



# 『賦稅資訊系統整合再造更新整體實施計畫』 地方稅建置委外服務案

執行雲端運算資料搜尋、分析與備份

教育訓練教材

# 課程大綱

## 雲端運算的三大關鍵技術

Part 1 : Overview of Cloud Computing Core Technologies

## 深入解析大量資料分析技術

Part 2 : Deep Dive into Data Science Technologies

## 巨量資料分析處理平台：Hadoop

Part 3 : Introduction to Hadoop

## 打造內網搜尋引擎：Nutch & Crawlzilla

Part 4 : Introduction to Nutch and Crawlzilla

## 資料異機同步儲存的機制：DropBox

Part 5 : Introduction to DropBox



# 雲端運算的三大關鍵技術

Part 1 : Overview of Cloud Computing Core Technologies

**Jazz Wang**  
**Yao-Tsung Wang**  
**jazz@nchc.org.tw**



Powered by DRBL

# National Definition of Cloud Computing

## 美國國家標準局 NIST 給雲端運算所下的定義

### 5 Characteristics

五大基礎特徵

### 4 Deployment Models

四個佈署模型

### 3 Service Models

三個服務模式

#### 1. On-demand self-service.

隨需自助服務

#### 2. Broad network access

隨時隨地用任何網路裝置存取

#### 3. Resource pooling

多人共享資源池

#### 4. Rapid elasticity

快速重新佈署靈活度

#### 5. Measured Service

可被監控與量測的服務

# 4 Deployment Models of Cloud Computing

## 雲端運算的四種佈署模型

Public Cloud

公用雲端



Microsoft

Google

**Dynamic Resource Provisioning  
between public and private cloud**

私有雲端動態根據計算需求  
調用公用雲端的資源

Target Market

is **S.M.B.**

主要客戶為  
中小企業

Hybrid  
Cloud

以大型企業  
為主要客戶

**Enterprise is  
key market**

Community Cloud

社群雲端



私有雲端

Private Cloud

Academia 學術為主



# 3 Service Models of Cloud Computing

## 雲端運算的三種服務模式 (市場區隔)

### IaaS

Infrastructure as a Service

架構即服務

### PaaS

Platform as a Service

平台即服務

### SaaS

Software as a Service

軟體即服務



# 2 perspectives : Services vs Technologies

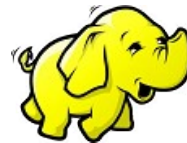
您想聽的是「雲端服務」還是「雲端技術」？

Google YouTube e W



雲端服務

Microsoft



雲端技術



Cloud computing hype spurs confusion, Gartner says

<http://www.computerworld.com/s/article/print/9115904>

淺談雲端運算 (Cloud Computing)

[http://www.cc.ntu.edu.tw/chinese/epaper/0008/20090320\\_8008.htm](http://www.cc.ntu.edu.tw/chinese/epaper/0008/20090320_8008.htm)

# 1 key spirit of Cloud Computing

用一句話說明雲端運算！服務才是王道！

Anytime 隨時

Anywhere 隨地

With Any Devices 使用任何裝置

Accessing Services 存取各種服務

Cloud Computing =~ Network Computing

雲端運算 =~ 網路運算

Key spirit of Cloud ~

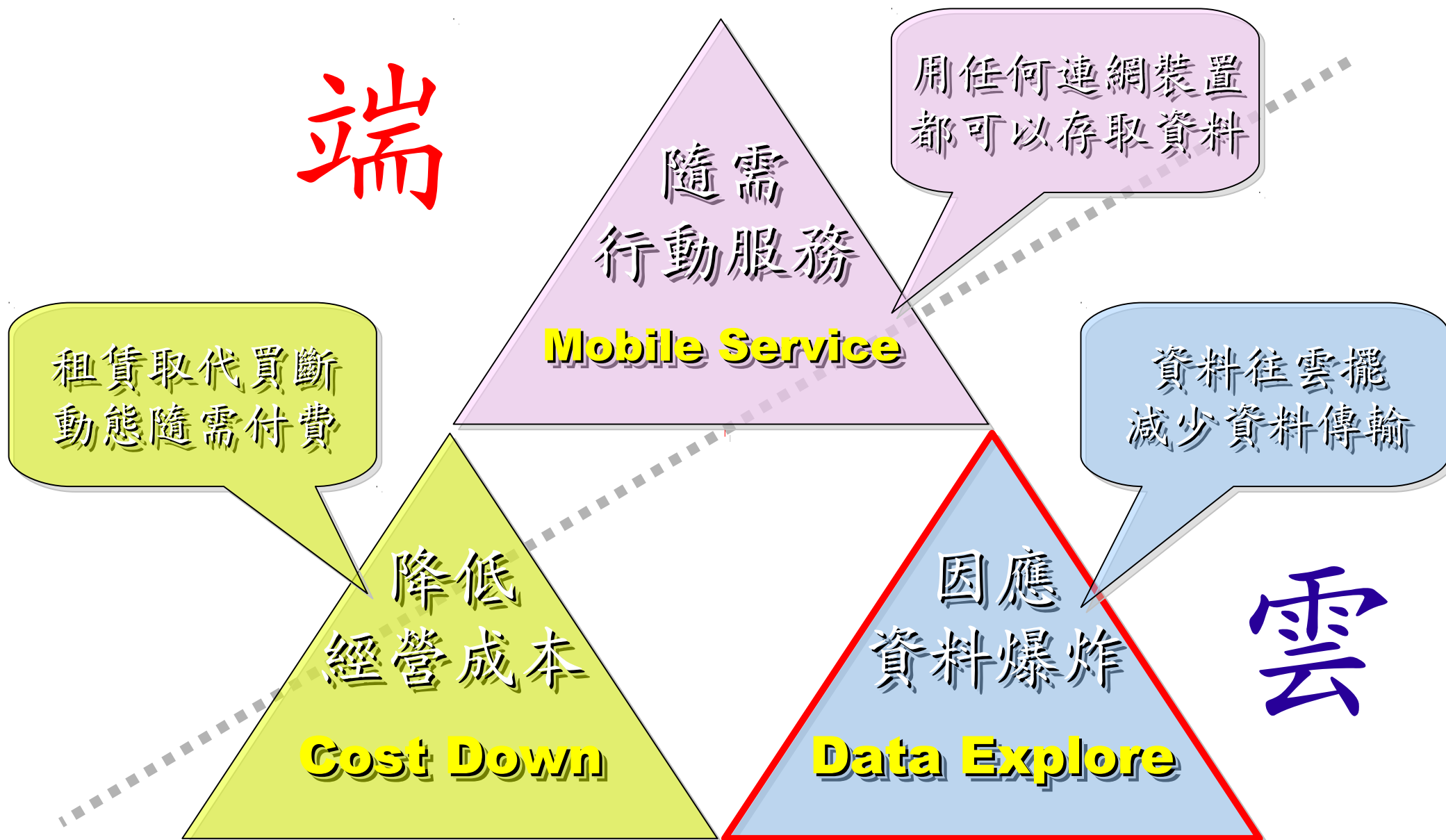
形成服務才是重點！！

Everything as a Service !!

