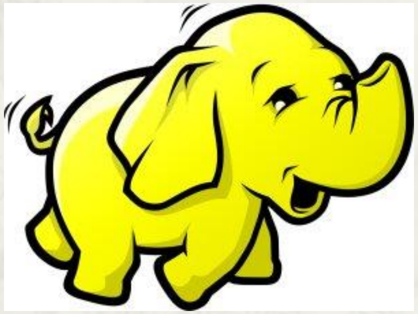
Hadoop+R應用於資料分析

Outline

- Hadoop 介紹
- R 介紹
- Hadoop+R 介紹
- Mongodb
- Fluentd
- Drill
- Demo

Hadoop



What is Hadoop?

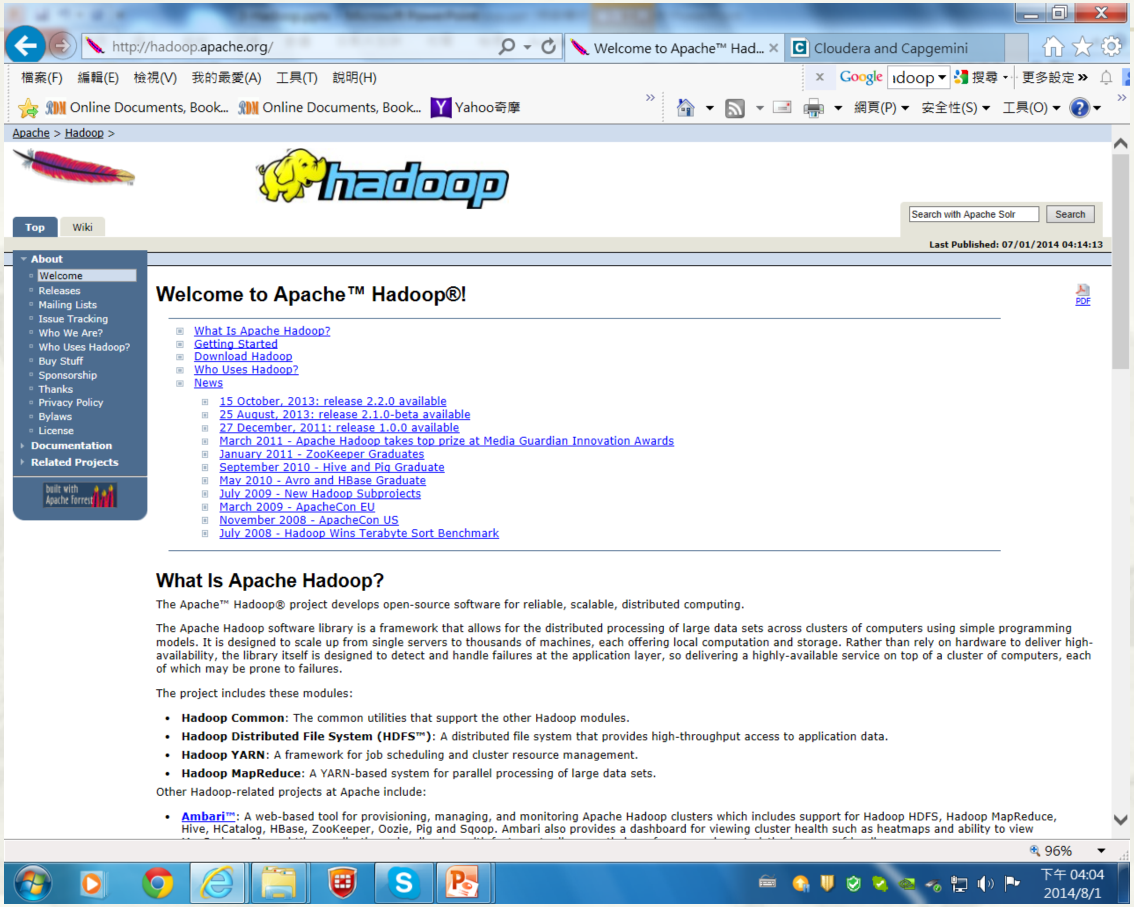
- Hadoop is a software platform that lets one easily write and run applications that process vast amount of data.
  - ✓ 它是軟體平台用來處理程式具巨量資料
- Hadoop can reliably store and process petabytes.
  - ✓ 它可用來可靠地儲存和處理PB級巨量資料
- It distributes the data and processing across clusters of commonly available computers. These clusters can number into the thousands of nodes.
  - ✓ 它將資料和處理程序分散到可以使用的電腦上，而且這些電腦的數量可以達到上千台之多

What is Hadoop?

- By distributing the data, Hadoop can process it in parallel on the nodes where the data is located. This make it extremely rapid.
  - ✓ 藉由分散資料的處理，Hadoop可以平行的運算這些資料，使得處理速度變得非常快速

- Hadoop automatically maintains multiple copies of data and automatically redeploys computing tasks based on failures.
  - ✓ Hadoop可以將運算的程式和放置的資料在每一個可以運行的節點間進行複製和自動化的備份，可以避免執行中的程式或存放的資料，因為電腦的硬體或系統的上的損壞而使程式無法執行或檔案損毀

Apache Hadoop



Last Published: 07/01/2014 04:14:13

Welcome to Apache™ Hadoop®!

- What Is Apache Hadoop?
- Getting Started
- Download Hadoop
- Who Uses Hadoop?
- News
  - 15 October, 2013: release 2.2.0 available
  - 25 August, 2013: release 2.1.0-beta available
  - 27 December, 2011: release 1.0.0 available
  - March 2011 - Apache Hadoop takes top prize at Media Guardian Innovation Awards
  - January 2011 - ZooKeeper Graduates
  - September 2010 - Hive and Pig Graduate
  - May 2010 - Avro and HBase Graduate
  - July 2009 - New Hadoop Subprojects
  - March 2009 - ApacheCon EU
  - November 2008 - ApacheCon US
  - July 2008 - Hadoop Wins Terabyte Sort Benchmark

What Is Apache Hadoop?

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

Other Hadoop-related projects at Apache include:

- **Ambari™**: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view

http://hadoop.apache.org/

Hadoop

- 免費軟體
- 利用MapReduce作為分散式處理技術
- 利用HDFS作為分散式檔案系統

MapReduce

MapReduce

- 是一種軟體框架(Software Framework)
- 可在不同電腦組成的叢集(Clusters)上執行
- 能為巨量資料(Big Data)做分散運算處理
- 此框架的功能概念主要是映射(Map)和化簡(Reduce)兩種
- 實作上可用JAVA、R或其他程式語言來達成

MapReduce

- Map
  - ✓ 從主節點(Master Node)輸入一組Input，此Input是一組 key/value序對，將這組輸入切分成好幾個小的子部分，分散到各個工作節點(Slave Nodes)去做運算
  - ✓ 輸入是一組 Key/Value 序對 ，輸出則為另一組中間過程(Intermediate)的 key/value 序對
    - $(K_{in}, V_{in}) \rightarrow \text{list}(K_{inter}, V_{inter})$

