

# Hadoop Lab 實作課程

2016-03-21

蔡秉文

Cookie Tsai

# Resources

- CookeTsai 的手記
  - <http://tsai-cookie.blogspot.tw/>

The screenshot shows a web browser displaying the Blogger page for 'CookieTsai 的手記'. The browser's address bar shows the URL 'http://tsai-cookie.blogspot.tw/'. The page header includes the Blogger logo, a search bar, and the user's email 'mcacookie@gmail.com' with links for '資訊主頁' and '登出'. The main content area features the title 'CookieTsai 的手記' and the subtitle 'Cookie's Blogger'. Below this, the date '2015年9月1日 星期二' is displayed. The main post title is 'Hadoop + HBase + Hive 建置手冊 (完全分布式)'. A '注意事項' (Notes) box contains three points: 1. All commands must be run as root for practice. 2. Pre-setup steps must be done on master, slaver1, and slaver2. 3. PDF format may cause text distortion. A '目錄' (Table of Contents) section lists: Hadoop + HBase + Hive 建置手冊 (完全分布式), 目錄, 套件清單, 環境配置, and 前置步驟. The right sidebar includes a '關於我自己' (About Me) section with a profile picture and name '蔡秉文', a '追蹤' (Follow) button, and a '搜尋此網誌' (Search this blog) section with a search bar. At the bottom, a '網誌存檔' (Blog archive) section shows posts for '2015 (3)' and '九月 (1)'. The footer text is 'Hadoop + HBase + Hive 建置手冊'.

CookieTsai 的手記

Cookie's Blogger

2015年9月1日 星期二

Hadoop + HBase + Hive 建置手冊 (完全分布式)

注意事項：

1. 所有指令皆使用 root 身份執行，僅供練習使用。
2. 前置步驟在 master, slaver1, slaver2 都必須做一遍。
3. PDF 格式部分文字會失真，輸入時請注意符號是否正確。

目錄

- Hadoop + HBase + Hive 建置手冊 (完全分布式)
  - 目錄
  - 套件清單
  - 環境配置
  - 前置步驟

關於我自己

蔡秉文

追蹤 0

檢視我的完整簡介

搜尋此網誌

搜尋

網誌存檔

▼ 2015 (3)

▼ 九月 (1)

