DOS2:~/dos$ **sudo nano /opt/hadoop-2.5.2/etc/hadoop/hadoop-env.sh**

```
# The java implementation to use.
export JAVA_HOME=/opt/jdk1.7.0_71
                                ::
 The maximum amount of heap to use, in MB. Default is 1000.
export HADOOP_HEAPSIZE="256"
                                        ::
export HADOOP_CLIENT_OPTS="-Xmx256m $HADOOP_CLIENT_OPTS"
                                ::
export HADOOP_LOG_DIR=/tmp
                                ::
```

4. 設定 DataNode

student@UB14DOS2:~/dos$ **sudo nano /opt/hadoop-2.5.2/etc/hadoop/slaves**

```
dn252a
dn252b
```

5. 初始化 HDFS 系統

5-1 設定 nn252

student@UB14DOS2:~/dos$ **sudo lxc-console -n nn252**
ubuntu@nn252:~$ **mkdir nn**
ubuntu@nn252:~$ **mkdir sn**

ubuntu@nn252:~$ **hdfs namenode -format**

```
14/11/12 16:17:15 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = nn252/172.16.1.200
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.5.2
                                    ::
14/11/12 16:17:34 INFO util.ExitUtil: Exiting with status 0
14/11/12 16:17:34 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at nn252/172.16.1.200
************************************************************/
```

按 **Ctrl + a q** 離開 Console 模式

5-2 設定 dn252a 及 dn252b

student@UB14DOS2:~/dos$ **sudo lxc-console -n dn252a**
ubuntu@dn252a:~$ **mkdir dn**
按 **Ctrl + a q** 離開 Console 模式

student@UB14DOS2:~/dos$ **sudo lxc-console -n dn252b**
ubuntu@dn252b:~$ **mkdir dn**
按 **Ctrl + a q** 離開 Console 模式

6. 啟動 **HDFS** 分散檔案系統

student@UB14DOS2:~/dos$ **sudo lxc-console -n nn252**
ubuntu@nn252:~$ **start-dfs.sh**
Starting namenodes on [nn252]
nn252: starting namenode, logging to /tmp/hadoop-ubuntu-namenode-nn252.out
dn252b: starting datanode, logging to /tmp/hadoop-ubuntu-datanode-dn252b.out
dn252a: starting datanode, logging to /tmp/hadoop-ubuntu-datanode-dn252a.out
Starting secondary namenodes [nn252]
nn252: starting secondarynamenode, logging to /tmp/hadoop-ubuntu-secondarynamenode-nn252.out

7. 檢視 **HDFS** 分散檔案系統狀態

ubuntu@nn252:~$ **hdfs dfsadmin -report**
Configured Capacity: 122331324416 (113.93 GB)
Present Capacity: 93389774848 (86.98 GB)
DFS Remaining: 93389725696 (86.98 GB)
DFS Used: 49152 (48 KB)
DFS Used%: 0.00%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0

Live datanodes (2):

Name: 172.16.1.210:50010 **(dn252a)**
Hostname: dn252a
Decommission Status : Normal
::

Name: 172.16.1.211:50010 **(dn252b)**
Hostname: dn252b
Decommission Status : Normal
::

當你執行 Hadoop Cluster，並偵測監控介面時，以下那些異常狀態可以被偵測到?（請選擇兩個）
A. HDFS 空間不足
B. NameNode 斷線
C. Map 異常造成堆疊的無限循環（infinite loop）
D. DataNode 與 Cluster 多重連結

答案：

請問以下那一個異常情況會造成 HDFS的不可用 ？
Resource Manager 故障
NodeManager 故障
DataNode 故障
NameNode 故障

答案：

請問以下對 HDFS（分散式檔案系統）的描述何者為正確 ？（請選擇兩個答案）
A. 於分散式的儲存環境中，提供單一的目錄系統（Single Namespace）
B. 資料存取的特性為 Write Once Read Many 存取模式
C. 每個區塊的大小通常為 16 MB
D. 檔案複本預設是每個檔案儲存 2 份
E. Server/Client 架構

答案：

若要格式化 HDFS 的檔案系統， /usr/local/hadoop/bin/hadoop namenode －_____ 請問上述的空隔要填充以下那一個指令 ？
A. format
B. formated
C. formating
D. mat

答案：

HDFS 分散檔案系統權限設定

建立系統帳號

1. 登入 nn252

student@UB14DOS2:~/dos$ **sudo lxc-console -n nn252**

2. 產生 pig, hive 帳號

ubuntu@nn252:~$ **sudo useradd -m -s /bin/bash pig**
ubuntu@nn252:~$ **sudo useradd -m -s /bin/bash hive**

3. 建立 biguser 群組

ubuntu@nn252:~$ **sudo groupadd biguser**

4. 將 pig, hive 加入 biguser 群組

ubuntu@nn252:~$ **sudo usermod -a -G biguser pig**
ubuntu@nn252:~$ **sudo usermod -a -G biguser hive**
ubuntu@nn252:~$ **sudo usermod -a -G biguser ubuntu**