MapReduce

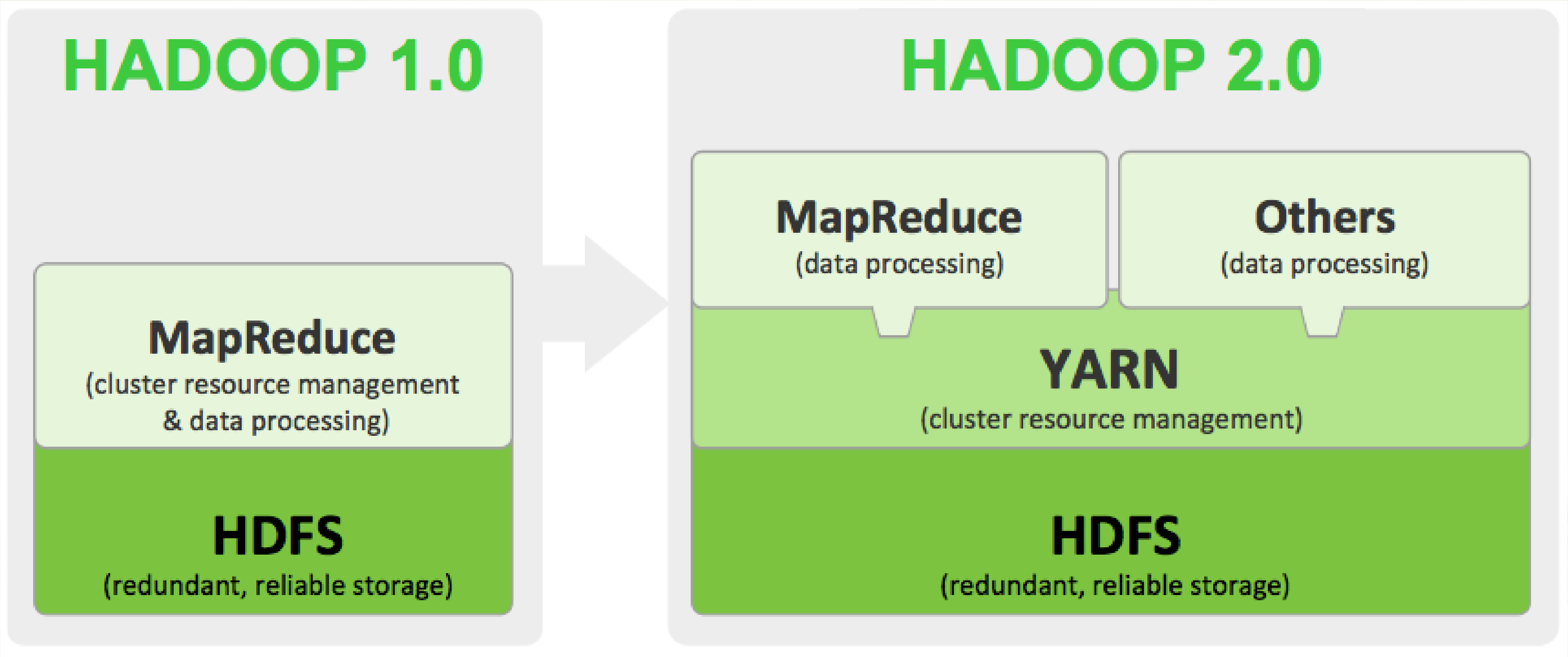
- Reduce
  - ✓ 負責針對相同的中間過程 key 合併其所有相關聯的 Value，並產生輸出結果的 key/value 序對
  - ✓ 將多對具相同 Key但不同 Value 的資料，結合為多對的 Key/Value
    - ◆$(K_{inter}, list(V_{inter}))$ ➔ $list(K_{out}, V_{out})$

HDFS

Hadoop Distributed File System (HDFS)

- 在分散式儲存環境中，提供單一的目錄系統
- 資料以 Write-once-read-many 方式存取
- 每個檔案被分割成許多Block，每個Block複製許多複本(Replica)，並分散儲存於不同的DataNode上
  - ✓ NameNode：負責維護HDFS的檔案名稱空間 (File System Namespace)
  - ✓ DataNode：實際儲存檔案區塊(Blocks)的伺服器

Hadoop 2.X


HADOOP 1.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS
(redundant, reliable storage)

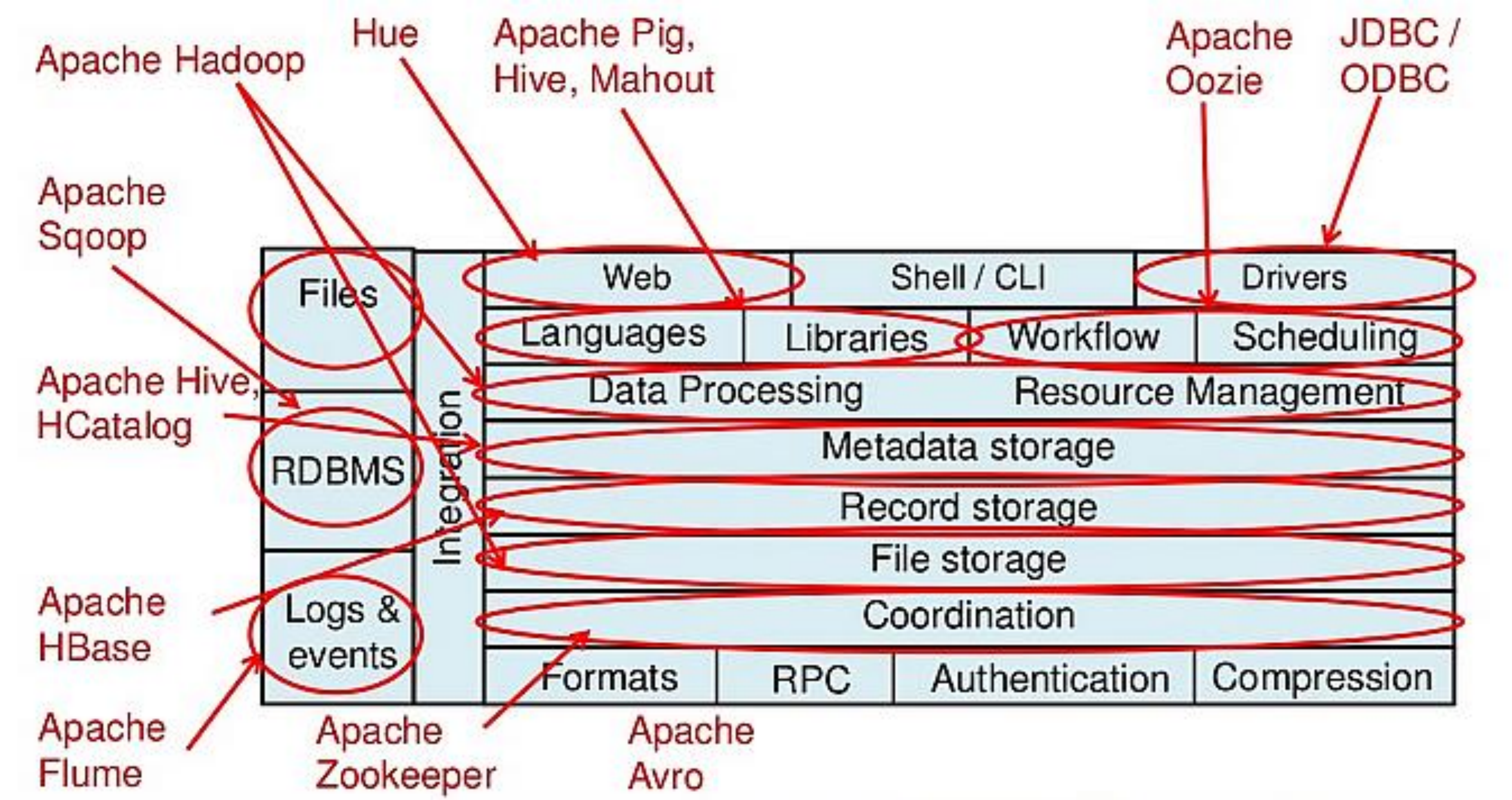
http://hortonworks.com/blog/office-hours-qa-on-yarn-in-hadoop-2/