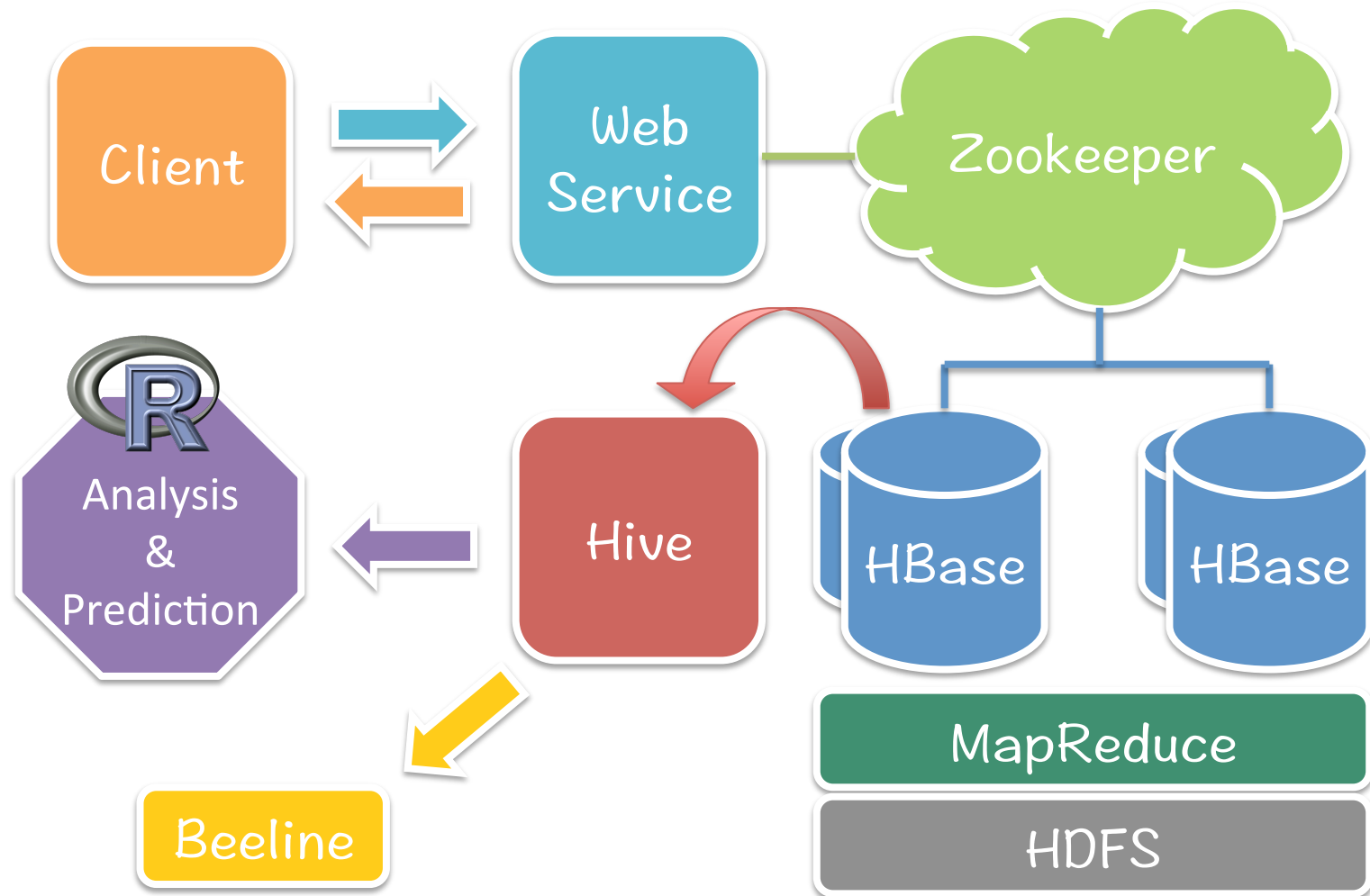
Hadoop + HBase + Hive 建置手冊

# About Me

- Education
  - III
  - NUTC
- Experience
  - Mitake
  - JC Software Services
- Honors & Awards
  - The Winner of Etu Hadoop Competition 2015
  - 2011 電信創新應用大賽 智慧家庭組 優選
  - 2010 電信奧斯卡 MOD應用組 佳作



# Lab deployment



# System architecture

- Virtual Box

Host Name	IP	OS
master	192.168.60.101	CentOS 6.7
slaver1	192.168.60.102	CentOS 6.7
slaver2	192.168.60.103	CentOS 6.7

- Packages

Package	Package Name	Version
Apache Hadoop	hadoop-2.4.1.tar.gz	2.4.1
Apache HBase	hbase-0.98.13-hadoop2-bin.tar.gz	0.98.13
Apache Hive	apache-hive-1.2.1.tar.gz	1.2.1
Apache Zookeeper	zookeeper-3.4.6.tar.gz	3.4.6

# Setup for testing hosts (3 VMs)

- Install Virtual Box
- Import Virtual Box VM
- Modify to the static IP and try a test

# You will learn

- Basic hadoop
  - HDFS, MapReduce, HBase(NoSQL)
- Basic hadoop ecosystem
  - Hive, R
- Back end
  - Web Service, Shell Script
- Front end
  - HTML, CSS and JQuery

# What is hadoop

- A big-data platform for data manipulation
- Store data in distributed repositories
- Distributed job process to deal with big-data
- Dig out the data insight and data analytics
- High availability and stabilized
- Many ecosystems supports