設定 HDFS 分散檔案系統權限

1. 在 nn252 虛擬主機, 設定 HDFS 分散檔案系統權限

執行以下命令前, 請確認 HDFS 分散檔案系統要啟動 (start-dfs.sh)

ubuntu@nn252:~$ **hdfs dfs -chown -R ubuntu:biguser /**
ubuntu@nn252:~$ **hdfs dfs -chmod -R 770 /**
ubuntu@nn252:~$ **hdfs dfs -mkdir /tmp**
ubuntu@nn252:~$ **hdfs dfs -mkdir /user**

2. 建立所有使用者的 HDFS 家目錄

ubuntu@nn252:~$ **hdfs dfs -mkdir /user/hive**
ubuntu@nn252:~$ **hdfs dfs -mkdir /user/pig**
ubuntu@nn252:~$ **hdfs dfs -mkdir /user/ubuntu**

3. 檢視根目錄權限

ubuntu@nn252:~$ **hdfs dfs -ls -d /**

```
Found 1 items
drwxrwx---   - ubuntu biguser          0 2014-05-24 23:14 /
```

檢視 /tmp 及 /user 目錄權限, 命令如下：

ubuntu@nn252:~$ **hdfs dfs -ls /user**

```
Found 3 items
drwxrwxr-x   - ubuntu biguser          0 2014-10-13 16:24 /user/hive
drwxrwxr-x   - ubuntu biguser          0 2014-10-13 16:24 /user/pig
drwxrwxr-x   - ubuntu biguser          0 2014-10-13 16:25 /user/ubuntu
```

在 HDFS 內，使用者看見一個檔案的權限為 rw-r--r--， 請問以下對此檔案權限的描述何者為正確 ？
A. 此檔案只能被檔案擁有者刪除 (delete)。
B. 任何人都可以刪除此檔案。
C. 此檔案不能當成 MapReduce job 來執行。
D. 只有檔案擁有者可以修改 (modify) 檔案內容。
E. 任何人都可以修改此檔案內容。

答案：

當使用者要儲存檔案於 HDFS 時， 請問使用者對此檔案具有以下那些權限 ？(請選擇三個)
A. 具有索引 (index) 權限。
B. 具有更新 (update) 權限。
C. 具有更新名稱 (rename) 權限。
D. 具有刪除 (delete) 權限。
E. 具有搬移 (move) 權限。

答案：

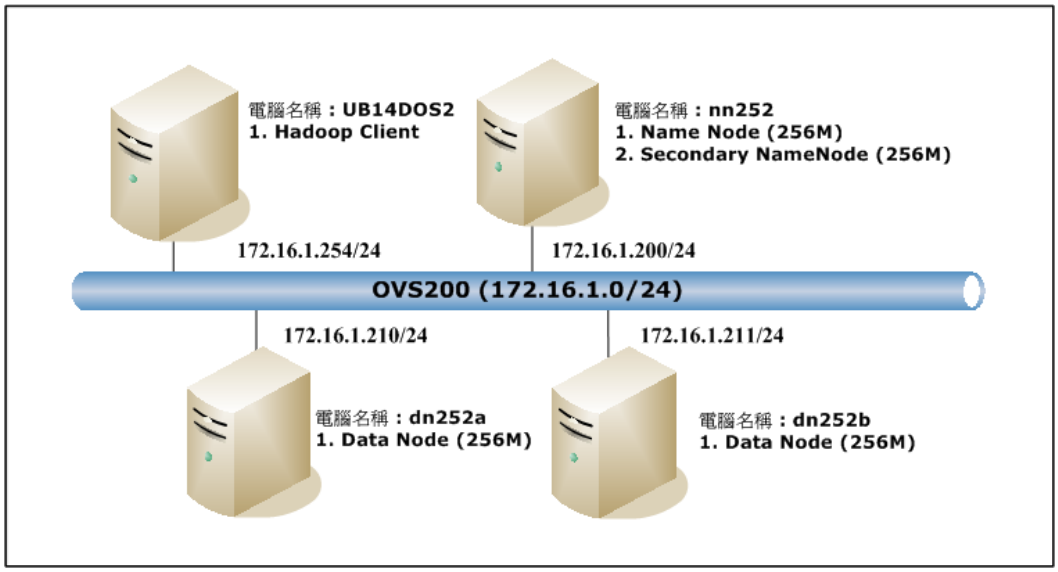
HDFS 資料出現錯誤時。 請問可以使用 Hadoop 所提供那一個指令來找出 HDFS資料的錯誤 ？
A. dskchk
B. du
C. fsck
D. map-only 偵測工具

答案：

請問以下對 Secondary NameNode 的描述何者為正確?(請選擇兩個)
A. 主要功能是幫助 NameNode 管理 Metadata 資料
B. 為 Hadoop 第二个 NameNode
C. 提供 NameNode 服务
D. NameNode 的一種管理工具

答案：

HDFS 分散檔案系統 - 基本操作命令



1. 建立目錄

ubuntu@nn252:~$ **hdfs dfs -mkdir /test**
ubuntu@nn252:~$ **hdfs dfs -ls /**

```
Found 3 items
drwxrwxr-x   - ubuntu biguser          0 2014-07-18 15:25 /test
drwxrwxr-x   - ubuntu biguser          0 2014-07-18 15:22 /tmp
drwxrwxr-x   - ubuntu biguser          0 2014-07-18 15:23 /user
```

[註] 如 Name Node 還在 **Safe Mode**, 這時會出現以下錯誤訊息：
mkdir: org.apache.hadoop.hdfs.server.namenode.SafeModeException: Cannot create directory /user. Name node is in safe mode.

