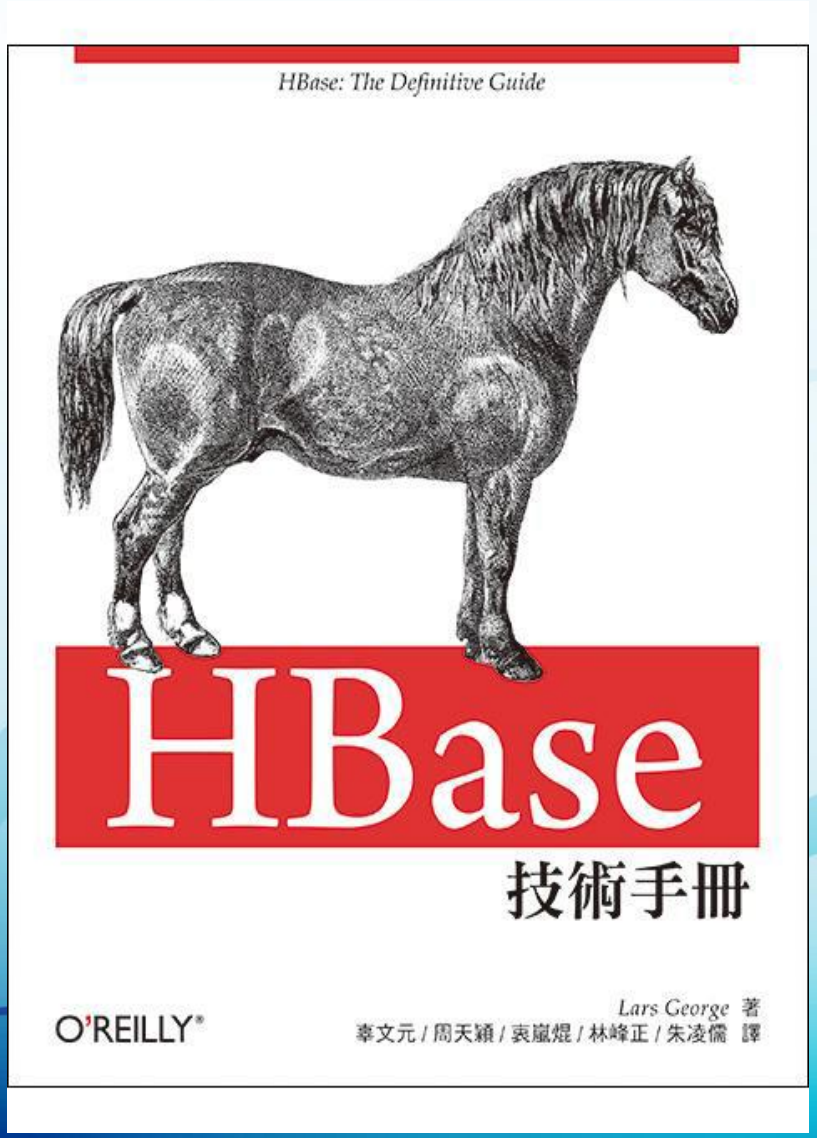
Apache Hadoop
專案經驗&教學經驗分享

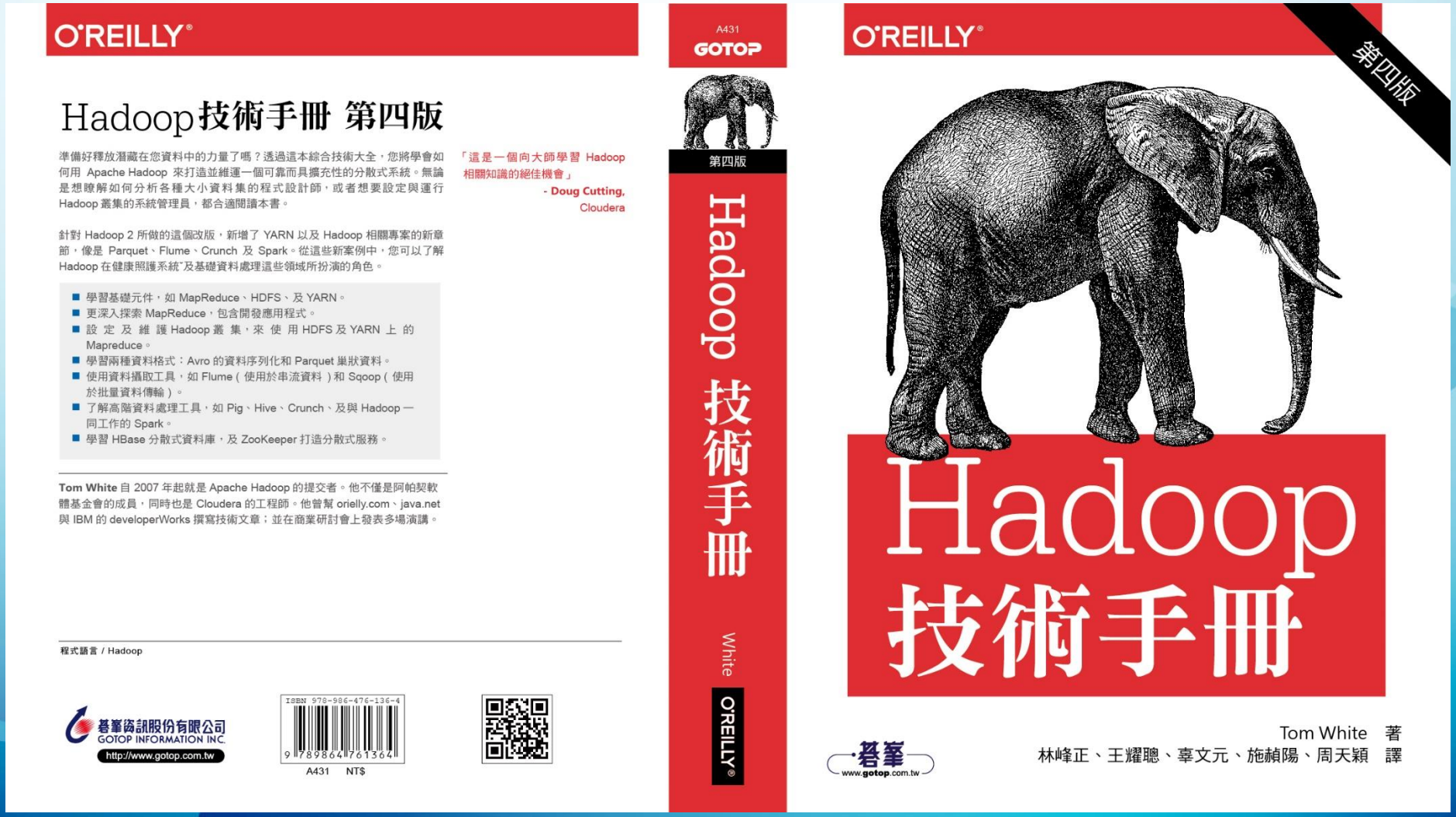
逢甲大學資訊工程系
林峰正 助理教授
fclin@fcu.edu.tw

2018/3/31

工商服務


HBase: The Definitive Guide
HBase
技術手冊
O'REILLY
Lars George 著
辜文元 / 周天穎 / 袁嵐煜 / 林峰正 / 朱凌儒 譯