YARN

- Yet Another Resource Negotiator

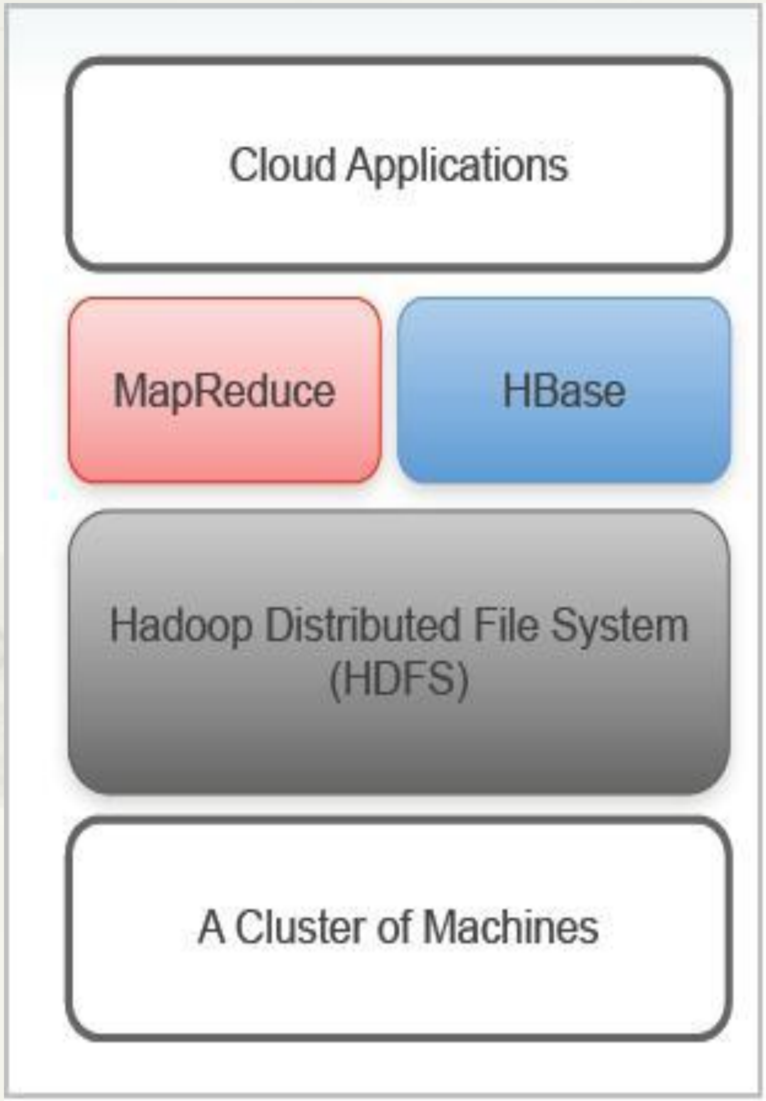
- YARN is a more general purpose framework of which classic MapReduce is one application.
  - ✓ YARN 是更通用的軟體框架，而 MapReduce 只是其中的一個應用

Hadoop Ecosystem



http://www.slideshare.net/cloudera/the-hadoop-stack-then-now-and-in-the-future-eli-collins-charles-zedlewski-cloudera

HBase



HBase:

- 是HDFS上的資料庫。
- 沒有正規化與Join的觀念
- 利用Family Columns將相似的欄位群聚在一起，用於強化效率。

Hive

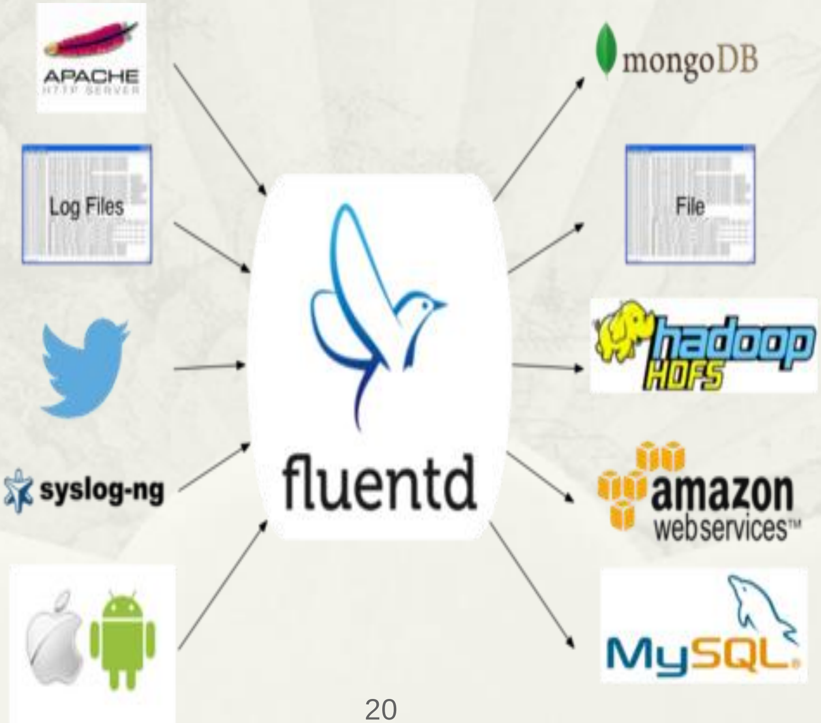
- Developed at Facebook
- “Relational database” built on Hadoop
  - ✓ Maintains list of table schemas
  - ✓ SQL-like query language (HiveQL)
  - ✓ Can call Hadoop Streaming scripts from HiveQL

Sqoop

- 是一個用來將 Hadoop 和關聯式資料庫中的資料相互轉移的工具，可以將一個關聯式資料庫（MySQL, SQL Server等）中的資料導入到 Hadoop 的 HDFS 中，也可以將 HDFS 的資料導入到關聯式資料庫中

