Hadoop Lab 實作課程

2016-03-21

蔡秉文

Cookie Tsai

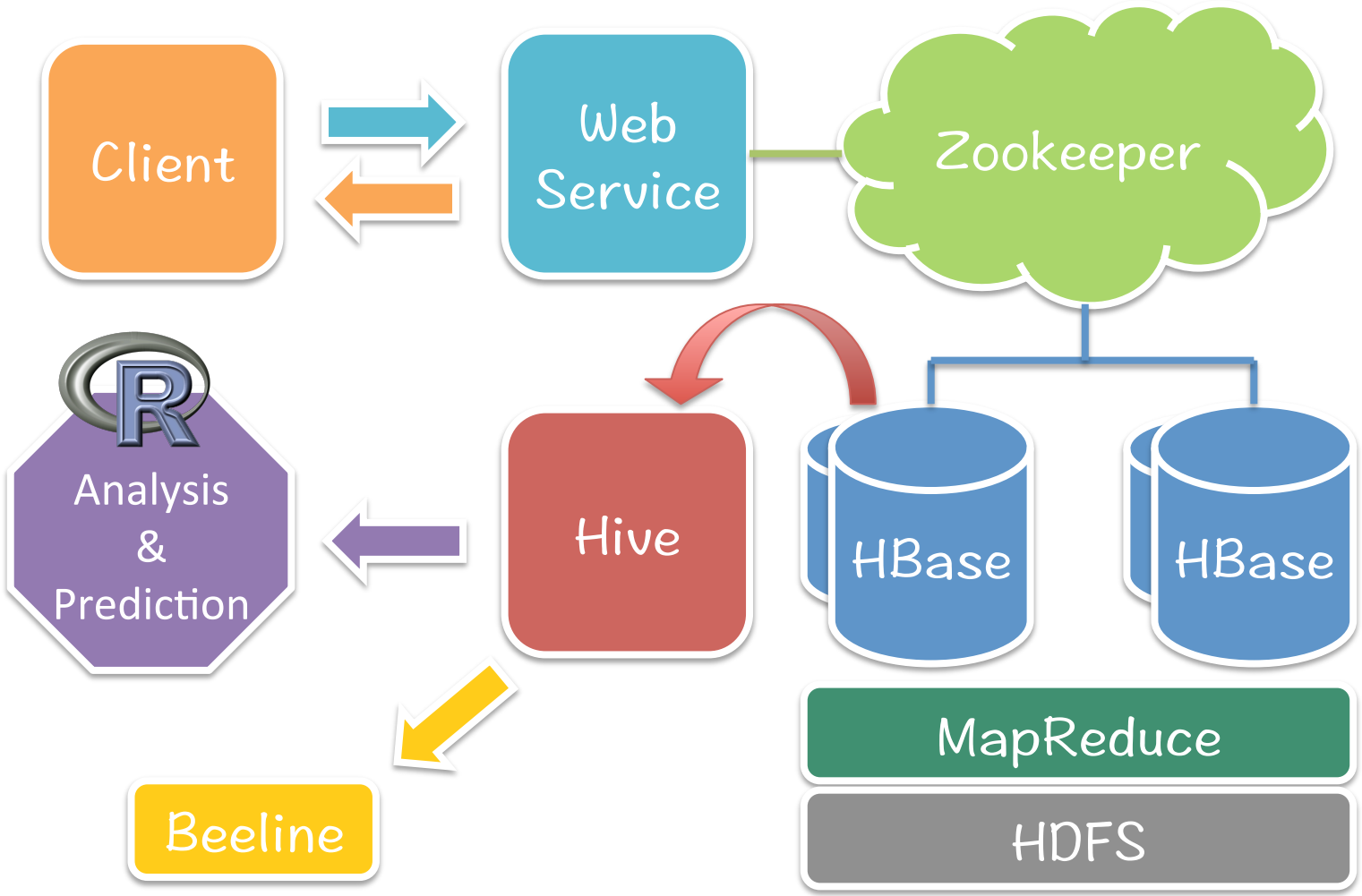
Resources

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

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 deploment