工商服務


O'REILLY®
Hadoop技術手冊 第四版
準備好釋放潛藏在您資料中的力量了嗎？透過這本綜合技術大全，您將學會如何用 Apache Hadoop 來打造並維運一個可靠而具擴充性的分散式系統。無論是想瞭解如何分析各種大小資料集的程式設計師，或者想要設定與運行 Hadoop 叢集的系統管理員，都合適閱讀本書。
針對 Hadoop 2 所做的這個改版，新增了 YARN 以及 Hadoop 相關專案的新章節，像是 Parquet、Flume、Crunch 及 Spark。從這些新案例中，您可以了解 Hadoop 在健康照護系統"及基礎資料處理這些領域所扮演的角色。
學習基礎元件，如 MapReduce、HDFS、及 YARN。
更深入探索 MapReduce，包含開發應用程式。
設定及維護 Hadoop 叢集，來使用 HDFS 及 YARN 上的 Mapreduce。
學習兩種資料格式：Avro 的資料序列化和 Parquet 巢狀資料。
使用資料攝取工具，如 Flume（使用於串流資料）和 Sqoop（使用於批量資料傳輸）。
了解高階資料處理工具，如 Pig、Hive、Crunch、及與 Hadoop 一同工作的 Spark。
學習 HBase 分散式資料庫，及 ZooKeeper 打造分散式服務。
「這是一個向大師學習 Hadoop 相關知識的絕佳機會」
- Doug Cutting, Cloudera
Tom White 自 2007 年起就是 Apache Hadoop 的提交者。他不僅是阿帕契軟體基金會的成員，同時也是 Cloudera 的工程師。他曾幫 oreilly.com、java.net 與 IBM 的 developerWorks 撰寫技術文章；並在商業研討會上發表多場演講。
程式語言 / Hadoop
碁峰資訊股份有限公司
GOTOP INFORMATION INC.
http://www.gotop.com.tw
ISBN 978-986-476-136-4
A431 NT$
A431
GOTOP
第四版
Hadoop 技術手冊
White
O'REILLY®
O'REILLY®
第四版
Hadoop
技術手冊
碁峰
www.gotop.com.tw
Tom White 著
林峰正、王耀聰、辜文元、施賦陽、周天穎 譯