Fluentd

- 是一個日誌收集系統，它的特點是可通過簡單的配置，將日誌收集到不同的地方
- 使用 Fluentd + MongoDB 構建即時日誌收集系統

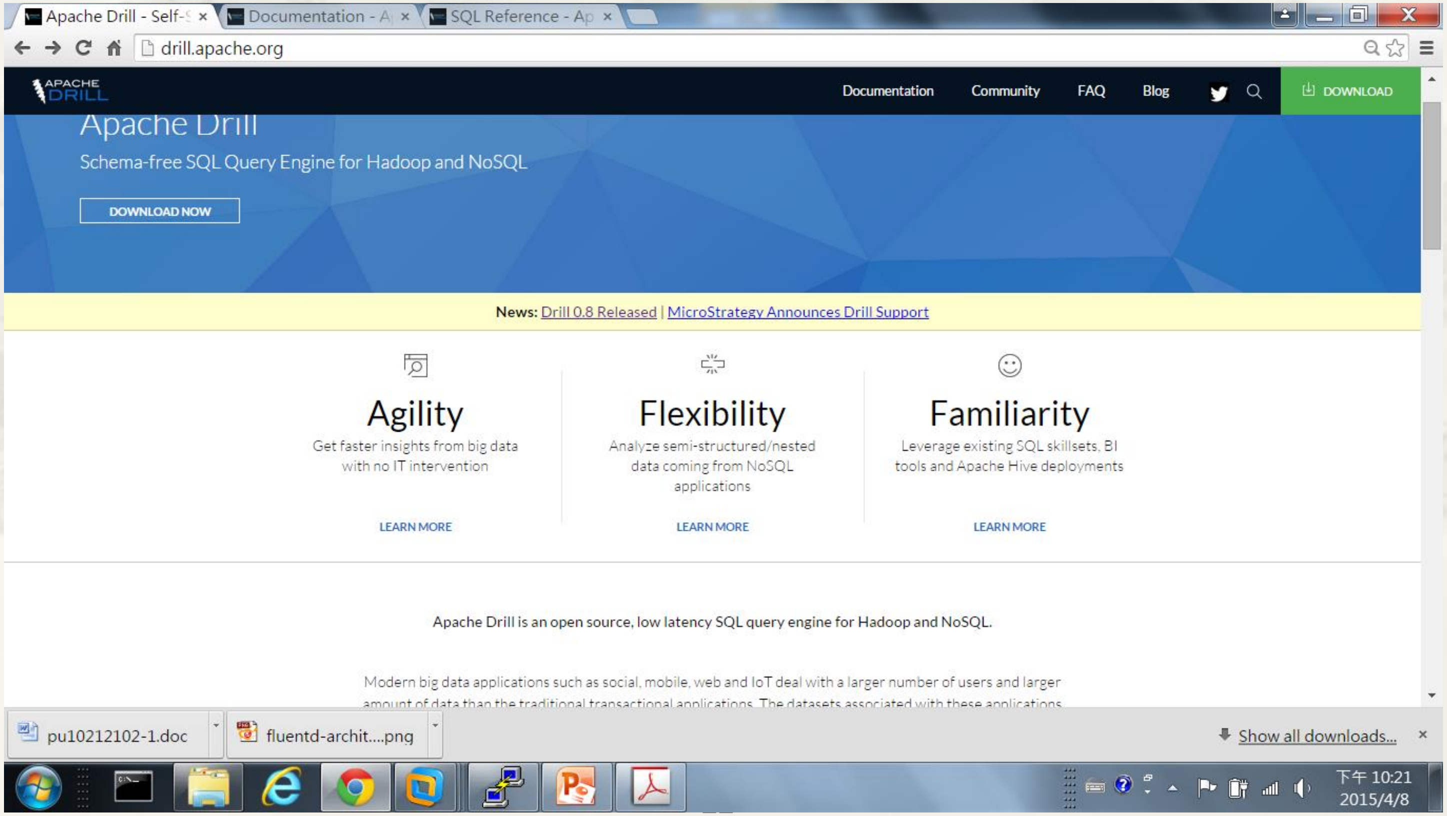


MongoDB

- 隨著資訊爆炸時代的來臨，RDBMS 的效能與彈性遭遇相當大的瓶頸，因此我們開始需要一個可以處理巨量資料(Big Data)的儲存方式，這時 NoSQL 已經悄悄誕生
- Mongodb是由10gen團隊所開發的一套NoSQL程式，它是文件式的資料庫系統，也就是說你的每一個Mongo資料庫都會以「檔案」的形式存放在資料夾中，如果要刪除資料庫，就直接把檔案刪掉就可以了

Drill

- Apache Drill 將有助於Hadoop用戶實現以SQL指令更快查詢巨量資料



Apache Drill

Schema-free SQL Query Engine for Hadoop and NoSQL

News: Drill 0.8 Released | MicroStrategy Announces Drill Support

Agility

Get faster insights from big data with no IT intervention

Flexibility

Analyze semi-structured/nested data coming from NoSQL applications

Familiarity

Leverage existing SQL skillsets, BI tools and Apache Hive deployments

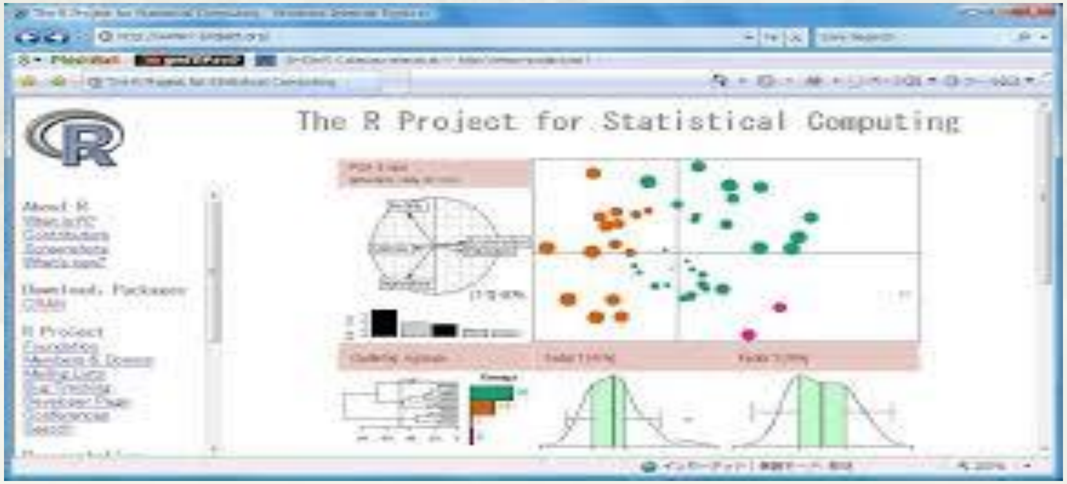
Apache Drill is an open source, low latency SQL query engine for Hadoop and NoSQL.

