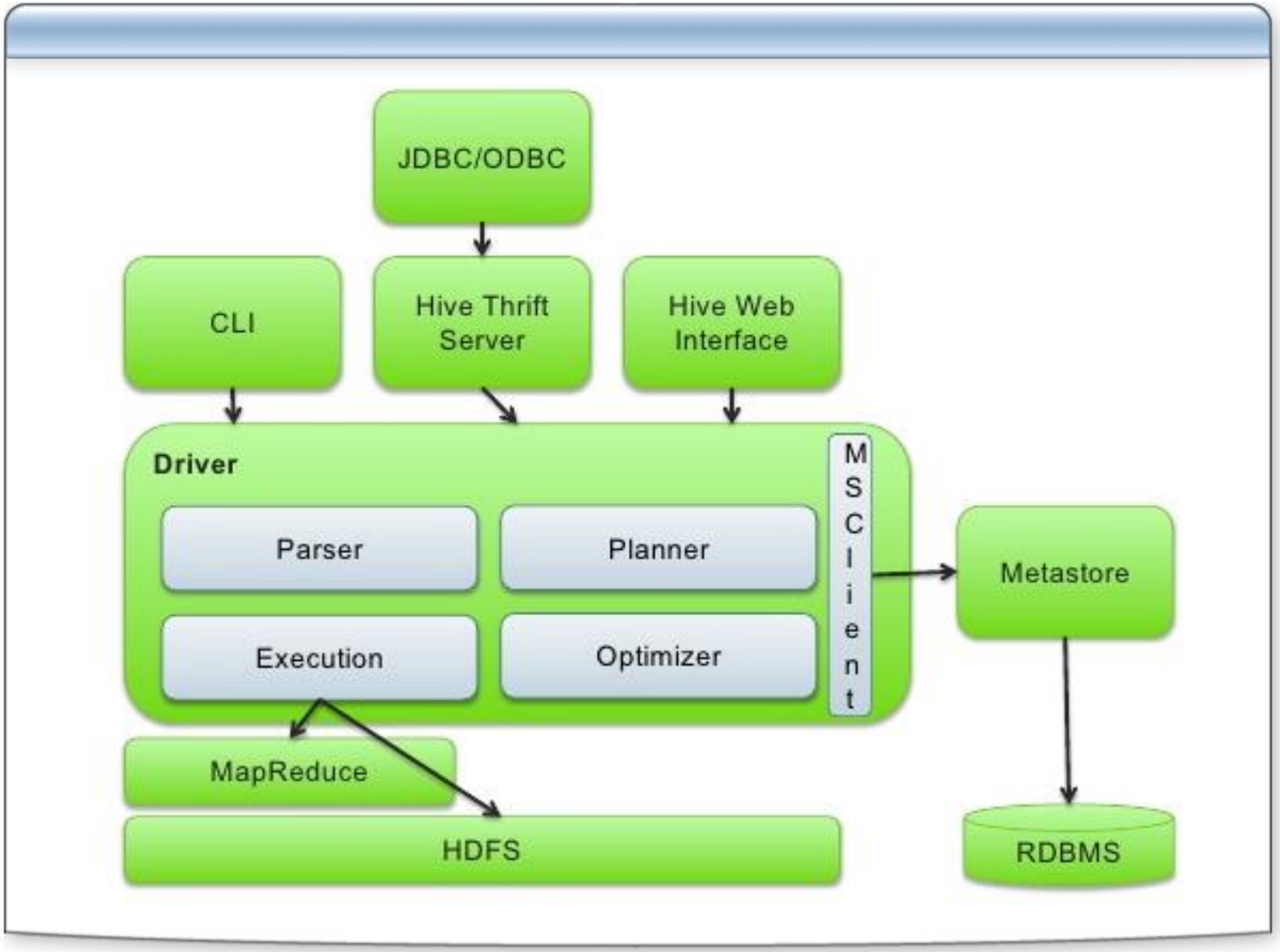
# Hive 資料倉儲工具


1. Data Warehousing - Schemas
http://www.tutorialspoint.com/dwh/dwh_schemas.htm

# Apache Hive 運作架構圖



# 啟動 Hadoop 核心系統 (HDFS,YARN)

```
$ starthd  a
[Cluster A]
start Application Container ok
start HDFS ok
start YARN ok
```