Hadoop

- 於專案
- 於教學
- 結論

Hadoop 於專案

建立簡易雲端應用介面

1.雲端運算模版

2.水利雲端開發插件

3.SaaS案例-水利署公文系統

建立簡易雲端應用介面

雲端運算模版架構圖

GIS.FCU

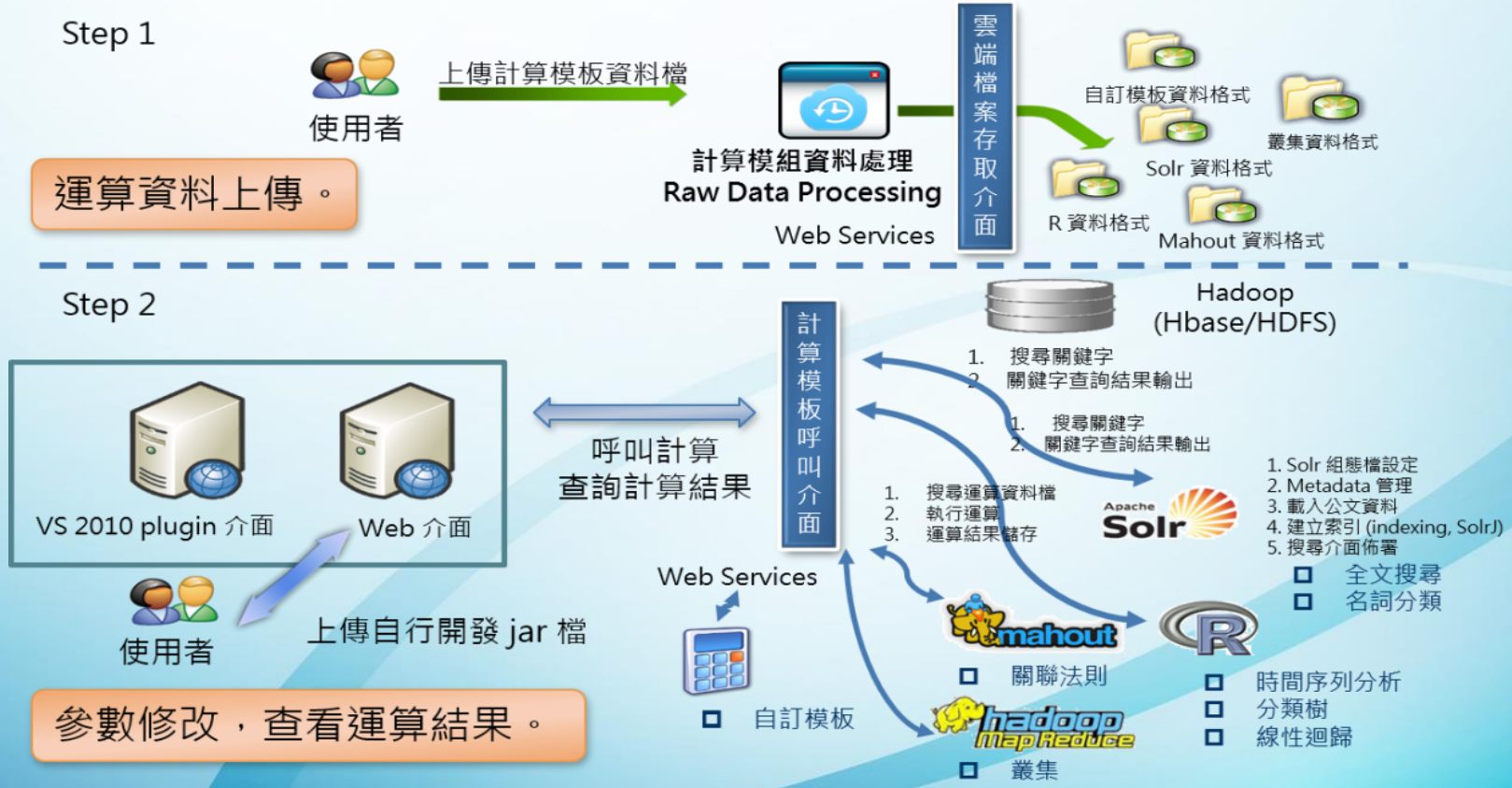
2014 協助經濟部水利署

建立簡易雲端應用介面

運算模版功能介紹

開發模板	作用	演算法	公文案例使用模板	引用模板範例
分類樹 (decision tree)	分類樹中的每個分支即為判斷準則，每個葉節點就是一連串法則後的分類結果，分類樹的結果也就是決策樹，可以用來做為預測使用。	forest decision tree with Mahout	利用訓練樣本訓練出決策樹，當一新公文需求丟入該模型，輸入承辦人員、來文機關、回應機關、與辦文期限要求(最速件：一日、速件：三日、普通件：六日)，即可預測公文回應時間。	**預測公文處理速度**
叢集 (Clustering)	將資料分為數群，其目的是要將群與群之間的差異找出來，同時也要將一個群之中特徵的相似性找出來。	K-Means with Mahout	萃取公文屬性，並給予給每一屬性類別編號，根據分群的參數(*n*群)，執行多維度的叢集運算，將同性質公文歸類一群。	**公文分群**
關聯 (Association)	找出在某一事件或是資料中會同時出現項目的關聯性。	Frequent Pattern Mining with Mahout	找到某一個關鍵字出現在某一個公文主旨中，同時也會出現在其他公文主旨的關聯性。	**類似查詢(more like this) – 依文找文**
時間序列分析 (Time series Analysis)	時間序列是依事件或資料發生的先後次序排列的一群統計數據，而時間數列分析的目的在於觀察及分析過去的資料，是一種預測未來的分析方法。	Time series Analysis with R	以公文案例來說，可以輸入累積週數的公文量，以預測未來一周的公文量。	**預測公文量**
線性迴歸分析 (linear regression)	迴歸分析是一種分析資料的方法，目的在於瞭解兩個或多個變數間相關性及強度。線性迴歸是迴歸分析其中的一種，主要用於單一變數預測及判斷兩變數間的相關性。	Simple linear regression model with R	暫無適用情境	---
全文搜尋 (Full Text Search)	提供使用者透過關鍵字，找出文件中是否與關鍵字相符，並找出關鍵字在文件中的位置。	Hit highlighting search component with Solr	檢索會根據索引的辭庫，支援自動提示（autocomplete），例如輸入保，可以自動提示保育、保固或是保養，並提供命中標示(Hit highlighting search)。	**公文搜尋**
名詞分類 (Noun classification)	經由使用者事先定義好之名詞分類規則，並將大量檔案依其分類規則進行分析，找出所有檔案對映之類別。	Facets search component with Solr	利用名詞分類，經由使用者事先定義好之名詞分類規則，並將大量公文依其分類規則進行分析，可以縮小搜尋範圍。	**公文分類搜尋**

建立簡易雲端應用介面



運算模版運作示意圖

建立簡易雲端應用介面

- LinearRegression.jar (線性回歸運算模板)

指令說明:
1. 登錄用戶
2. 切換目錄目錄至/home/wra；
3. 輸入以下指令(指令一): java -Djava.library.path=/home/wra/R/x86_64-pc-linux-gnu-library/2.15/rJava/jri/ -cp /home/wra/R/x86_64-pc-linux-gnu-library/2.15/rJava/jri -jar LinearRegression.jar -sm TRUE -x /user/wra/x.csv -y /user/wra/y.csv -o /user/wra/linear -d -w 1024 -h 768
3.1. 結果輸出到 /user/wra/linear，輸出圖檔

4. 輸入以下指令(指令二): java -Djava.library.path=/home/wra/R/x86_64-pc-linux-gnu-library/2.15/rJava/jri/ -cp /home/wra/R/x86_64-pc-linux-gnu-library/2.15/rJava/jri -jar LinearRegression.jar -dm TRUE -x1 /user/wra/x1.csv -x2 /user/wra/x2.csv -y /user/wra/y2.csv -o /user/wra/linear2
4.1. 結果輸出到 /user/wra/linear2，不輸出圖檔

輸入運算資料

指令一運算檔案CSV:
x.csv
2,3,4,8
y.csv
5,9,8,9

指令二運算檔案CSV:
x1.csv
2,3,4,8
x2.csv
6,6,5,2
y2.csv
1,3,4,13

線性迴歸分析運算模板指令說明

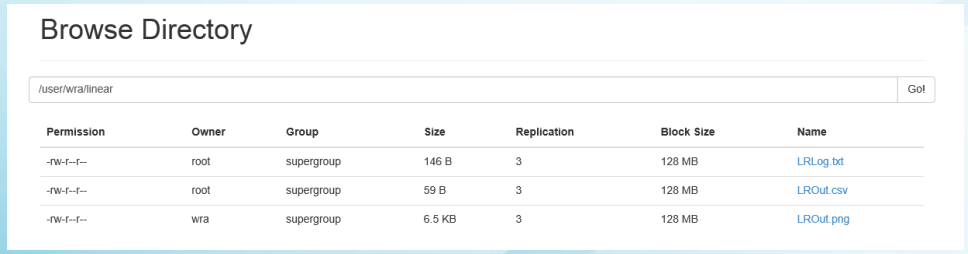
Usage:
[--input <input> --output <output> --frequency <frequency> --gamma <gamma> --forecasting <forecasting> --draw <draw> --width <width>|--height <height> --help]

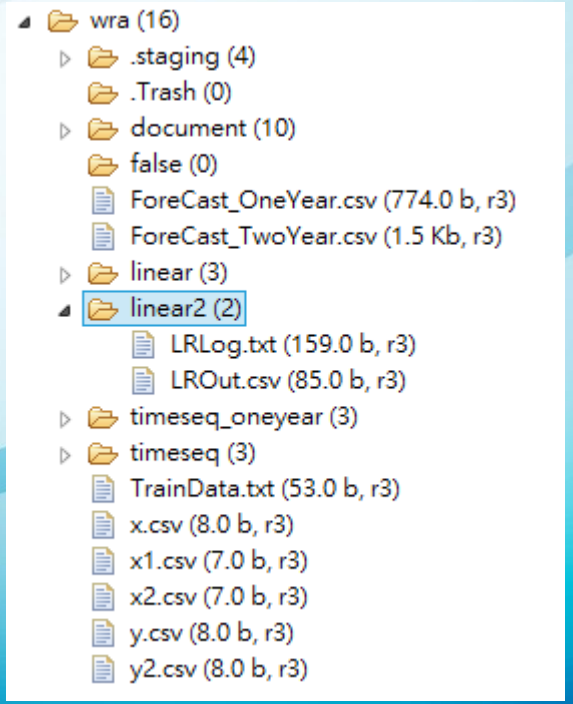
建立簡易雲端應用介面

- LinearRegression.jar (線性回歸運算模板)