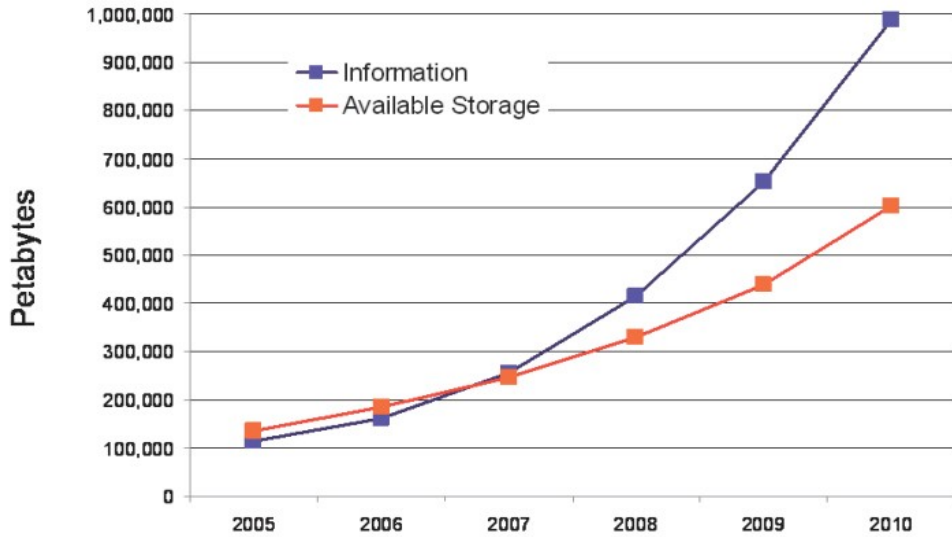


# Key Driving Forces of Cloud Computing

## 雲端運算的關鍵驅動力



# Information Versus Available Storage



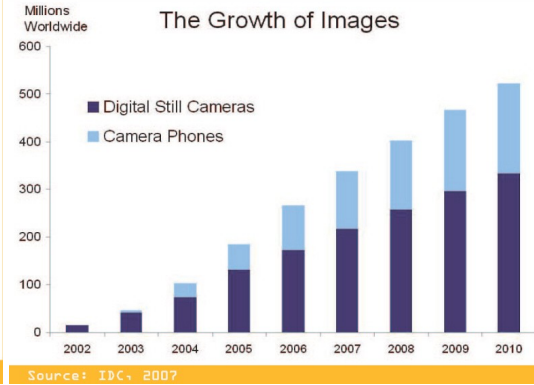
Source: IDC, 2007

# 2007 Data Explore

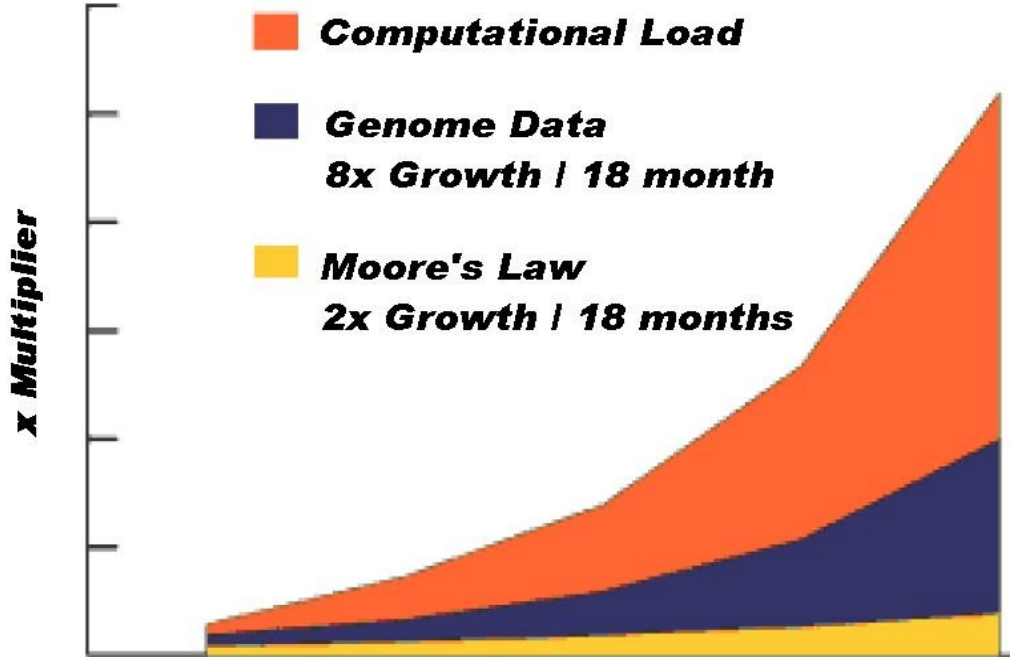
**Top 1 : Human Genomics - 7000 PB / Year**  
**Top 2 : Digital Photos - 1000 PB+ / Year**  
**Top 3 : E-mail (no Spam) - 300 PB+ / Year**



Source: IDC, 2007



Source: IDC, 2007



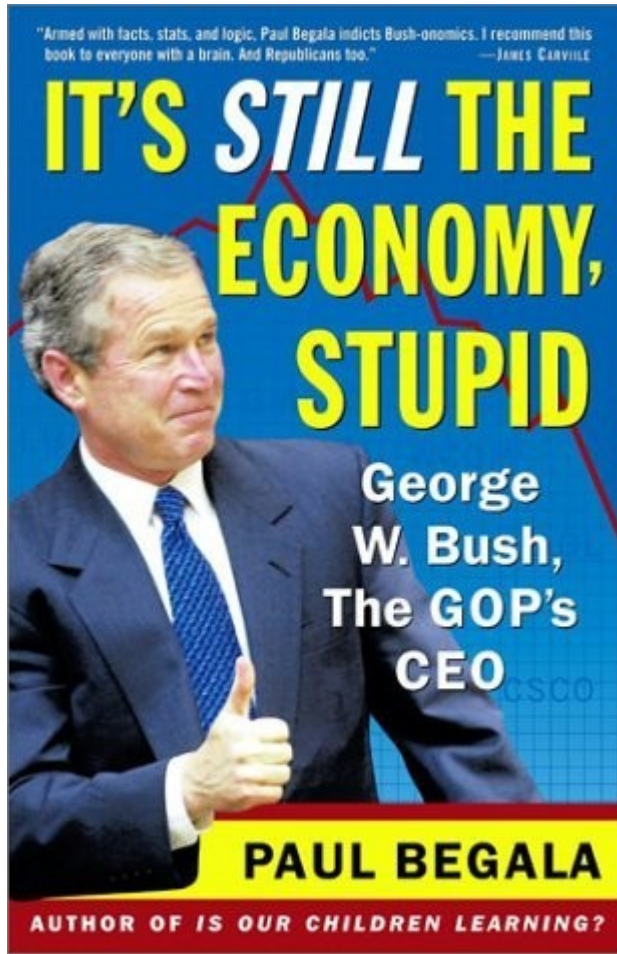
Particle Physics Large Hadron Collider (15PB)	Human Genomics (7000PB) 1GB / person 200PB+ captured 200% CAGR	World Wide Web (~1PB)	Wikipedia (10GB) 100% CAGR
Annual Email Traffic, no spam (300PB+)	Internet Archive (1PB+)	Estimated On-line RAM in Google (8PB)	Personal Digital Photos (1000PB+) 100% CAGR
200 of London's Traffic Cams (8TB/day)	2004 Walmart Transaction DB (500TB)	Typical Oil Company (350TB+)	Merck Bio Research DB (1.5TB/qtr)
UPMC Hospitals Imaging Data (500TB/yr)	MIT Babytalk Speech Experiment (1.4PB)	Terashake Earthquake Model of LA Basin (1PB)	One Day of Instant Messaging in 2002 (750GB)
Total digital data to be created this year <b>270,000PB</b> (IDC)			

Phillip B. Gibbons, Data-Intensive Computing Symposium

Source: <http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf>

Source: [http://lib.stanford.edu/files/see\\_pasig\\_dic.pdf](http://lib.stanford.edu/files/see_pasig_dic.pdf)

# IT'S THE DATA, STUPID!



「笨蛋！重點在經濟」

( **"It's the economy, stupid"** )

卡維爾 ( **James Carville** ) 自創這句標語，  
促使柯林頓當上美國第 **42** 屆總統。

- **1992** 年

「笨蛋！重點還是在經濟」

( **"It's STILL the economy, stupid"** )

卻讓小布希嘲笑是幼稚的總統。

- **2002** 年

雲端時代，谷歌會說：「笨蛋！重點在資料」

( **"It's the data, stupid"** )

誰掌握了你的資料，就有機會掌握你的荷包  
想想看，電腦、手機掉了，您心疼的是甚麼呢？

- **2007** 年



## 善用雲端架構 打造企業人才庫

對於雲端的運用，多半仍停留在創造新商機的層次，然而善用雲端運算，可以替組織創造更多業務、行銷和人才培訓的機會。

作者：麥肯錫 出處：天下雜誌

過去五年，麥肯錫觀察重要科技發展，其中雲端、大量資訊 (big data)、智慧裝置 (smart assets) 三項，以超乎想像的速度發展。這三大技術，帶來五大趨勢，可被應用在企業營運及組織運作。先分別來看這三項技術：

第一、雲端運算。「雲端」在台灣已被一般民眾熟知。但我認為大家多半仍停留在雲端運算如何能創造新商機，卻很少好好思索，該怎麼運用雲端運算來替組織創造更多機會。特別是服務提供者，譬如電信業者、有線電視業者等，都應更有效應用雲端運算，為業務帶來更多機會。

第二、大量資訊。目前，絕大多數台灣企業，分析大量龐雜資料，仍使用類似微軟工具如 excel 等來整理。事實上，大量資訊經過快速運算分析，能更省時、省費用、有效的進行行銷活動。

第三、智慧裝置。如何善用監控器、智慧電表這類智慧裝置，來更優化公司營運。

參考來源：善用雲端架構 打造企業人才庫，作者：麥肯錫，出處：天下雜誌 455 期 (2010/09)  
<http://www.cw.com.tw/article/print.jsp?id=41776>

### 雲端運算

### 大量資訊分析

### 智慧裝置

# New Data Science : Social Network + Realtime Search

## 當「社交網路」遇上「即時搜尋」 = 即時市場行銷分析

### 創意行銷 / 臉書行銷 每天400萬顧客在線上

【經濟日報/潘俊琳】

2010.10.11 02:20 am

社交網站臉書Facebook的興起，重新定義了網路行銷的概念，大量的人潮讓業者彷彿看到滾滾錢潮，但臉書「開放平台」的模式，讓習慣有規則可循的行銷業者，必須開始學習全新的社群行銷，試著擁抱這項利器並串連消費者。

根據美國comScore的統計，美國網友8月分共花了1,140萬分鐘在臉書上，首次超越停留在Google旗下網站的時間，而臉書全球已經有超過5億的使用者，其中有35%的人每天登入。

#### 快速分享 即時知道顧客反應

聖洋科技執行長邱繼弘表示，台灣臉書每個月約有700萬的累計使用人次，以60%每天上臉書的人口來算，就有420 萬人天天上線。

邱繼弘指出，臉書最大的行銷價值在於「開放平台」，只要符合它的基本規範，任何人、任何公司都可以在上面「免費」發揮自己的行銷創意。過去想要利用網路行銷，企業必須自己架站，林林總總的後台建設非常繁瑣，有多少人會來也是個問號？

但臉書幫企業解決了後台建設以及人潮，不論是企業或個人，只要成立自己的「粉絲專頁」，然後發揮行銷創意，回收可能比自己架站還更豐碩。因為臉書玩家只要在粉絲專頁按「讚」，就成為「粉絲團」的一員，往後企業發布在粉絲專頁的訊息，所有粉絲團成員都會收到，如果粉絲團的成員覺得某個行銷訊息不錯，只要按「分享」這個訊息就會出現在粉絲個人的臉書上，他所有的朋友就會看到這則行銷訊息，這是目前最高明的病毒式行銷。