# 檢視 Hadoop 核心系統 (HDFS,YARN)

```
$ dkls a
Docker Utility 0.3.0 (2016/08/02)

[Container]
--------------------------------------------------------------------------------
wka02(a402e38b3de5) 172.17.8.11 Running ( NodeManager DataNode )
wka01(9ad97f9f042f) 172.17.8.10 Running ( NodeManager DataNode )
rma(f010967f8db2) 172.17.6.12 Running ( ResourceManager JobHistoryServer )
nna(b72af384c024) 172.17.6.10 Running ( NameNode SecondaryNameNode )
cla01(bea796f6fdd7) 172.17.2.11 Running (CVBG:22101->cla01:22, user:dsa01)
cla00(9891609dd106) 172.17.2.10 Running (CVBG:22100->cla00:22, user:dsa00)

[Images]
--------------------------------------------------------------------------------
dafu/worker        272             6fa11bb3425a        4 days ago          922.4 MB
```

# 資料科學家上工

# 資料科學家登入 Hadoop Client

```
$ ssh dsa01@cla01
dsa01@cla01's password: dsa01
Welcome to Ubuntu 14.04.3 LTS (GNU/Linux 3.16.0-46-generic x86_64)

 * Documentation:  https://help.ubuntu.com/
Last login: Tue Sep  1 20:10:04 2015 from 172.17.42.1
```

# Hive 首部曲


## Hive 基本命令操作

# 建立 Hive Schema Database (Derby)

```
$ schematool -initSchema -dbType derby
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/apache-hive-2.1.0-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Metastore connection URL:
jdbc:derby:;databaseName=metastore_db;create=true
Metastore Connection Driver :  org.apache.derby.jdbc.EmbeddedDriver
Metastore connection User:     APP
Starting metastore schema initialization to 2.1.0
Initialization script hive-schema-2.1.0.derby.sql
Initialization script completed
schemaTool completed
```

**[注意] 如 metastore_db/ 目錄已存, 請先移除它, 命令如下：**

```
$ rm -r metastore_db/
```

# 檢視 Hive 版本及內定設定值

```
$ hive  --version
Hive 1.2.1
Subversion git://localhost.localdomain/home/sush/dev/hive.git -r
243e7c1ac39cb7ac8b65c5bc6988f5cc3162f558
Compiled by sush on Fri Jun 19 02:03:48 PDT 2015
From source with checksum ab480aca41b24a9c3751b8c023338231

$ hive -S -e 'set -v' | grep 'fs.defaultFS'
fs.defaultFS=hdfs://nna:8020
mapreduce.job.hdfs-servers=${fs.defaultFS}
```

# 設定使用者專屬 Hive 資料倉儲目錄

Hive 內定所有使用者的**資料庫**與**資料表**, 存在 HDFS 分散檔案系統的 /user/hive/warehouse/ 這目錄, 如多人同時使用 Hive, 勢必造成資料衝突.

**在 使用者 家目錄中, 編輯 .hiverc 設定檔**

```
$ nano .hiverc
set hive.cli.print.current.db=true;
set hive.metastore.warehouse.dir=/user/dsa01/hive;
set hive.exec.scratchdir=/user/dsa01/tmp;
```