2. 上載檔案

ubuntu@nn252:~$ **hdfs dfs -copyFromLocal .bash* /test**
ubuntu@nn252:~$ **hdfs dfs -ls /test**

```
Found 3 items
-rw-rw-r--   2 ubuntu biguser      16031 2014-07-18 15:26 /test/.bash_history
-rw-rw-r--   2 ubuntu biguser        220 2014-07-18 15:26 /test/.bash_logout
-rw-rw-r--   2 ubuntu biguser       3861 2014-07-18 15:26 /test/.bashrc
```

3. 顯示檔案內容

ubuntu@nn252:~$ **hdfs dfs -cat /test/.bashrc | head -n 5**

```
# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples

# If not running interactively, don't do anything
```

4. 檢視檔案儲存資訊

ubuntu@nn252:~$ **hdfs fsck /test/.bashrc -files -blocks -locations**
Connecting to namenode via http://nn252:50070
FSCK started by ubuntu (auth:SIMPLE) from /172.16.1.200 for path /test/.bashrc at Fri Jul 18 15:28:04 CST 2014
/test/.bashrc 3861 bytes, 1 block(s): OK
0. BP-1366195422-127.0.1.1-1405609369773:blk_1073741827_1003 len=3861 **repl=2**
[**172.16.1.210:50010, 172.16.1.211:50010**]
::
The filesystem under path '/test/.bashrc' is HEALTHY

5. 取回檔案

ubuntu@nn252:~$ **hdfs dfs -get /test/.bashrc /tmp/test.txt**
ubuntu@nn252:~$ **cat /tmp/test.txt | head -n 5**

~/.bashrc: executed by bash(1) for non-login shells.
see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
for examples

If not running interactively, don't do anything

6. 刪除檔案

ubuntu@nn252:~$ **hdfs dfs -rm /test/.bash***

14/07/18 15:29:44 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /test/.bash_history
14/07/18 15:29:44 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /test/.bash_logout
14/07/18 15:29:44 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /test/.bashrc

7. 刪除 **/test** 目錄

ubuntu@nn252:~$ **hdfs dfs -rm -r /test**

14/07/18 15:30:25 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /test

ubuntu@nn252:~$ **hdfs dfs -ls /**

Found 2 items
drwxrwx--- - ubuntu biguser 0 2014-05-24 23:29 /tmp
drwxrwx--- - ubuntu biguser 0 2014-05-24 23:14 /user

當你要偵錯（Debug）Map/Reduce程式時， 請問那一個指令可以例出 JAVA 程序（processes）？
A. jps
B. process
C. show
D. debug

答案：

以下對 HDFS 架構提供的資料備份（redundancy），何者描述為正確 ？（請選擇兩個）
A. 資料會被存儲為照順序的 block，除了最後一個 block 其他的 size 都相等。
B. 將資料重複儲存於不同的 DataNodes 中。
C. 使用在 DataNodes 複製時使用 RAID 架構。
D. 使用在 DataNodes 複製時使用 SAN 設備。

答案：

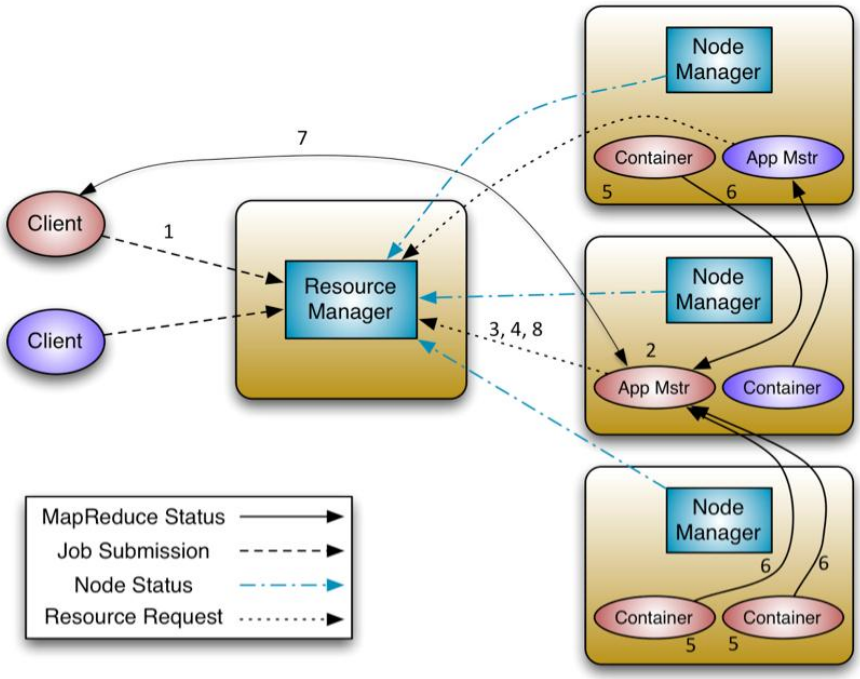
您在設定兩個 Hadoop cluster（cluster1, cluster2）， 你想要將 /home/foo/cluster1內的資料 複製到 cluster2的 /home/bar/資料夾內 請問以下何者指令可完成此工作?
A. distCp HDFS://cluster1/home/foo hdfs://cluster2/home/bar/
B. distCp cluster1:/home/foo cluster2:/home/bar/
C. Hadoop distCp cluster1:/home/foo cluster2:/home/bar/
D. hadoop distCp HDFS://cluster1/home/foo hdfs://cluster2/home/bar/

答案：

當管理員在設定 Slave node 的 cluster 為 4 x 2 TB 時， 請問建議使用以下何者方式 ？
A. RAID 0
B. JBOD
C. RAID 5
D. RAID 1+0

答案：

YARN 分散運算原理



1. A client program submits the application, including the necessary specifications to launch the application-specific **ApplicationMaster** itself.

