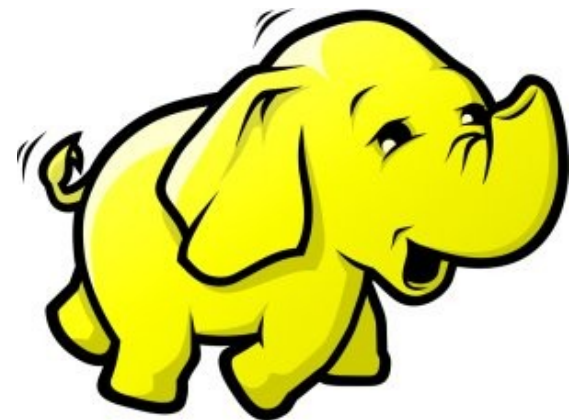




# 海量資料處理平台 Hadoop 與抓抓龍

## Introduction to Hadoop and Crawlzilla

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





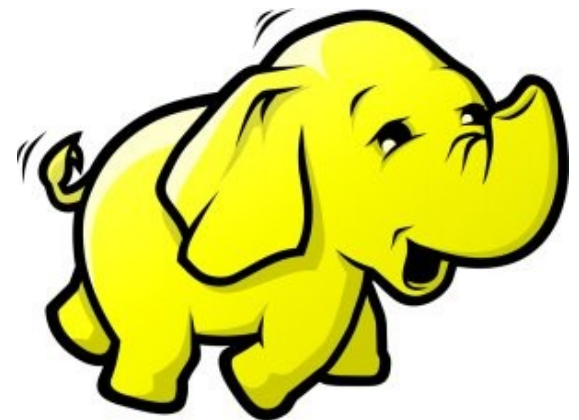
# 海量資料的趨勢與挑戰

## Trends and Challenges of Big Data

**Jazz Wang**

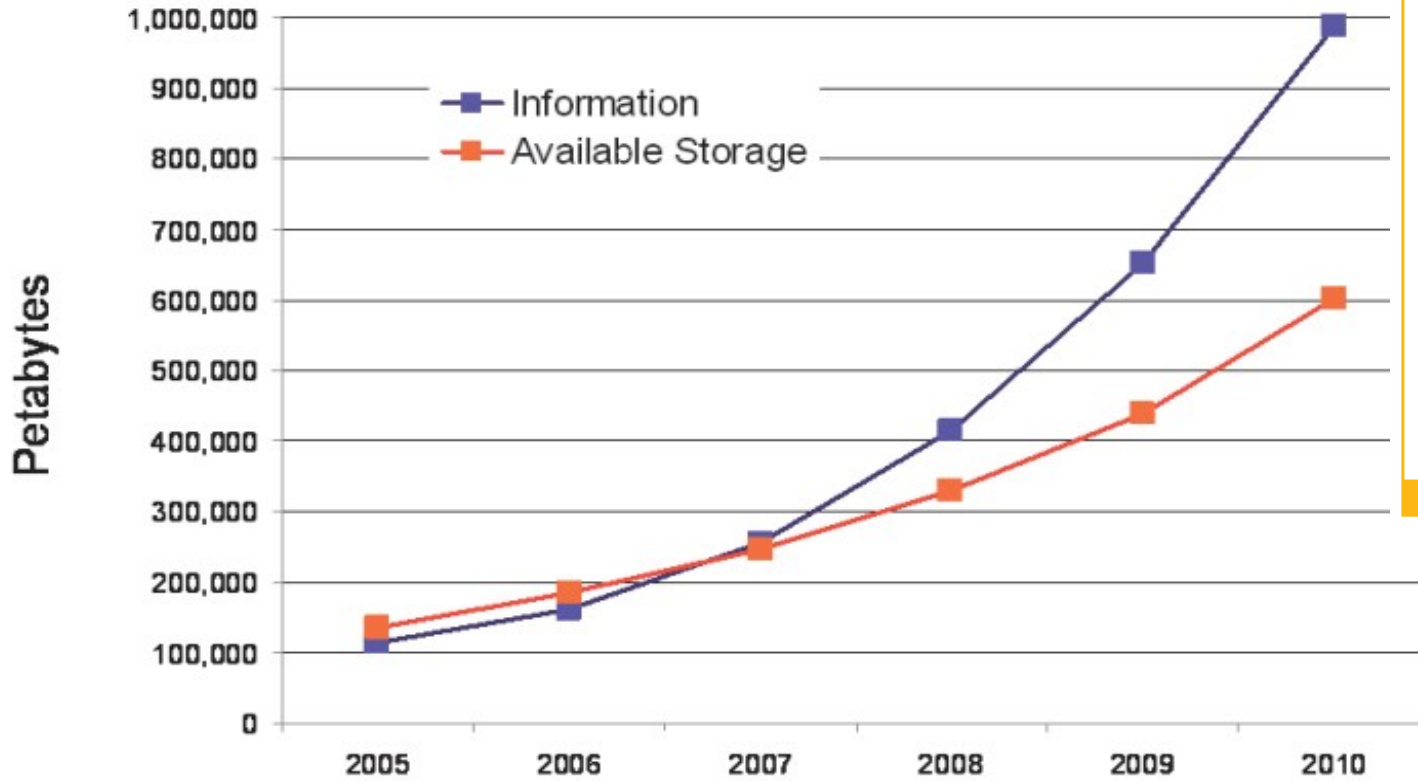
**Yao-Tsung Wang**

**[jazz@nchc.org.tw](mailto:jazz@nchc.org.tw)**



# Data Explosion!! 始於 2007 的「資料大爆炸」時代

## Information Versus Available Storage

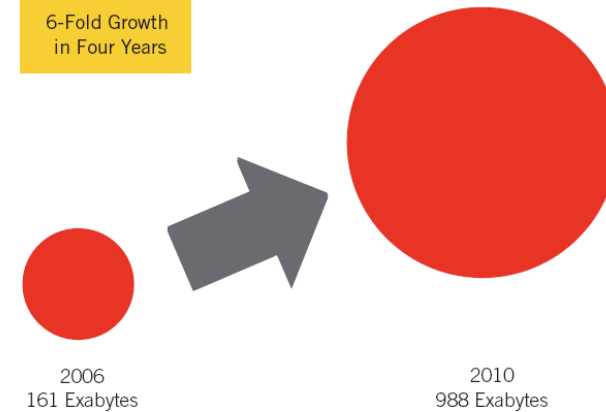


Source: IDC, 2007

Figure 1

### Information Created, Captured and Replicated

6-Fold Growth  
in Four Years



Source: IDC, 2007

2007 年，IDC 預估  
2010 年會成長**六倍**！  
(相較 2006 年)

2006 161 EB  
2010 988 EB (預測)

出處：The Expanding Digital Universe,  
A Forecast of Worldwide Information Growth Through 2010,  
March 2007, An IDC White Paper - sponsored by EMC

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

Data expanded 1.5x each year !! 每年約略 1.5 倍



追蹤歷年的 IDC 數據：

2006	161	EB	
2007	281	EB	
2008	487	EB	