社交網路

即時搜尋

評價排行榜



參考來源：創意行銷 / 臉書行銷 每天 400 萬顧客在線上

【經濟日報 / 潘俊琳】

<http://udn.com/NEWS/FINANCE/FIN11/5901891.shtml>



## 2011 年 10 大策略科技

科技	影響
雲端運算	大型企業將會在 2012 年成立動態採購小組，專門負責雲端運算相關的決定以及管理。
媒體平板以及行動應用	2010 年將會有 12 億人使用具備上網能力的手機。隨著行動上網裝置以及應用程式日趨普及，與地點(location)、動作(motion)相關的應用軟體，可望進一步推動裝置的銷售。
社交溝通以及協作 (collaboration)	多數的公司在 2016 年已經把社交科技整合至多數的企業應用中，整合的範圍包含內部社交 CRM、溝通及協作以及外部社交網站。
影片	2013 年每位工作者看到的內容中，將有 25% 都是照片、影音。
次世代分析	隨著電腦、行動裝置運算能力、連結能力更強，影響企業如何決策，SAS 是長期領導廠商，IBM 以及甲骨文(Oracle) 事後起之秀。
社交分析	衡量人、主題以及想法的關係，範圍不限於社交網路，IBM 預計在 2011 年成為該領域的主要廠商之一。
情境感知運算 (context-aware computing)	較人工智慧更為寬廣，預計在 2013 年時 Fortune 500 大企業中超過半數會有相關採用方案。
儲存等級記憶體 (storage class memory)	快閃記憶體在消費性裝置、娛樂設備中的使用更多。
無所不在的運算 (ubiquitous computing)	儘管 Gartner 已經提及這個概念許多年，但隨著手機、射頻晶片更為普及，越多的物件可以連上網路。
架構化(fabric-based) 的基礎建設以及電腦	運算能力模組化，系統可以透過不同的模組來建構，可望提升效能。

資料來源：DIGITIMES 整理，2010/10

製表：雷佳宜、李盈瑩

雲端運算

平板行動應用

社交溝通協作

多媒體內容

次世代分析

社交分析

情境感知運算

儲存等級記憶體

無所不在的運算

模組化基礎建設

Source : <http://www.gartner.com/it/page.jsp?id=1454221>

Source : [http://www.digitimes.com.tw/tw/dt/n/shwnws.asp?CnId=4&cat=400&cat1=20&id=0000205798\\_CUZ63ZS3LCRY7E7UBK6V8](http://www.digitimes.com.tw/tw/dt/n/shwnws.asp?CnId=4&cat=400&cat1=20&id=0000205798_CUZ63ZS3LCRY7E7UBK6V8)

# 端

平板行動應用

社交溝通協作

多媒體內容

次世代分析

社交分析

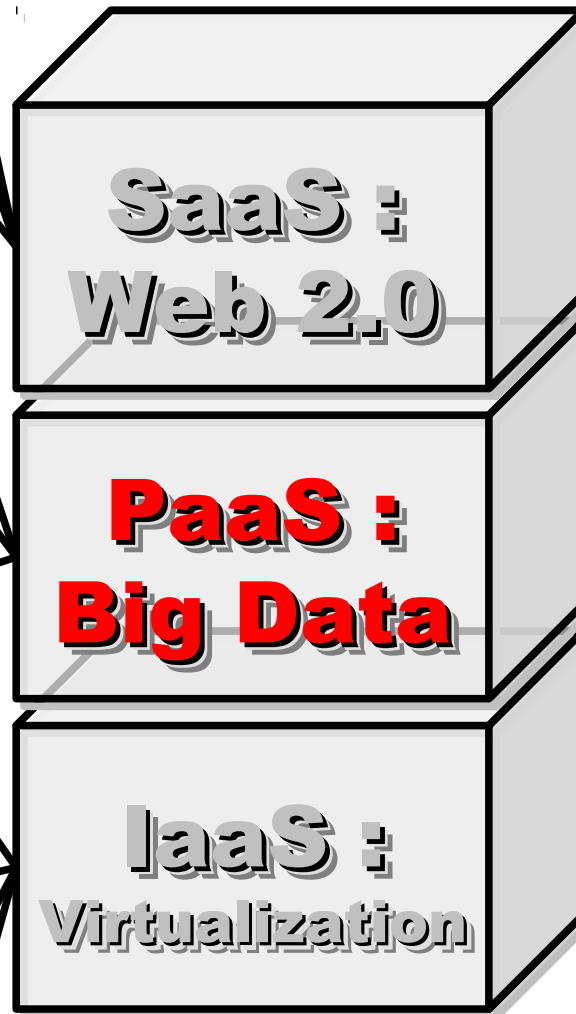
情境感知運算

儲存等級記憶體

無所不在的運算

模組化基礎建設

雲端運算



社交網路

評價排行榜

即時搜尋

智慧裝置

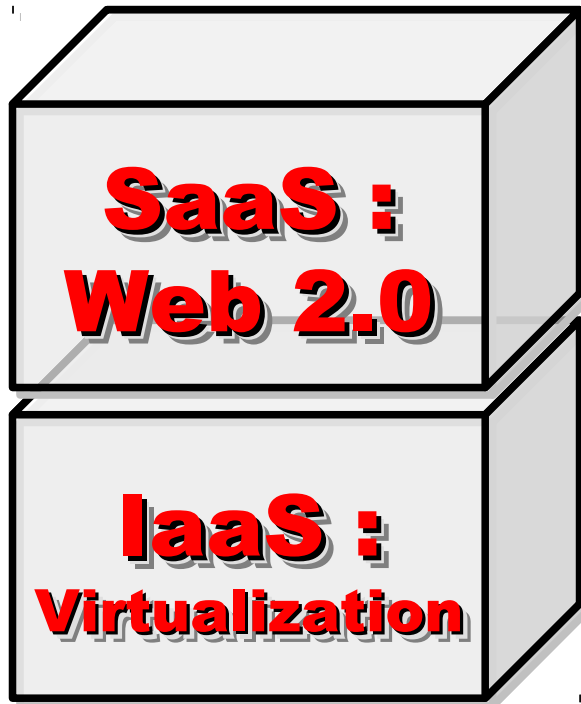
大量資訊分析

雲端運算

# 雲

# Two Type of Cloud Architecture ?

雲端架構的兩大陣營？



想盡辦法誘你用計算跟網路  
**Computing Intensive**



想盡辦法誘你提供資料作分析  
**Data Intensive**

# Reference Cloud Architecture

## 雲端運算的參考架構

### 應用軟體 Application

Social Computing, Enterprise, ISV, ...

### 程式語言 Programming

Web 2.0 介面, Mashups, Workflows, ...

### 控制管理 Control

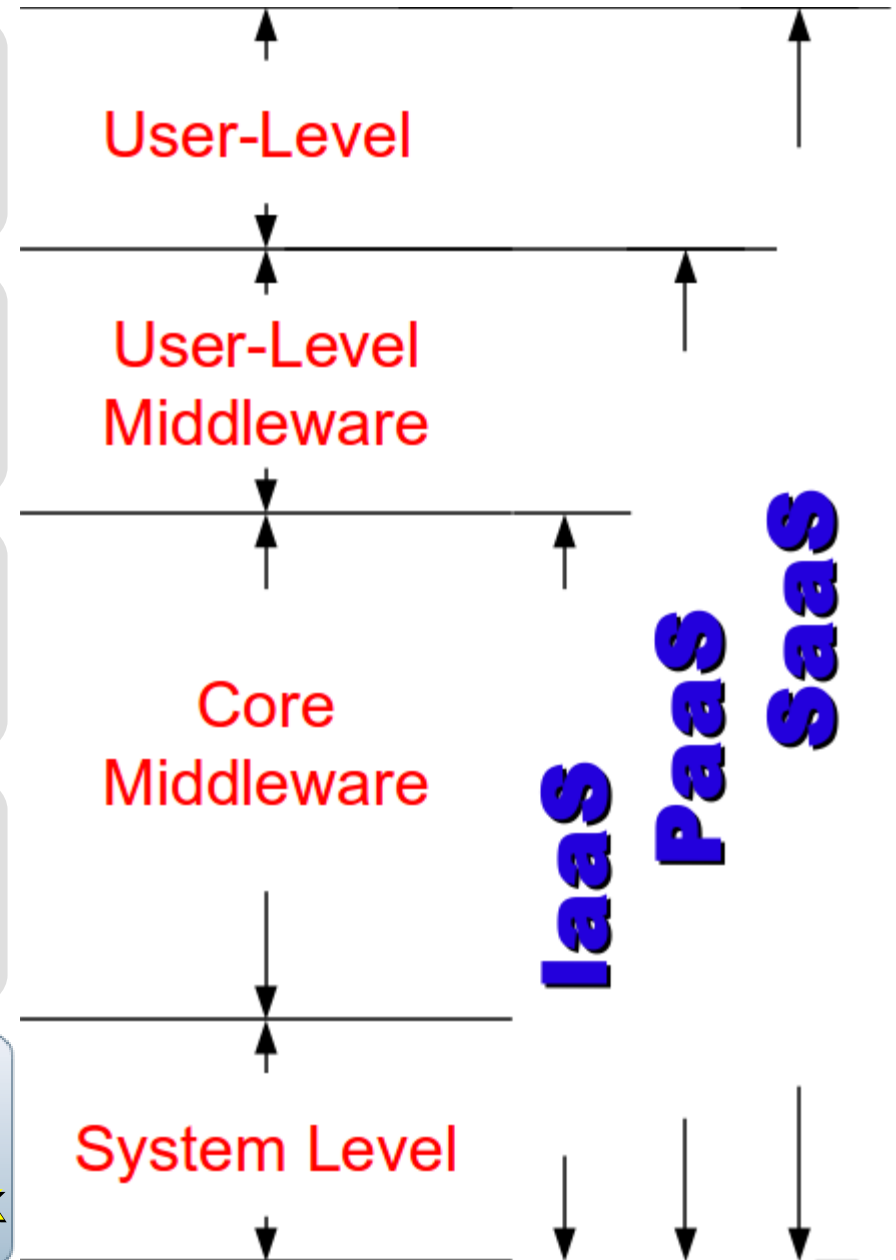
Qos Negotiation, Admission Control, Pricing, SLA Management, Metering...

### 虛擬化 Virtualization

VM, VM management and Deployment

### 硬體設施 Hardware

Infrastructure: Computer, Storage, Network





# Open Source to build Cloud Service

## 建構雲端服務的 自由軟體

### 應用軟體 Application

Social Computing, Enterprise, ISV, ...

eyeOS, Nutch, ICAS,  
X-RIME, ...

### 程式語言 Programming

Web 2.0 介面, Mashups, Workflows, ...

Hadoop (MapReduce),  
Sector/Sphere, AppScale

### 控制管理 Control

Qos Negotiation, Admission Control,  
Pricing, SLA Management, Metering...

OpenNebula, Enomaly,  
Eucalyptus, OpenQRM, ...