# 啟動 Hive 並建立資料表

產生 dummy.txt 資料檔

```
$ echo "x" > dummy.txt

$ hive -S
hive (default)>
```

建立 dummy 表格

```
hive (default)> create table dummy (value string);
```

第一次建立表格, Hive 會在 HDFS 建立 /user/dsa01/hive 目錄, 用來儲存資料表的資料, 檢視命令如下：

```
hive (default)> dfs  -ls  /user/dsa01/hive;
Found 1 items
drwxr-xr-x   - dsa01 biguser    0 2016-02-02 23:41 /user/dsa01/hive/dummy
```

# 表格目錄

此時 Hive 會在本機目前的目錄, 建立 metastore_db 子目錄, 用來儲存資料庫的 Meta 資料(不包含資料庫的資料), 執行以下命令：

```
hive> ! tree -L 2 metastore_db;
metastore_db
├── dbex.lck
├── db.lck
├── log
│   ├── log1.dat
│   ├── log.ctrl
│   └── logmirror.ctrl
├── seg0
│   ├── c101.dat
│   ├── c10.dat
│   ├── c111.dat
                    :
                    :
│   ├── cc0.dat
│   ├── cd1.dat
│   ├── ce1.dat
│   └── cf0.dat
├── service.properties
└── tmp
```

# Hive 基本命令（一）

顯示目前所有表格名稱

```
hive> show tables;
dummy
```

匯入 dummy.txt 資料至 dummy 表格

```
hive> load data local inpath 'dummy.txt' into table dummy;
```

[註] 上述 load 命令會將本機 dummy.txt 移至 HDFS 的 /user/hive/warehouse/dummy 目錄中

```
hive> dfs -ls /user/dsa01/hive/dummy;
Found 1 items
-rwxr-xr-x  2 dsa01 biguser  2 2015-09-09 01:30 /user/hive/warehouse/dummy/dummy.txt
```

查詢 dummy 表格資料

```
hive> select * from dummy;
x
```

# Hive 基本命令（二）

增加資料

```
hive (default)> insert into table dummy select value from dummy;

hive (default)> insert into dummy values('abc');

hive (default)> select * from dummy;
x
abc
x
```

刪除 dummy 表格

```
hive> drop table dummy;
hive> quit;
```

# Hive 二部曲


## Hive 分析實作

實作網址：http://hive.3du.me/Lab-009.html

# 分析大專校院校別學生數（一）

**下載與處理 大專校院校別學生數檔**

```
$ wget --no-check-certificate
https://stats.moe.gov.tw/files/detail/103/103_student.txt
```

**轉換編碼**

```
$ iconv -f UCS-2 -t utf8 103_student.txt -o temp.txt
$ sed 's/\"//g' < temp.txt >student.txt
```

**將 ’總計’ 欄位資料的 ’,’ 字元刪除**

```
$ sed 's/,//g' < student.txt >student1.txt
```

```
$ head -n 4 student1.txt
大專校院校別學生數
103 學年度 SY2014-2015
學校代碼 學校名稱 日間/進修別 等級別 總計 男生計 女生計 一年級男生 一年級女生 二年級男生二年級女生 三年級男生 三年級女生 四年級男生 四年級女生 五年級男生 五年級女生 六年級男生六年級女生 七年級男生 七年級女生 延修生男生 延修生女生 縣市名稱 體系別
0001 國立政治大學 D 日 D 博士 973 583 390 117 76 79 62 94 58 98 5775 53 61 43 59 41 - - 30 臺北市 1 一般
```

# 分析大專校院校別學生數（二）

建立 **student** 資料表

```
$ hive -S
hive (default)> CREATE TABLE student (code string, name
string, type string, class string, total int) row format
delimited fields terminated by '\t' stored as textfile;
```

載入資料

```
hive (default)> load data local inpath 'student1.txt' into table
student;
```

顯示資料

```
hive (default)> select code,name,total from student limit 8;
大專校院校別學生數                NULL
101 學年度 SY2012-2013                    NULL
學校代碼    學校名稱    NULL
0001    國立政治大學    973
0001    國立政治大學    3816
0001    國立政治大學    9639
0001    國立政治大學    1625
0002    國立清華大學    1786
```