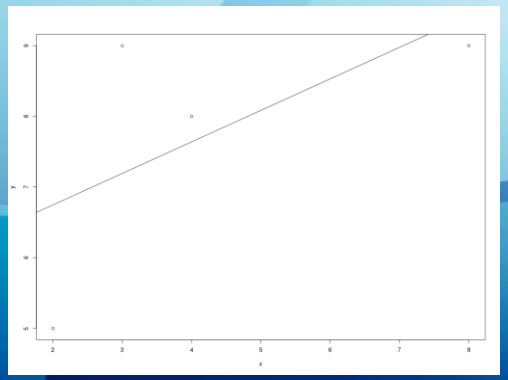
測試指令一的產出，到網頁:
http://10.0.0.52:50070/explorer.html#/user/wra/linear
產生三個檔案，其中 **LROut.csv** 如下，代表截距與x係數。
"","x"
"(Intercept)",5.85542168674699
"x",0.44578313253012
LROut.png 如下，匯出單一自變數的圖式，及下圖佐證執行成功
(linear目錄中有三個檔案)。

測試指令二的產出，到網頁:
http://10.0.0.52:50070/explorer.html#/user/wra/linear2
找到其中之一檔案 **LROut.csv** 如下，代表截距與x1及x2係數。
"","x"
"(Intercept)",-1.18181818181817
"x1",1.81818181818182
"x2",-0.272727272727274
下圖佐證有執行成功 (linear2 目錄夾有兩個檔案):





LROut.png 如下，匯出單一自變數的圖



建立簡易雲端應用介面

- SaaS 整體流程

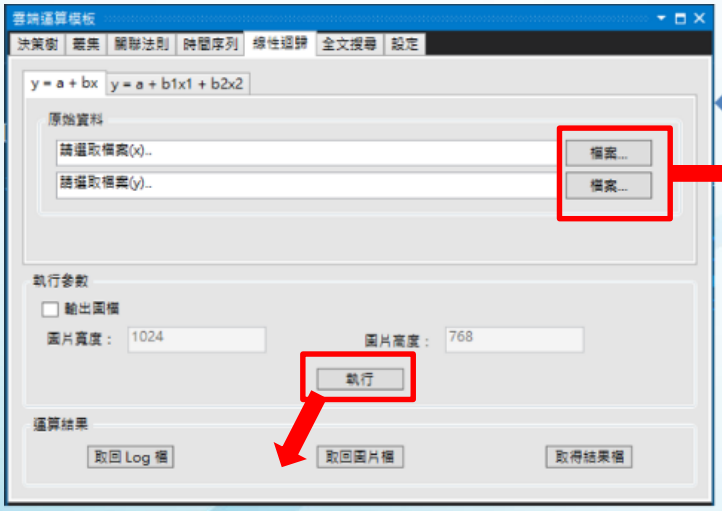
建立簡易雲端應用介面

- 公文系統資料再活化運作介面

介面展示

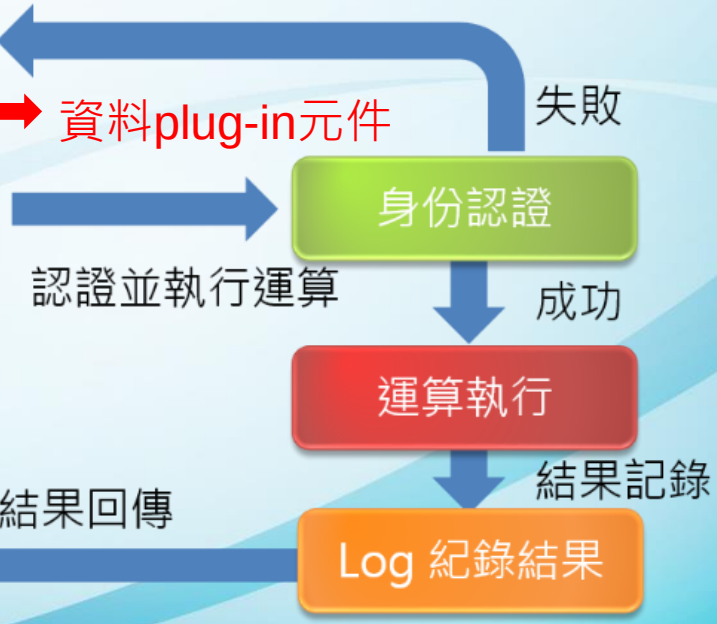
建立簡易雲端應用介面

整合插件(plug-in)處理流程，開發人員完成插件安裝後，會自動將插件加入至Visual Studio的工具箱中，開發人員可以從工具箱中選擇要使用的插件。



資料plug-in元件

運算plug-in元件



依據模版設定相關參數後，即可於開發環境使用該運算模版，並可透過自己的開發環境，撰寫讀取 log 檔以及運算結果的程式片段。
運算成功，傳回運算結果，失敗，log取回。