### 虛擬化 Virtualization

VM, VM management and Deployment

Xen, KVM, VirtualBox,  
QEMU, OpenVZ, ...

### 硬體設施 Hardware

Infrastructure: Computer, Storage,  
Network





# 深入解析雲端大量資料分析技術

Part 2 : Deep Dive into Data Science Technologies

**Jazz Wang**  
**Yao-Tsung Wang**  
**jazz@nchc.org.tw**



Powered by DRBL

# Big Data Analysis : Social Computing & Business Intelligence

## 「社交運算」與「商業智慧」均仰賴大量資料分析

**DIGITIMES** 網站內容的著作權為大椽股份有限公司 (DIGITIMES Inc.) 所有, 或其他授權DIGITIMES使用的內容提供者所有。

使用者下載或拷貝網站的內容或服務僅限於供個人、非商業用途之使用, 但不得以任何形式傳輸、重製、散布或提供予公眾。使用人利用時必須遵守著作權法的所有相關規定, 不可變更、發行、播送、轉賣、重製、改作、散布、表演、展示或利用DIGITIMES所屬網站上局部或全部內容及服務以賺取利益。

### 提升商業分析效果 資料倉儲業提倡資料社交化

2010/10/27 - DIGITIMES 馬培治 / 台北

社交運算(social computing)隨著包括Facebook在內的各式社交網絡服務持續發燒, 也成為企業資訊系統發展的重點之一, 繼IBM、微軟(Microsoft)與甲骨文(Oracle)等大廠提倡在應用軟體功能上支援社交功能之後, 資料倉儲(Data Warehouse)業者Teradata則提倡企業資料分析, 應納入包括社交資訊在內的多元因子, 讓不同資料源間的資料「社交化」(socialization of data), 以增強商業分析效果, 提高掌握用戶行為並輔助商業決策。

Teradata業務發展暨行銷執行副總裁Darryl D. McDonald於25日在自家全球合作夥伴與使用者大會上表示, 除了傳統企業營運資料, 各種可用來擷取資訊的資料源, 如RFID、智慧型裝置、社交網路, 乃至各種感應器, 將會對現今的企業分析帶來龐大的衝擊, 他建議企業可以開始著手思考, 如何將這些新興資料源的資料與傳統商業智慧分析的資料進行整合, 以期從更豐富的資料中, 找出過去商業分析方法看不到的隱性資訊。

McDonald表示, Facebook目前已經擁有超過5億個註冊用戶, 而推特(Tweeter)每天也有超過8,500萬條訊息產生, 若企業能夠將自身的用戶資訊或營運資料與這些龐大的資訊源進行有意義的分析, 將能夠激發在商業分析領域的創新應用。

他以參加Teradata全球合作夥伴暨使用者大會的3,000多名與會者為例進行分析, 發現這些與會者代表的公司總計具有9兆美元的資本額, 以及合計達230萬個線上社交網路服務的人際連結數, McDonald說, 這些資訊代表龐大的商機, 以及可供未來利用在業務推廣、行銷等目的使用。

參考來源：提升商業分析效果 資料倉儲業提倡資料社交化 (2010/10/27)

<http://goo.gl/2GoMo>



# 中華電信用 Hadoop 技術分析通話明細

iThome online

找資料 >

請輸入關鍵字

全站文章

IT邦幫忙

搜尋

[訂閱電子報](#) [RSS訂閱](#)

首頁

新聞

技術

IT管理

研討會

IT邦幫忙

IT邦部落格

小7聚樂部

iThome Download

個資法

| 手機版



hicloud 雲端運算伺服器

郵件、資料庫、防火牆...  
輕鬆解決企業 IT 資源需求

[iThome週週為IT人打氣!](#)

[雲端伺服器首選, 半年免費](#)

[企業選平板? 選最相容的!](#)

新聞

新聞專題

即時新聞

新聞簡訊

技術

產品報導

技術專題

IT書訊

IT管理

CIO

IT人物

專欄

新聞總覽

業界動態

訂閱電子報

## 中華電信用Hadoop技術分析通話明細

文/辜雅菴 2011-06-12

+1 0



62 人說讚。快免費註冊來查看你的朋友對什麼說讚。

我要收藏

中華電信利用自行開發的Hadoop大資料運算平臺，找出非結構化資料中的結構性，精簡資料後再置於資料倉儲運算，節省儲存空間

面對資料快速成長以及非結構性資料的增加，中華電信資訊處第四科科長楊秀一表示，中華電信近來利用Hadoop雲端運算技術自行開發了一個專門用來分析非結構化資料的巨量資料 (Big Data) 運算平臺，嘗試在資料進到資料倉儲系統之前，先進行資料的分析與處理以減少資料倉儲的資料量。

近年來行動語音市場趨於飽和，為了掌握用戶特性進行客製化行銷，一份資料要進行分析，就會被多次複製，因此即使用戶增加趨緩，但中華電信擁有的資料量仍快速暴增。

### 研討會訊息

· [Websense TRITON電子郵件資料安全解決方案研討會](#)

· [2011 JavaTWO專業技術大會](#)

[+更多研討會](#)

▼ ADVERTISEMENT ▼

Microsoft

# Three Core Technologies of Google ....

## Google 的三大關鍵技術 ....

- Google 在一些會議分享他們的三大關鍵技術
- Google shared their design of web-search engine
  - SOSP 2003 :
    - “The Google File System”
    - <http://labs.google.com/papers/gfs.html>
  - OSDI 2004 :
    - “MapReduce : Simplified Data Processing on Large Cluster”
    - <http://labs.google.com/papers/mapreduce.html>
  - OSDI 2006 :
    - “Bigtable: A Distributed Storage System for Structured Data”
    - <http://labs.google.com/papers/bigtable-osdi06.pdf>



# Open Source Mapping of Google Core Technologies

## Google 三大關鍵技術對應的自由軟體

### BigTable

A huge key-value datastore

HBase, Hypertable  
Cassandra, ....

### MapReduce

To parallel process data

Hadoop MapReduce API  
Sphere MapReduce API, ...

### Google File System

To store petabytes of data

Hadoop Distributed File System (HDFS)  
Sector Distributed File System

更多不同語言的 MapReduce API 實作：

<http://trac.nchc.org.tw/grid/intertrac/wiki%3Ajazz/09-04-14%23MapReduce>

其他值得觀察的分散式檔案系統：

- IBM GPFS - <http://www-03.ibm.com/systems/software/gpfs/>
- Lustre - <http://www.lustre.org/>
- Ceph - <http://ceph.newdream.net/>



# Building PaaS with Open Source

## 用自由軟體打造 PaaS 雲端服務

應用軟體 Application  
Social Computing, Enterprise, ISV, ...

eyeOS, Nutch, ICAS,  
X-RIME, ...

程式語言 Programming  
Web 2.0 介面, Mashups, Workflows, ...

Hadoop (MapReduce),  
Sector/Sphere, AppScale

控制管理 Control  
Qos Negotiation, Admission Control,  
Pricing, SLA Management, Metering...

OpenNebula, Enomaly,  
Eucalyptus, OpenQRM, ...

虛擬化 Virtualization  
VM, VM management and Deployment

Xen, KVM, VirtualBox,  
QEMU, OpenVZ, ...

硬體設施 Hardware  
Infrastructure: Computer, Storage, Network

# Hadoop

- <http://hadoop.apache.org>
- Hadoop 是 Apache Top Level 開發專案
- **Hadoop is Apache Top Level Project**
- 目前主要由 Yahoo! 資助、開發與運用
- **Major sponsor is Yahoo!**
- 創始者是 Doug Cutting，參考 Google Filesystem
- **Developed by Doug Cutting, Reference from Google Filesystem**
- 以 Java 開發，提供 HDFS 與 MapReduce API。
- **Written by Java, it provides HDFS and MapReduce API**
- 2006 年使用在 Yahoo 內部服務中
- **Used in Yahoo since year 2006**
- 已佈署於上千個節點。
- **It had been deploy to 4000+ nodes in Yahoo**
- 處理 Petabyte 等級資料量。
- **Design to process dataset in Petabyte**



**Facebook, Last.fm,  
Joost, Twitter  
are also powered  
by Hadoop**

# Sector / Sphere

- <http://sector.sourceforge.net/>
- 由美國資料探勘中心研發的自由軟體專案。
- **Developed by National Center for Data Mining, USA**
- 採用 C/C++ 語言撰寫，因此效能較 Hadoop 更好。
- **Written by C/C++, so performance is better than Hadoop**
- 提供「類似」Google File System 與 MapReduce 的機制
- **Provide file system similar to Google File System and MapReduce API**
- 基於 [UDT 高效率網路協定](#) 來加速資料傳輸效率
- **Based on UDT which enhance the network performance**
- [Open Cloud Testbed](#) 有提供測試環境，並開發 [MalStone 效能評比軟體](#)
- **Open Cloud Consortium provide Open Cloud Testbed and develop MalStone toolkit for benchmark**



National Center for Data Mining  
University of Illinois at Chicago



Open Data Group  
<http://www.opendatagroup.com/>



# 巨量資料分析處理平台：Hadoop

Part 3 : Introduction to Hadoop

**Jazz Wang**  
**Yao-Tsung Wang**  
**jazz@nchc.org.tw**



Powered by DRBL

# What is Hadoop ?

用一句話解釋 **Hadoop** 是什麼 ??

