目錄

1. Application Starter - Cloudera 產品的 CaaS 規格有哪些？

答：支援的 OS 為 Centos 6.4，其規格如下：

- Application Starter - Hadoop 0.20 (Cloudera CDH3u5)：
  - 支援 Big Data 標準型 S、M、L、XL。
- Application Starter - Hadoop 2.00 (Cloudera CDH4.3)
  - 支援 Big Data 標準型 S、M、L、XL。

2. Application Starter - Cloudera 預載的服務有哪些？

答：

- Application Starter - Hadoop 0.20 (Cloudera CDH3u5)主要預載 Apache Hadoop 0.20 與 Hbase 0.90.6 軟體，其他預載軟體說明描述如下：

服務名稱	版本資訊	網頁
HDFS	0.20.2+923.421-1	http://Master_IP:50070
MapReduce	0.20.2+923.421	http://Master_IP:50030
HBase	0.90.6+84.73-1	http://Mster_IP:60010
Ganglia	3.1.7-3	http://Master_IP/ganglia/
Oozie	2.3.2+27.23-1	http://Master_IP:11000/oozie
Hue	1.2.0.0+114.42-1	http://Master_IP:8088
Zookeeper	3.3.5+19.5-1	
Hive	0.7.1+42.56-2	
HDFS over FTP	CDH3u5	ftp://Master_IP:2222 (預設帳密：hdfs / 密碼可參考本文件使用說明 10)
Flume	0.9.4+25.46-1	

- Application Starter - Hadoop 2.00 (Cloudera CDH4.3)主要預載 Apache Hadoop 2.00 與 Hbase 0.94.6 軟體，其他預載軟體說明描述如下：

服務名稱	版本資訊	網頁
HDFS	2.0.0+1357-1	http://Master_IP:50070
MapReduce	2.0.0+1357-1	http://Master_IP:50030
HBase	0.94.6+96-1	http://Mster_IP:60010
Ganglia	3.1.7-3	http://Master_IP/ganglia/
Oozie	3.3.2+49-1	http://Master_IP:11000/oozie

Hue	2.3.0+136-1	http://Master_IP:8088
Zookeeper	3.4.5+19-1	
Hive	0.10.0+121-1	
HDFS over FTP	CDH4.3	ftp://Master_IP:2222 (預設帳密：hdfs / 密碼可參考本文件使用說明 10)
Flume	1.3.0+159-1	
ClouderaSearch(Solr)	4.4.0+69-1	

註 1: Master_IP 為 Hadoop 平台中 master 主機的 IP，主機資訊的取得方法可參考 Big Data 叢集安裝服務使用說明手冊的 Hadoop 平台環境說明章節。
註 2:各服務詳細資訊可參考官方文件：

- CDH3u5

http://www.cloudera.com/content/support/en/documentation/cdh3-documentation/cdh3-documentation-v3u5.html

- CDH4.3

http://www.cloudera.com/content/support/en/documentation/cdh4-documentation/cdh4-documentation-v4-3-0.html

3. Application Starter - Cloudera 預載的服務安裝於何處？

答：各服務的安裝目錄及參數的設定檔位置描述如下：

- CDH3u5

服務名稱	安裝目錄	設定檔位置
HDFS	/usr/lib/hadoop	/etc/hadoop
MapReduce	/usr/lib/hadoop	/etc/hadoop
HBase	/usr/lib/hbase	/etc/hbase
Ganglia	/etc/ganglia	/etc/ganglia
Oozie	/usr/lib/oozie	/etc/oozie
Hue	/use/share/hue	/etc/hue
Zookeeper	/usr/lib/zookeeper	/etc/zookeeper
Hive	/usr/lib/hive	/etc/hive
HDFS over FTP	/usr/lib/hdfs-over-ftp	/usr/lib/hdfs-over-ftp
Flume	/usr/lib/flume	/etc/flume

- CDH4.3

服務名稱	安裝目錄	設定檔位置
HDFS	/usr/lib/hadoop-hdfs/	/etc/hadoop
MapReduce	/usr/lib/hadoop-0.20-mapreduce/	/etc/hadoop
HBase	/usr/lib/hbase	/etc/hbase
Ganglia	/etc/ganglia	/etc/ganglia
Oozie	/usr/lib/oozie	/etc/oozie
Hue	/use/share/hue	/etc/hue
Zookeeper	/usr/lib/zookeeper	/etc/zookeeper
Hive	/usr/lib/hive	/etc/hive
HDFS over FTP	/usr/lib/hdfs-over-ftp	/usr/lib/hdfs-over-ftp
Flume	/usr/lib/ flume-ng	/etc/flume
ClouderaSearch(Solr)	/usr/lib/solr	/etc/solr

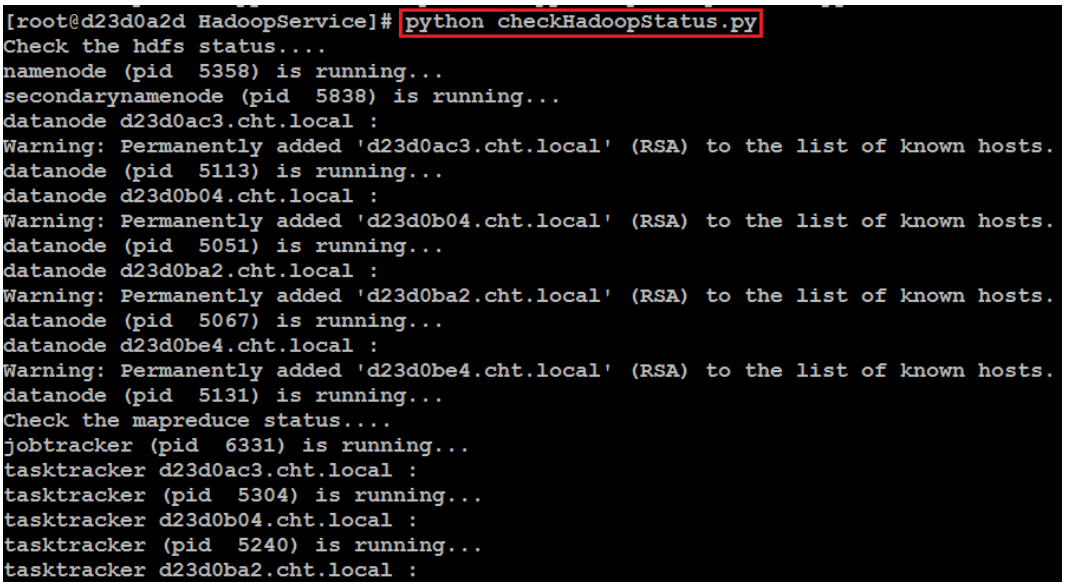
4. 如何查看 Apache HDFS、MapReduce、HBase、Ganglia、Solr 服務狀態？