# Install Hadoop

# What is Zookeeper

- Used for message management in distributed system, such like naming, synchronization service, clustering management
- Considering to HA, ZK also provides clustering mode
- In Hadoop, it manages Namenode, HBase... for message passing and sync

# Install Zookeeper

# What is HBase

- A kind of NoSQL
- Manipulation in HDFS
- Using column family qualifier
- Each Row-Key is also a indexed column

Row-Key	Column		Timestamp	Value
	Family	Qualifier		
row1	cf	name	1442053885486	Tom
row2	cf	name	1442053885487	Mary
row2	cf	phone	1442053885487	0999XXXXXX
row3	cf	name	1442053885486	John

Install HBase

# What is Hive

- Data warehouse software facilitates querying and managing large datasets residing in distributed storage.
- SQL-like language called HiveQL
- At the same time this language also allows traditional map/reduce

Install Hive

# What is R

- R is a free software environment for statistical computing and graphics.
- It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.
- It provides an unparalleled platform for programming new statistical methods in an easy and straightforward manner.



Install R Lib & RStudio

# Web Service

- What is myApp
  - It's a Simple Java Project
  - It's a RESTful Service
  - Using Jersey
- Install myApp
  - `$ tar -zxvf /tmp/myApp.tar.gz`
  - `$ java -jar myApp/application-1.0-SNAPSHOT.jar`

# What WampServer

- WampServer is a Windows web development environment.
- It allows you to create web applications with Apache2, PHP and a MySQL database. Alongside, PhpMyAdmin allows you to manage easily your databases.

Install WampServer

# Using Web Client



角色：工程師, 剩餘可消費金額：65000

A. Car



120,000 \$

[view](#) [like](#) [order](#)

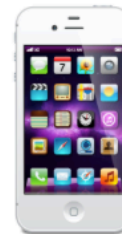
B. Notebook



40,000 \$

[view](#) [like](#) [order](#)

C. iPhone



20,000 \$

[view](#) [like](#) [order](#)

角色：工程師, 剩餘可消費金額：65000

D. Apple



E. Book



F. Book



# Using HBase Shell

```
[root@localhost ~]# hbase shell
2015-09-18 16:25:08,056 INFO [main] Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 0.98.13-hadoop2, r8f54f8daf8cf4d1a629f8ed62363be29141c1b6e, Wed Jun 10 23:01:33 PDT 2015

hbase(main):001:0> list
TABLE
2015-09-18 16:25:15,237 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... us
count
log
2 row(s) in 2.4250 seconds

=> ["count", "log"]
hbase(main):002:0> scan 'count'
ROW                                COLUMN+CELL
1                                  column=cf:likeCnt, timestamp=1442564540456, value=\x00\x00\
1                                  column=cf:orderAmount, timestamp=1442564541430, value=\x00\
1                                  column=cf:orderCnt, timestamp=1442564541422, value=\x00\x00\
1                                  column=cf:viewCnt, timestamp=1442564548292, value=\x00\x00\
2                                  column=cf:likeCnt, timestamp=1442495816850, value=\x00\x00\
2                                  column=cf:orderAmount, timestamp=1442495817486, value=\x00\
2                                  column=cf:orderCnt, timestamp=1442495817480, value=\x00\x00\
2                                  column=cf:viewCnt, timestamp=1442495816189, value=\x00\x00\
3                                  column=cf:likeCnt, timestamp=1442495819729, value=\x00\x00\
```