*Hadoop is a **software platform** that lets one easily write and run applications that **process vast amounts of data.***

**Hadoop** 是一個讓使用者簡易撰寫並執行處理海量資料應用程式的軟體平台。

亦可以想像成一個處理海量資料的生產線，只須學會定義 **map** 跟 **reduce** 工作站該做哪些事情。



# Two Key Elements of Operating System

## 作業系統兩大關鍵組成元素

**Scheduler**  
程序排程

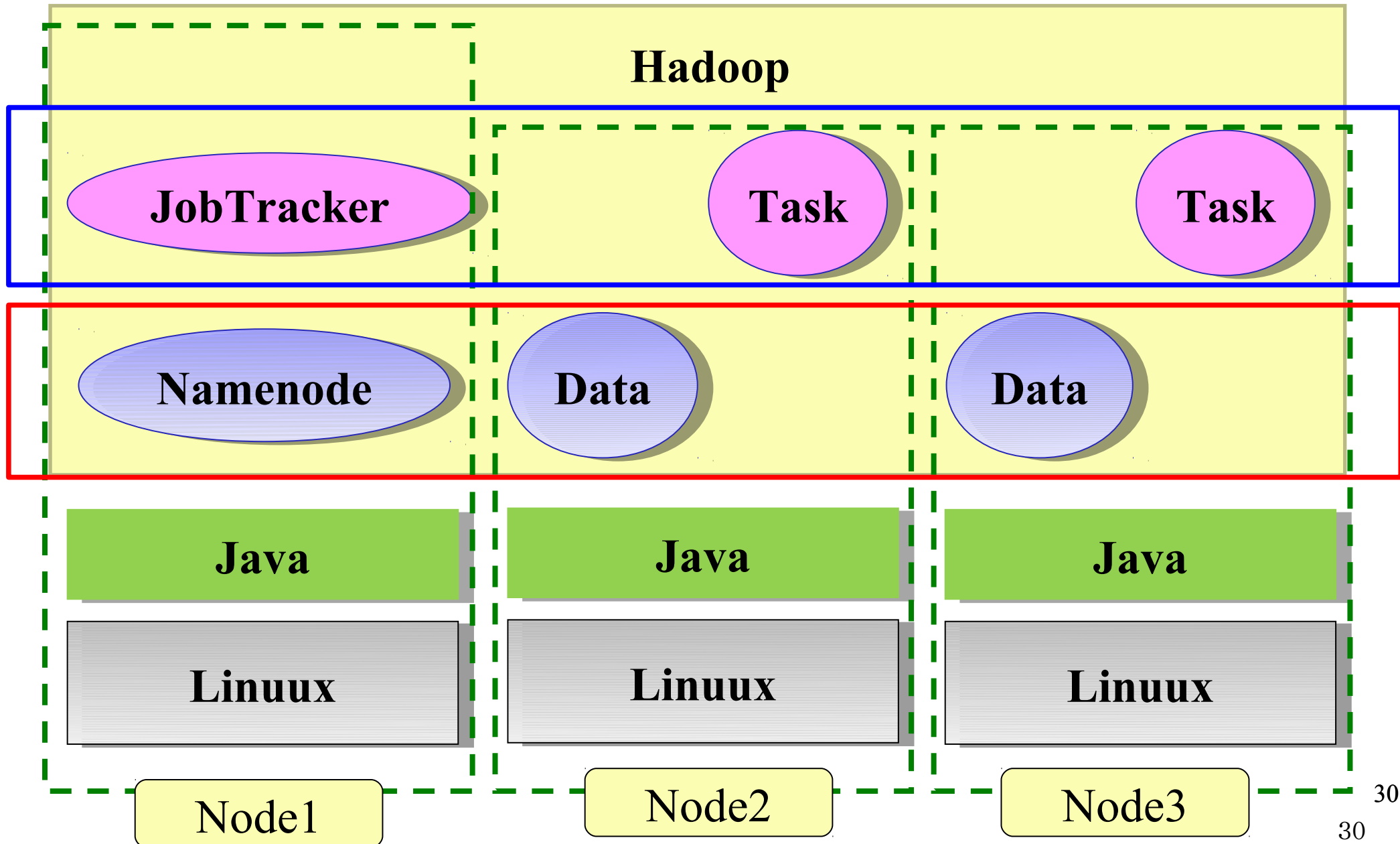


**File System**  
檔案系統



# Distributed Operating System of Hadoop

**Hadoop** 建構成一個分散式作業系統



# Features of Hadoop ...

## Hadoop 這套軟體的特色是 ...

- **海量 Vast Amounts of Data**
  - 擁有儲存與處理大量資料的能力
  - Capability to **STORE** and **PROCESS** vast amounts of data.
- **經濟 Cost Efficiency**
  - 可以用在由一般 PC 所架設的叢集環境內
  - Based on large clusters built of **commodity hardware**.
- **效率 Parallel Performance**
  - 透過分散式檔案系統的幫助，以致得到快速的回應
  - With the help of HDFS, Hadoop **have better performance**.
- **可靠 Robustness**
  - 當某節點發生錯誤，能即時自動取得備份資料及佈署運算資源
  - Robustness to add and remove computing and storage resource without shutdown entire system.

# Founder of Hadoop – Doug Cutting

**Hadoop** 這套軟體的創辦人 **Doug Cutting**

Doug Cutting Talks About The Founding Of Hadoop

clouderahadoop

9 部影片

編輯訂閱項目



Doug Cutting Talks About The Founding Of Hadoop

<http://www.youtube.com/watch?v=qxC4urJOchs>



# History of Hadoop ... 2002~2004

## Hadoop 這套軟體的歷史源起 ... 2002~2004



- Lucene

- <http://lucene.apache.org/>
- 用 Java 設計的高效能文件索引引擎 API
- a high-performance, full-featured **text search engine library** written entirely in **Java**.
- 索引文件中的每一字，讓搜尋的效率比傳統逐字比較還要高的多
- Lucene create an **inverse index** of every word in different documents. It enhance performance of text searching.



# History of Hadoop ... 2002~2004

## Hadoop 這套軟體的歷史源起 ... 2002~2004

- Nutch
  - <http://nutch.apache.org/>
  - Nutch 是基於開放原始碼所開發的網站搜尋引擎
  - Nutch is open source **web-search** software.
  - 利用 Lucene 函式庫開發
  - It builds on **Lucene and Solr**, adding web-specifics, such as a **crawler**, a **link-graph database**, parsers for HTML and other document formats, etc.



# History of Hadoop ... 2004 ~ Now

## *Hadoop* 這套軟體的歷史源起 ... 2004 ~ Now

- Nutch 後來遇到儲存大量網站資料的瓶頸，剛好看到 Google 在一些會議分享他們的三大關鍵技術 ...
- Added DFS & MapReduce implement to Nutch
- According to **user feedback** on the mail list of Nutch ....
- Hadoop became separated project **since Nutch 0.8**
- Nutch DFS → Hadoop Distributed File System (HDFS)
- **Yahoo** hire Dong Cutting to build a team of web search engine at **year 2006**.
  - Only **14 team members** (engineers, clusters, users, etc.)
- Dong Cutting joined Cloudera at year 2009.

**YAHOO!**

 cloudera



# Who Use Hadoop ??

有哪些公司在用 **Hadoop** 這套軟體 ??

- **Yahoo** is the key contributor currently.
- **IBM** and **Google** teach Hadoop in universities ...
- [http://www.google.com/intl/en/press/pressrel/20071008\\_ibm\\_univ.html](http://www.google.com/intl/en/press/pressrel/20071008_ibm_univ.html)
- **The New York Times** used **100 Amazon EC2 instances** and a Hadoop application to process **4TB of raw image TIFF data** (stored in S3) into **11 million finished PDFs** in the space of **24 hours** at a computation cost of about **\$240** (not including bandwidth)
  - from <http://en.wikipedia.org/wiki/Hadoop>
- <http://wiki.apache.org/hadoop/AmazonEC2>
- <http://wiki.apache.org/hadoop/PoweredBy>
  - A9.com
  - ADSDAQ by Contextweb
  - EHarmony
  - Facebook
  - Fox Interactive Media
  - IBM
  - ImageShack
  - ISI
  - Joost
  - Last.fm
  - Powerset
  - The New York Times
  - Rackspace
  - Veoh
  - Metaweb

# Performance improvement of Hadoop

## Hadoop 過去幾年的效能改進 (from Yahoo)

年份	日期	節點數	耗時 ( 小時 )
2006	四月	188	47.9
2006	五月	500	42
2006	十一月	20	1.8
2006	十一月	100	3.3
2006	十一月	500	5.2
2006	十一月	900	7.8
2007	七月	20	1.2
2007	七月	100	1.3
2007	七月	500	2
2007	七月	900	2.5

Sort benchmark, every nodes with terabytes data.



# Hadoop in production run ....

## 商業運轉中的 **Hadoop** 應用 ....

- February 19, 2008
- Yahoo! Launches World's Largest Hadoop Production Application
- <http://developer.yahoo.net/blogs/hadoop/2008/02/yahoo-worlds-largest-production-hadoop.html>

Number of links between pages in the index	roughly 1 trillion links
Size of output	over 300 TB, compressed!
Number of cores used to run single Map-Reduce job	over 10,000
Raw disk used in the production cluster	over 5 Petabytes

# Hadoop in production run ....

## 商業運轉中的 **Hadoop** 應用 ....

- **September 30, 2008**
- **Scaling Hadoop to 4000 nodes at Yahoo!**
- [http://developer.yahoo.net/blogs/hadoop/2008/09/scaling\\_hadoop\\_to\\_4000\\_nodes\\_a.html](http://developer.yahoo.net/blogs/hadoop/2008/09/scaling_hadoop_to_4000_nodes_a.html)

<b>Total Nodes</b>	<b>4000</b>
<b>Total cores</b>	<b>30000</b>
<b>Data</b>	<b>16PB</b>

	<b>500-node cluster</b>		<b>4000-node cluster</b>	
	<b>write</b>	<b>read</b>	<b>write</b>	<b>read</b>
<b>number of files</b>	990	990	14,000	14,000
<b>file size (MB)</b>	320	320	360	360
<b>total MB processes</b>	316,800	316,800	5,040,000	5,040,000
<b>tasks per node</b>	2	2	4	4
<b>avg. throughput (MB/s)</b>	<b>5.8</b>	<b>18</b>	<b>40</b>	<b>66</b>