答：

(1) 以 root 身分登入 master 主機，解壓縮檢查服務工具包「tar zxvf /root/manage_cluster.tgz」

(2) 執行查看服務狀態程式

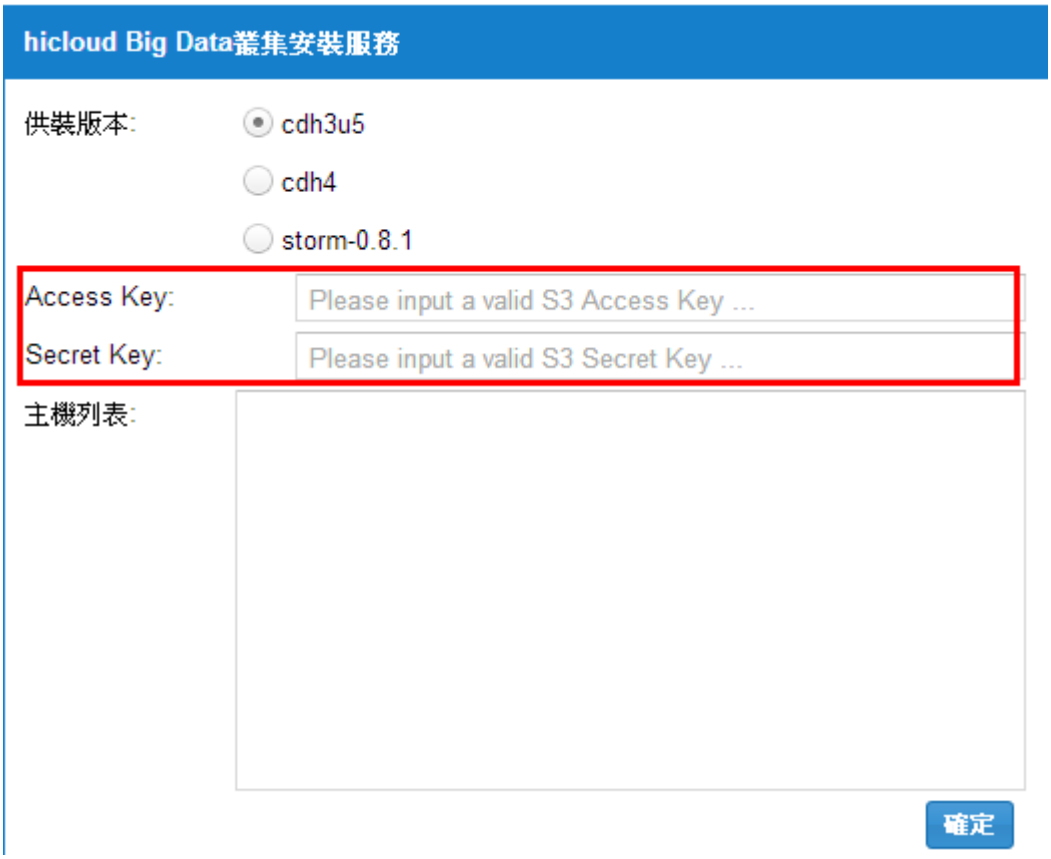
i. 執行「python /root/manage_cluster/HadoopService/checkHadoopStatus.py」



5. 如何整合使用 hicloud S3？

答：有兩種方式提供給使用者使用 hicloud S3

1. 於 Hadoop 平台服務安裝時加入 S3 相關資訊
   在供裝 Hadoop 平台時預先填入 S3 Access Key 與 Secret Key，平台安裝完成後即可使用 S3 服務



2. 手動登入 Hadoop 平台進行設定

以 root 身份分別登入 Hadoop 虛擬主機，並依照下列步驟進行設定

(1) 修改**所有 Hadoop 平台主機**的 S3 設定，參考指令如下：

vim /etc/hadoop/conf/core-site.xml

```
<property>
   <name>fs.s3n.awsAccessKeyId</name>
   <value>於此處填上系統配發給您的 access key</value>
</property>
<property>
   <name>fs.s3n.awsSecretAccessKey</name>
   <value>於此處填上系統配發給您的 secrect key </value>
</property>
```

完成上述步驟後您就可以經由 Hadoop 指令存取 hicloud S3 的資源。下面列出常用的指令範例：

(1) 在 S3 建立目錄
hadoop fs –mkdir s3n://**S3 bucket 名稱**/**目錄名稱**

(2) 上傳本機檔案至 S3 目錄
hadoop fs –put **本機檔案完整路徑** s3n://**S3 bucket 名稱**/**目錄名稱**/**檔案名稱**

(3) 下載 S3 檔案至本機
hadoop fs –get s3n://**S3 bucket 名稱**/**目錄名稱**/**檔案名稱** **本機檔案完整路徑**

(4) 刪除 S3 檔案
hadoop fs –rm s3n://**S3 bucket 名稱**/**目錄名稱**/**檔案名稱**

(5) 刪除 S3 目錄
hadoop fs –rm s3n://**S3 bucket 名稱**/**目錄名稱**

(6) 將多個檔案從 S3 複製到 HDFS
hadoop fs -cp s3n://**S3 bucket 名稱**/**目錄名稱**/* **HDFS 目錄名稱**/

(7) 將多個檔案從 HDFS 複製到 S3
hadoop fs -cp **HDFS 目錄名稱**/* s3n://**S3 bucket 名稱**/**目錄名稱**/

(8) 利用 distcp 進行 S3→HDFS 檔案複製平行處理
hadoop distcp -p s3n://**S3 bucket 名稱**/**目錄名稱**/ hdfs://**NN 主機名稱**/**HDFS 目錄名稱**/

(9) 利用 distcp 進行 HDFS→S3 檔案複製平行處理
hadoop distcp -p hdfs://**NN 主機名稱**/**HDFS 目錄名稱**/ s3n://**S3 bucket 名稱**/**目錄名稱**/

(10) 您可以輸入 hadoop fs –help 以及 hadoop distcp 取得完整說明

6. 如何檢查 Apache HDFS、MapReduce、HBase 服務是否正常運作？

答：

(1) 以 root 身分登入 master 主機

(2) 解壓縮檢查服務工具包「tar zxvf /root/HDP-TEST.tgz」，確定服務狀態為啟動，執行檢查服務程式「sh /root/HDP-TEST/hadoop-test-main.sh」，檢查結果將會顯示在螢幕上。

```
[root@d23d0a2d ~]# sh HDP-TEST/hadoop-test-main.sh
Security is not enable : start normal function testing
Starting hadoop function testing...
Please input the number of Data Node (HDFS) in Cluster
4   輸入slave個數
Please input the number of Task Tracker (MapReduce) in Cluster
4   輸入slave個數
Please input the number of Region Server (HBase) in Cluster
4   輸入slave個數
You input: 4 Data Nodes, 4 Task Trackers, 4 Region Servers
Press any key to begin test...
```

```
Start hdfs-t01-1....Checking Available Data Nodes..........[pass]
Start hdfs-t01-2....Creating Directory.....................[pass]
Start hdfs-t01-3....Modifying HDFS Directory...............[pass]
Start hdfs-t01-4....Creating HDFS File.....................[pass]
Start hdfs-t01-5....Modifying HDFS File....................[pass]
Start hdfs-t01-6....Removing HDFS File.....................[pass]
Start hdfs-t01-7....Removing HDFS Directory................[pass]
Start mr-t02-1......Checking Available Task Trackers.......[pass]
Start mr-t02-2......Running MapReduce Benchmark............[pass]
Start hbase-t03-1...Checking Active Servers................[pass]
Start hbase-t03-2...Creating HBase Table...................[pass]
Start hbase-t03-3...Putting HBase Record...................[pass]
Start hbase-t03-4...Getting HBase Record...................[pass]
Start hbase-t03-5...Scaning HBase Record...................[pass]
Start hbase-t03-6...Adding HBase Column Family.............[pass]
Start hbase-t03-7...Removing HBase Column Family...........[pass]
Start hbase-t03-8...Removing HBase Record..................[pass]
Start hbase-t03-9...Removing HBase Table...................[pass]
```