# Learning HBase Shell

# Using Beeline

```
[root@localhost ~]# beeline -u jdbc:hive2://master:10000
```

```
Connecting to jdbc:hive2://master:10000
```

```
Connected to: Apache Hive (version 1.2.1)
```

```
Driver: Hive JDBC (version 1.2.1)
```

```
Transaction isolation: TRANSACTION_REPEATABLE_READ
```

```
Beeline version 1.2.1 by Apache Hive
```

```
0: jdbc:hive2://master:10000> show tables;
```

```
+-----+---+
| tab_name |
+-----+---+
| count    |
| log      |
+-----+---+
```

```
2 rows selected (0.295 seconds)
```

```
0: jdbc:hive2://master:10000> select * from count;
```

```
+-----+-----+-----+-----+-----+---+
| count.key | count.likecnt | count.viewcnt | count.ordercnt | count.orderamount |
+-----+-----+-----+-----+-----+---+
| 1          | 3              | 4              | 3              | 30000              |
| 2          | 1              | 1              | 1              | 10000              |
| 3          | 1              | 1              | 4              | 40000              |
| 4          | 2              | 2              | 1              | 10000              |
| 5          | 3              | 3              | 2              | 20000              |
| 6          | 1              | 1              | 1              | 10000              |
+-----+-----+-----+-----+-----+---+
```

```
6 rows selected (0.643 seconds)
```

```
0: jdbc:hive2://master:10000> █
```



# Learning HiveQL

# Using RStudio

RStudio

Project: (None)

```
49 ### viewcnt
50 vector <- data$count.viewcnt
51 names(vector) <- data$count.key
52 barplot(vector, main = "barplot of viewcnt")
53
54 ## plot and model
55
56 ### orderamount ~ ordercnt
57 plot(data$count.orderamount ~ data$count.ordercnt)
58 data.lm = lm(data$count.orderamount ~ data$count.ordercnt)
59 abline(data.lm, col="red")
60 summary(data.lm)
61
62 ## Pie Chart
63 |
64 ### likecnt
65 slices <- data$count.likecnt
```

63:1 (Top Level) R Script

Warning message:  
In summary.lm(data.lm) : essentially perfect fit: summary may be unreliable

```
>
> ## Pie Chart
>
> ### likecnt
> slices <- data$count.likecnt
> lbls <- data$count.key
> pie(slices, labels = lbls, main="Pie Chart of likecnt")
>
> dbDisconnect(conn)
[1] TRUE
>
```

Environment History

Global Environment

Data

data	6 obs. of 5 variables
myData	6 obs. of 5 variables

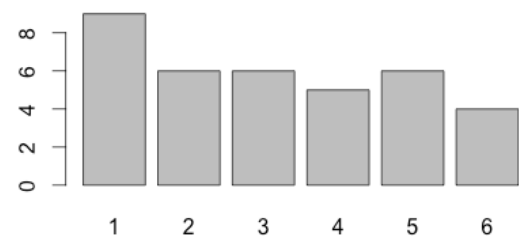
Values

class.path	"/Users/Mitake/Cookie/git/LabHadoop/R/hive...
conn	Formal class JDBCConnection
current.path	"/Users/Mitake/Cookie/git/LabHadoop/R"
data.lm	List of 12
driver.jar.name	"HiveJDBC4.jar"
driver.package...	"com.cloudera.hive.jdbc4.HS2Driver"

Files Plots Packages Help Viewer

Zoom Export Clear All

barplot of viewcnt



Category	Value
1	8.5
2	6.0
3	6.0
4	5.0
5	6.0
6	4.0

Learning R

# Download

- Hadoop-2.5.2
  - <http://apache.stu.edu.tw/hadoop/common/hadoop-2.5.2/hadoop-2.5.2.tar.gz>
- Zookeeper-3.4.6
  - <http://apache.stu.edu.tw/zookeeper/zookeeper-3.4.6/zookeeper-3.4.6.tar.gz>
- HBase-0.98.13
  - <http://ftp.tc.edu.tw/pub/Apache/hbase/0.98.13/hbase-0.98.13-hadoop2-bin.tar.gz>
- Hive-1.2.1
  - <http://apache.stu.edu.tw/hive/hive-1.2.1/apache-hive-1.2.1-bin.tar.gz>
- R-3.1.3
  - <http://cran.r-project.org/src/base/R-3/R-3.1.3.tar.gz>

Thank you for your listening