2009	800	EB	(0.8 ZB)
2010	988	EB	(預測)
2010	1200	EB	(1.2 ZB)
2011	1773	EB	(預測)
2011	1800	EB	(1.8 ZB)

景氣差而成長趨緩？  
或受新技術抑制？

出處：[Extracting Value from Chaos](#),  
June 2011, An IDC White Paper - sponsored by EMC

<http://www.emc.com/collateral/about/news/idc-emc-digital-universe-2011-infographic.pdf>

# What is Big Data?! 何謂『海量資料』？

海量資料泛指資料大小已無法用一般軟體擷取、管理與處理；  
單一資料集大小介於數十 TB 至數 PB 的資料。

'Big Data' = few dozen TeraBytes to PetaBytes in single data set.

## Definition

[edit]

Big data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set.

In a 2001 research report<sup>[14]</sup> and related conference presentations, then META Group (now Gartner) analyst, Doug Laney, defined data growth challenges (and opportunities) as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in/out), and variety (range of data types, sources). Gartner continues to use this model for describing big data.<sup>[15]</sup>

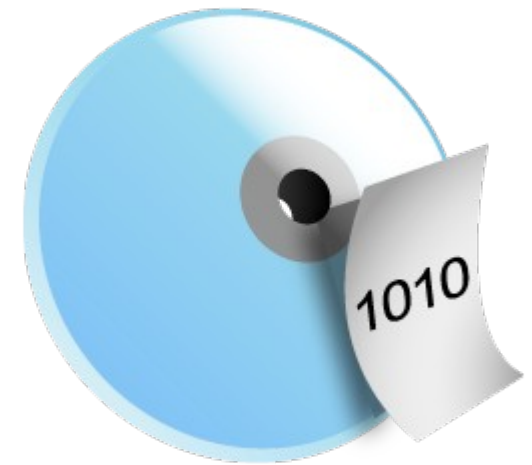
出處：[http://en.wikipedia.org/wiki/Big\\_data](http://en.wikipedia.org/wiki/Big_data)