(3) 檢查項目如下表：

編號	服務	名稱	描述
1-1	HDFS	Checking Available Data Nodes	檢查 Data Node 個數是否如預期，通過顯示 pass，沒通過則顯示 failed。
1-2	HDFS	Creating Directory	可否在 HDFS 上建立目錄，通過顯示 pass，沒通過則顯示 failed。
1-3	HDFS	Modifying HDFS Directory	可否修改 HDFS 上的目錄，通過顯示 pass，沒通過則顯示 failed。
1-4	HDFS	Creating HDFS File	可否在 HDFS 上建立檔案，通過顯示 pass，沒通過則顯示 failed。
1-5	HDFS	Modifying HDFS File	可否修改 HDFS 上的檔案，通過顯示 pass，沒通過則顯示 failed。
1-6	HDFS	Removing HDFS File	可否移除 HDFS 上的檔案，通過顯示 pass，沒通過則顯示 failed。

1-7	HDFS	Removing HDFS Directory	可否移除 HDFS 上的目錄，通過顯示 pass，沒通過則顯示 failed。
2-1	MapReduce	Checking Available Task Trackers	檢查 Task Tracker 個數是否如預期，通過顯示 pass，沒通過則顯示 failed。
2-2	MapReduce	Running MapReduce Benchmark	是否可以順利執行 MapReduce job，通過顯示 pass，沒通過則顯示 failed。
3-1	HBase	Checking Active Servers	檢查 Region server 個數是否如預期，通過顯示 pass，沒通過則顯示 failed。
3-2	HBase	Creating HBase Table	是否可以建立 HBase table，通過顯示 pass，沒通過則顯示 failed。
3-3	HBase	Putting HBase Record	是否可以 put HBase record，通過顯示 pass，沒通過則顯示 failed。
3-4	HBase	Getting HBase Record	是否可以 get HBase record，通過顯示 pass，沒通過則顯示 failed。
3-5	HBase	Scaning HBase Record	是否可以 scan HBase record，通過顯示 pass，沒通過則顯示 failed。
3-6	HBase	Adding HBase Column Family	是否可以增加 HBase column family，通過顯示 pass，沒通過則顯示 failed。
3-7	HBase	Removing HBase Column Family	是否可以移除 HBase column family，通過顯示 pass，沒通過則顯示 failed。
3-8	HBase	Removing HBase Record	是否可以移除 HBase record，通過顯示 pass，沒通過則顯示 failed。
3-9	HBase	Removing HBase Table	是否可以移除 HBase table，通過顯示 pass，沒通過則顯示 failed。

7. 如何關閉 Apache HDFS、MapReduce、HBase、Ganglia、Solr 服務？

答：

(1) 以 root 身分登入 master 主機，解壓縮檢查服務工具包「tar zxvf /root/

manage_cluster.tgz」

(2) 執行停止服務程式，執行時須輸入欲停止的服務名稱(HBase=hbase、MapReduce=mapred、HDFS=hdfs、Ganglia=ganglia、Solr=solr)，「python /root/manage_cluster/HadoopService/stopHadoopService.py <hbase/mapred/hdfs>」，服務停止順序必須為 HBase→MapReduce→HDFS。

(3) 執行結果可參考如下：

```
[root@d23d0a2d HadoopService]# python stopHadoopService.py hbase
Stop the Hbase service....
stopping regionserver d23d0ac3.cht.local :
stopping regionserver...
stopping regionserver d23d0b04.cht.local :
stopping regionserver...
stopping regionserver d23d0ba2.cht.local :
stopping regionserver....
stopping regionserver d23d0be4.cht.local :
stopping regionserver.
Stopping hbasemaster d23d0a2d.cht.local :
stopping master.
Stopping zookeeper0 :
JMX enabled by default
Using config: /etc/zookeeper/zoo.cfg
Stopping zookeeper ... STOPPED
```

```
[root@d23d0a2d HadoopService]# python stopHadoopService.py mapred
Stop the mapreduce service....
Stopping tasktracker d23d0ac3.cht.local :
Stopping Hadoop tasktracker daemon (hadoop-tasktracker): stopping tasktracker
Stopping tasktracker d23d0b04.cht.local :
Stopping Hadoop tasktracker daemon (hadoop-tasktracker): stopping tasktracker
Stopping tasktracker d23d0ba2.cht.local :
Stopping Hadoop tasktracker daemon (hadoop-tasktracker): stopping tasktracker
Stopping tasktracker d23d0be4.cht.local :
Stopping Hadoop tasktracker daemon (hadoop-tasktracker): stopping tasktracker
Stopping jobtracker d23d0a2d.cht.local :
Stopping Hadoop jobtracker daemon (hadoop-jobtracker): stopping jobtracker
```

```
[root@d23d0a2d HadoopService]# python stopHadoopService.py hdfs
Stop the hdfs service....
Stopping datanode d23d0ac3.cht.local :
Stopping Hadoop datanode daemon (hadoop-datanode): stopping datanode
datanode is stopped
Stopping datanode d23d0b04.cht.local :
Stopping Hadoop datanode daemon (hadoop-datanode): stopping datanode
datanode is stopped
Stopping datanode d23d0ba2.cht.local :
Stopping Hadoop datanode daemon (hadoop-datanode): stopping datanode
datanode is stopped
Stopping datanode d23d0be4.cht.local :
Stopping Hadoop datanode daemon (hadoop-datanode): stopping datanode
datanode is stopped
Stopping namenode d23d0a2d.cht.local :
Stopping Hadoop namenode daemon (hadoop-namenode): stopping namenode
Stopping secondarynamenode d23d0a2d.cht.local :
Stopping Hadoop secondarynamenode daemon (hadoop-secondarynamenode): stopping secondarynamenode
```

註 1: Solr(cloudera-Search)僅 CDH4.3 支援。

8. 如何啟動 Apache HDFS、MapReduce、HBase、 Ganglia、Solr 服務？

答：

(1) 以 root 身分登入 master 主機，解壓縮檢查服務工具包「tar zxvf /root/manage_cluster.tgz」

(2) 執行啟動服務程式，執行時須輸入欲啟動的服務名稱(HDFS=hdfs、MapReduce=mapred、HBase=hbase、Ganglia=ganglia、Solr=solr)，「python /root/manage_cluster/HadoopService/startHadoopService.py <hdfs/mapred/hbase>」，服務停止順序必須為 HDFS→MapReduce→HBase。