2. The **ResourceManager** assumes the responsibility to negotiate a specified container in which to start the ApplicationMaster and then launches the ApplicationMaster.

3. The ApplicationMaster, on boot-up, registers with the ResourceManager – the registration allows the client program to query the ResourceManager for details, which allow it to directly communicate with its own ApplicationMaster.

4. During normal operation the ApplicationMaster negotiates appropriate resource containers via the resource-request protocol.

5. On successful container allocations, the ApplicationMaster launches the container by providing the container launch specification to the NodeManager. The launch specification, typically, includes the necessary information to allow the container to communicate with the ApplicationMaster itself.

6. The application code executing within the container then provides necessary information (progress, status etc.) to its ApplicationMaster via an application-specific protocol.

7. During the application execution, the client that submitted the program communicates directly with the ApplicationMaster to get status, progress updates etc. via an application-specific protocol.

8. Once the application is complete, and all necessary work has been finished, the ApplicationMaster deregisters with the ResourceManager and shuts down, allowing its own container to be repurposed.

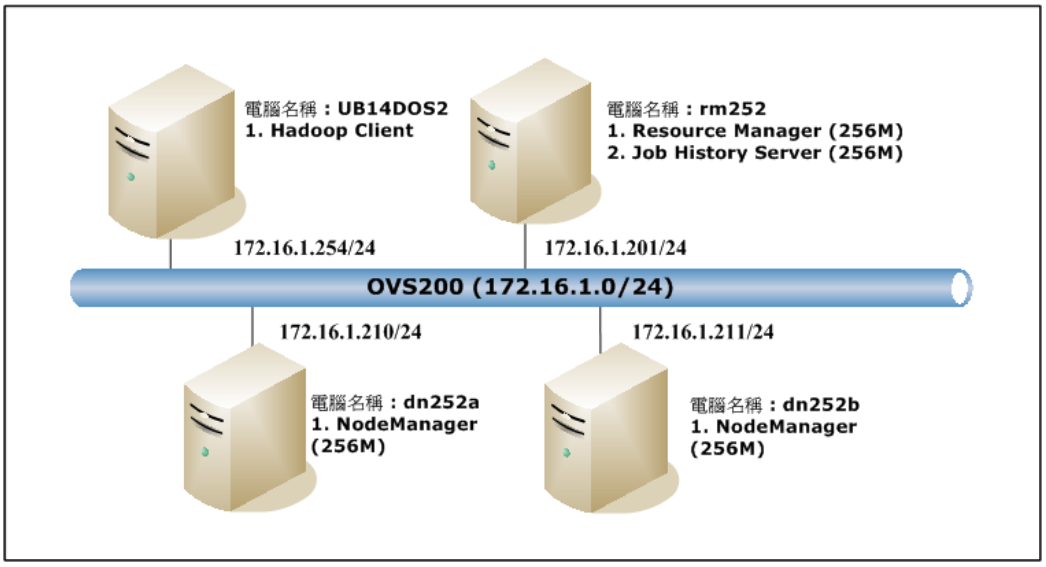
請問以下對 MapReduce 框架的描述何者為正確 ？(請選擇兩個答案)
A. Google 所發展的軟體框架。
B. 協調機器資源分配及處理的程式輸入與執行。
C. 常用於地圖資料處理
D. 可減少地圖檔案大小

答案：

在 Hadoop 環境內，管理員若要使用 JVM 執行及管理 Mapreduce 程序 (MRv1) 請問以下何者元件可以完成此工作 ？
A. Nodemanager
B. applicationmanager
C. Applicationmaster
D. Tasktracker
E. Jobtracker
F. Datanode
G. Namenode
H. Resourcemanager

答案：

YARN 分散運算系統 – 設定與管理



1. 回到 **UB14DOS2** 虛擬電腦, 設定 **mapred-site.xml**

student@UB14DOS2:~/dos$ **sudo nano /opt/hadoop-2.5.2/etc/hadoop/mapred-site.xml**

```
<?xml version="1.0"?>
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
```

2. 設定 **yarn-site.xml**

student@UB14DOS2:~/dos$ **sudo nano /opt/hadoop-2.5.2/etc/hadoop/yarn-site.xml**

```
 <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.webapp.address</name>
   <value>rm252:8088</value>
  </property>
  <property>
   <name>yarn.resourcemanager.hostname</name>
   <value>rm252</value>
  </property>
  <property>
   <name>yarn.nodemanager.local-dirs</name>
   <value>/home/ubuntu/yarn</value>
  </property>
  <property>
   <name>mapreduce.jobhistory.address</name>
   <value>rm252:10020</value>
  </property>
```