Modern big data applications such as social, mobile, web and IoT deal with a larger number of users and larger amount of data than the traditional transactional applications. The datasets associated with these applications

R 語言

R 簡介

- Ross Ihaka 與 Robert Gentleman (1966) 所開發出來之相似於 AT & T 貝爾實驗室所開發之 S 語言
- R 有 Windows、Unix、Linux 及 Apple MacOS 等不同作業系統的版本
- 免費軟體，其網站位於 http://www.r-project.org



R 語言

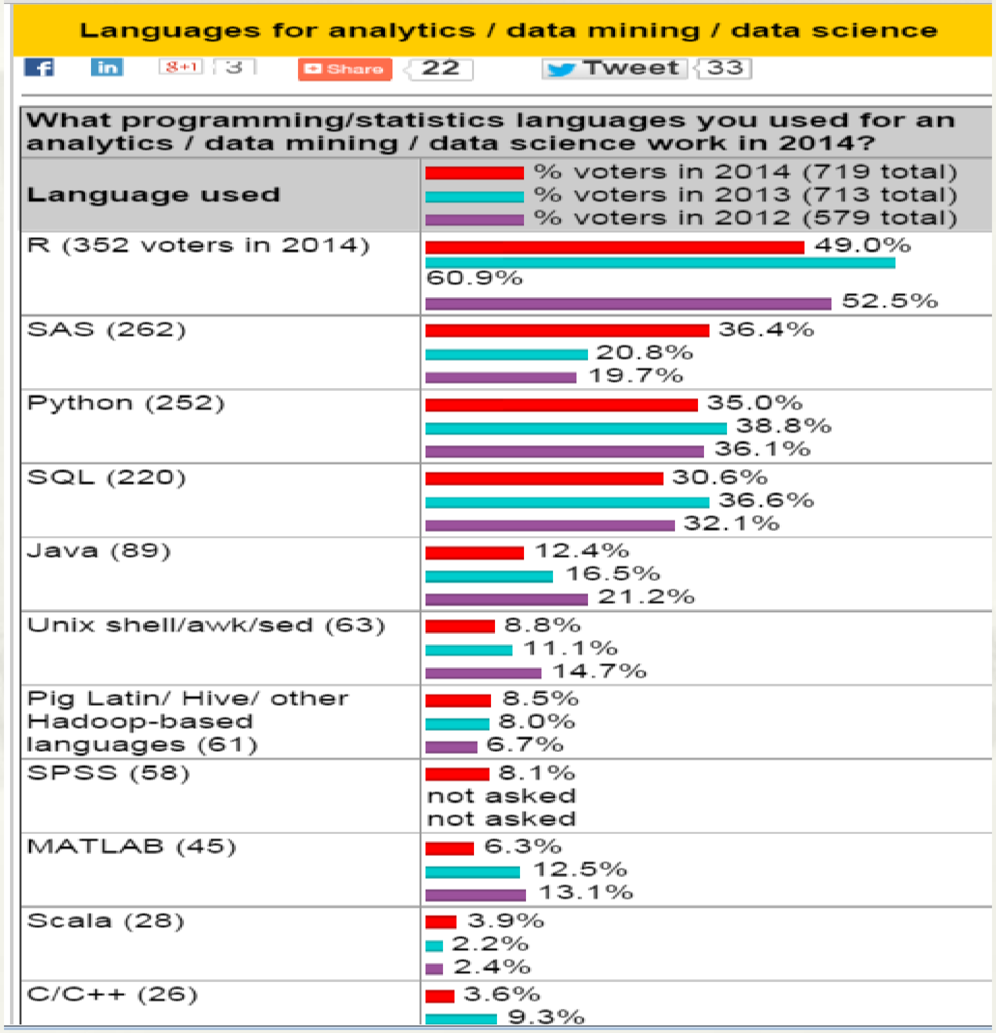
- 內建許多函式(Function)及約5000多個免費套件

- R 是直譯式語言 (Interpreted Language)
  - 一行行執行，可直接看到執行結果

- R 是物件導向語言 (Object Oriented Language)

常用的資料分析語言

最近最受歡迎的資料分析語言 R



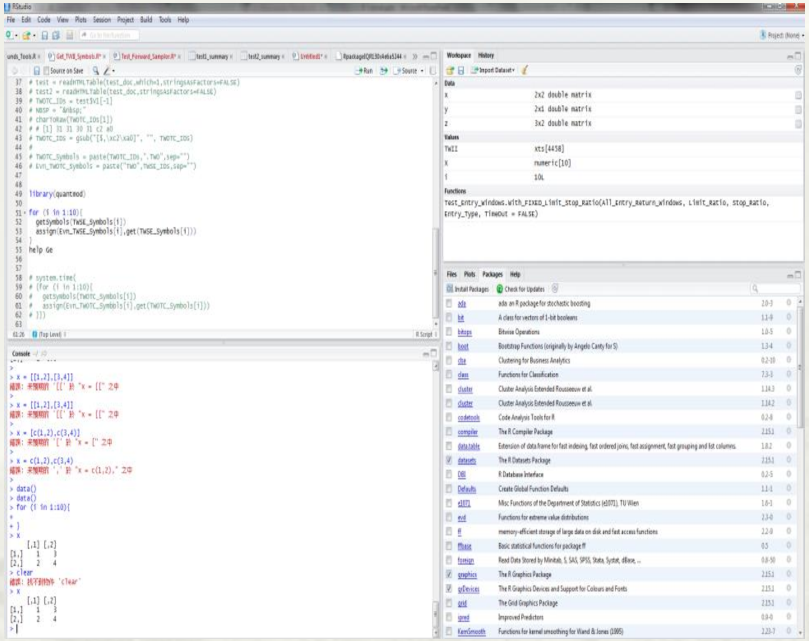
Languages for analytics / data mining / data science

What programming/statistics languages you used for an analytics / data mining / data science work in 2014?

Language used	🟥 % voters in 2014 (719 total)	🟦 % voters in 2013 (713 total)	🟪 % voters in 2012 (579 total)
R (352 voters in 2014)	49.0%	60.9%	52.5%
SAS (262)	36.4%	20.8%	19.7%
Python (252)	35.0%	38.8%	36.1%
SQL (220)	30.6%	36.6%	32.1%
Java (89)	12.4%	16.5%	21.2%
Unix shell/awk/sed (63)	8.8%	11.1%	14.7%
Pig Latin/ Hive/ other Hadoop-based languages (61)	8.5%	8.0%	6.7%
SPSS (58)	8.1%	not asked	not asked
MATLAB (45)	6.3%	12.5%	13.1%
Scala (28)	3.9%	2.2%	2.4%
C/C++ (26)	3.6%	9.3%	

http://www.kdnuggets.com/polls/2014/languages-analytics-data-mining-data-science.html

整合式開發環境 IDE



```r
# test = readHTMLTable(test_doc,which=1,stringsAsFactors=FALSE)
# test2 = readHTMLTable(test_doc,stringsAsFactors=FALSE)
# TWOTC_IDs = test$V1[-1]
# NBSP = "&nbsp;"
# charToRaw(TWOTC_IDs[1])
# # [1] 31 31 30 31 c2 a0
# TWOTC_IDs = gsub("[$,\xc2\xa0]", "", TWOTC_IDs)
#
# TWOTC_Symbols = paste(TWOTC_IDs,".TWO",sep="")
# Evn_TWOTC_Symbols = paste("TWO",TWSE_IDs,sep="")


library(quantmod)

for (i in 1:10){
  getSymbols(TWSE_Symbols[i])
  assign(Evn_TWSE_Symbols[i],get(TWSE_Symbols[i]))
}
help Ge


# system.time(
# {for (i in 1:10){
#   getSymbols(TWOTC_Symbols[i])
#   assign(Evn_TWOTC_Symbols[i],get(TWOTC_Symbols[i]))
# }})