(3) 執行結果可參考如下：

```
[root@d23d0a2d HadoopService]# python startHadoopService.py hdfs
Start the hdfs service....
Starting namenode d23d0a2d.cht.local :
Starting Hadoop namenode daemon (hadoop-namenode): starting namenode, logging to /usr/lib/hadoop-0.20/logs/hadoop-hadoop-namenode-d23d0a2d.out
Starting secondarynamenode d23d0a2d.cht.local :
Starting Hadoop secondarynamenode daemon (hadoop-secondarynamenode): starting secondarynamenode, logging to /usr/lib/hadoop-0.20/logs/hadoop-hadoop-secondarynamenode-d23d0a2d.out
Starting datanode d23d0ac3.cht.local :
Starting Hadoop datanode daemon (hadoop-datanode): starting datanode, logging to /usr/lib/hadoop-0.20/logs/hadoop-hadoop-datanode-d23d0ac3.out
datanode (pid  8896) is running...
Starting datanode d23d0b04.cht.local :
Starting Hadoop datanode daemon (hadoop-datanode): starting datanode, logging to /usr/lib/hadoop-0.20/logs/hadoop-hadoop-datanode-d23d0b04.out
datanode (pid  8831) is running...
Starting datanode d23d0ba2.cht.local :
Starting Hadoop datanode daemon (hadoop-datanode): starting datanode, logging to /usr/lib/hadoop-0.20/logs/hadoop-hadoop-datanode-d23d0ba2.out
datanode (pid  8861) is running...
Starting datanode d23d0be4.cht.local :
Starting Hadoop datanode daemon (hadoop-datanode): starting datanode, logging to /usr/lib/hadoop-0.20/logs/hadoop-hadoop-datanode-d23d0be4.out
datanode (pid  8915) is running...
```

```
[root@d23d0a2d HadoopService]# python startHadoopService.py mapred
Start the mapreduce service....
Starting jobtracker d23d0a2d.cht.local :
Starting Hadoop jobtracker daemon (hadoop-jobtracker): starting jobtracker, logging to /usr/lib/hadoop-0.20/logs/hadoop-hadoop-jobtracker-d23d0a2d.out
Starting tasktrackerd23d0ac3.cht.local :
Starting Hadoop tasktracker daemon (hadoop-tasktracker): starting tasktracker, logging to /usr/lib/hadoop-0.20/logs/hadoop-hadoop-tasktracker-d23d0ac3.out
Starting tasktrackerd23d0b04.cht.local :
Starting Hadoop tasktracker daemon (hadoop-tasktracker): starting tasktracker, logging to /usr/lib/hadoop-0.20/logs/hadoop-hadoop-tasktracker-d23d0b04.out
Starting tasktrackerd23d0ba2.cht.local :
Starting Hadoop tasktracker daemon (hadoop-tasktracker): starting tasktracker, logging to /usr/lib/hadoop-0.20/logs/hadoop-hadoop-tasktracker-d23d0ba2.out
Starting tasktrackerd23d0be4.cht.local :
Starting Hadoop tasktracker daemon (hadoop-tasktracker): starting tasktracker, logging to /usr/lib/hadoop-0.20/logs/hadoop-hadoop-tasktracker-d23d0be4.out
```

```
[root@d23d0a2d HadoopService]# python startHadoopService.py hbase
Start the Hbase service....
Starting zookeeper0 :
Starting hbasemaster d23d0a2d.cht.local :
starting master, logging to /var/log/hbase/hbase-hbase-master-d23d0a2d.out
Starting regionserver d23d0ac3.cht.local :
starting regionserver, logging to /var/log/hbase/hbase-hbase-regionserver-d23d0ac3.out
Starting regionserver d23d0b04.cht.local :
starting regionserver, logging to /var/log/hbase/hbase-hbase-regionserver-d23d0b04.out
Starting regionserver d23d0ba2.cht.local :
starting regionserver, logging to /var/log/hbase/hbase-hbase-regionserver-d23d0ba2.out
Starting regionserver d23d0be4.cht.local :
starting regionserver, logging to /var/log/hbase/hbase-hbase-regionserver-d23d0be4.out
```

註 1: Solr(cloudera-Search)僅 CDH4.3 支援。

9. 如何設定各服務防火牆規則？

答：主機作業系統內的防火牆預設為關閉，為了安全考量，建議利用防火牆設定工具，設定防火牆規則，設定步驟如下。

(1) 以 root 身分登入 master 主機，解壓縮檢查服務工具包「tar zxvf /root/manage_cluster.tgz」

(2) 執行建立防火牆規則程式「python /root/manage_cluster/FW/genFWrule.py」，防火牆配置只針對平台主機做設定，如需要加入平台外的主機，必須將規則加入到「/root/firewall/iptables.allow」。

```
[root@d23d0a2d FW]# cd /root/manage_cluster/FW/
[root@d23d0a2d FW]# python genFWrule.py
FireWall rule generated!!
You need to add your IP on
/root/fireWall/iptables.allow before starting firewall
```

(3) 執行佈署防火牆設定程式「python /root/manage_cluster/FW/syncFireWall.py」。

```
[root@d23d0a2d FW]# python syncFireWall.py
deploy firewall on d23d0ac3.cht.local
startSlaveFireWall.sh                                 100% 1713     1.7KB/s   00:00
iptables.allow                                        100%  364     0.4KB/s   00:00
startMasterFireWall.sh                                100% 3967     3.9KB/s   00:00
deploy firewall on d23d0b04.cht.local
startSlaveFireWall.sh                                 100% 1713     1.7KB/s   00:00
iptables.allow                                        100%  364     0.4KB/s   00:00
startMasterFireWall.sh                                100% 3967     3.9KB/s   00:00
deploy firewall on d23d0ba2.cht.local
startSlaveFireWall.sh                                 100% 1713     1.7KB/s   00:00
iptables.allow                                        100%  364     0.4KB/s   00:00
startMasterFireWall.sh                                100% 3967     3.9KB/s   00:00
deploy firewall on d23d0be4.cht.local
startSlaveFireWall.sh                                 100% 1713     1.7KB/s   00:00
iptables.allow                                        100%  364     0.4KB/s   00:00
startMasterFireWall.sh                                100% 3967     3.9KB/s   00:00
```

(4) 執行啟動平台防火牆程式「python /root/manage_cluster/FW/startClusterFW.py」。

```
[root@d23d0a2d FW]# python startClusterFW.py
Starting firewall on d23d0ac3.cht.local
iptables: Saving firewall rules to /etc/sysconfig/iptables: [  OK  ]
Starting firewall on d23d0b04.cht.local
iptables: Saving firewall rules to /etc/sysconfig/iptables: [  OK  ]
Starting firewall on d23d0ba2.cht.local
iptables: Saving firewall rules to /etc/sysconfig/iptables: [  OK  ]
Starting firewall on d23d0be4.cht.local
iptables: Saving firewall rules to /etc/sysconfig/iptables: [  OK  ]
Starting firewall on d23d0a2d.cht.local
iptables: Saving firewall rules to /etc/sysconfig/iptables:[  OK  ]
```

(5) master/slave 主機防火牆規則如下

- master 主機

Service	Chain	Proto	Interface	Source port	Destination port	Target
SSH	INPUT	TCP	eth0	1024:65534	22	ACCEPT