3. 設定 MapReduce,YARN 及 JobHistory

```
student@UB14DOS2:~/dos$ sudo nano /opt/hadoop-2.5.2/etc/hadoop/mapred-env.sh
                                        ::
export HADOOP_JOB_HISTORYSERVER_HEAPSIZE=256
                                        ::
# JobHistory log file path, 會自動產生目錄
export HADOOP_MAPRED_LOG_DIR="/home/ubuntu/jhslog"

student@UB14DOS2:~/dos$ sudo nano /opt/hadoop-2.5.2/etc/hadoop/yarn-env.sh
                                      ::
YARN_HEAPSIZE=256
                                      ::
# default log directory & file
export YARN_LOG_DIR=/tmp (加入這一行)
if [ "$YARN_LOG_DIR" = "" ]; then
  YARN_LOG_DIR="$HADOOP_YARN_HOME/logs"
fi                                    ::
```

4. 登入 rm252, 完成 SSH 自動登入 rm252, dn252a 及 dn252b 運算主機

```
student@UB14DOS2:~/dos$ sudo lxc-console -n rm252
ubuntu@rm252:~$ ssh rm252
The authenticity of host 'rm252 (172.16.1.201)' can't be established.
ECDSA key fingerprint is 7d:9f:23:df:47:d8:14:6d:f6:42:ae:74:0a:eb:ce:b0.
Are you sure you want to continue connecting (yes/no)? yes
                                      ::
ubuntu@rm252:~$ exit
ubuntu@rm252:~$ ssh dn252a
The authenticity of host 'dn252a (172.16.1.210)' can't be established.
ECDSA key fingerprint is 7d:9f:23:df:47:d8:14:6d:f6:42:ae:74:0a:eb:ce:b0.
Are you sure you want to continue connecting (yes/no)? yes
                                        ::
ubuntu@dn252a:~$ exit
ubuntu@rm252:~$ ssh dn252b
The authenticity of host 'dn252b (172.16.1.211)' can't be established.
ECDSA key fingerprint is 7d:9f:23:df:47:d8:14:6d:f6:42:ae:74:0a:eb:ce:b0.
Are you sure you want to continue connecting (yes/no)? yes
                                                        ::
ubuntu@dn252b:~$ exit

ubuntu@rm252:~$ mkdir jhslog
```

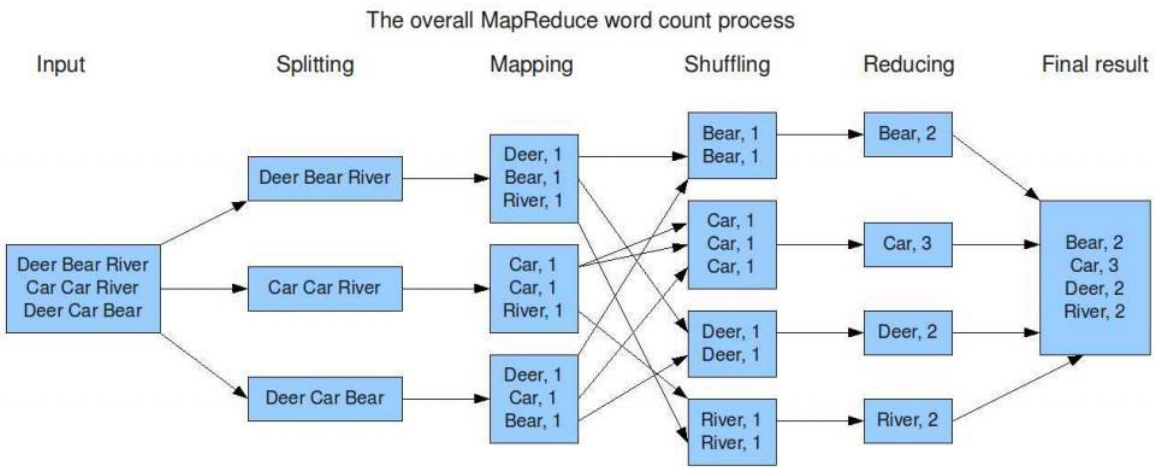
5. 啟動 **YARN** 系統

```
ubuntu@rm252:~$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /tmp/yarn-ubuntu-resourcemanager-rm252.out
dn252a: starting nodemanager, logging to /tmp/yarn-ubuntu-nodemanager-dn252a.out
dn252b: starting nodemanager, logging to /tmp/yarn-ubuntu-nodemanager-dn252b.out
```

6. 檢視 **YARN** 分散運算系統狀態

```
ubuntu@rm252:~$ yarn node  -all  -list
14/11/12 16:54:52 INFO client.RMProxy: Connecting to ResourceManager at rm252/172.16.1.201:8032
Total Nodes:2
        Node-Id         Node-State   Node-Http-Address                Number-of-Running-Containers
   dn252a:47689                    RUNNING      dn252a:8042                                    0
   dn252b:40709                    RUNNING      dn252b:8042                                    0
```