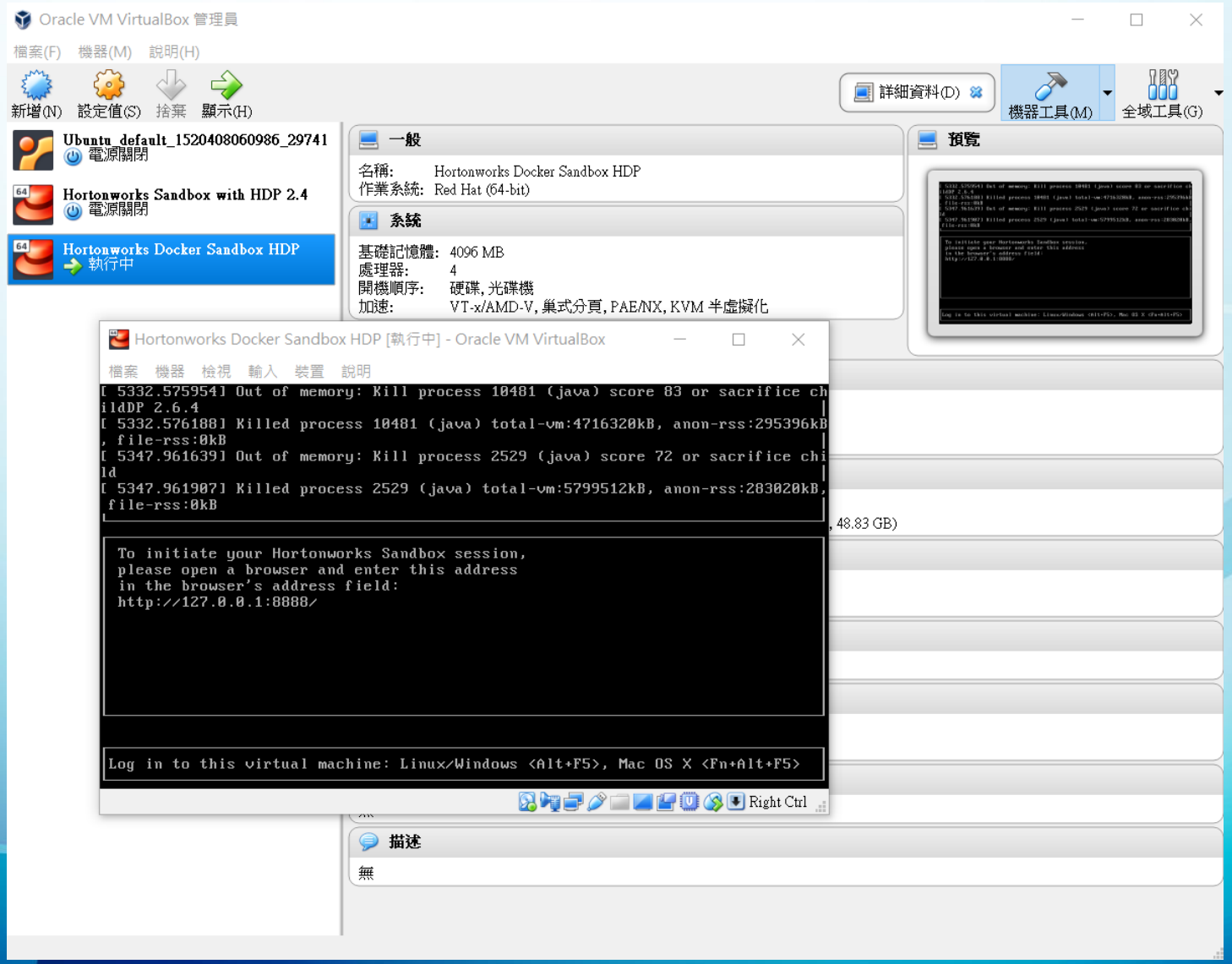
Hadoop 於教學

18 週課程的其中一部份

今天要學會		
課程單元	課程內容	時數
Hadoop 平台簡介與安裝	(1) 開發及管理環境簡介	X
	(2) Hortonworks Sandbox 介紹與操作	
	(3) Vagrant 管理環境建置與操作	
	(4) Aambari 管理監控平台介紹+安裝 (3 VMs in One PC)	
Spark 初步簡介	(1) Spark 單機運作	X

逢甲資工系大三選修 – 雲端應用系統開發

執行單機版 HDP (Hortonworks)



製造一個 HelloWorld.java 檔案

```
public class HelloWorld {
    public static void main(String[] args) {
        System.out.println("Hello! World!");
    }
}
```

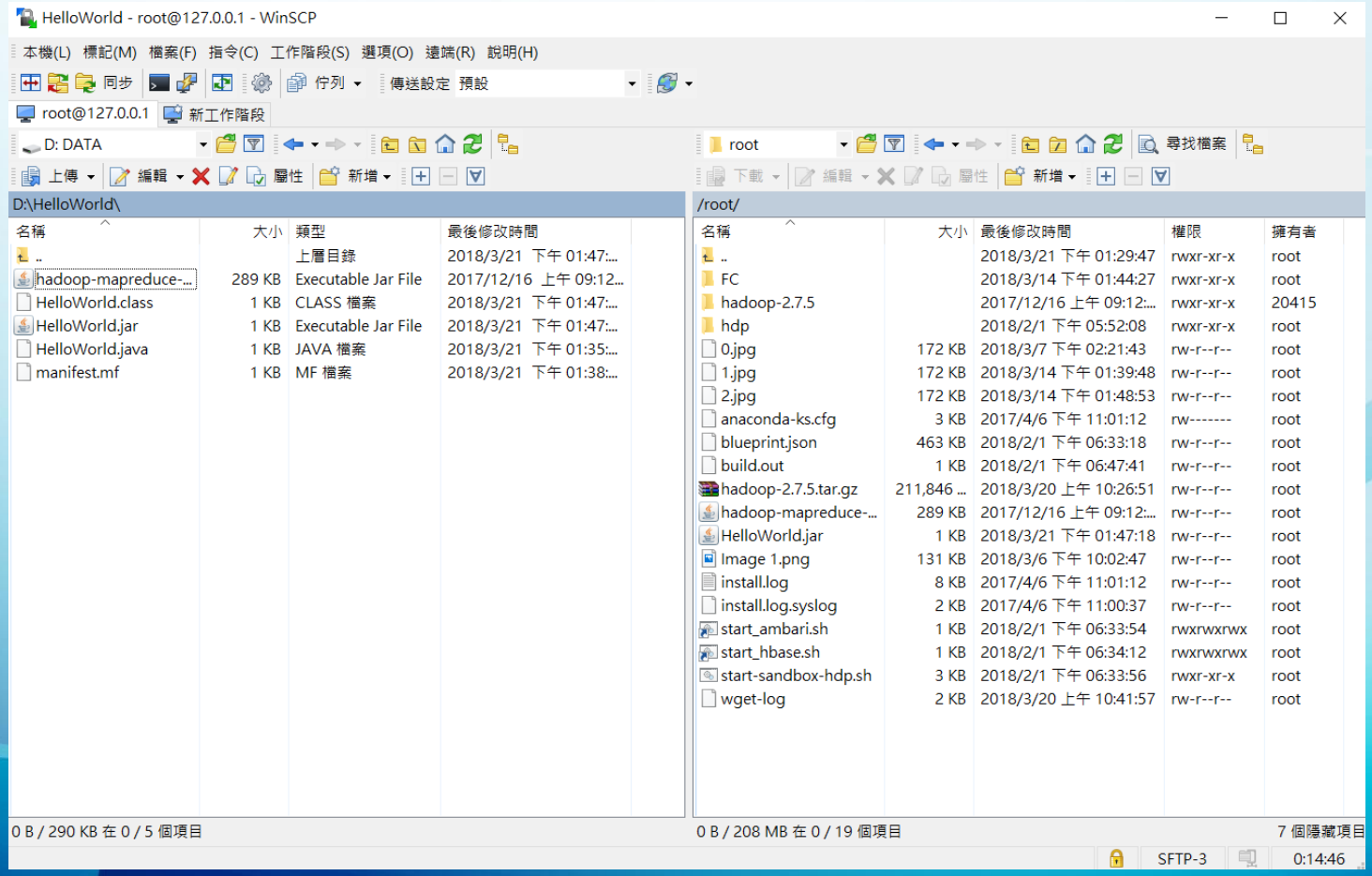
執行包裝 (包裝為 HelloWorld.jar)