HDFS	INPUT	TCP	eth0	1024:6553	50070	ACCEPT
MapReduce	INPUT	TCP	eth0	1024:6553	50030	ACCEPT
HBase	INPUT	TCP	eth0	1024:6553	60010	ACCEPT
Hive Metastore	INPUT	TCP	eth0	1024:6553	9038	ACCEPT
Hive Server	INPUT	TCP	eth0	1024:6553	10000	ACCEPT
Zookeeper	INPUT	TCP	eth0	1024:6553	2181	ACCEPT
Module	Chain	State				Target
state	INPUT	NEW, INVALID				REJECT

- slave 主機

Service	Chain	Proto	Interface	Source port	Destination port	Target
HDFS	INPUT	TCP	eth0	1024:6553	50075	ACCEPT
MapReduce	INPUT	TCP	eth0	1024:6553	50060	ACCEPT
HBase	INPUT	TCP	eth0	1024:6553	60020, 60030	ACCEPT
Module	Chain	State				Target
state	INPUT	NEW, INVALID				REJECT

10. 如何取得 Hadoop 平台服務相關密碼？

答：各項服務的密碼查詢方式如下

■ oozie

/etc/oozie/conf/oozie-site.xml 或 /etc/oozie/oozie-site.xml

參數 oozie.service.StoreService.jdbc.password 值即為 oozie 密碼

■ hive

/etc/hive/conf/hive-site.xml

參數 javax.jdo.option.ConnectionPassword 值即為 hive 密碼

■ mysql

密碼檔置於 master 主機/root/mysql.pw，解密方式:

i. 以 root 身分登入 master 主機

ii. openssl enc -aes-256-cbc -in /root/mysql.pw -d

iii. 輸入 Master 主機登入密碼後, mysql root 密碼顯示於 console

```
openssl enc -aes-256-cbc -in /root/mysql.pw -d
enter aes-256-cbc decryption password:
1QC9KJRR
```

- hdfsOverFtp

密碼檔置於 master 主機/root/hdfsOverFtp.pw，解密方式:

i. 以 root 身分登入 master 主機
ii. 執行 openssl enc -aes-256-cbc -in /root/hdfsOverFtp.pw -d
iii. 輸入 Master 主機登入密碼後, ftp 密碼顯示於 console

```
openssl enc -aes-256-cbc -in /root/ hdfsOverFtp.pw-d
enter aes-256-cbc decryption password:
2EC4KJII
```

11. 如何新增 Data Node 至既有的 Hadoop 平台？

答：請進入「hicloud CaaS 雲運算」申裝欲新增至既有 Hadoop 平台的主機，產品規格處「1-1 作業系統類型」選擇 Linux，「1-2 映像檔」選擇與叢集相對應的 hadoop 版本 Application Starter - Hadoop 0.20 (Cloudera CDH3u5)或是 Application Starter - Hadoop 2.00 (Cloudera CDH4.3)，開機後自行登入主機，並執行以下步驟

- Hadoop 服務

i. 以 root 身分登入 master 主機
ii. 新增欲加入節點 IP 至/etc/hadoop/conf/includes
iii. 更新 namenode 的 datanode 清單，執行以下指令
sudo –u hdfs hadoop dfsadmin –refreshNodes
iv. 新增欲加入節點 IP 至/etc/hadoop/conf/slaves
v. 將修正後的 hadoop 設定檔同步至所有節點(包含新增節點)，**同時將/etc/hosts 檔案同步至新增節點中**
vi. 以 root 身分連線至**新增節點**並行以下指令

```
alternatives --install /etc/hadoop-0.20/conf hadoop-0.20-conf
/etc/hadoop-0.20/hadoop 60
mkdir -p /opt/hadoop/hdfs
mkdir -p /opt/hadoop/mapred
mkdir -p /var/run/hadoop-0.20/
mkdir -p /var/lib/hadoop-0.20/cache/
chown -R hdfs:hadoop /opt/hadoop
chown -R mapred:mapred /opt/hadoop/mapred
chown -R hdfs:hadoop /var/run/hadoop-0.20/
chgrp -R hadoop /var/lib/hadoop-0.20/cache/
chmod -R 777 /var/lib/hadoop-0.20/cache/
```

vii. 啟動 datanode 服務
service hadoop-0.20-datanode start

viii. 重新啟動 mapreduce 服務
i. 可參考本文件使用說明 8
ii. 可透過以下指令避免每次操作皆須輸入密碼
ssh-copy-id -i ~/.ssh/id_dsa.pub 新增節點 IP

ix. 檢查新增節點是否有出現在 UI 網頁中 ((http://Master_IP:50070) (http://Master_IP:50030))

x. 以 root 身分登入 master 主機，並執行以下指令進行 hdfs balance
sudo –u hdfs hadoop balancer

可參考 Hadoop FAQ「I have a new node I want to add to a running Hadoop cluster; how do I start services on just one node?」的說明，將主機新增至既有的 Hadoop 平台。

- Ganglia 監控服務
  i. 以 root 身分登入 master 主機
  ii. 複製 master 主機中的 gmond 設定檔至新增節點

```
scp /etc/ganglia/gmond.conf   $NewAddNodeIP:/etc/ganglia/
```

  iii. 登入新增主機並執行下列指令，啟用 gmond 服務

```
chkconfig --add gmond
chkconfig gmond on
service gmond start
```

- 補充 : passwordless 設定
  i. 以 root 身分登入 master 主機，執行以下指令，輸入新增節點 root 密碼後即可完成 master 主機到新增節點 ssh passwordless 設定

```
ssh-copy-id -i ~/.ssh/id_rsa.pub $NewAddNodeIP
```

12. 如何從既有的 Hadoop 平台移除 Data Node？

答：須注意移除節點會因 Hadoop 平台所存的資料量影響移除節點所花費的時間，移除節點執行步驟如下

i. 以 root 身分登入 master 主機
ii. 新增欲移除節點 IP 至/etc/hadoop/conf/excludes
iii. 重新啟動 mapreduce 服務
  i. 可參考本文件使用說明 8
iv. 更新 namenode 的 datanode 清單，執行以下指令

sudo –u hdfs hadoop dfsadmin –refreshNodes

v. 到 hdfs UI(http://Master_IP:50070)查看移除節點狀態是否為 **Decommission In Progress**

vi. 當節點狀態顯示 **Decommissioned**，可以將移除集點關機退租 (可能需花費相當時間)

vii. 從/etc/hadoop/conf/includes 移除節點資訊(master 主機)

viii. 更新 namenode 的 datanode 清單，執行以下指令

sudo –u hdfs hadoop dfsadmin –refreshNodes

ix. 從/etc/hadoop/conf/slaves 移除節點資訊(master 主機)

x. 登入移除節點並完成關機作業，完成後進行退租

可參考 Hadoop FAQ「I want to make a large cluster smaller by taking out a bunch of nodes simultaneously. How can this be done?」。

13. 如何使用 HDFS over FTP 上傳檔案至 HDFS？

答：用一般的 FTP client 工具(ex:FileZilla)連至平台 HDFS over FTP server，即可以一般 FTP 的操作方法將本機的檔案上傳到 HDFS。HDFS over FTP server 會裝在平台的 Master 主機上，port 是 2222，帳號是 hdfs，密碼是 Master 主機登入密碼，前述資訊皆可在「**hicloud Big Data 叢集安裝紀錄**」頁面取得。