多個檔案，容量 100TB



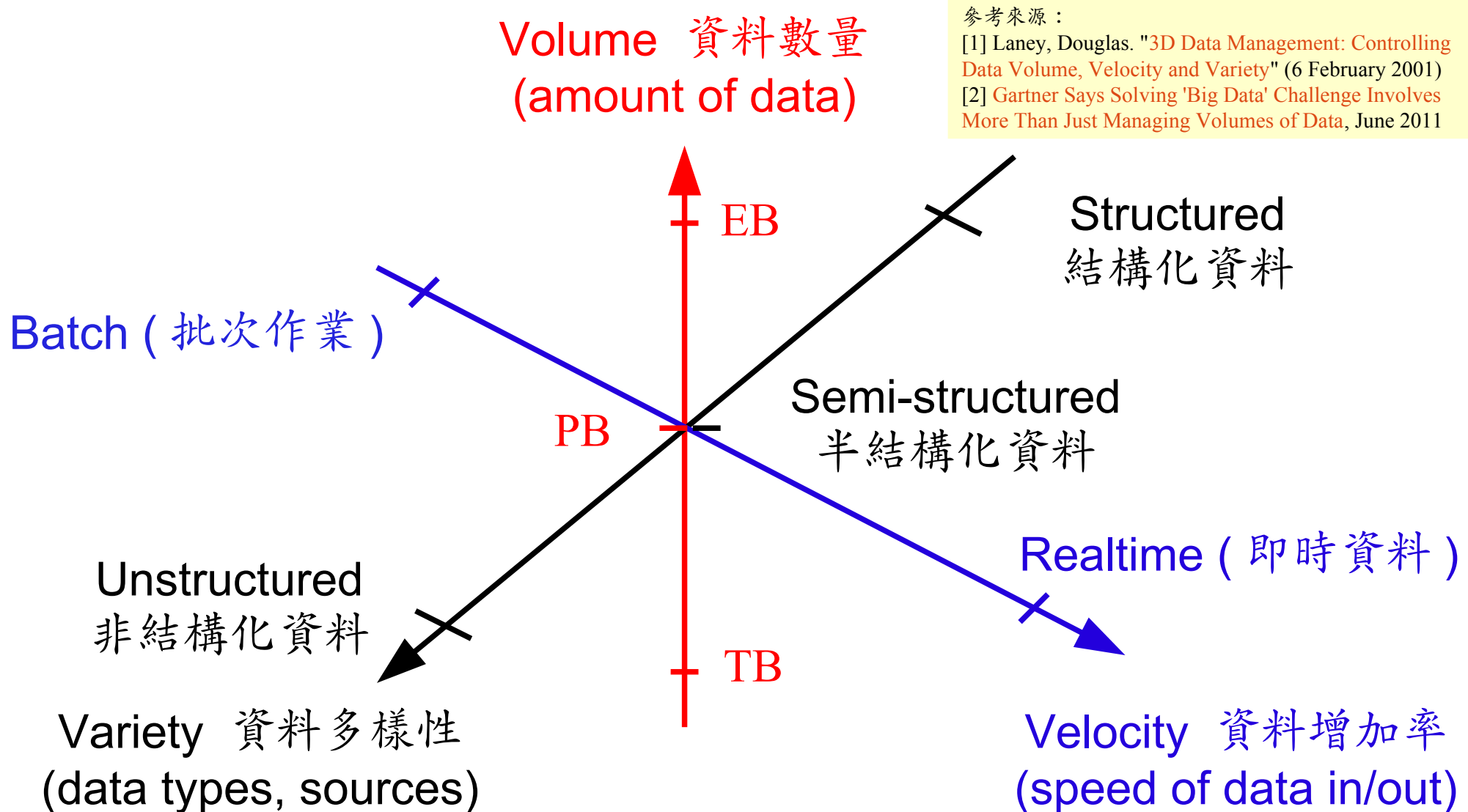
一個資料庫，容量 100TB



一個檔案，容量 100TB

# Gartner Big Data Model? 海量資料的模型?

海量資料的挑戰在於如何管理「數量」、「增加率」與「多樣性」



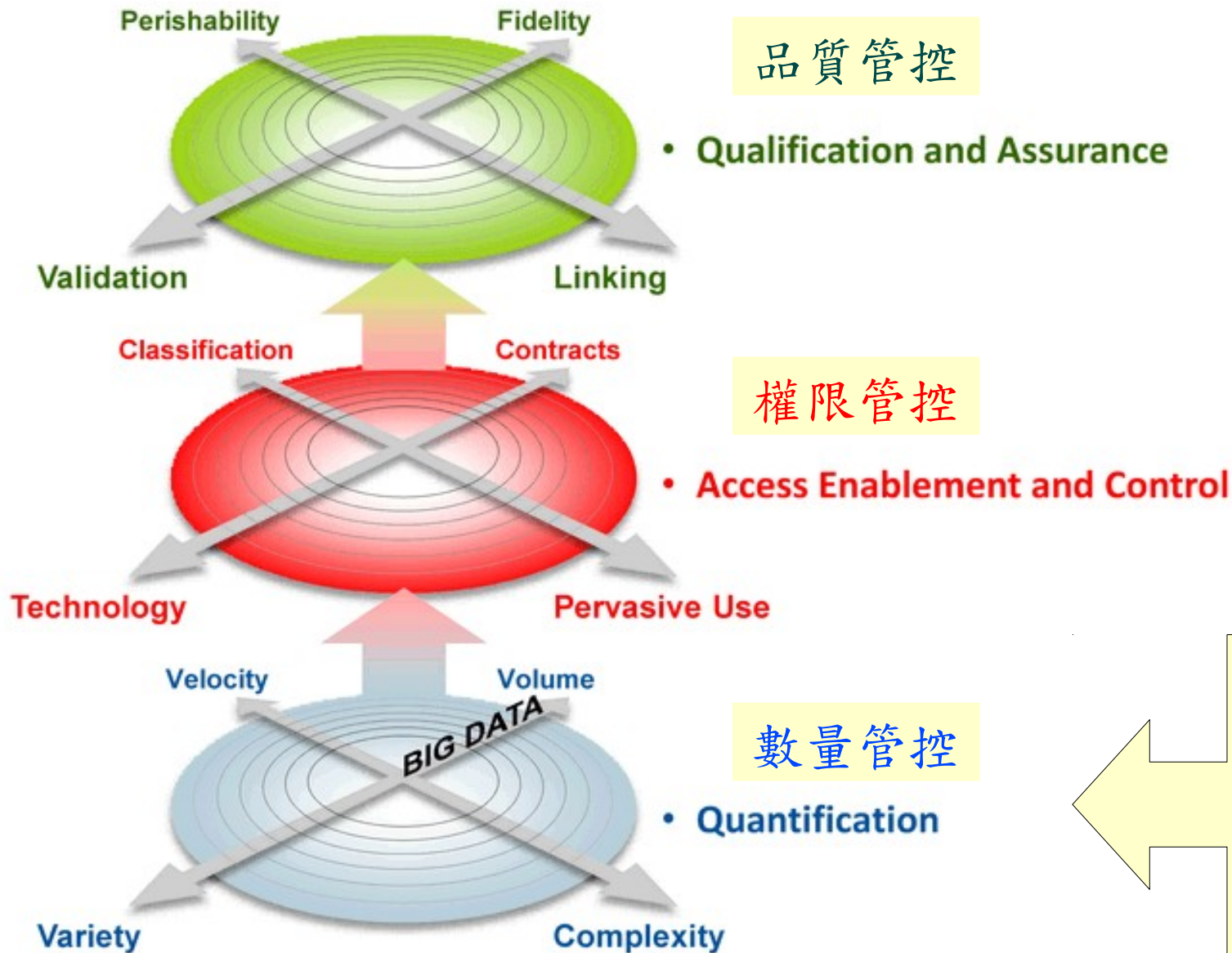
參考來源:

[1] Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety" (6 February 2001)

[2] Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data, June 2011



# 12D of Information Management? 12 個維度?



Big Data  
只是終極  
資訊管理  
的開端！

Source: Gartner (March 2011), 'Big Data' Is Only the Beginning of Extreme Information Management, 7 April 2011, <http://www.gartner.com/id=1622715>



