

大數據資料處理 – Hadoop MapReduce Programming

課程總結 與 Hadoop計算環境改進



授課教授：王經篤 (Jing-Doo Wang)

Email: jdwang@asia.edu.tw

亞洲大學 資訊工程系 （CSIE, Asia University）

目的 (Goal)

- 快速進入Hadoop開發練習 (壓縮檔案下載解開，即可使用)
  - Windoop 程式開發環境(MS-Window)
    - Hadoop
    - Java
    - Eclipse
- 利用Hadoop MapReduce程式，練習巨量資料整理。
  - 利用線上開放資料-『交通部高速公路閘道資料』
  - 完整的資料前處理(Data Preprocessing)與資訊擷取(Information Extraction)

課程大綱 （Outline)

- Hadoop 介紹 (Hadoop Introduction)
- Windoop 開發環境介紹
  - Windoop 安裝 (MS-Windows + Hadoop)
  - Hadoop File System (HDFS) 系統操作 (HDFS Operations)
- Windoop 練習
  - WordCount 程式載入修改與執行
- 網路資料自動下載 (Web Robot)
  - 交通部高速公路閘道資料（http://tisvcloud.freeway.gov.tw/）
  - Google Map 匝道座標匯入
- 巨量資料處理案例 (Case Study)
  - 高速公路閘道資料之資訊擷取 (Information Extraction from Freeway Gantry Records)

作業與評分 (Assignment and Score)

1. (10%) Hadoop MapReduce 基本測驗測驗卷
2. (10%) Windoop 單機系統安裝 +Hadoop Java 程式載入 修改與執行
3. (10%) 網路資料自動下載
4. (30%) 專題：高速公路閘道資料之資訊擷取

Apache Hadoop Releases (Check Date:2017/11/13)



Cloudera (CDH)
https://www.cloudera.com/

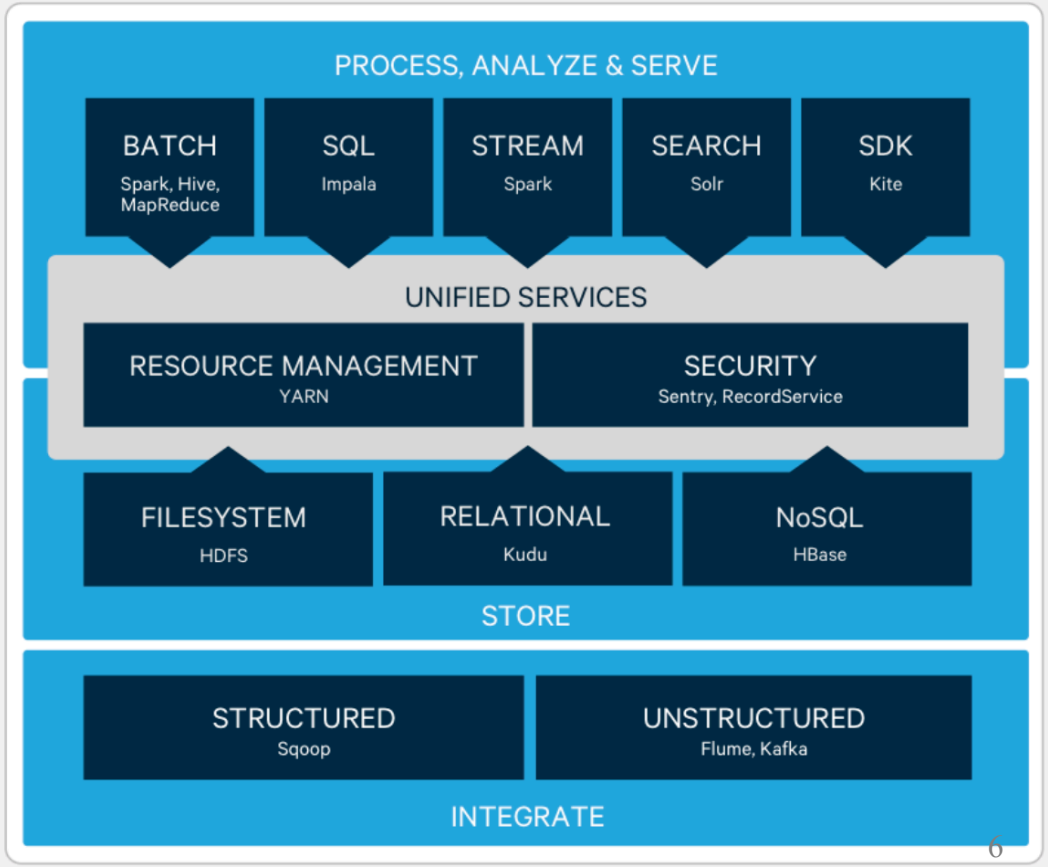
cloudera Products Solutions Downloads More