14. 如何執行 word count 範例程式？

答：word count 是 Apache Hadoop 預設提供的數個範例程式之一，您可以執行這個程式來計算檔案內單字出現的次數。執行 word count 程式的步驟如下：

i. 以 root 身分登入 master 主機

ii. 您可以將檔案存放在 HDFS 既有的目錄，或是建立新的目錄來存放檔案。建立目錄的指令範例如下：

```
sudo –u hdfs hadoop fs –mkdir HDFS 目錄名稱
```

iii. 將檔案從 master 主機上傳至 HDFS，操作指令範例如下：

```
sudo -u hdfs hadoop fs -put 本機來源檔案完整路徑 HDFS 目錄名稱/
```

iv. 檢視剛剛上傳檔案的資訊，操作指令範例如下：

```
sudo -u hdfs hadoop fs -ls HDFS 目錄名稱/
Found 1 items
-rw-r--r-- 3 hdfs supergroup 13366 2013-XX-XX XX:XX HDFS 目錄名稱/來源檔案名稱
```

v. 開始執行程式，操作指令範例如下：

- CDH3u5

```
sudo -u hdfs hadoop jar /usr/lib/hadoop/hadoop-examples.jar wordcount
HDFS 目錄名稱/來源檔案名稱  HDFS 結果存放目錄路徑
```

- CDH4.3

```
sudo -u hdfs hadoop jar /usr/lib/hadoop-0.20-mapreduce/hadoop-examples.jar
wordcount
HDFS 目錄名稱/來源檔案名稱  HDFS 結果存放目錄路徑
```

請注意結果存放目錄不能事先存在，否則執行時會出現下列的錯誤訊息：

```
…
XX/XX/XX XX:XX:XX ERROR security.UserGroupInformation:
PriviledgedActionException as:hdfs (auth:SIMPLE)
cause:org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory
HDFS 結果存放目錄路徑 already exists
```

vi. 若程式正常執行，console 將會顯示 Job 目前執行進度，畫面範例如下：

```
XX/XX/XX XX:XX:07 INFO input.FileInputFormat: Total input paths to process : 1
XX/XX/XX XX:XX:07 INFO mapred.JobClient: Running job: job_201309261422_0002
XX/XX/XX XX:XX:08 INFO mapred.JobClient:    map 0% reduce 0%
XX/XX/XX XX:XX:13 INFO mapred.JobClient:    map 100% reduce 0%
XX/XX/XX XX:XX:22 INFO mapred.JobClient:    map 100% reduce 33%
XX/XX/XX XX:XX:23 INFO mapred.JobClient:    map 100% reduce 100%
```

vii. 您可以經由 Hadoop 指令檢視執行結果，操作指令範例如下：

```
sudo -u hdfs hadoop fs –cat HDFS 結果存放目錄/part-r-00000|more
"AS        3
"Contribution"    1
"Contributor"     1
```

或是下載結果檔案至本機端，操作指令範例如下：

```
sudo -u hdfs hadoop fs –get HDFS 結果存放目錄/part-r-00000 本機端完整路徑
```

15. 如何在 Excel 連結 Apache Hive？

答：您可以經由 Hive ODBC 驅動程式，在微軟 Excel 讀取儲存在 Apache Hive 裡的資料。下面將以 Windows 7 與 Excel 2007 為例，說明如何進行安裝與設定：

i. 啟動 Hadoop 平台的 Hive Server 服務

A. 以 root 身分登入 master 主機

B. 執行下列指令以啟動 Hive Server 服務，預設使用 TCP PORT **10000**；您可以自行修改成其它 Port

```
su - hdfs
cd /usr/lib/hive/bin
export HIVE_PORT=10000
./hive --service hiveserver > /tmp/hive.log 2>&1 &
```

C. 確認 Hive Server 服務是否開始運作

```
ps -ef|grep java|grep hive-service|more
```

D. 檢視 Log 內容查看 Hive Server 的運作記錄

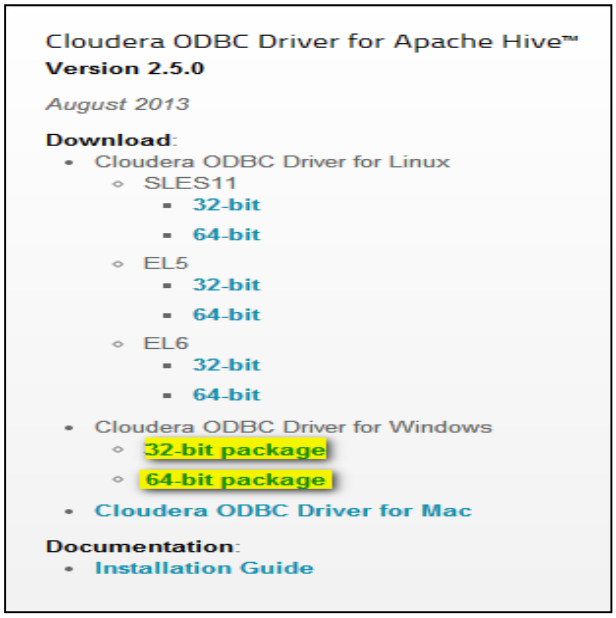
```
vim /tmp/hive.log
```

E. 您可以參考 Apache Hive Wiki 網頁取得更完整的說明
https://cwiki.apache.org/confluence/display/Hive/HiveServer

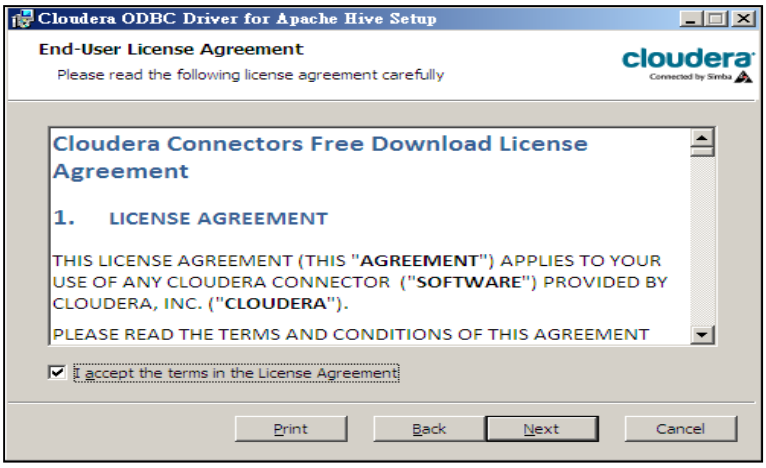
ii. 在個人電腦安裝 Apache Hive ODBC 驅動程式

A. 連結至 Cloudera ODBC Drivers for Apache Hive 網頁
http://www.cloudera.com/content/support/en/downloads/download-components/download-products/downloads-listing/connectors/cloudera-odbc-drivers.html