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

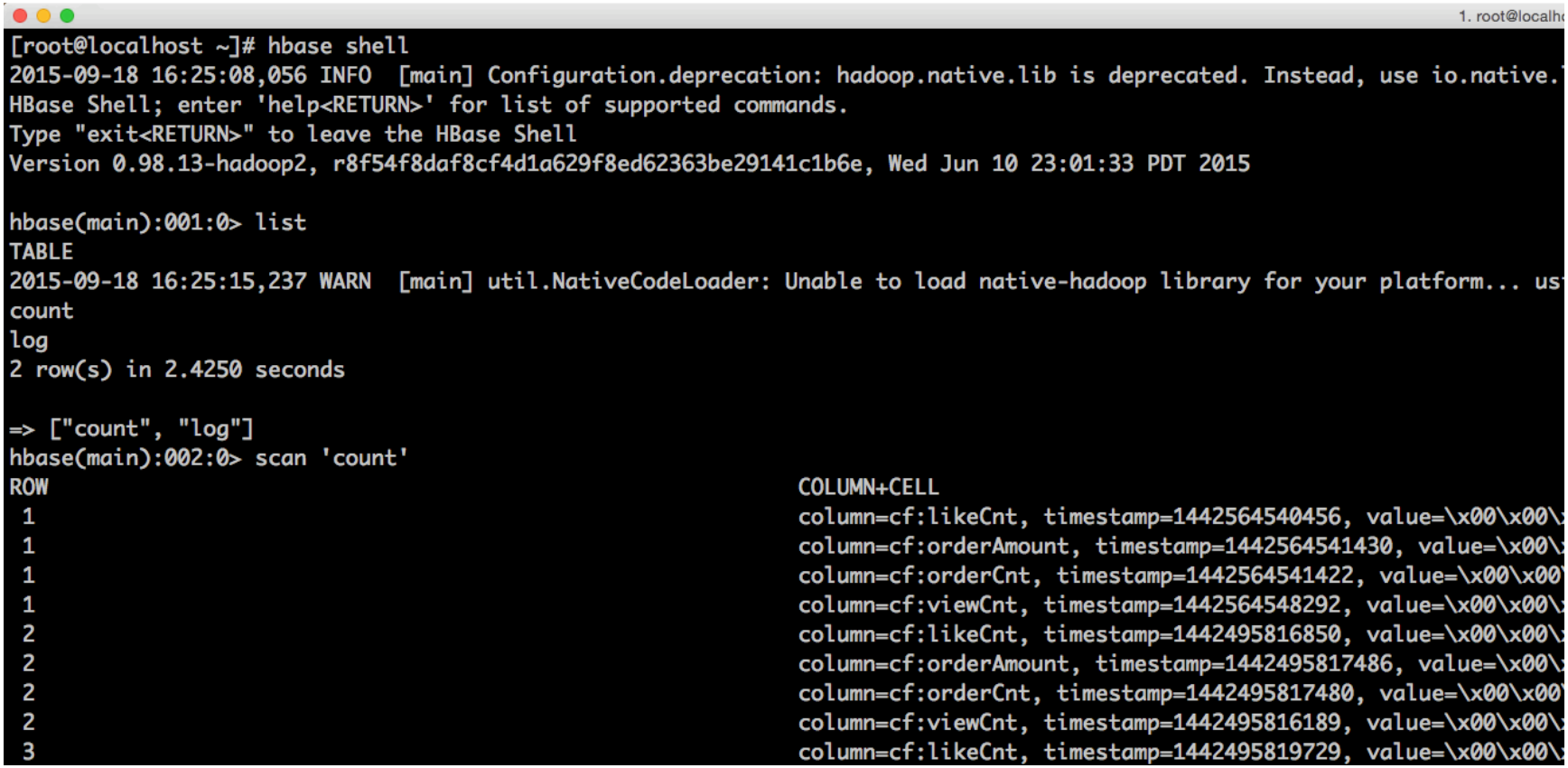
What WampServer

- 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