YARN 分散運算系統 – MapReduce 運算



執行以下命令, 請務必先啟動 HDFS 及 YARN 系統

1. 登入 rm252 虛擬電腦

student@UB14DOS2:~/dos$ **sudo lxc-console -n rm252**

2. 執行 MapReduce 程式

ubuntu@rm252:~$ **hadoop jar /opt/hadoop-2.5.2/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.5.2.jar pi 1 1**

```
Number of Maps  = 1
Samples per Map = 1
Wrote input for Map #0
Starting Job
14/10/14 01:24:34 INFO client.RMProxy: Connecting to ResourceManager at rm252/172.16.1.201:8032
14/10/14 01:24:37 INFO input.FileInputFormat: Total input paths to process : 1
14/10/14 01:24:37 INFO mapreduce.JobSubmitter: number of splits:1
14/10/14 01:24:41 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1413220919522_0001
14/10/14 01:24:45 INFO impl.YarnClientImpl: Submitted application application_1413220919522_0001
14/10/14 01:24:46 INFO mapreduce.Job: The url to track the job:
http://rm252:8088/proxy/application_1413220919522_0001/
14/10/14 01:24:46 INFO mapreduce.Job: Running job: job_1413220919522_0001
14/10/14 01:25:43 INFO mapreduce.Job: Job job_1413220919522_0001 running in uber mode : false
14/10/14 01:25:43 INFO mapreduce.Job:  map 0% reduce 0%
                                        ::
14/10/14 01:28:18 INFO mapred.ClientServiceDelegate: Application state is completed.
FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
Job Finished in 253.721 seconds
Estimated value of Pi is 4.00000000000000000000
```

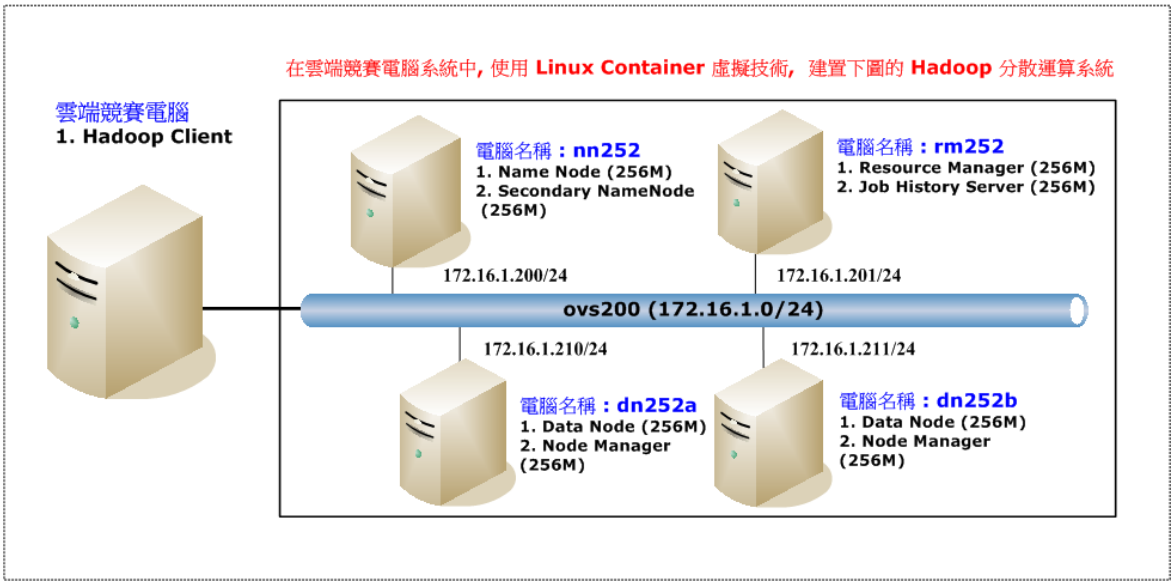
請問以下對 HBase（儲存系統）的描述何者為正確 ？(請選擇兩個答案)
A. 為 Hadoop 所使用的資料庫
B. 適用於隨機且即時的讀寫超大資料集時所使用
C. 為關聯式資料庫（relational database）
D. 只是單純的資料庫，不具程式設計能力

答案:

您已經安裝好single-node cluster，你要啟動此功能。 請問要使用以下那一個指令?
A. /usr/local/hadoop/bin/start-all.sh
B. /usr/local/hadoop/bin/start.sh
C. /usr/local/hadoop/bin/run.sh
D. /usr/local/hadoop/bin/restart.sh

答案:

Hadoop Client 系統設定



1. 在 UB14DOS2 虛擬電腦, 修改 /opt/hadoop25.sh

student@UB14DOS2:~/dos$ **sudo nano /opt/hadoop25.sh**
export JAVA_HOME**=/opt/jdk1.7.0_71/**
export HADOOP_HOME=**/opt/hadoop-2.5.2**

export PATH=**$PATH:$JAVA_HOME/bin**
export PATH=**$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin**

2. 修改 ~/.bashrc

student@UB14DOS2:~/dos$ **nano ~/.bashrc**
::
export PATH=$PATH:~/dos
. /opt/hadoop25.sh

3. 設定 /etc/Hosts

student@UB14DOS2:~/dos$ **sudo nano /etc/hosts**
127.0.0.1 localhost UB14DOS2
172.16.1.200 nn252
172.16.1.201 rm252
172.16.1.210 dn252a
172.16.1.211 dn252b