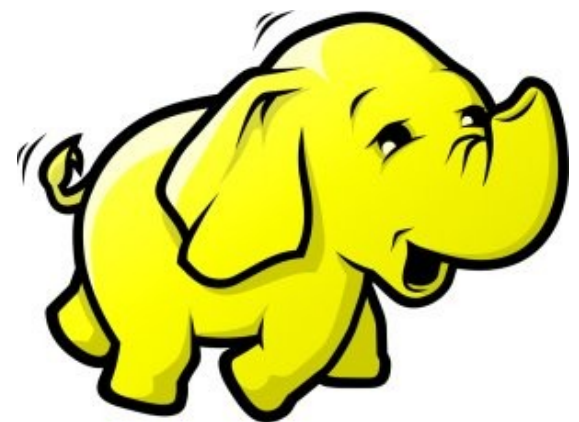
# 處理海量資料的資訊架構與關鍵技術

Technologies to build IT Stack for Big Data

**Jazz Wang**

**Yao-Tsung Wang**

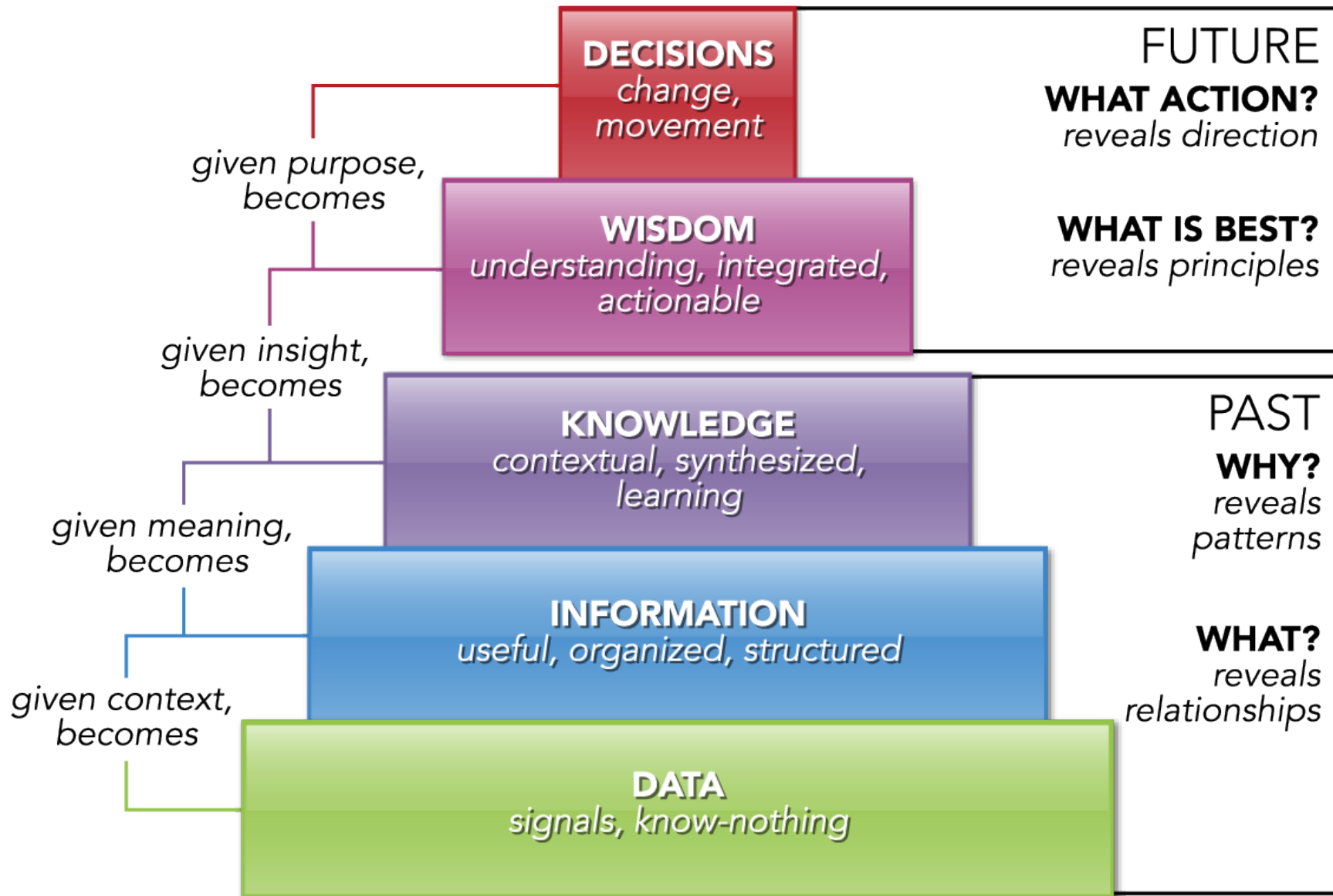
**[jazz@nchc.org.tw](mailto:jazz@nchc.org.tw)**