# 打造內網搜尋引擎：Nutch & Crawlzilla

Part 4 : Introduction to Nutch and Crawlzilla

**Jazz Wang**  
**Yao-Tsung Wang**  
**jazz@nchc.org.tw**



Powered by DRBL

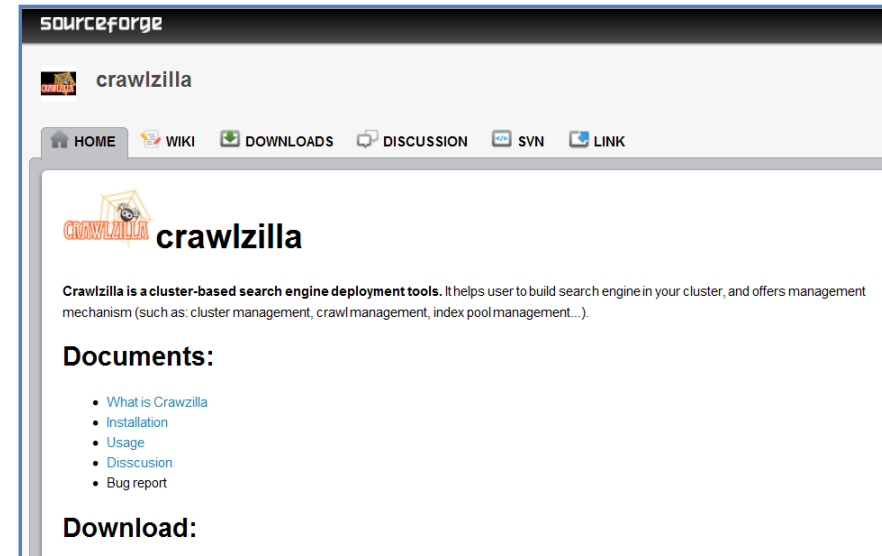
# Crawlzilla ?

- 建構專屬於己的搜尋引擎  
自由軟體專案

- 快速找到資訊所在
- 保障資料機敏性
- 提供索引庫統計資訊
- 會操作滑鼠就會使用

- 專案網站

- 中文：<http://crawlzilla.info/>
- 英文：<http://sf.net/p/crawlzilla>





# Crawlzilla @

## 建立防火牆內的 搜尋

- 現有的公開搜尋引擎無法、也不可穿透防火牆，搜尋內部網路的資料

## 在正確的資料 內搜尋

- 減少廣告、不必要的內容、不當的資訊

## 破除資料庫內 搜尋的限制

- 使用者上傳的附件檔（ doc, ppt, pdf... ） 、或超連結到站外的網站資訊

# Crawlzilla !

## Admin

```

正在重建相依關係
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成
沒有套件將會被安裝、升級或移除。
0 個套件升級, 0 個新安裝, 0 個將移除且 13 個不會升級。
需要下載 0B 的解壓縮檔案。解裝後將用去 0B。
正在編輯延伸狀態訊息... 完成
正在讀取套件清單... 完成
正在重建相依關係
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成

系統有 Sun Java 1.6 以上版本
系統已有 ssh。
系統已有 ssh Server (sshd)。
系統已有 dialog。
歡迎使用 Crawlzilla, 此安裝程序會為您新建一個 crawler 帳號並協助您設定密碼。
請輸入欲設定的 crawler 密碼:
password:
請再輸入一次確認密碼:
password:

Master 網域 IP 位址為: 146.110.138.186
Master 的 MAC 為: 08:00:27:99:4d:09
請確認上述的安裝資訊: 1. 正確 2. 不正確
    
```

```

[ Crawlzilla 管理介面 ] -by NCHC =
請選擇:
[ 管理功能選項 ]
luster_status 檢查 Cluster 狀態
fast_manage 快速啟動/關閉所有服務及 Tomcat
cluster_setup 設定 datanode & tasktracker
server_setup 設定 namenode & jobtracker
tomcat_switch 啟動/停止/重新啟動 Tomcat
tomcat_port 更改 Tomcat port
lang_switch 更換語言
client_install Client 安裝步驟
exit 結束
    
```

建立搜尋環境、選擇佈署叢集

## IT

### Crawl-建立搜尋引擎

#### Crawl 爬取設定

索引庫名稱:

輸入欲爬取的網址(可多行):

爬取深度設定:

#### 排程設定(Option)

#### 基本資訊

索引庫名稱: narl\_3  
 搜尋引擎連結位置: /home/crawler/crawlzilla/user/admin/IDB/narl\_3/index  
 搜尋引擎狀態: OK  
 爬取深度: 3  
 建立時間: 20110503-12:15:19  
 執行時間: 0:55:53  
 超始連結: http://www.narl.org.tw/tw/

#### 立即觀看及網頁嵌入語法

### 索引庫內容 - narl\_3

#### 資料總覽:

#### 被搜尋分析到的網址:

Order	Contents	Counts	Order	Contents	Counts
0	site:www.narl.org.tw	248	1	site:www.stb.org.tw	20
2	site:www.nspo.org.tw	3	3	site:conf.ncrec.org.tw	3
4	site:www.niac.narl.org.tw	2	5	site:i-one.org.tw	2
6	site:www.cic.narl.org.tw	1	7	site:www.mirdc.org.tw	1
8	site:web1.nsc.gov.tw	1	9	site:www.itri.org.tw	1
	site:www.ttfri.narl.org.tw	1	11	site:www.itrc.narl.org.tw	1

建立索引庫、瀏覽索引庫統計資訊

## User

Crawlzilla 管理介面

搜索

[首頁](#) | [關於](#) | [公告](#) | [新聞](#) | [服務](#) | [支援](#) | [聯繫](#) | [隱私](#) | [安全](#) | [法律](#) | [其他](#)

國家高速網路與計算中心  
 National Center for High-Performance Computing  
 Better HPC Better Living

科技貢獻獎

Hits 1-11 (out of about 11 total matching pages):

**國網中心公告系統**  
 ... 關「行政院傑出科技貢獻獎實施要點」、「行政院 ... 理行政院傑出科技貢 ...  
[https://intra.nchc.org.tw/HCMS/itr/inform\\_info.php?post=1302156081](https://intra.nchc.org.tw/HCMS/itr/inform_info.php?post=1302156081) (cached)

**ISO文件::管理規範專區**  
 ... 驗研究院傑出科技貢獻獎作業要點 TOP ...  
<http://iso.nchc.org.tw/document/> (cached) (explain) (anchors)

**重要記事::國研院國網中心**  
 ... 年度行政院傑出科技貢獻獎 2007年6月 國 ... 年度行政院傑出科技貢 ...  
<http://www.nchc.org.tw/tw/about/history.php> (cached) (explain) (anchors)

享受搜尋效益

# 技術的突破性

- **Crawlzilla** 被打造成企業或個人都可以輕鬆擁有專屬的搜尋引擎，也是目前沒有任何軟體 / 搜尋引擎可以取代的。
- 以自由軟體為巨人的肩膀，讓使用者有使用、複製、修改與再散播的自由
- 化繁為簡，透過簡明的介面完成建構需複雜資訊技術的搜尋引擎，是本專案最大的突破

# 技術的突破性

使用最新的視覺化網頁介面：**Web 2.0**

- AJAX 技術, W3C standard

整合最熱門的雲端運算演算法：**MapReduce**

- Google like
- 高效率、高容錯、高平行化

依循最穩固的程式開發架構：**Model-View-Control**

- 單元開發、程式再利用
- 全球化



# 2010 開放原始碼創新應用開發大賽 職業組冠軍



# 來自世界各地的下載

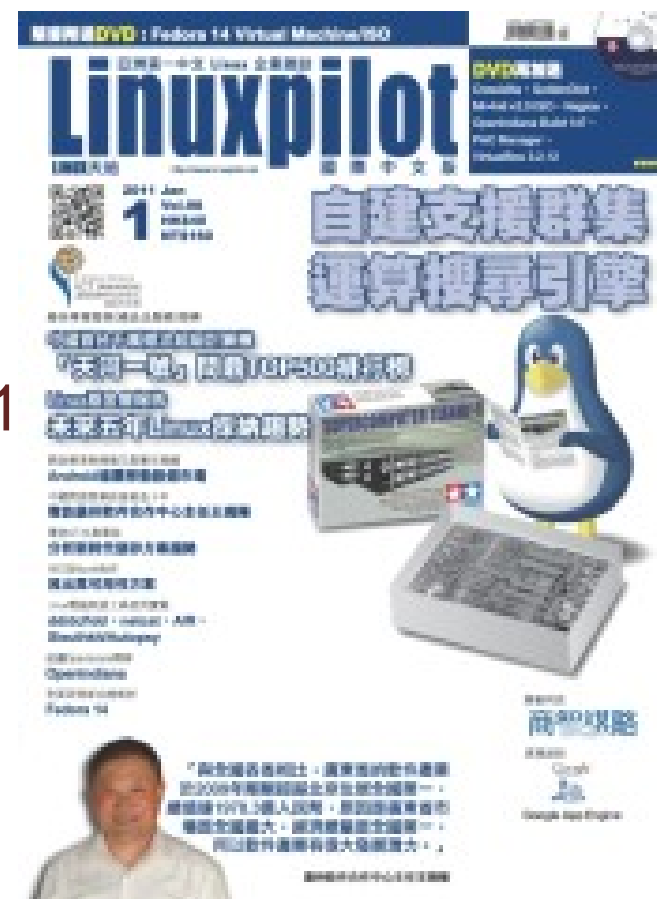
Visits ▾



- 來自 33 個國家， 1,397 次下載量
- 來自 53 個國家， 4479 次造訪紀錄
- 統計區間： 2010-08-17~2011-03-25

# 媒體報導

- 專案文獻：
  - TANET 2010：快速佈署叢集式搜尋引擎 - Crawlzilla
  - INTENSIVE 2011：Crawlzilla - A Toolkit for Deploying Cluster Search Engine Quickly and Easily
- 媒體報導：
  - 自建支援群集運算搜尋引擎
    - LinuxPilot Taiwan Vol.98 2011/01/13
  - 開放原始碼創新應用開發大賽
    - 創新發現誌 2011/01/25 Vol.20110
  - Three little zillas from Taiwan - iTWire 201