- HelloWorld.java ➔ HelloWorld.class ➔ HelloWorld.jar
- javac HelloWorld.java
- java HelloWorld
- jar cvfm HelloWorld.jar manifest.mf HelloWorld.class
- java -jar HelloWorld.jar

```
D:\HelloWorld>jar cvfm HelloWorld.jar manifest.mf HelloWorld.class
已新增資訊清單
新增: HelloWorld.class (讀=427)(寫=289)(壓縮 32%)
```

winscp 丟檔案到主機端

- yarn jar HelloWorld.jar

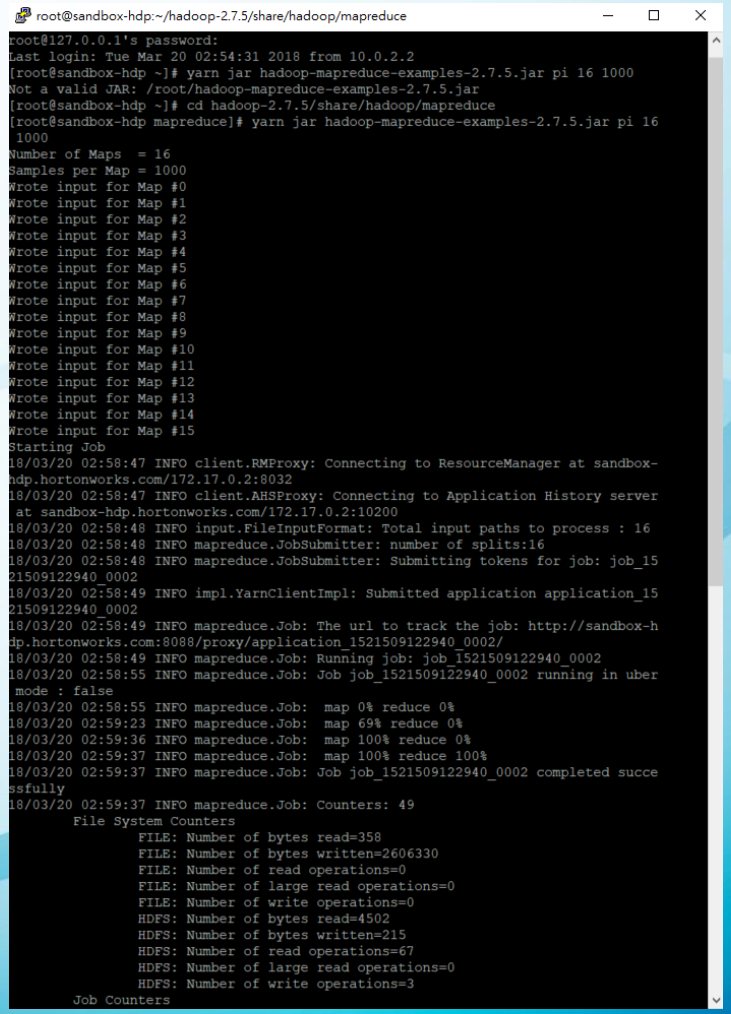


執行成功 by Putty

yarn jar hadoop-mapreduce-examples-2.7.5.jar pi 16 1000

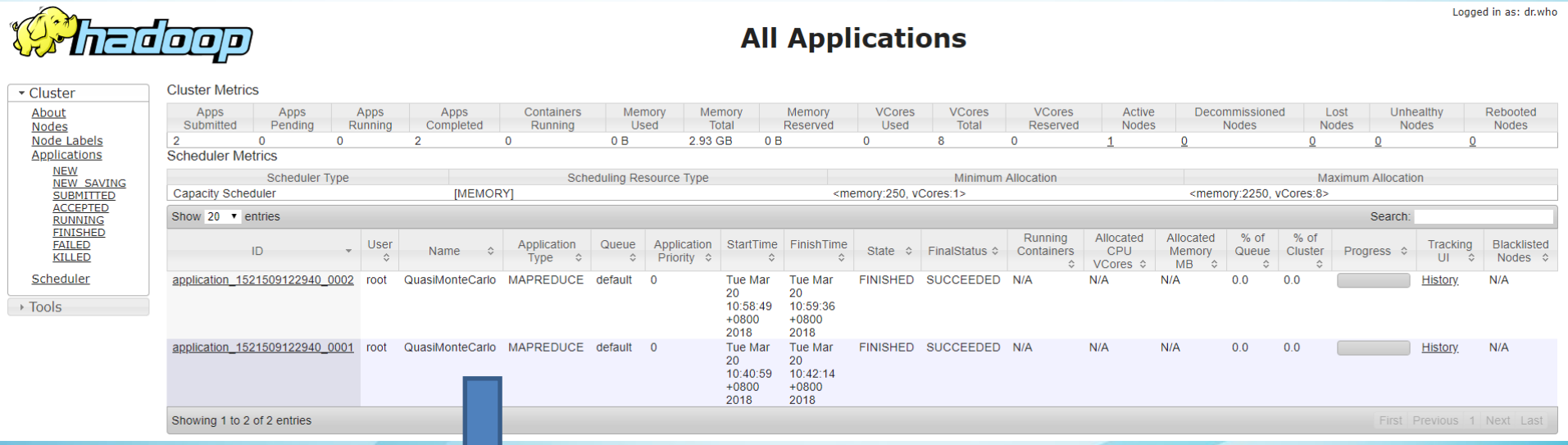
yarn jar HelloWorld.jar

```
[root@sandbox-hdp ~]# yarn jar HelloWorld.jar
Hello! World!
[root@sandbox-hdp ~]#
```