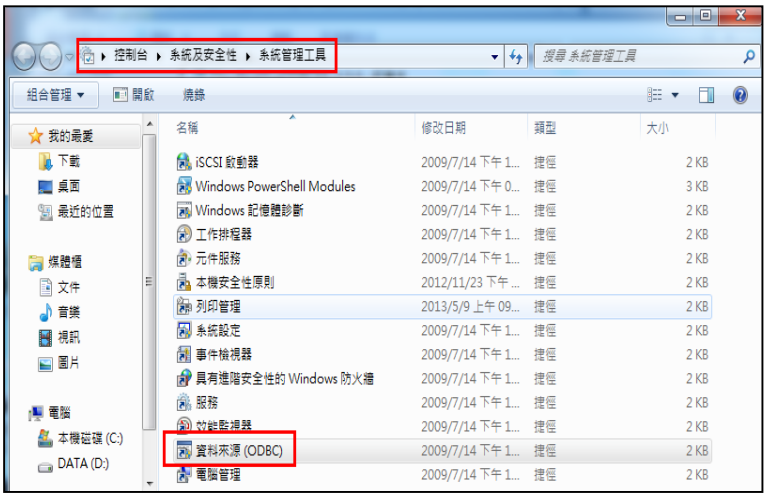
B. 請根據您的 Windows 環境下載 32-bit 或 64-bit 的驅動程式



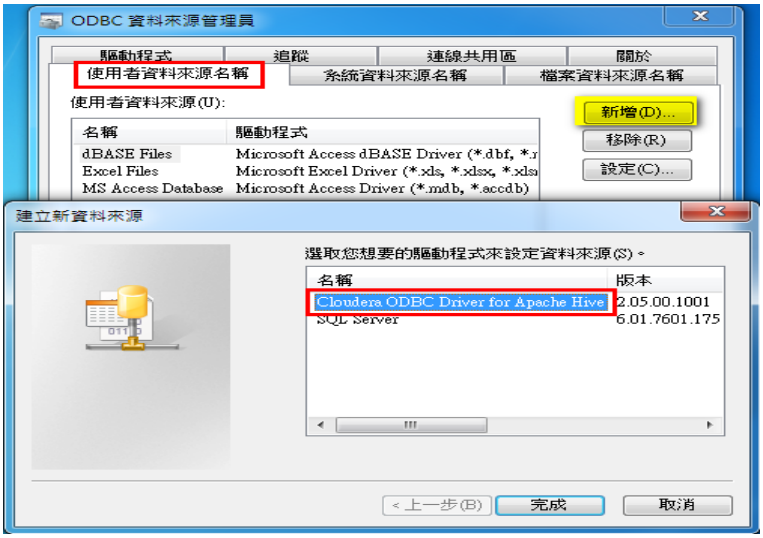
C. 執行前一步驟下載的驅動程式，點選「Next」，勾選「I accept the terms in the License Agreement」，然後選擇安裝路徑開始進行安裝



D. 安裝完成後，請至「開始」→「控制台」→「系統與安全性」→「系統管理工具」，點選「資料來源(ODBC)」進行設定

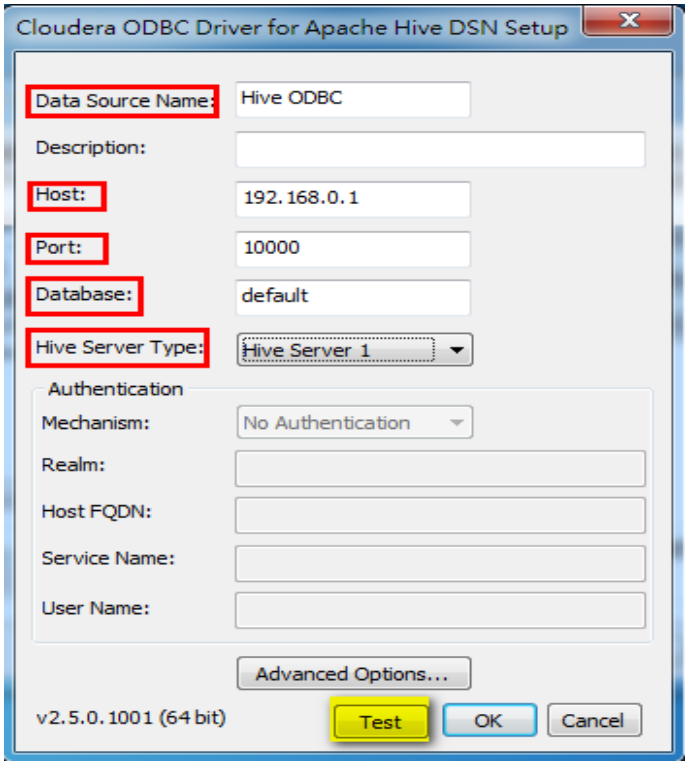


E. 選擇「使用者資料來源名稱」後點選新增，在「建立新資料來源」視窗選擇「**Cloudera ODBC Driver for Apache Hive**」，然後點選完成

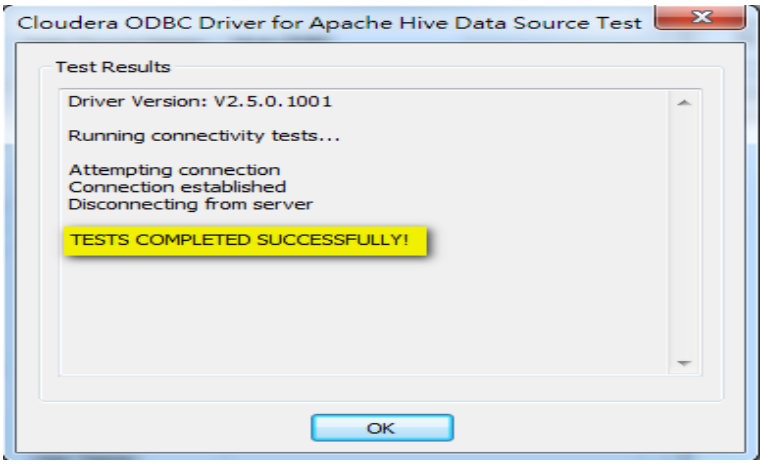


F. 在設定視窗請依序輸入：

- Data Source Name：您可以自訂資料來源名稱
- Host：參考前述步驟(i.A) 的 master 主機 IP 地址
- Port：參考前述步驟(i.B)的 Port，預設是 **10000**
- Database：請輸入 **default**
- Hive Server Type：請選擇「**Hive Server 1**」



G. 完成輸入後您可以點選下方「Test」進行測試，若是沒有問題將會顯示連線測試正常訊息



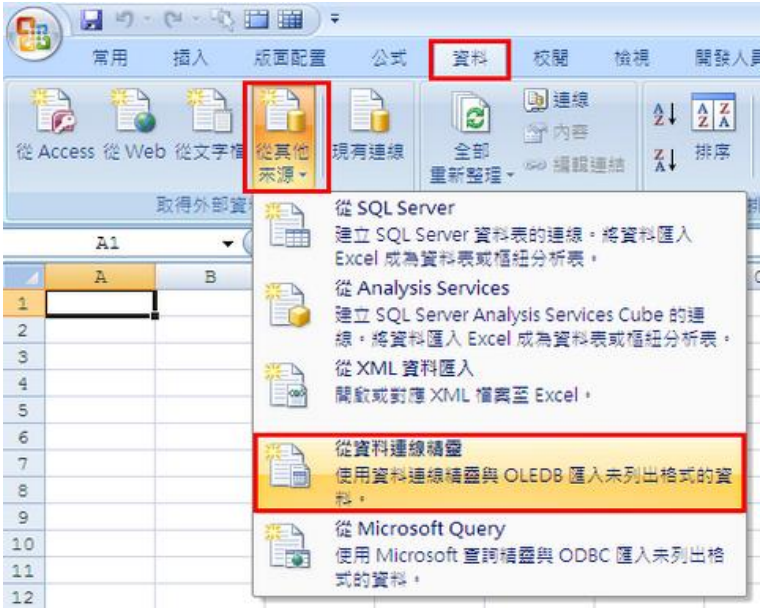
H. 您可以參考前述步驟(ii.A)網頁下方的「Installation Guide」取得完整說明