# 分析大專校院校別學生數（三）

**列印各校總人數**

```
hive> select name,sum(total) from student group by name;
                    ::
長庚科技大學    7595
長榮大學        10474
開南大學        9742
靜宜大學        12249
馬偕醫學院      530
馬偕醫護管理專科學校    4160
高美醫護管理專科學校    850
高苑科技大學    7723
高雄醫學大學    6981
黎明技術學院    4602
龍華科技大學    11254
```

# 檢視資料表儲存目錄

```
hive (default)> dfs -ls /user/dsa01/hive;
Found 1 items
drwxr-xr-x   - dsa01 aplpha  0 2015-11-28 02:37
/user/dsa01/hive/student

hive (default)> dfs -ls /user/dsa01/hive/student;
Found 1 items
-rwxr-xr-x  2 dsa01  alpha  104643 2015-11-28 02:37
/user/dsa01/hive/student/student1.txt

hive (default)> quit;
```

# 實作練習


1. 博士班人數小於50 的學校

2. 博士班人數小於50 的學校有幾所

# Hive 資料倉儲工具


## Hive 外部資料表

# 下載 Dataset, 並上載至 Hive 資料倉儲目錄

```
$ wget
http://community.jaspersoft.com/sites/default/files/wiki_attachments/accounts.csv

$ head -n 1 accounts.csv
a69dae1f-b2ee-1257-3895-438dfb8ea964;2005-11-30 19:19:03;2005-11-30 19:19:03;1;beth_id;1;Alpha-Murraiin Communications, Inc;;Manufacturing;Communications;;;5423 Camby Rd.;La Mesa;CA;35890;USA;;;612-555-4878;;;;www.alpha-murraiincommunications,inc.com;;;;;5423 Camby Rd.;La Mesa;CA;35890;USA;0
```

上載至使用者專屬 Hive 資料倉儲目錄

```
$ hdfs dfs -mkdir /user/dsa01/hive/myacc
$ hdfs dfs -put accounts.csv /user/dsa01/hive/myacc
```

# 建立 accounts 外部資料表

撰寫 **createtbl.q** 批次檔

```
$ nano createtbl.q
CREATE EXTERNAL TABLE accounts (
id STRING,
date_entered STRING,
                    ::
shipping_address_state STRING,
shipping_address_postalcode STRING,
shipping_address_country STRING,
deleted BOOLEAN
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\;'
STORED AS TEXTFILE LOCATION '/user/dsa01/hive/myacc';
```

**[註] /user/dsa01/hive/myacc 是目錄區, 不是資料集名稱**

```
$ hive -S -f createtbl.q
$ hive -S -e 'show tables'
accounts
student
```

# 管理 accounts 外部資料表（一）

顯示 accounts 資料表的第一筆資料

```
$ hive -S -e 'select * from accounts limit 1'
a69dae1f-b2ee-1257-3895-438dfb8ea964    2005-11-30 19:19:03     2005-11-30 19:19:03     1       beth_id 1       Alpha-Murraiin Communications, Inc      Manufacturing   Communications          5423 Camby Rd.  La Mesa CA      35890   USA             612-555-4878            www.alpha-murraiincommunications,inc.com        5423 Camby Rd.  La Mesa CA      35890   USA     NULL
```

顯示總筆數

```
$ hive -S -e 'select count(*) from accounts'
1201
```

# 管理 accounts 外部資料表（二）

**刪除 accounts 資料表**

```
$ hive -S -e 'drop table accounts'

$ hdfs dfs -ls hive/myacc
Found 1 items
-rw-r--r--   2 dsa01 biguser     357646 2015-11-28 23:34 hive/myacc/accounts.csv
```

[重點] accounts 資料表刪除後, 資料檔還是存在 (/user/dsa01/hive/myacc/accounts.csv)

# 資料科學家收工

# 離開工作主機

```
$ exit
logout
Connection to cla01 closed.
```