```

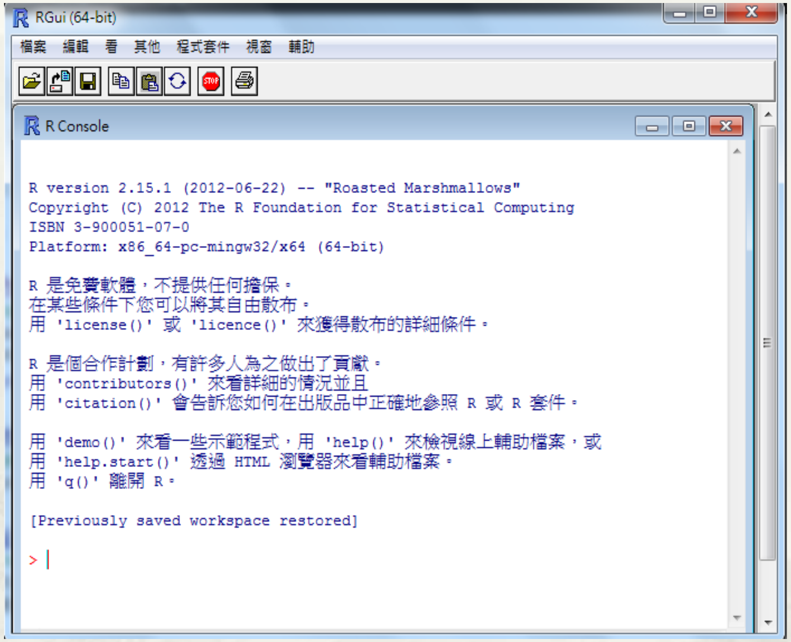
```
>
> x = [[1,2],[3,4]]
錯誤: 未預期的 '[[' 於 "x = [[" 之中
>
> x = [[1,2],[3,4]]
錯誤: 未預期的 '[[' 於 "x = [[" 之中
>
> x = [c(1,2),c(3,4)]
錯誤: 未預期的 '[' 於 "x = [" 之中
>
> x = c(1,2),c(3,4)
錯誤: 未預期的 ',' 於 "x = c(1,2)," 之中
>
> data()
> data()
> for (i in 1:10){
+
+ }
> x
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> clear
錯誤: 找不到物件 'clear'
> x
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> |
```

```
Data
x                    2x2 double matrix
y                    2x1 double matrix
z                    3x2 double matrix
Values
TWII                 xts[4458]
X                    numeric[10]
i                    10L
Functions
Test_Entry_Windows.with_FIXED_Limit_Stop_Ratio(All_Entry_Return_Windows, Limit_Ratio, Stop_Ratio,
Entry_Type, TimeOut = FALSE)
```

□	ada	ada: an R package for stochastic boosting	2.0-3
□	bit	A class for vectors of 1-bit booleans	1.1-9
□	bitops	Bitwise Operations	1.0-5
□	boot	Bootstrap Functions (originally by Angelo Canty for S)	1.3-4
□	cba	Clustering for Business Analytics	0.2-10
□	class	Functions for Classification	7.3-3
□	cluster	Cluster Analysis Extended Rousseeuw et al.	1.14.3
□	cluster	Cluster Analysis Extended Rousseeuw et al.	1.14.2
□	codetools	Code Analysis Tools for R	0.2-8
□	compiler	The R Compiler Package	2.15.1
□	data.table	Extension of data.frame for fast indexing, fast ordered joins, fast assignment, fast grouping and list columns.	1.8.2
☑	datasets	The R Datasets Package	2.15.1
□	DBI	R Database Interface	0.2-5
□	Defaults	Create Global Function Defaults	1.1-1
□	e1071	Misc Functions of the Department of Statistics (e1071), TU Wien	1.6-1
□	evd	Functions for extreme value distributions	2.3-0
□	ff	memory-efficient storage of large data on disk and fast access functions	2.2-9
□	ffbase	Basic statistical functions for package ff	0.5
□	foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase, ...	0.8-50
☑	graphics	The R Graphics Package	2.15.1
☑	grDevices	The R Graphics Devices and Support for Colours and Fonts	2.15.1
□	grid	The Grid Graphics Package	2.15.1
□	ipred	Improved Predictors	0.9-0
□	KernSmooth	Functions for kernel smoothing for Wand & Jones (1995)	2.23-7

R Studio

http://www.rstudio.com/



```
R version 2.15.1 (2012-06-22) -- "Roasted Marshmallows"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-mingw32/x64 (64-bit)

R 是免費軟體，不提供任何擔保。
在某些條件下您可以將其自由散布。
用 'license()' 或 'licence()' 來獲得散布的詳細條件。

R 是個合作計劃，有許多人為之做出了貢獻。
用 'contributors()' 來看詳細的情況並且
用 'citation()' 會告訴您如何在出版品中正確地參照 R 或 R 套件。

用 'demo()' 來看一些示範程式，用 'help()' 來檢視線上輔助檔案，或
用 'help.start()' 透過 HTML 瀏覽器來看輔助檔案。
用 'q()' 離開 R。

[Previously saved workspace restored]

> |
```