Apache Hadoop Ecosystem

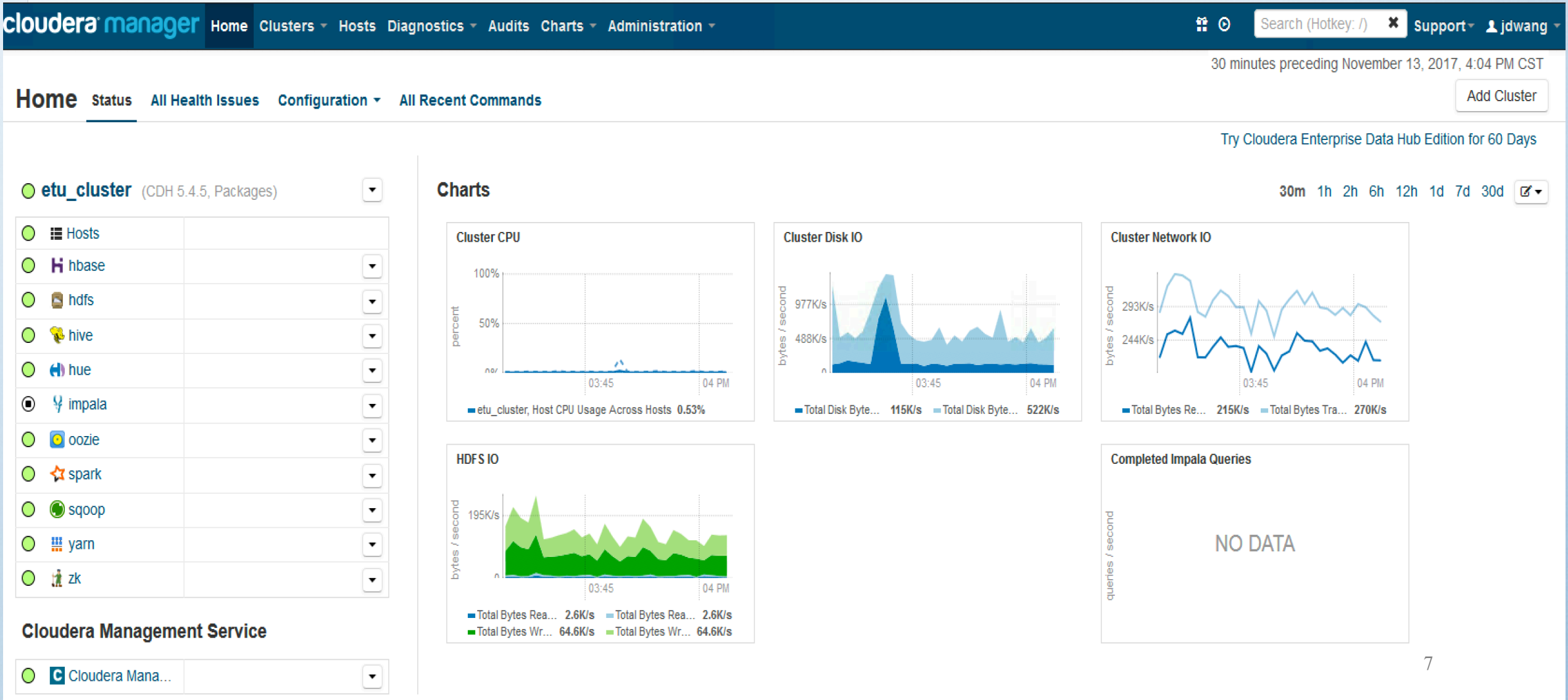
Hadoop is an ecosystem of open source components that fundamentally changes the way enterprises store, process, and analyze data. Unlike traditional systems, Hadoop enables multiple types of analytic workloads to run on the same data, at the same time, at massive scale on industry-standard hardware. CDH, Cloudera's open source platform, is the most popular distribution of Hadoop and related projects in the world (with support available via a Cloudera Enterprise subscription).