# 特色

## 雲端運算

- **Map Reduce** 演算法，分散工作量，整合運算結果

## 全系列Linux

- 單機、叢集上的任何**Linux** 套件版本，並自動解決軟體相依的問題

## 雲端介面

- 只需使用瀏覽器

## 統計管理

- 搜尋選項、瀏覽統計資料庫、叢集狀態

## 全文索引

- 全文索引引擎，並且能分析各種檔案格式（**html, txt, pdf, doc, ppt...**等）。

## 在地化MIT

- 提供完整的中英文操作設定，

## 多庫並存

- 可多個搜尋引擎庫同時上線使用



## 與其他國際知名的自由軟體比較

	Spidr	Nutch	Crawlzilla V 0.3	Crawlzilla V 1.0
安裝方式	Rube 套件 安裝	配置設定 檔	提供自動安裝 程式	提供自動安裝 程式
爬取網頁	O	O	O	O
分析內容	X	O	O	O
搜尋庫資訊	X	X	O	O
操作介面	指令	指令	Web-UI	Web-UI
中文最佳化	X	X	O	O
多人帳號, 排程機制	X	X	X	O

# 應用實例

- 嘉義縣網中心

- 將用於課堂教材的統籌搜尋，如校園資訊、教育部成語字典、..等網站為搜尋字庫的基礎，提供學生正確有益的關鍵字結果

- 慈濟 - 資訊處

- 將使用 **crawlzilla** 來對所有內部的文件，提供統一的搜尋服務，提供更快更便捷的方式找到資料

- 東海大學高效能實驗室

- 結合雲端分散儲存與 **Nutch** 搜尋引擎之影音網站



# 應用實例

## NCHC- 內網首頁

### 院部數位服務

- ▮ HRMS人資系統
- ▮ 電子表單
- ▮ 電子郵件服務
- ▮ 公文系統
- ▮ 工時系統



[簡介](#) [常見問題](#)

科技貢獻獎

Search [help](#)

Hits **1-11** (out of about 11 total matching pages):

#### [國網中心公告系統](#)

... 關「行政院傑出**科技貢獻獎**實施要點」、「行政院 ... 理行政院傑出**科技貢** ...  
[https://intra.nchc.org.tw/HCMS/itr/inform\\_info.php?post=1302156081](https://intra.nchc.org.tw/HCMS/itr/inform_info.php?post=1302156081) ([cached](#))

#### [ISO文件::管理規範專區](#)

... 驗研究院傑出**科技貢獻獎**作業要點 TOP ...  
<http://iso.nchc.org.tw/document/> ([cached](#)) ([explain](#)) ([anchors](#))

#### [重要記事::國研院國網中心](#)

... 年度行政院傑出**科技貢獻獎** 2007年6月 國 ... 年度行政院傑出**科技貢** ...  
<http://www.nchc.org.tw/tw/about/history.php> ([cached](#)) ([explain](#)) ([anchors](#))

# Crawlzilla 效益

- 節省建置的成本
  - 某商業版的搜尋引擎費用為 USD \$18000 (NTD \$57 萬) ，年費和客製費用另計，且不提供程式碼。
- 節省資料搜尋的時間
- 基於 Apache License 2.0 ，讓企業可客製化成自由軟體或商業獨家軟體
- 透過 開放源碼的搜尋引擎 Crawlzilla 激發未來更多學術和商業價值





# 資料異機同步儲存的機制：DropBox

Part 5 : Introduction to DropBox

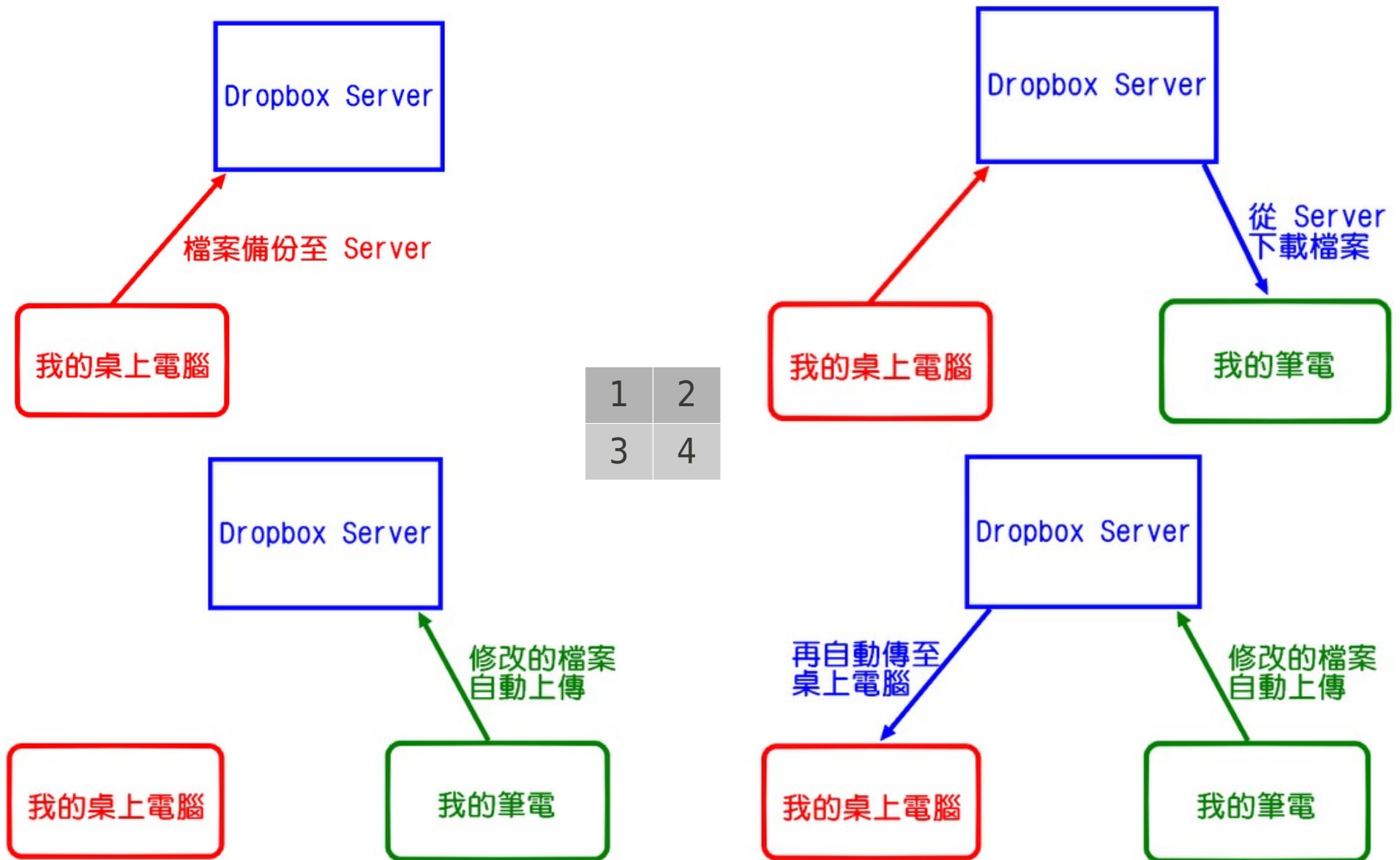
**Jazz Wang**  
**Yao-Tsung Wang**  
**jazz@nchc.org.tw**



Powered by DRBL

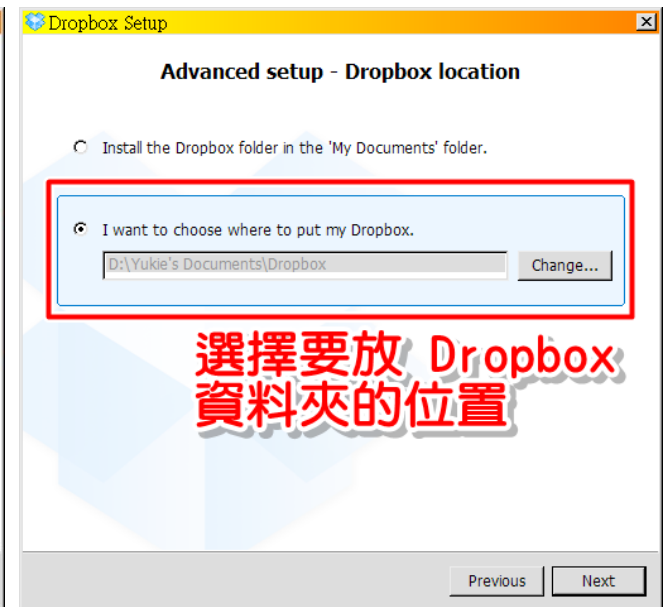
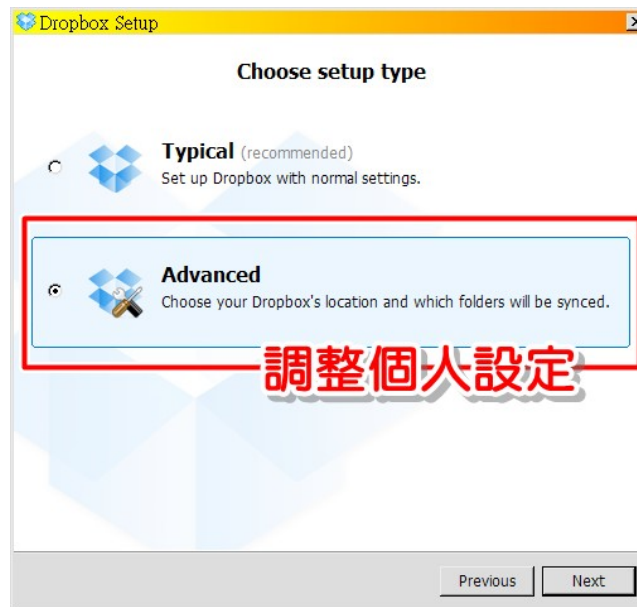
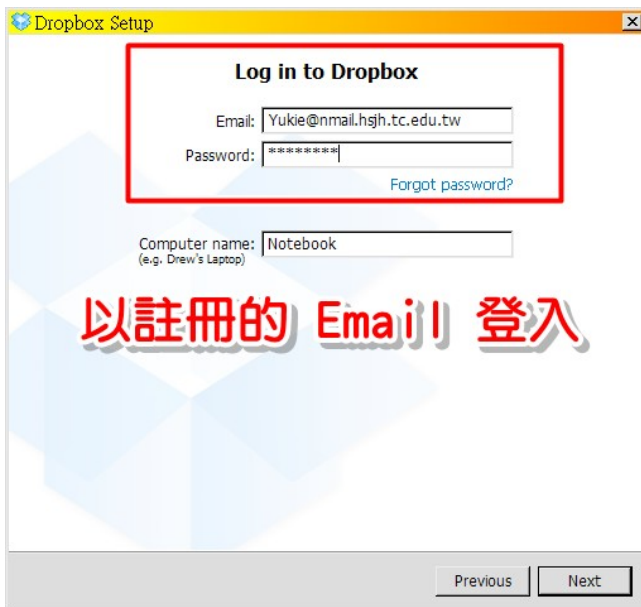
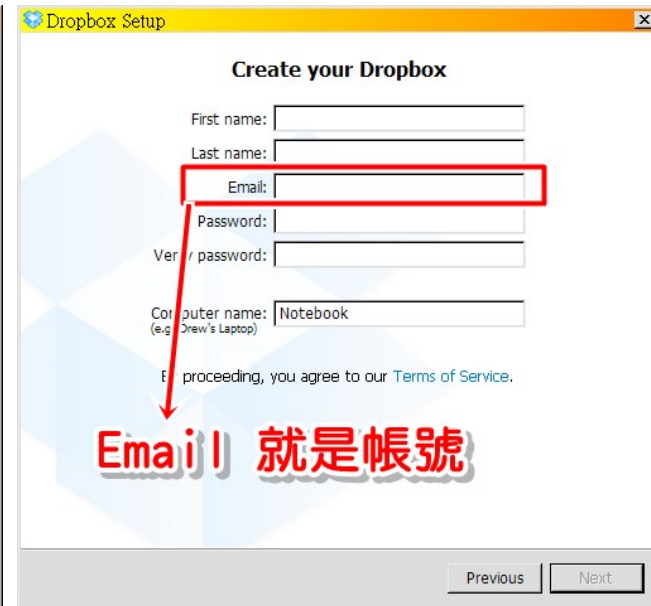
# DropBox 可以自動同步多台電腦的資料