iii. 在 Excel 讀取 Apache Hive 的資料

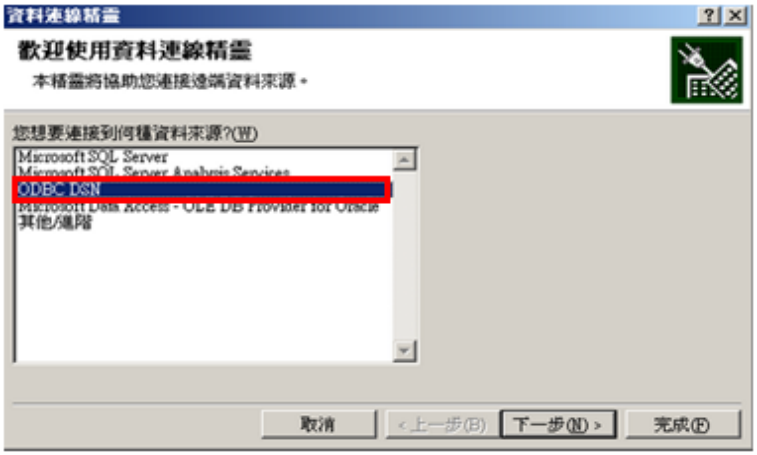
A. 開啟 Excel 2007

B. 選擇上方工具列「資料」→「從其他來源」→「從資料連線精靈」開啟

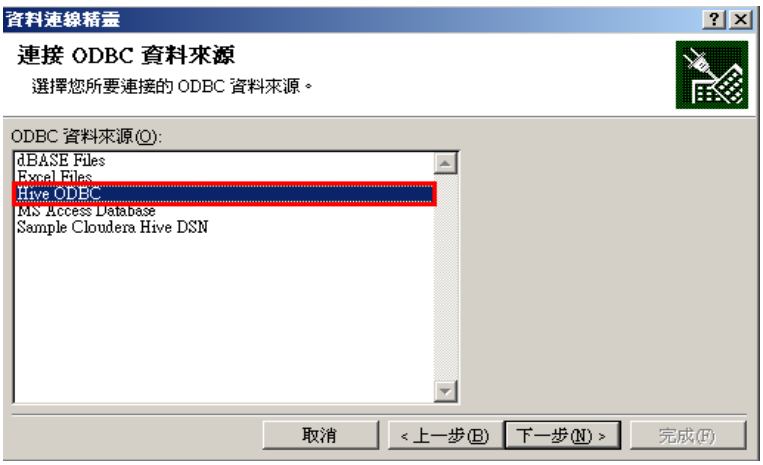
設定視窗



C. 在「歡迎」視窗選擇「ODBC DSN」並點選下一步



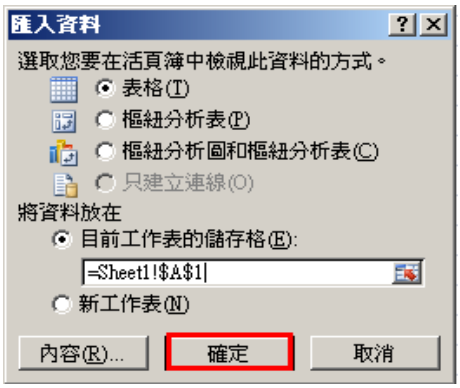
D. 在「連接 ODBC 資料來源」視窗選擇您在前述步驟(ii.F)輸入的「**Data Source Name**」，並點選下一步



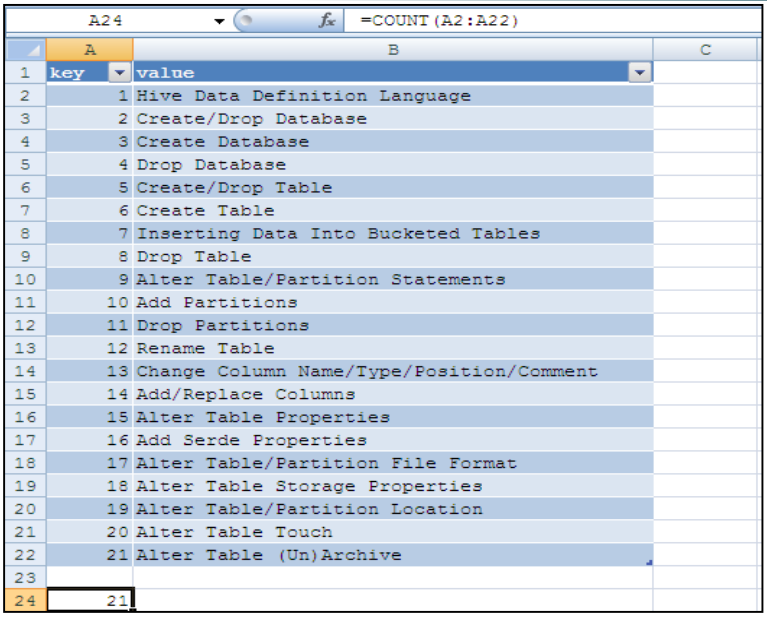
E. 在「選擇資料庫與資料表」視窗選擇您要匯入的表格，然後點選下一步

F. 在「儲存資料連線檔案」視窗選擇您要儲存的路徑，然後點選完成

G. 在「匯入資料」視窗選擇您要將資料存放在 Excel 工作表的位置，然後點選確定



H. 接下來資料將匯入 Excel 工作表，您可以對這些資料作進一步的計算或分析



A24 =COUNT(A2:A22)

	A	B	C
1	key	value	
2	1	Hive Data Definition Language	
3	2	Create/Drop Database	
4	3	Create Database	
5	4	Drop Database	
6	5	Create/Drop Table	
7	6	Create Table	
8	7	Inserting Data Into Bucketed Tables	
9	8	Drop Table	
10	9	Alter Table/Partition Statements	
11	10	Add Partitions	
12	11	Drop Partitions	
13	12	Rename Table	
14	13	Change Column Name/Type/Position/Comment	
15	14	Add/Replace Columns	
16	15	Alter Table Properties	
17	16	Add Serde Properties	
18	17	Alter Table/Partition File Format	
19	18	Alter Table Storage Properties	
20	19	Alter Table/Partition Location	
21	20	Alter Table Touch	
22	21	Alter Table (Un)Archive	
23			
24	21		

I. 您可以參考下列網頁取得完整的說明

- 「連接 (匯入) 外部資料」
  http://office.microsoft.com/zh-tw/excel-help/HP010089898.aspx
- 「使用 Microsoft Query 擷取外部資料」
  http://office.microsoft.com/zh-tw/excel-help/HA010099664.aspx