Try now >

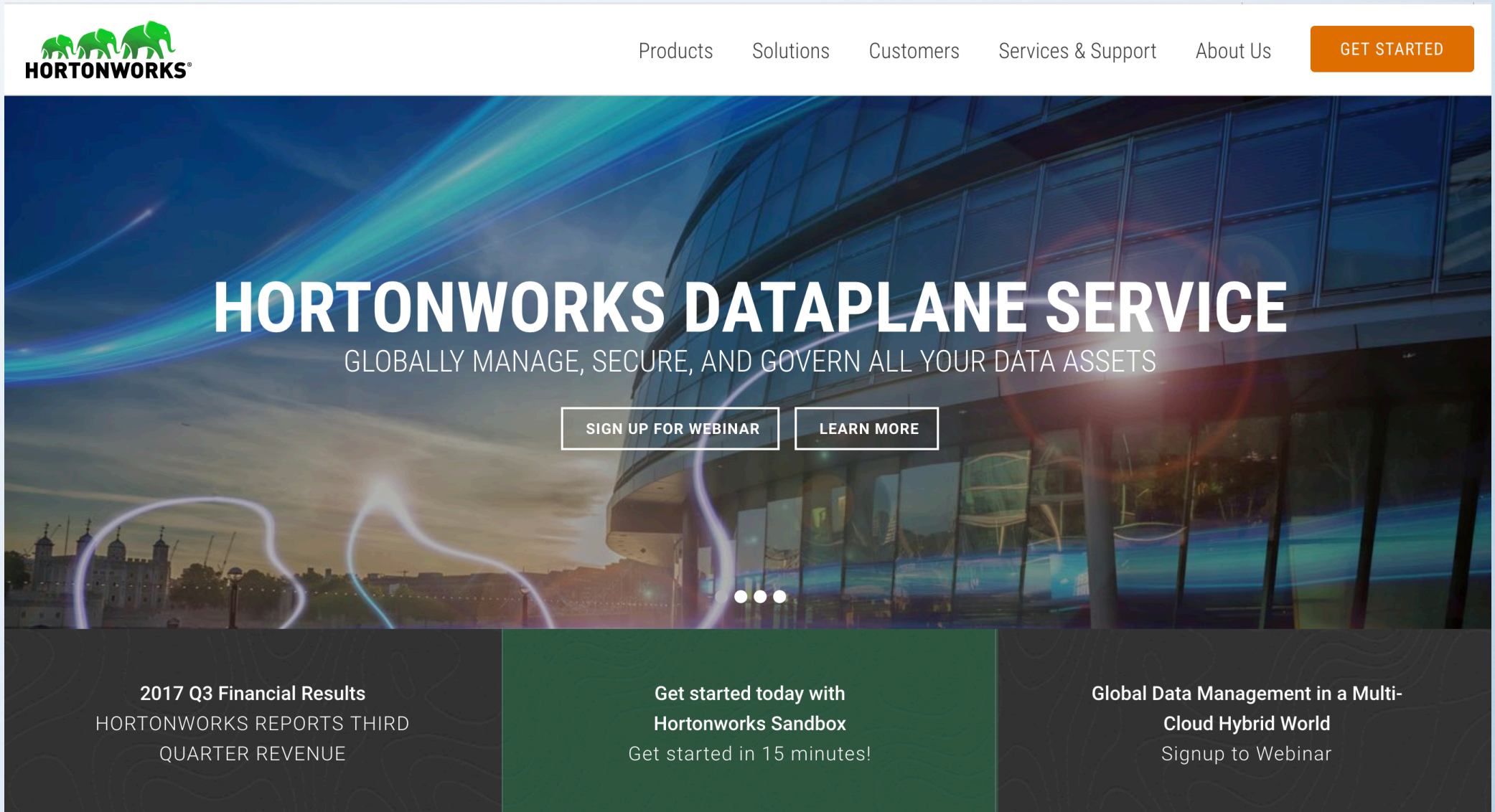
Hadoop ecosystem in the Engineering Blog >