## What is DropBox ?



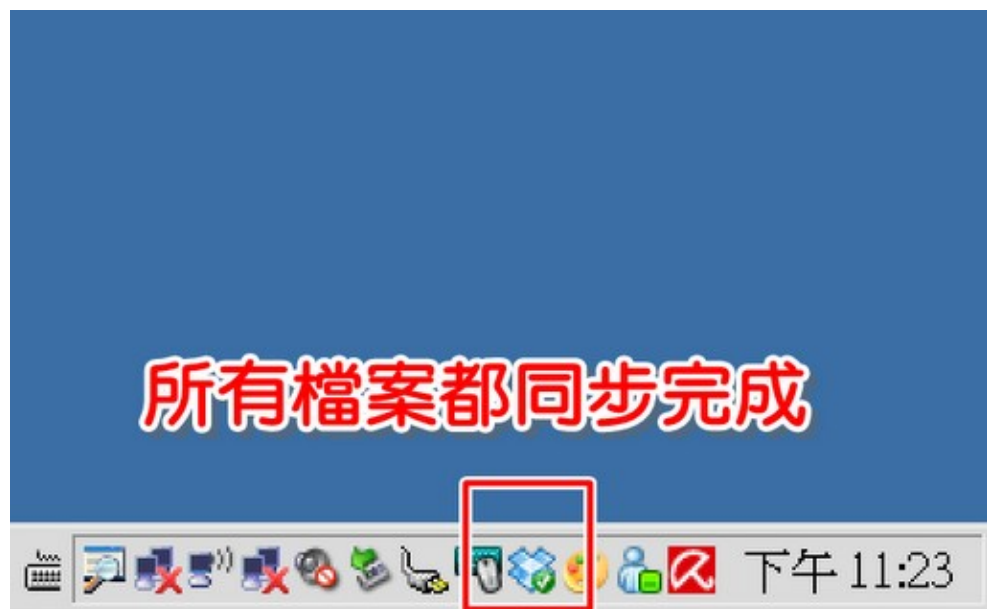
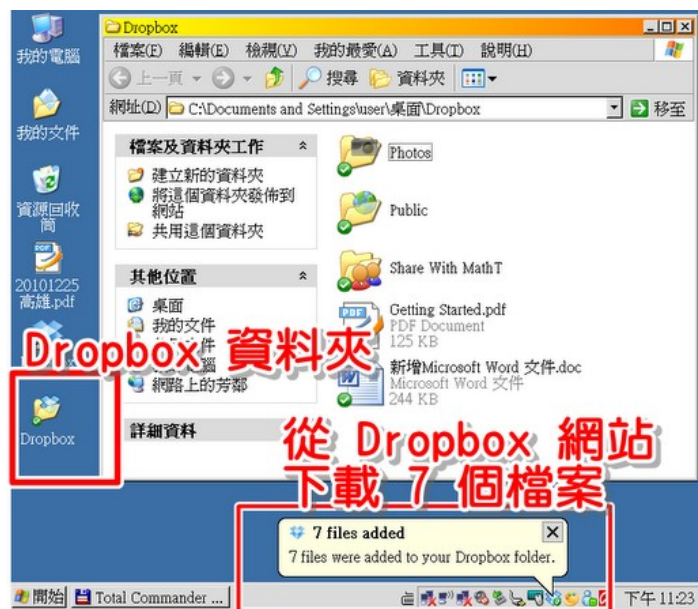
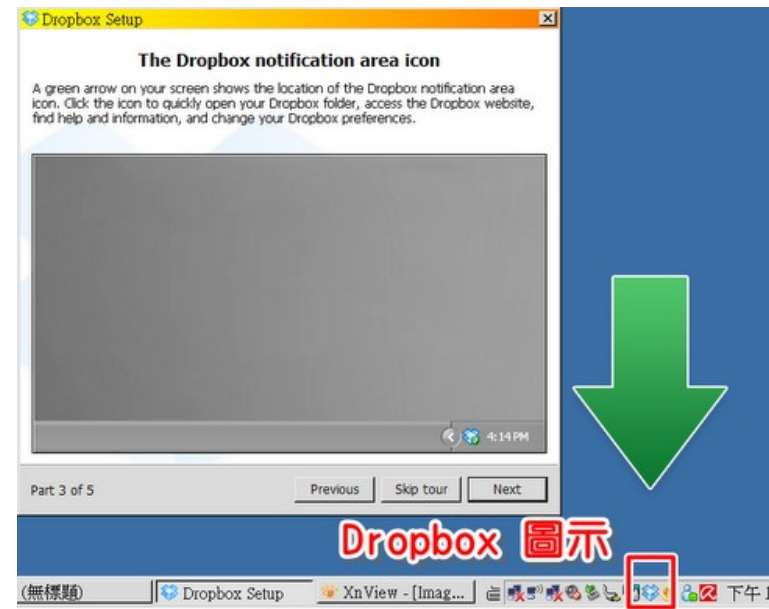
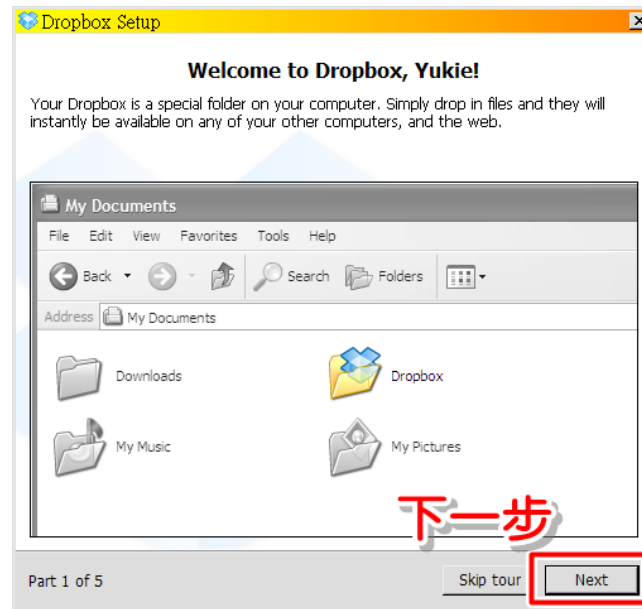
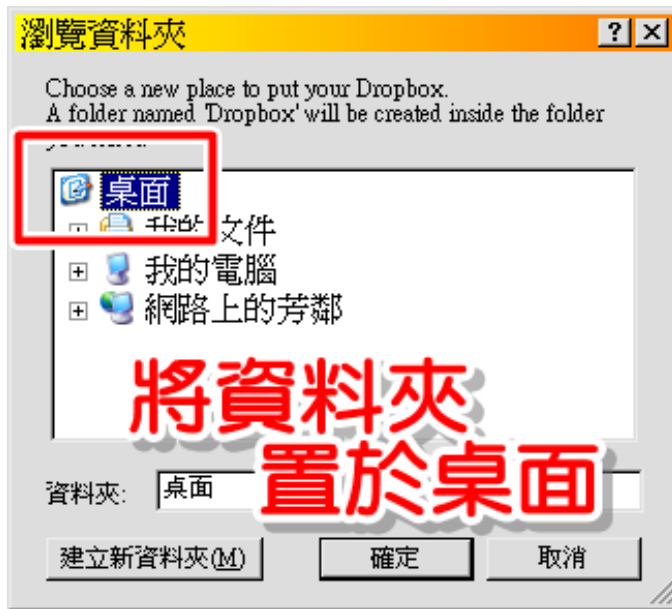
# 如何申請 DropBox ?

## How to apply DropBox accounts ?



# 如何設定與使用 DropBox ?

## How to setup and use DropBox ?







## Questions?

Slides - <http://trac.nchc.org.tw/cloud>

**Jazz Wang**  
**Yao-Tsung Wang**  
**jazz@nchc.org.tw**



Powered by DRBL