4. 離開終端機

student@UB14DOS2:~/dos$ **exit**

5. 檢視 JDK 及 Hadoop 版本

再次開啟終端機, 執行以下命令：
student@UB14DOS2:~/dos$ **java - version**
java version "1.7.0_71"
Java(TM) SE Runtime Environment (build 1.7.0_71-b14)
Java HotSpot(TM) 64-Bit Server VM (build 24.71-b01, mixed mode)

student@UB14DOS2:~/dos$ **hadoop version**
Hadoop 2.5.2
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r
2e18d179e4a8065b6a9f29cf2de9451891265cce

Hadoop Client 系統操作

第一次連接 HDFS (失敗)

student@UB14DOS2:~/dos$ hdfs dfs -ls /
ls: Permission denied: user=student, access=READ_EXECUTE, inode="/":ubuntu:biguser:drwxrwx---

設定 Hadoop Client 帳號權限

1. 登入 **nn252,** 產生 **student** 使用者帳號

student@UB14DOS2:~/dos$ **sudo lxc-console -n nn252**
ubuntu@nn252:~$ **sudo useradd -m -s /bin/bash student**
ubuntu@nn252:~$ **sudo usermod -a -G biguser student**
ubuntu@nn252:~$ **hdfs dfs -mkdir /user/student**

按 **Ctrl + a q** 離開 Console 模式

2. 連接 **HDFS** 分散檔案系統

student@UB14DOS2:~/dos$ **hdfs dfs -mkdir test**

student@UB14DOS2:~/dos$ **hdfs dfs -ls /user/student**

Found 1 items

drwxrwxr-x - student biguser 0 2014-10-15 00:50 /user/student/test

student@UB14DOS2:~/dos$ **hdfs dfs -rm -r /user/student/test**

第一次連接 YARN (失敗)

student@UB14DOS2:~/dos$ **hadoop jar /opt/hadoop-2.5.2/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.5.2.jar pi 1 1**
Number of Maps = 1
Samples per Map = 1
Wrote input for Map #0
Starting Job
14/11/12 17:25:14 INFO client.RMProxy: Connecting to ResourceManager at rm252/172.16.1.201:8032
org.apache.hadoop.security.AccessControlException: **Permission denied: user=student, access=EXECUTE, inode="/tmp/hadoop-yarn":ubuntu:biguser:drwx------**

設定 Hadoop Client 的目錄權限

1. 登入 **nn252**

student@UB14DOS2:~/dos$ **sudo lxc-console -n nn252**

2. 設定 **/tmp/hadoop-yarn** 目錄權限

ubuntu@nn252:~$ **hdfs dfs -chmod -R 770 /tmp**

按 **Ctrl + a q** 離開 Console 模式

3. 執行 **MapReduce** 程式

student@UB14DOS2:~/dos$ **hadoop jar /opt/hadoop-2.5.2/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.5.2.jar pi 1 1**

Pig 資料分析工具

安裝與設定 Pig

1. 下載 **Pig** 套件

student@UB14DOS2:~/dos$ **wget http://ftp.twaren.net/Unix/Web/apache/pig/pig-0.13.0/pig-0.13.0.tar.gz**

2. 將 **pig-0.13.0.tar.gz** 解壓縮至 **/opt** 目錄

student@UB14DOS2:~/dos$ **sudo tar -xzf pig-0.13.0.tar.gz -C /opt**

3. 設定 **PATH** 環境變數

student@UB14DOS2:~/dos$ **nano ~/.bashrc** ::

export PATH=$PATH:/opt/pig-0.13.0/bin

4. 重啟終端機

student@UB14DOS2:~/dos$ **exit**

啟動 Pig

student@UB14DOS2:~/dos$ **pig**

14/10/15 01:23:42 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL

::

2014-10-15 01:23:46,147 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://nn252:8020

2014-10-15 01:23:52,103 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS

grunt> **ls /**

hdfs://nn252:8020/tmp <dir>

hdfs://nn252:8020/user <dir>

grunt> **quit**

Pig 資料分析工具 -大專校院名錄分析 (TXT)

資料處理

1. 下載 102 年大專校院名錄檔

student@UB14DOS2:~$ **wget --no-check-certificate https://stats.moe.gov.tw/files/school/102/u1_new.txt**

2. 安裝檢測檔案編碼命令 - enca

student@UB14DOS2:~$ **sudo apt-get install enca**

3. 檢測 "大專校院名錄檔" 編碼

student@UB14DOS2:~$ **enca u1_new.txt**

Universal character set 2 bytes; UCS-2; BMP

Mixed line terminators

Byte order reversed in pairs (1,2 -> 2,1)

4. 轉換編碼

student@UB14DOS2:~$ **iconv -f UCS-2 -t utf8 u1_new.txt -o school.txt**
student@UB14DOS2:~$ **cat school.txt | head -n 5**

102學年度大專校院名錄

代碼	學校名稱	縣市名稱	地址	電話	網址	體系別
0001	國立政治大學	[38]臺北市	[116]臺北市文山區指南路二段64號	(02)29393091	http://www.nccu.edu.tw	[1]一般
0002	國立清華大學	[18]新竹市	[300]新竹市東區光復路二段101號	(03)5715131	http://www.nthu.edu.tw	[1]一般