Cluster Manager



HORTONWORKS (https://hortonworks.com/)



計算效能提升（單機=>多機）

- Windoop Cluster 架設
  - (感謝:賴敬勳,王俊平,楊松儒 環境測試)
  - Hadoop Cluster環境設定(Windoop 2. 0)
    - https://asiaedutw-my.sharepoint.com/personal/105121007_live_asia_edu_tw/_layouts/15/WopiFrame.aspx?docid=0f8c3d4e7a83d4d20aefb58706bcdaed7&authkey=AdjdyLa7-fUECVUm7Oj6ato&action=view

Hadoop (容量) Vs. Spark （速度）

Hadoop

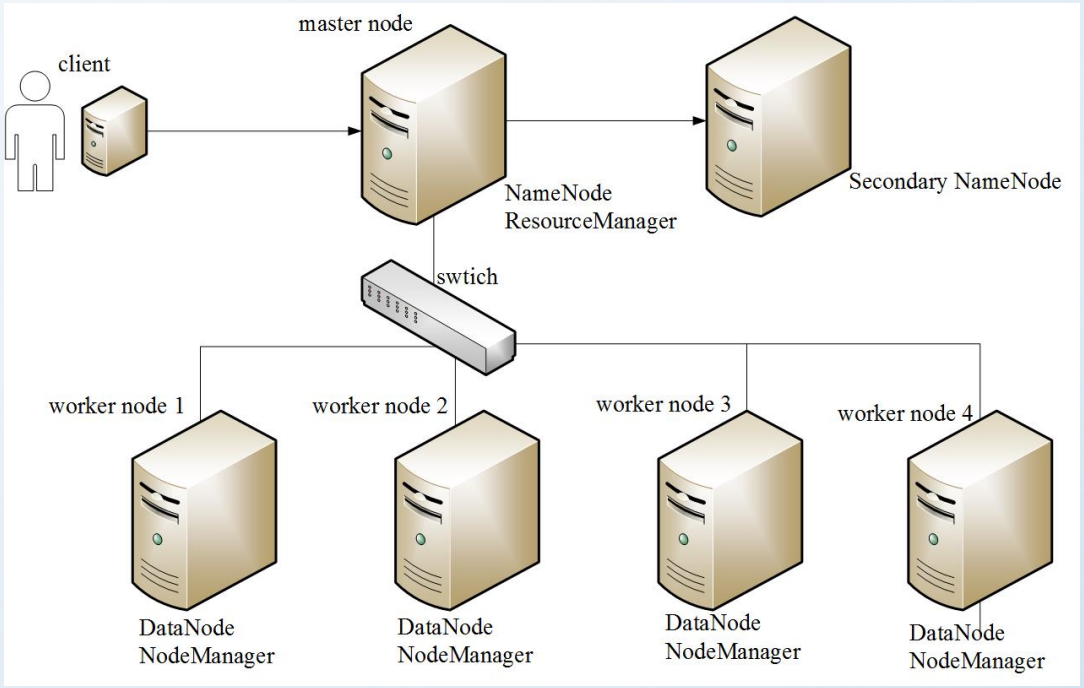


Spark

巨量資料處理與分析
教學環境-實習三階段

伺服器（低階）
CPU: 多核心
Memory: 64 GB
SDD: 512 GB
HD: 2T*2 (RAID)





Google Cloud Platform

amazon web services™

Windows Azure

單機
程式開發模擬
+
小量資料
（**課堂**等級）

多機
電腦叢集架設
+
中量資料
（**專題**等級）

雲端平台
商用雲端 服務訓練
+
巨量資料
（**企業**應用）

Google Cloud Platform- DATAPROC (https://cloud.google.com/dataproc/?hl=zh-tw)

安全 | https://cloud.google.com/dataproc/?hl=zh-tw

Google Cloud Platform 搜尋 控制台

為何選擇 Google 產品 | 產品 | 解決方案 | Launcher | 價格 | 客戶 | 說明文件 | 支援 | 合作夥伴 | 免費試用 | 與銷售人員聯絡