# Data, Information, Knowledge, Wisdom

## 知識管理模型：資料、資訊、知識與智慧



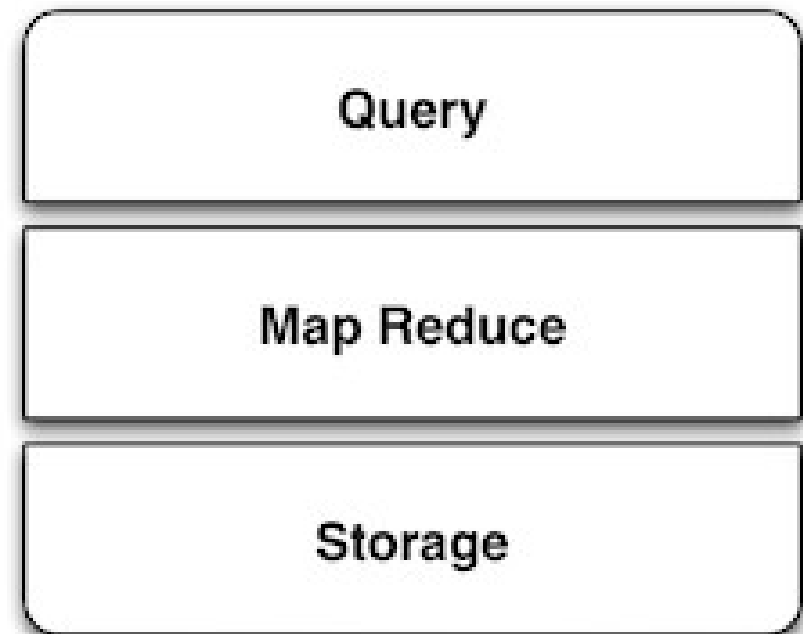
# The SMAQ stack for big data

## 海量資料處理的資訊架構

做網頁相關的人可能聽過 LAMP



未來處理海量資料的人必需知道  
SMAQ ( Storage, MapReduce and Query )



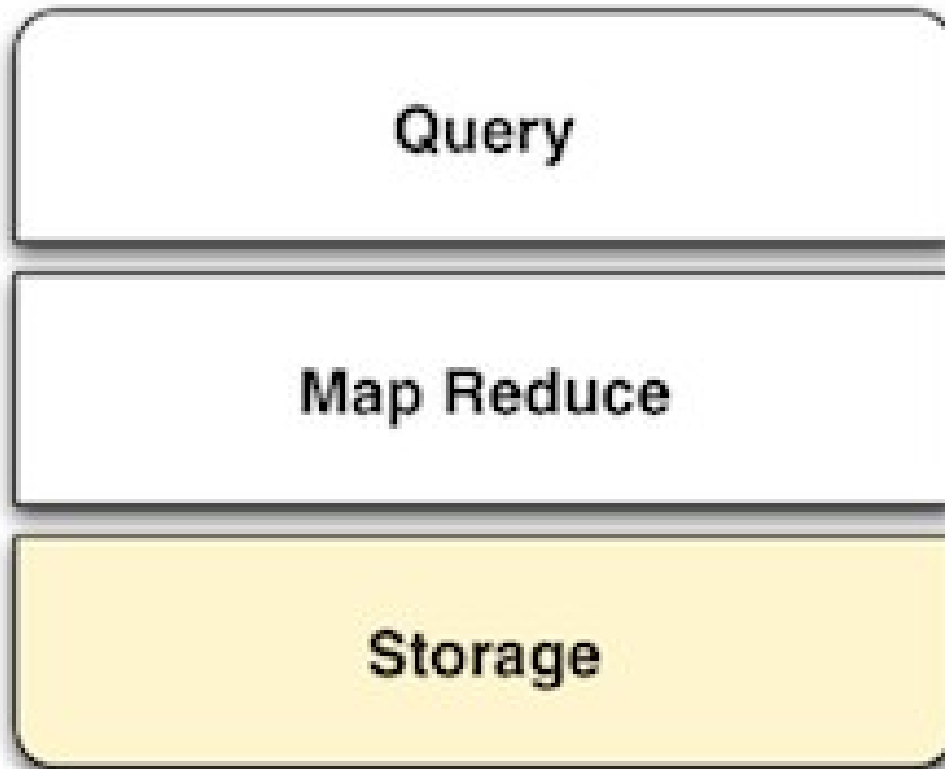
參考來源：The SMAQ stack for big data，Edd Dumbill，22 September 2010，

<http://radar.oreilly.com/2010/09/the-smaq-stack-for-big-data.html>

圖片來源：<http://smashingweb.ge6.org/wp-content/uploads/2011/10/apache-php-mysql-ubuntu.png> 10