資料分析

1. 進入Pig

student@UB14DOS2:~$ **pig**

2. 取出 大專校院名錄檔 的前二個欄位資料

grunt> copyFromLocal school.txt school.txt
grunt> school = LOAD 'school.txt' AS (sno:int,name:chararray);
grunt> dump school;

3. 依縣市分類

grunt> school = LOAD 'school.txt' AS (sno:int,name:chararray,city:chararray);
grunt> sdiv = GROUP school BY city;
grunt> describe sdiv
sdiv: {group: chararray,school: {(sno: int,name: chararray,city: chararray)}}

grunt> dump sdiv;

4. 統計各縣市大學數量

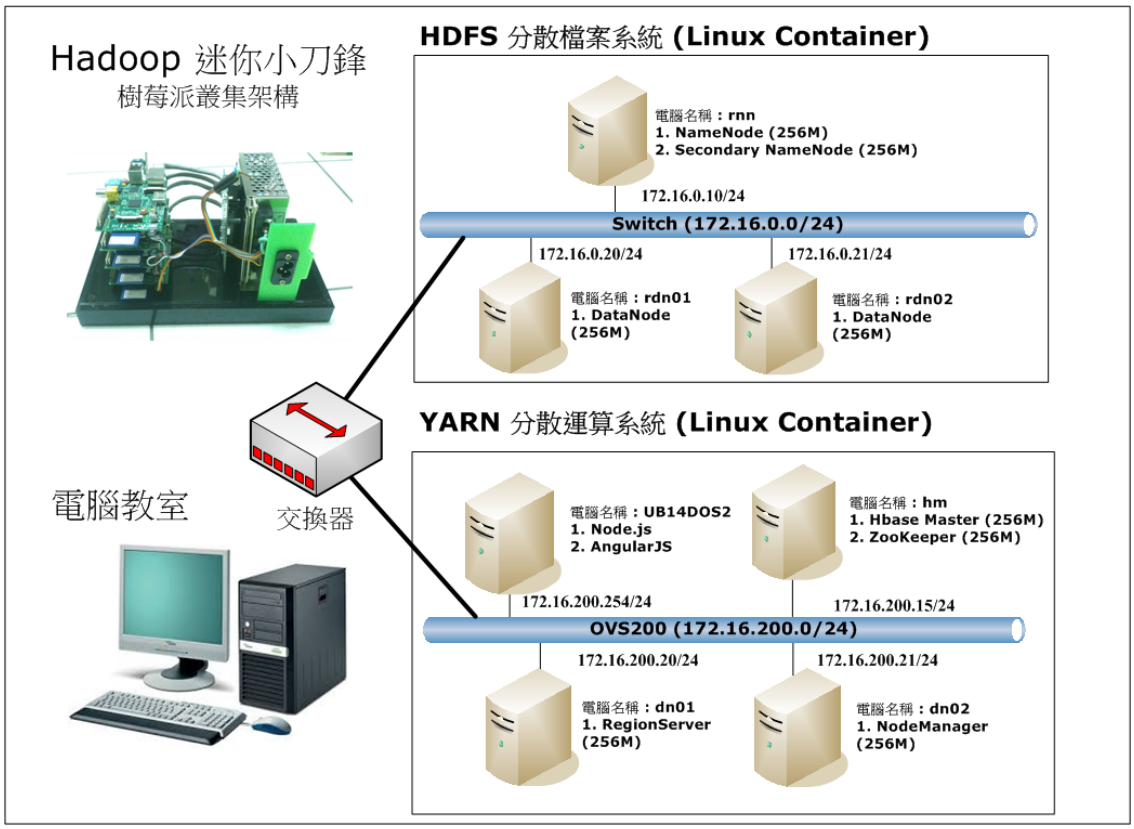
grunt> counts = foreach sdiv generate group,COUNT(school);
grunt> dump counts

5. 儲存上面命令的統計資料

grunt> store counts into 'sr';
grunt> ls
hdfs://nn.dos:8020/user/pig/excite-small.log<r 2> 208348 hdfs://nn.dos:8020/user/pig/school.txt<r 2> 21067
hdfs://nn.dos:8020/user/pig/sr <dir>

grunt> ls sr
hdfs://nn.dos:8020/user/pig/sr/_SUCCESS<r 2> 0 hdfs://nn.dos:8020/user/pig/sr/part-r-00000<r 2> 648

grunt> cat sr/part-r-00000


Hadoop 迷你小刀鋒
樹莓派叢集架構
HDFS 分散檔案系統 (Linux Container)
電腦名稱：rnn
1. NameNode (256M)
2. Secondary NameNode (256M)
172.16.0.10/24
Switch (172.16.0.0/24)
172.16.0.20/24
172.16.0.21/24
電腦名稱：rdn01
1. DataNode
(256M)
電腦名稱：rdn02
1. DataNode
(256M)
YARN 分散運算系統 (Linux Container)
電腦名稱：UB14DOS2
1. Node.js
2. AngularJS
電腦名稱：hm
1. Hbase Master (256M)
2. ZooKeeper (256M)
172.16.200.254/24
172.16.200.15/24
OVS200 (172.16.200.0/24)
172.16.200.20/24
172.16.200.21/24
電腦名稱：dn01
1. RegionServer
(256M)
電腦名稱：dn02
1. NodeManager
(256M)
電腦教室
交換器