CLOUD DATAPROC

Spark 和 Hadoop 管理化服務不但又快又好用，成本也十分低廉

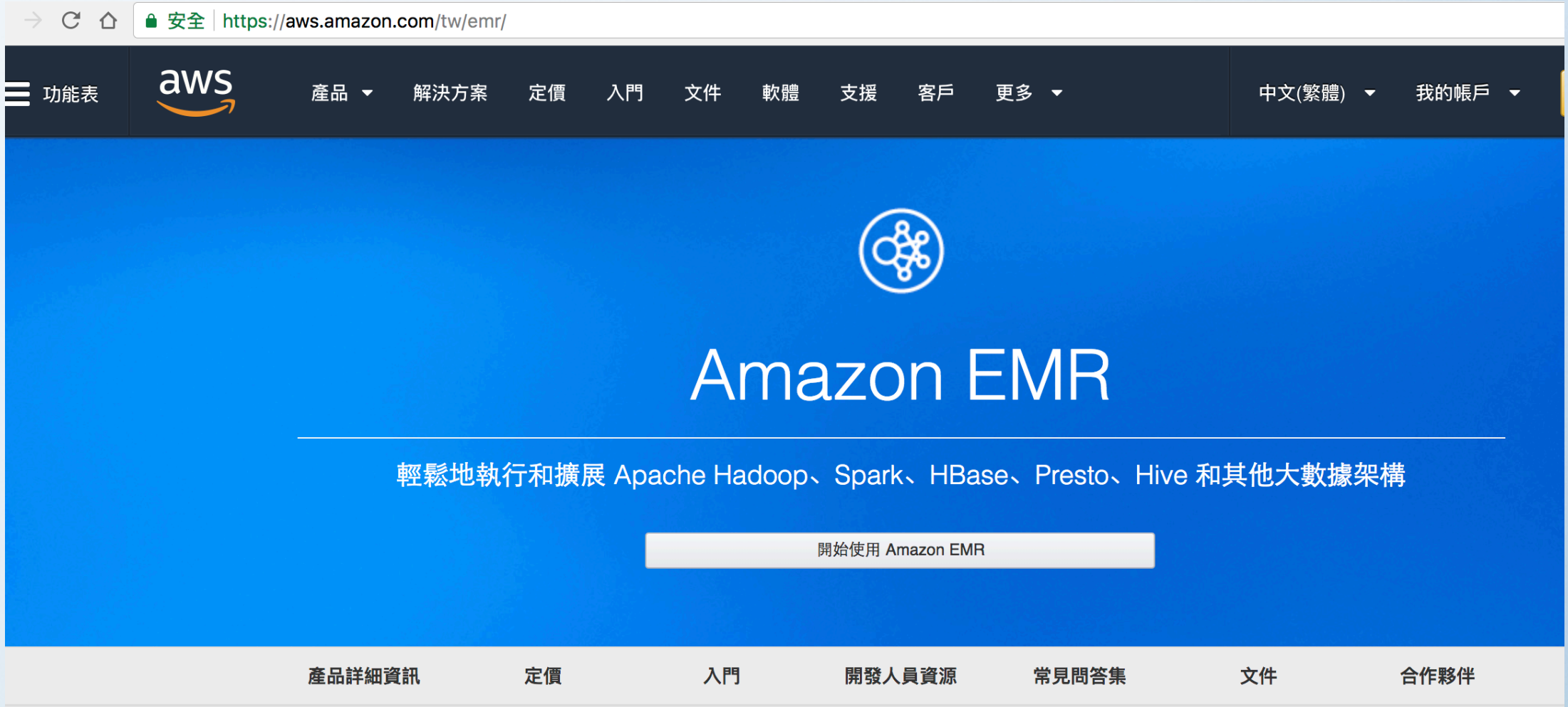
免費試用

Hadoop 和 Spark 管理化服務

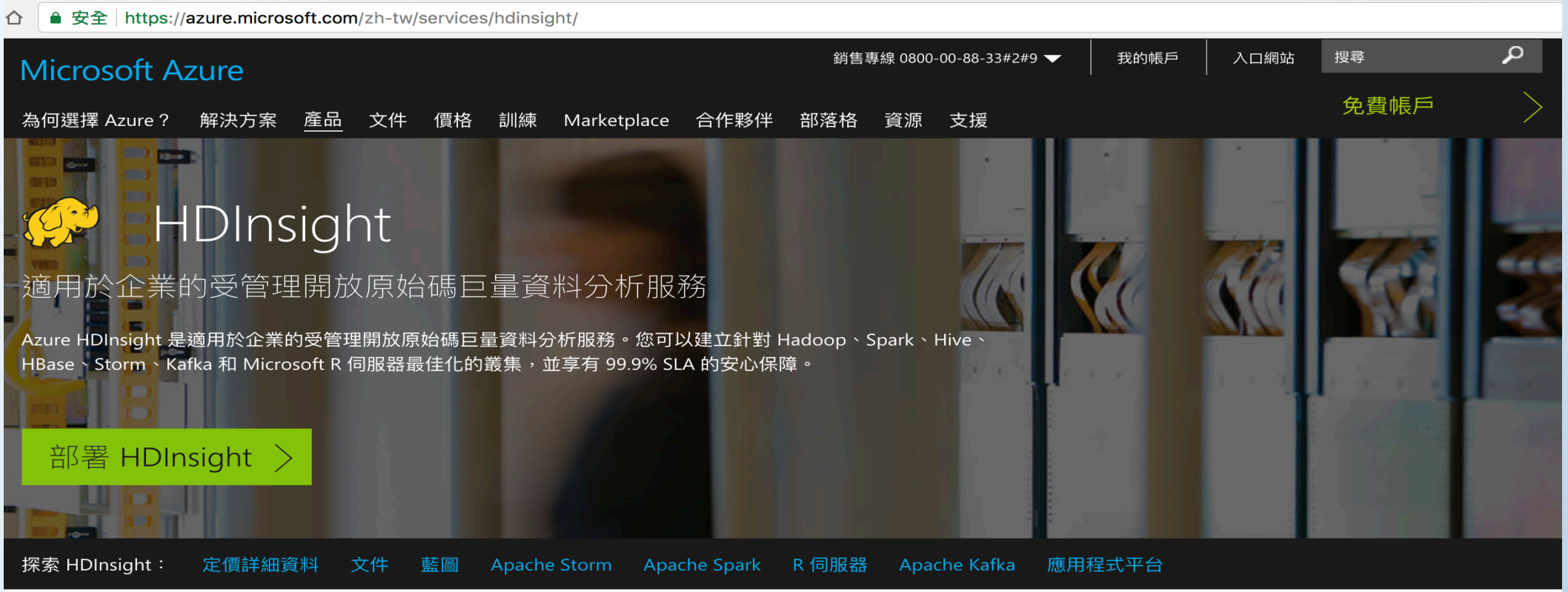
使用 Google Cloud Dataproc (**Apache Hadoop**、**Apache Spark**、**Apache Pig** 和 **Apache Hive** 服務) 處理大型資料集不僅簡便，花費的成本也十分低廉。您可以快速建立大小不限的管理化叢集，並且在用完後立即關閉，輕鬆控制成本。Cloud Dataproc 整合了多項 Google Cloud Platform 產品，是一款功能強大而完善的資料處理平台。