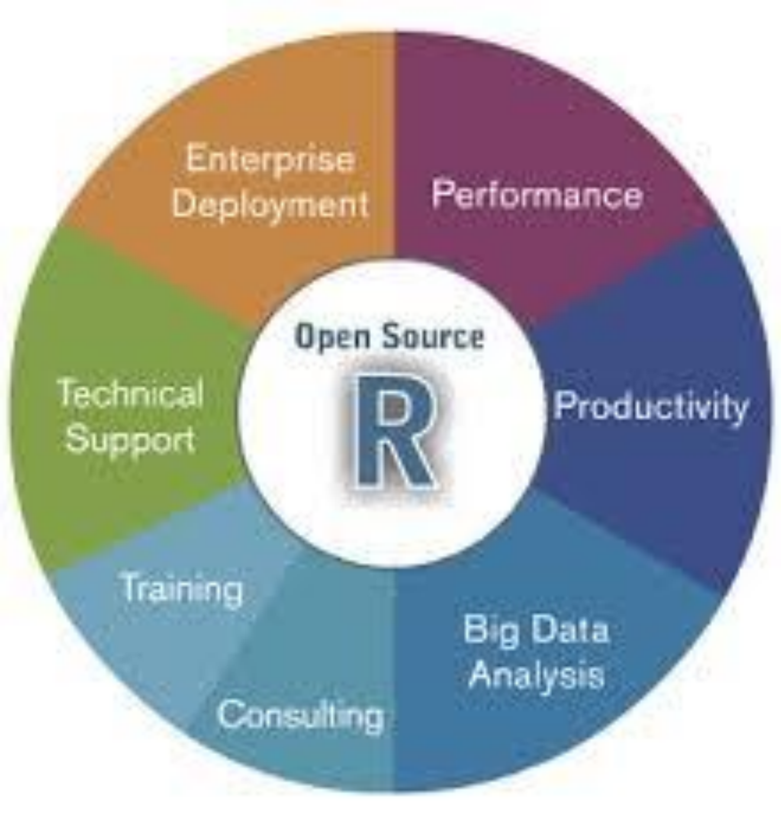
R

http://www.r-project.org/

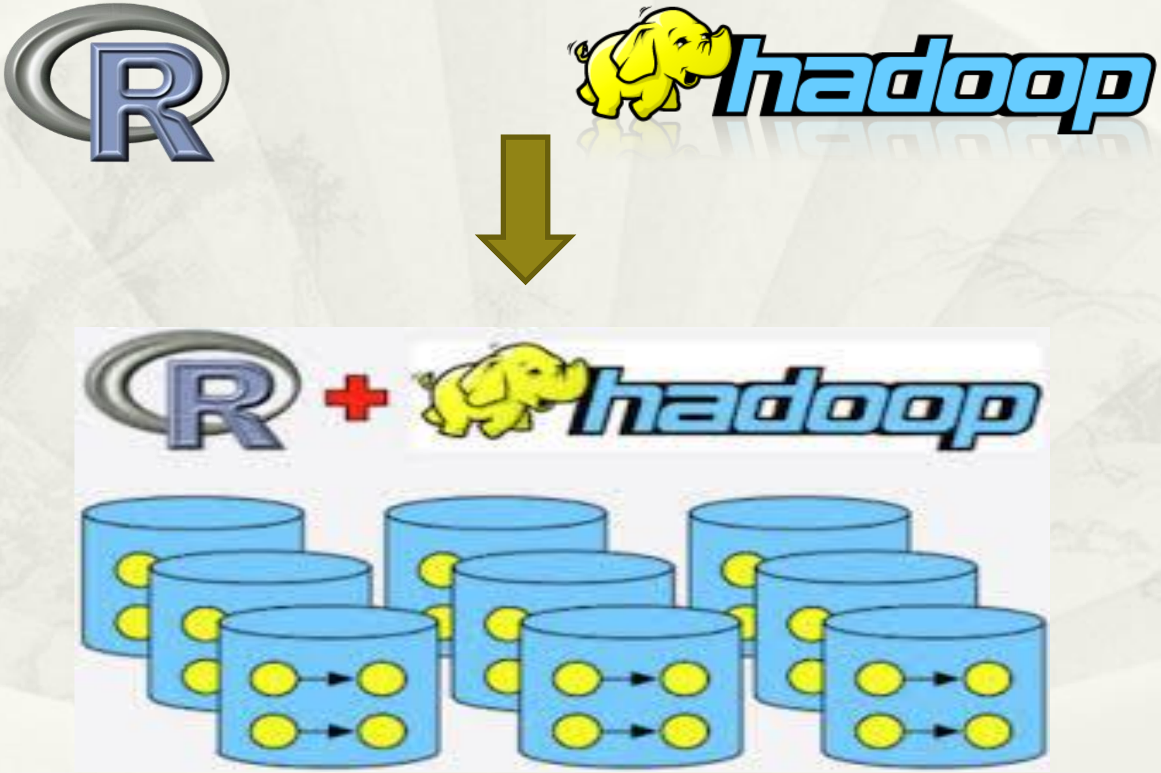
R 應用領域

- Big Data
- 統計分析
- 資料探勘
- 機器學習
- 推薦系統
- 文字探勘

…



Hadoop+R



Hadoop+R

- 擴大 R 處理資料能力
  - ✓ R 將資料全部讀進 Memory（無法讀入巨量資料）
  - ✓ Hadoop 讓 R 可以進行分散式運算
- 使用 R 語言就可輕易使用Hadoop功能

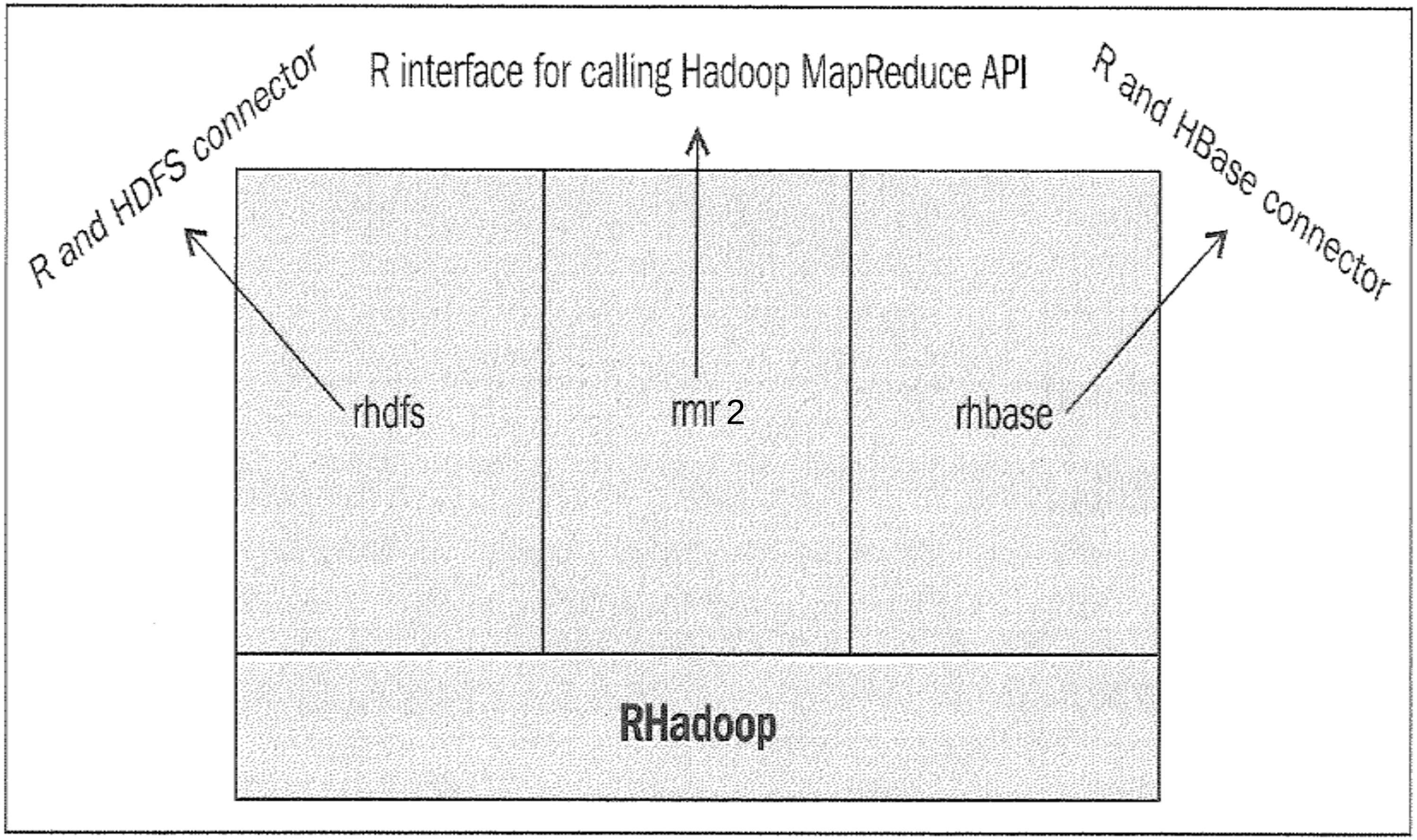
RHadoop 套件

- 由 Revolution Analytics 開發
- 針對 MapReduce、HDFS 及 HBase發展三個免費套件
  - ✓ rmr2
  - ✓ rhdfs
  - ✓ rhbase