# The SMAQ stack for big data

## 海量資料處理的資訊架構



用來儲存分散、沒有關聯  
的非結構化資料

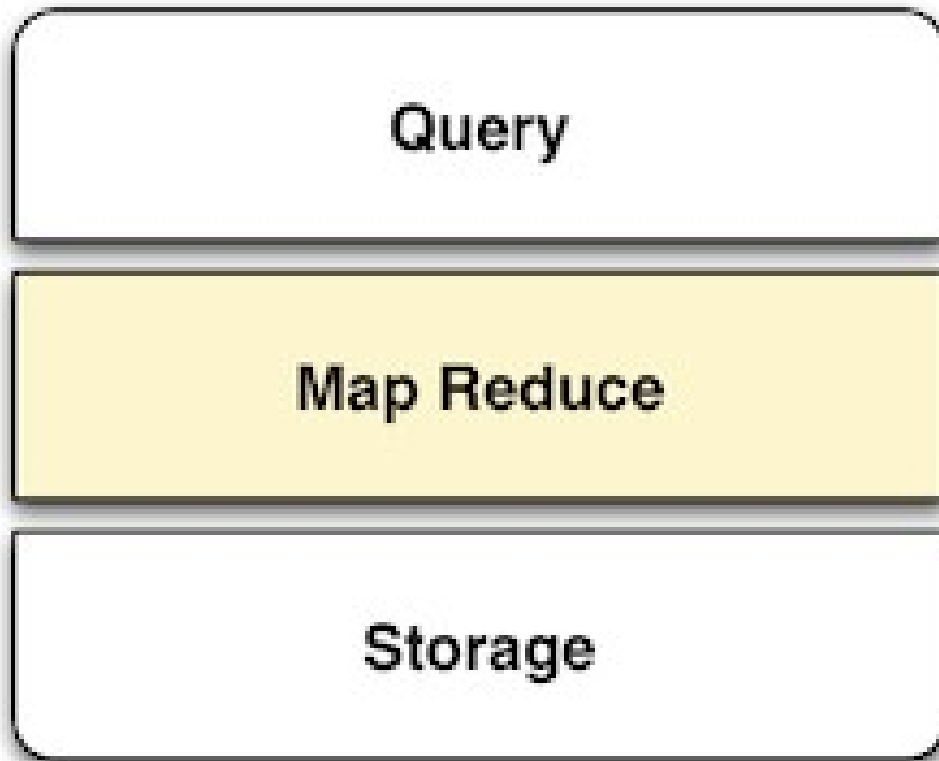
### Key features

- Distributed
- Non-relational or unstructured

# The SMAQ stack for big data

## 海量資料處理的資訊架構

運用批次處理的方式，將  
運算工作平均分散到許多  
的伺服器做運算。

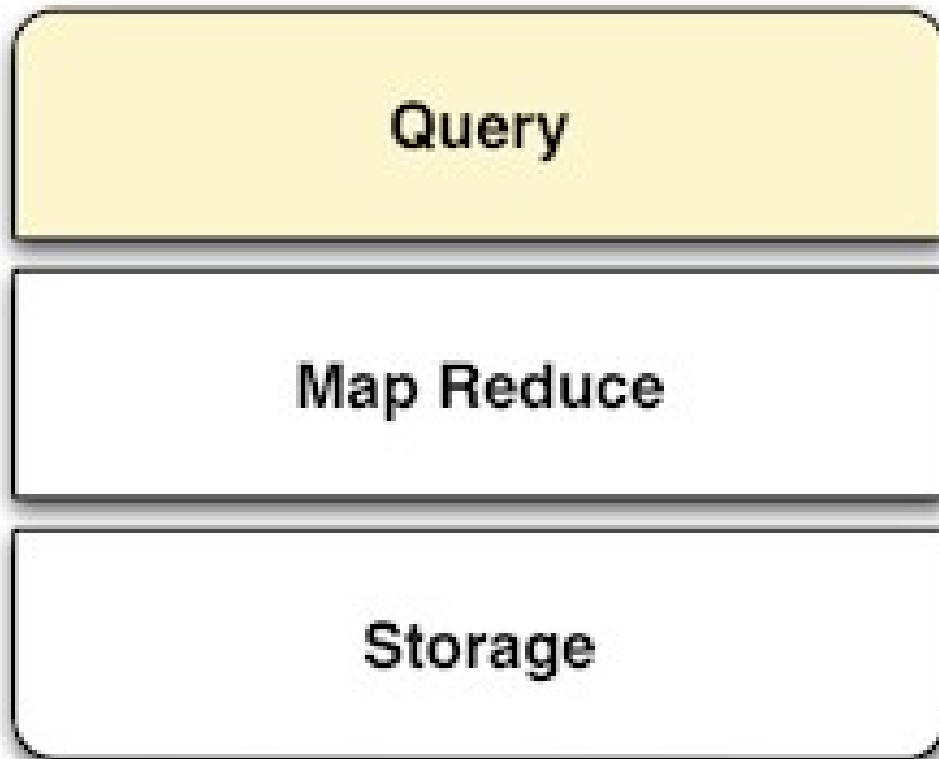


### Key features

- Distributes computation over many servers
- Batch processing model

# The SMAQ stack for big data

## 海量資料處理的資訊架構



### Key features

- Efficient way of defining computation
- Platform for user friendly analytical systems

將算完的結構化資料儲存到可供查詢的資料庫系統

# 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

Google 三大關鍵技術

自由軟體對應解決方案

**Q = Query**  
**BigTable**

A huge key-value datastore

**HBase**, Hypertable  
Cassandra, ....

**MapReduce**

To parallel process data