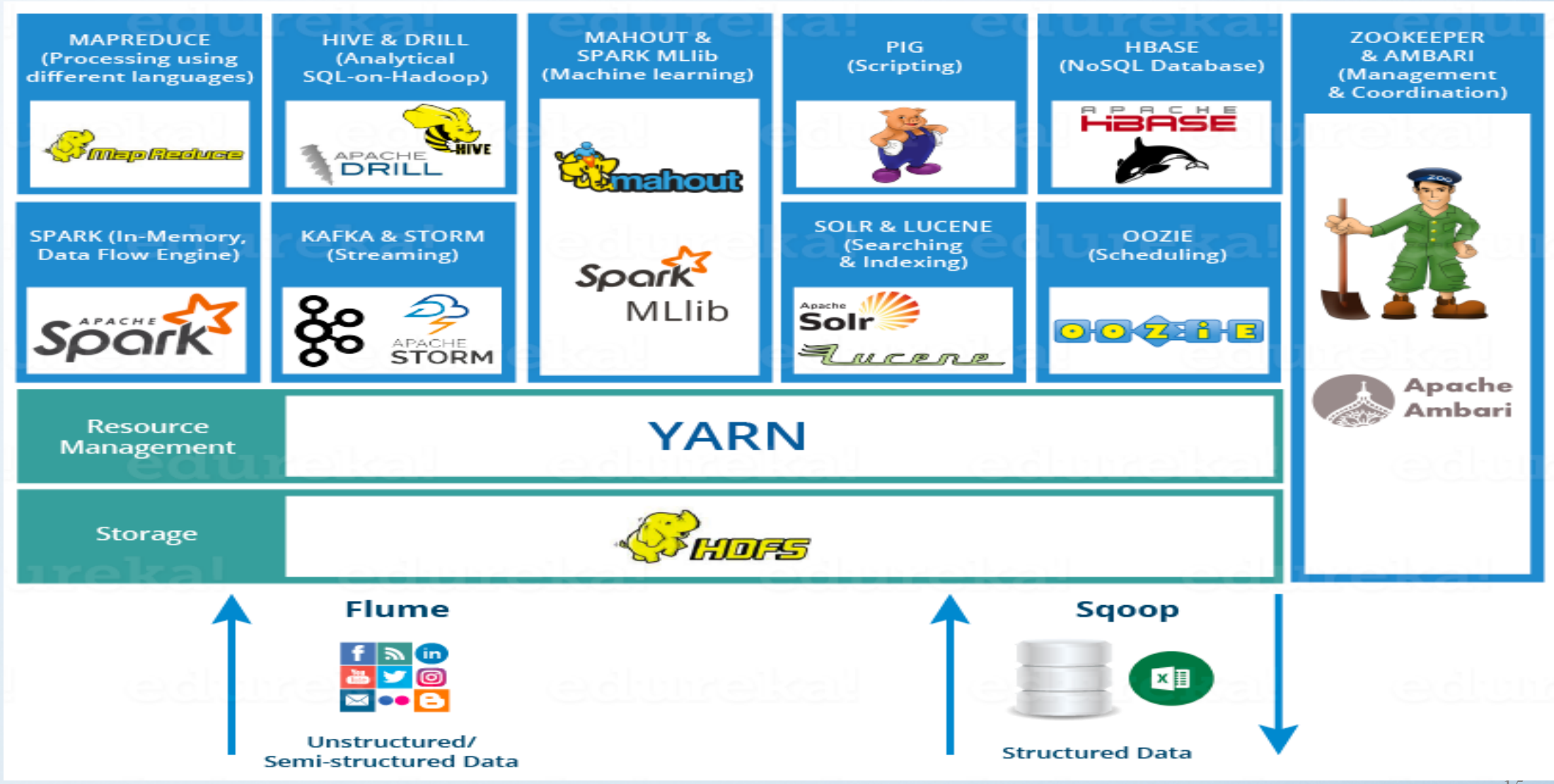
Amazon EMR (https://aws.amazon.com/tw/emr/)



Window Azure – HDInsight (https://azure.microsoft.com/zh-tw/services/hdinsight/)