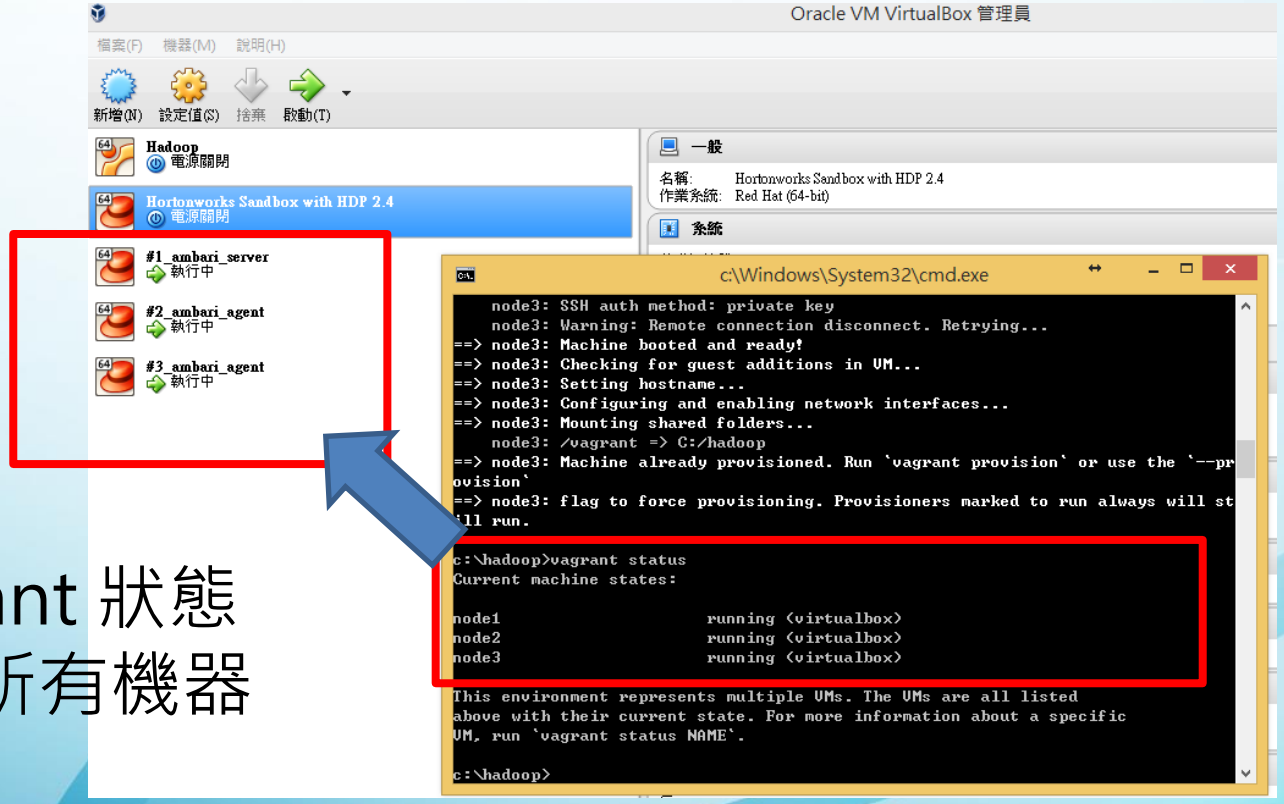
確定執行 mapreduce 成功

- http://127.0.0.1:8088



程式跑成功

多機板 by vagrant up



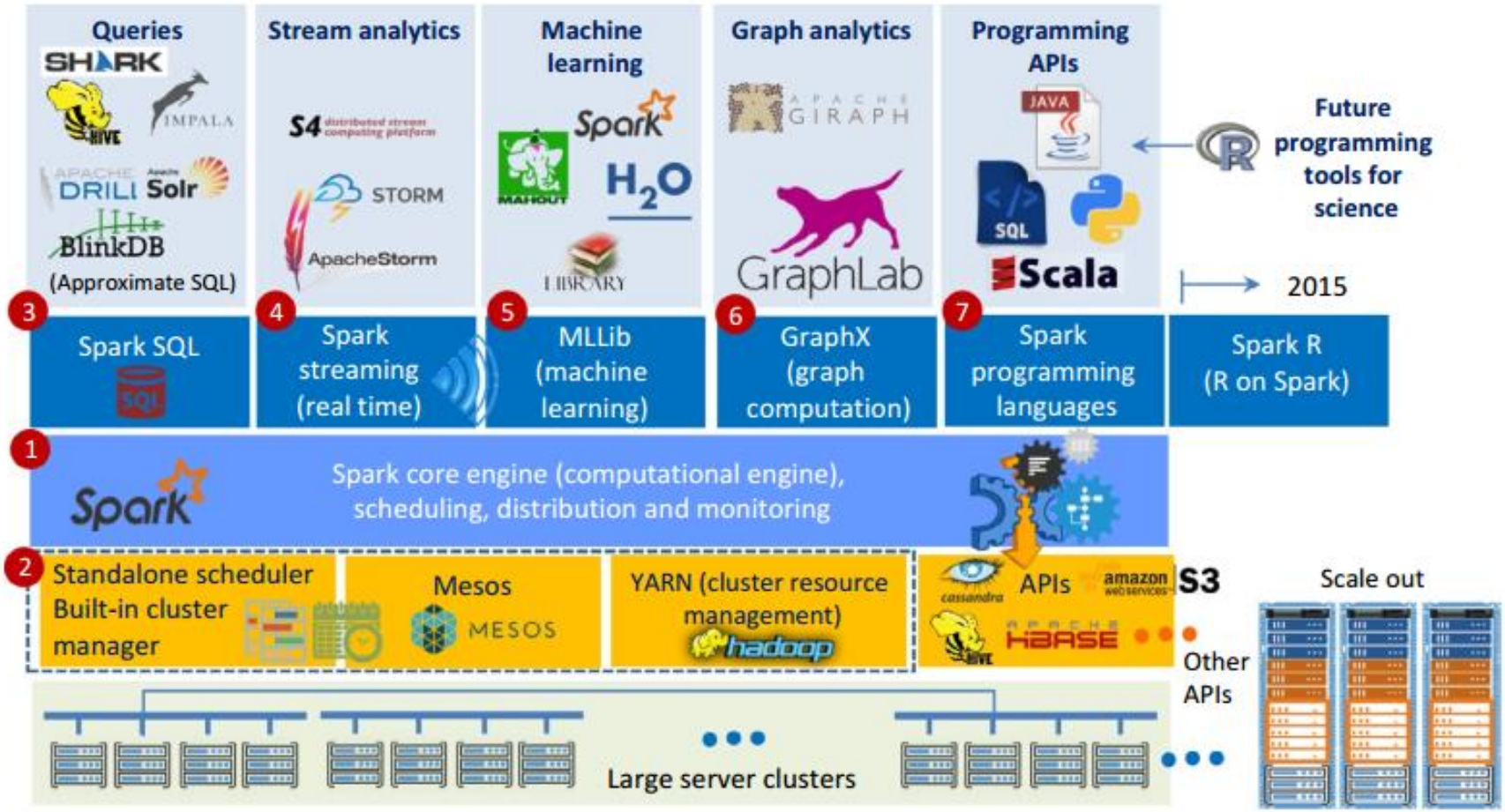
vagrant 狀態
查看所有機器

結論

滿滿的大平台

大哥，技術這種東西，
一山還有一山高！
只有三種層次：
(1) 「先」知道，重鑽研；
(2) 「後」知道，重應用；
(3) 不想知道，甘我屁事；
如此而已啊。
From 李智 老師