Using HBase Shell



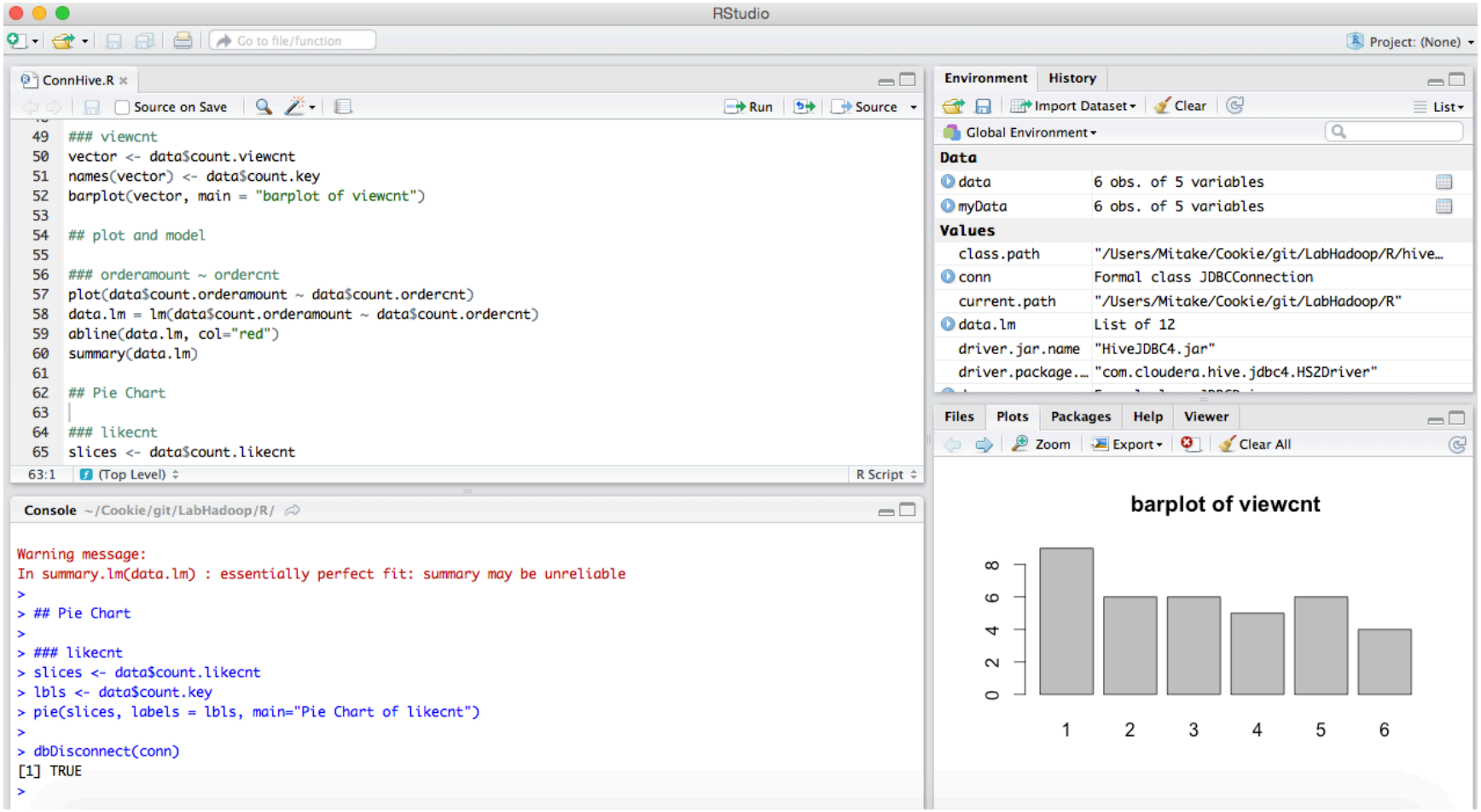
Learning HBase Shell

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



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