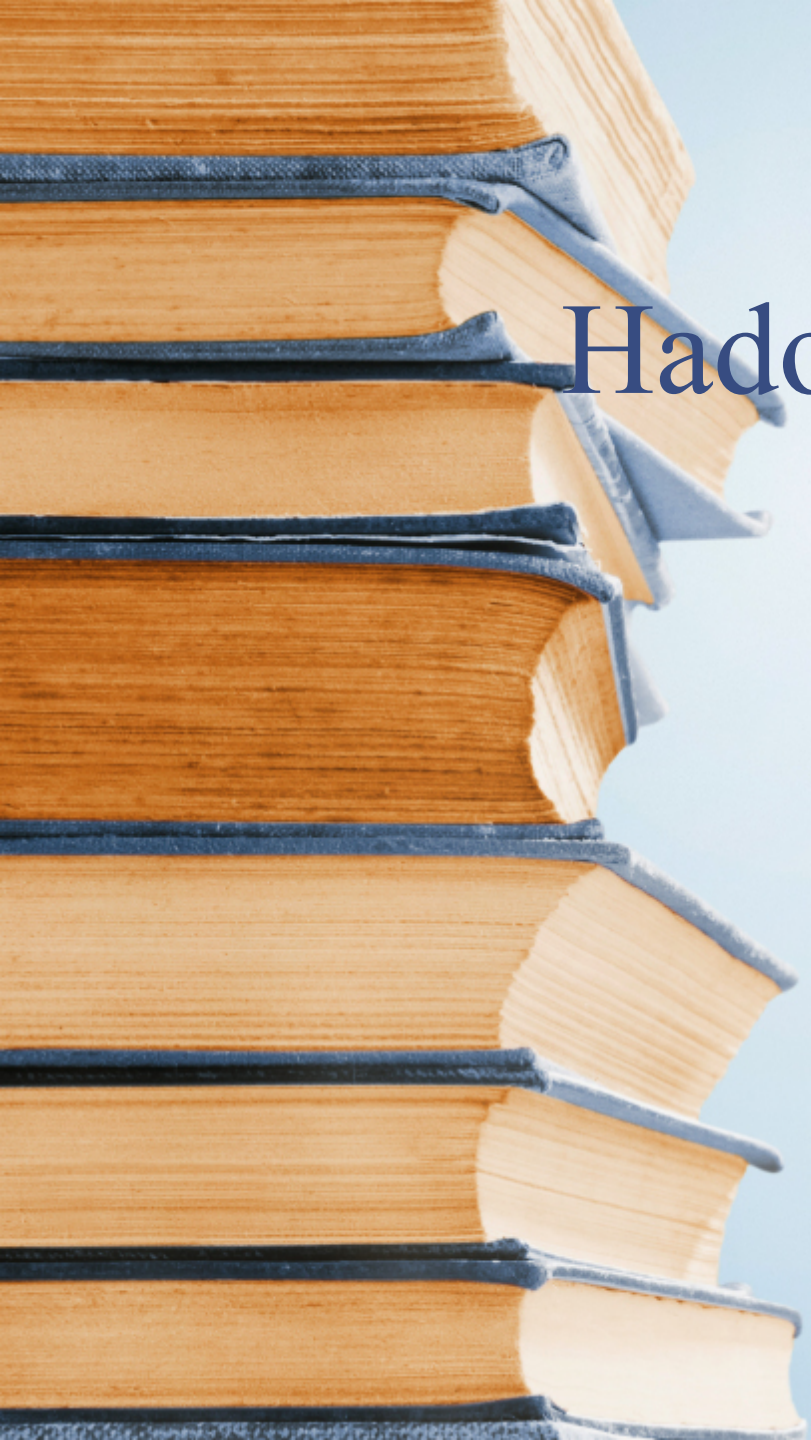
Hadoop Ecosystem



From: http://www.bogotobogo.com/Hadoop/BigData_hadoop_Ecosystem.php

感謝 （Acknowledgements）

- 亞洲大學-教學資源暨教師專業發展中心
  - **柯慧貞 副校長**
  - 磨課師計畫辦公室 教學設計助理
    - 楊美娟小姐 （MOOCs系統）
    - 朱凱聞先生 （錄影剪接）
- 林奇暻 先生 (Philip Lin)
  - 提供 Windoop 開發環境 （Hadoop + Eclipse + Present-Election）
- 王耀聰 先生 (Jazz Wang)
  - Hadoop 系統與MapReduce 程式問題
- 學生：王家恩,呂宏仁,王寬之,陳柏融



大數據資料處理

Hadoop MapReduce Programming (使用Windoop)

授課教授：王經篤 (Jing-Doo Wang)
Email: jdwang@asia.edu.tw
亞洲大學 資訊工程系 (CSIE, Asia University)



Data Visualization

授課教授：何承遠 (Cheng-Yuan Ho)
Email: tommyho@asia.edu.tw
亞洲大學 資訊工程系 兼 大數據研究中心
(CSIE and BDRC, Asia University)