目前 Spark 生態系統多樣性



發現，我們落後很多

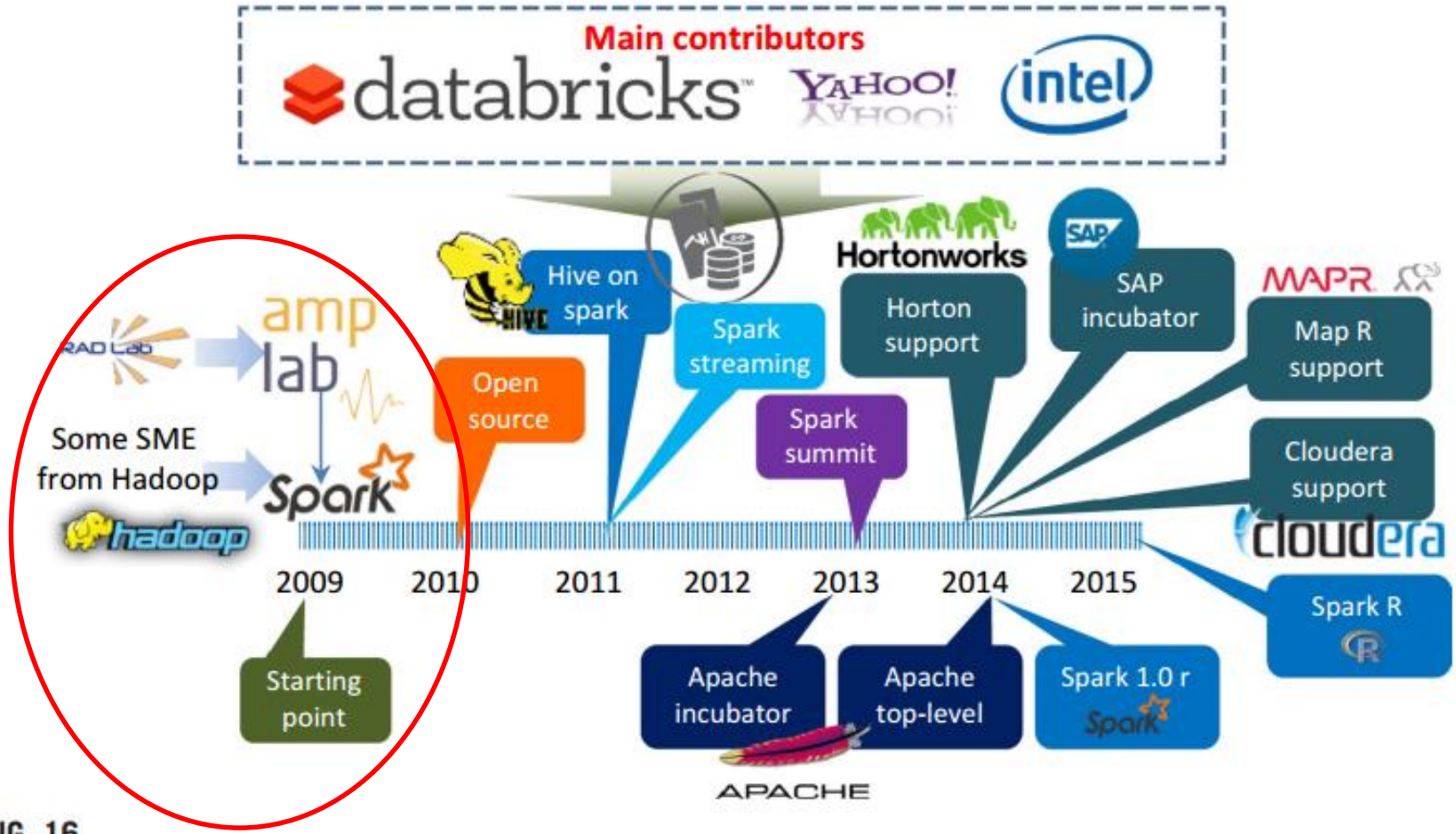


FIG. 16

Spark history.

其實我們應該要好好利用開源軟體的優勢

不能再一人挖坑

R

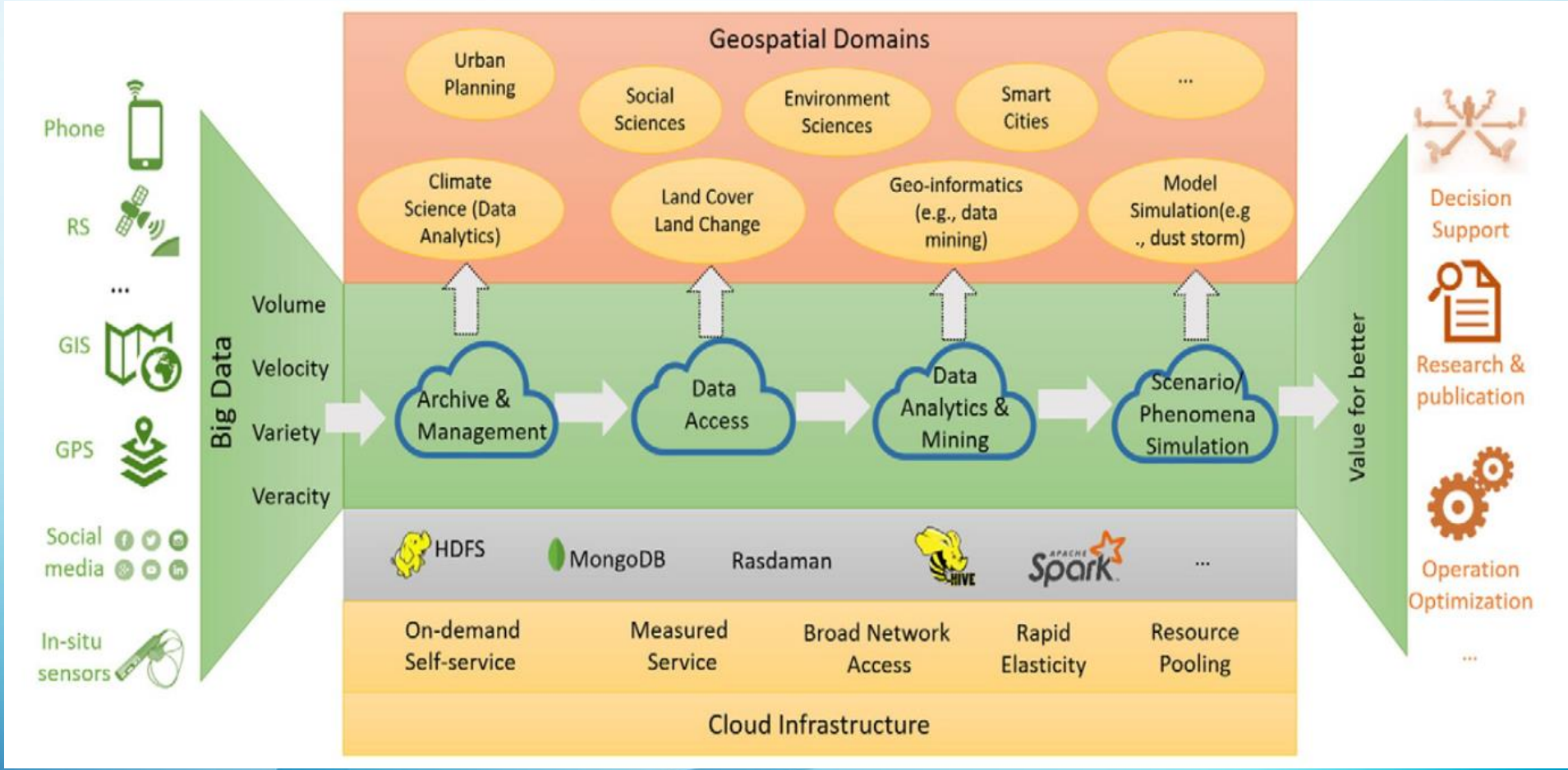
Java

Python

我們應該要把餅做大

文獻參考

學界 GeoSpatial 分析平台參考架構



Cloud Computing provides critical supports to the processing of Big Data to address the 4Vs to obtain Value for better decision support, research, and operations for various geospatial domains

C. Yang et al. /Utilizing Cloud Computing to address big geospatial data challenges
Computers, Environment and Urban Systems xxx (2016) xxx–xxx

結論

- 培養資料處理、平台基礎建設種子人員
- 不再強調大數據
- 不能一招半式走天下
- 多接觸不同程式語言 (R、Python、Java)
- 熱情、熱情、再熱情
- 技術分享

Q & A