RHadoop 套件功能

- rhdfs
  - ✓ 讓使用者可以透過 R 存取 HDFS
- rmr2
  - ✓ 可以讓使用者發展並呼叫 MapReduce 工作
- rhbase
  - ✓ 可以操作 HBase 資料

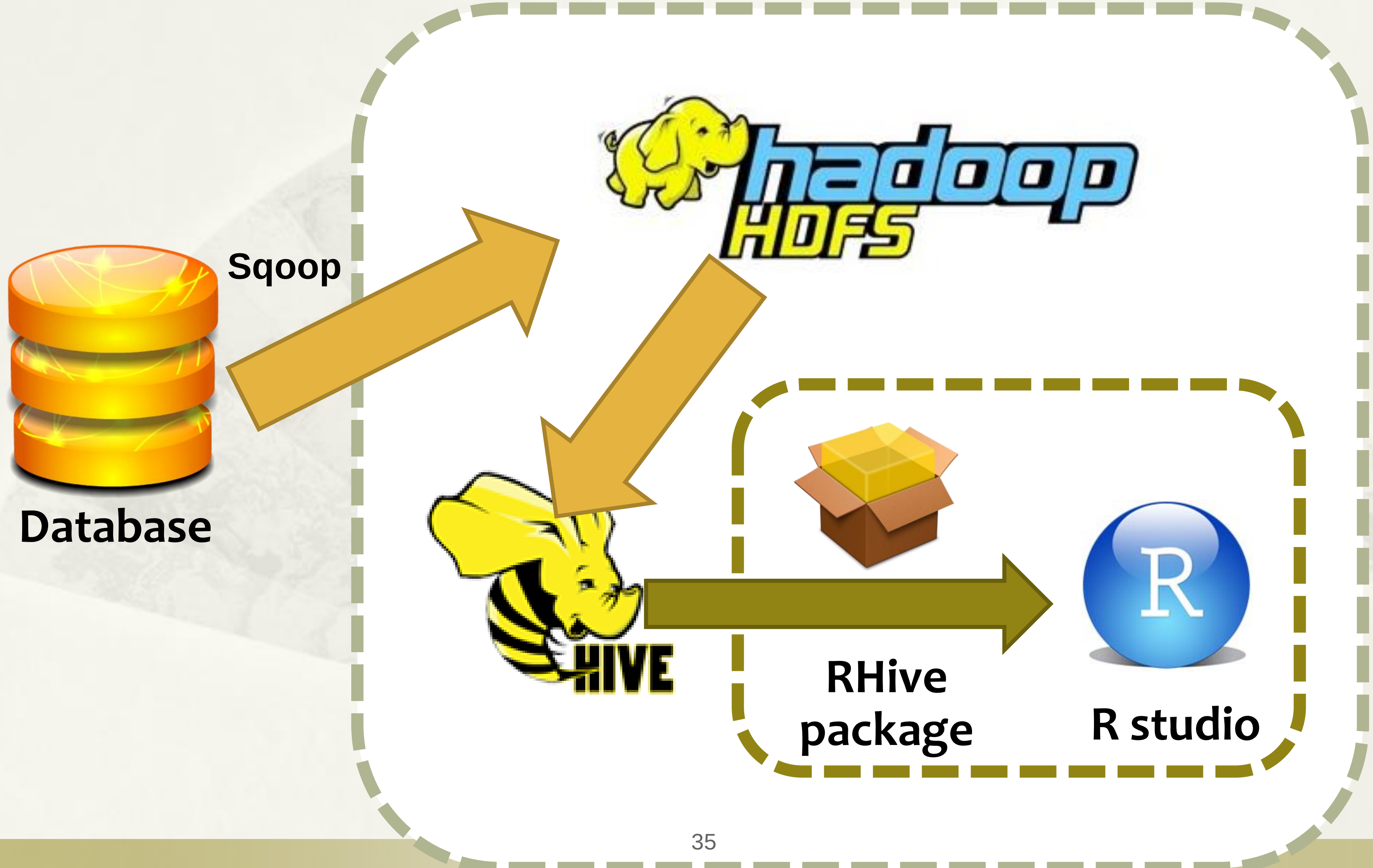
RHadoop 套件架構



RHadoop Ecosystem

Demo

執行流程

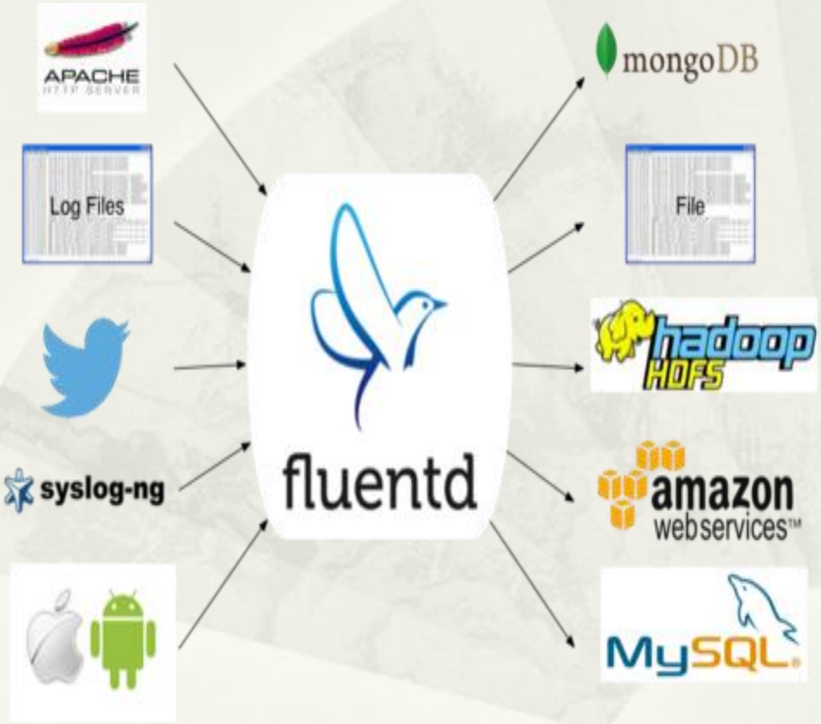


Sqoop 指令

```
sqoop import --connect
"jdbc:sqlserver://$SQL_SRV:$PORT;database=mitopac;username=$U
;password=$P" --hive-import -m 1 --table reader --warehouse-dir
$DEST_DIR --map-column-hive reader11=String --hive-overwrite
```

Demo

Log 資料收集及分析


APACHE
HTTP SERVER
Log Files
syslog-ng
fluentd
mongoDB
File
hadoop
HDFS
amazon
webservices™
MySQL

APACHE
DRILL

Fluentd+Mongodb(192.168.244.131)

```
1. # login root
2. # mongo httpd
3. > db.accesslog.count()
4. > db.accesslog.find()
5. quit()
```

Drill (192.168.244.131:8047, select * from mongo.httpd. accesslog limit 10)

```
1. bin/sqlline -u jdbc:drill:zk=local (./start)
2. use mongo.httpd;
3. show tables;
4. select * from accesslog; select host from accesslog;
5. !quit
```