**Hadoop MapReduce API**  
**Sphere MapReduce API**, ...

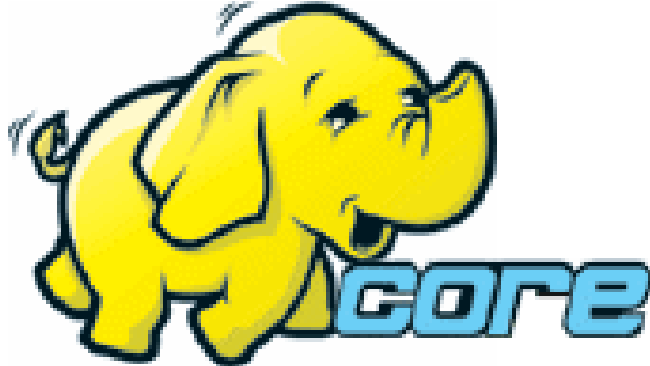
**S = Storage**

**Google File System**

To store petabytes of data

**Hadoop Distributed File System (HDFS)**  
**Sector Distributed File System**

# 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 are also  
powered by Hadoop**

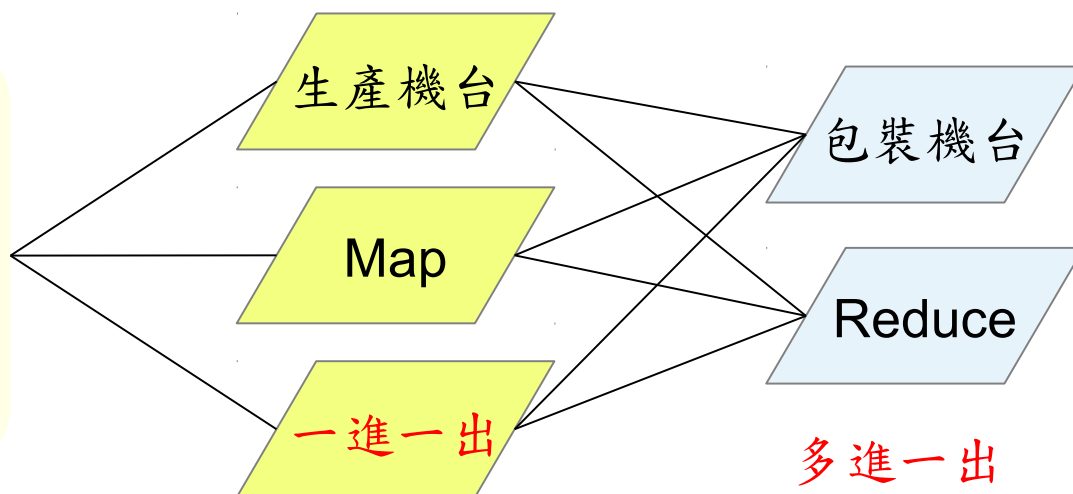
# Hadoop 簡介

**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。

就像工廠的倉庫  
存放生產原料跟待售貨物

HDFS 存放  
待處理的**非結構化資料**  
與處理後的**結構化資料**



# 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Ma1Stone效能評比軟體
- **Open Cloud Consortium provide Open Cloud Testbed and develop Ma1Stone toolkit for benchmark**

**Sector-Sphere**

National Center for Data Mining  
University of Illinois at Chicago

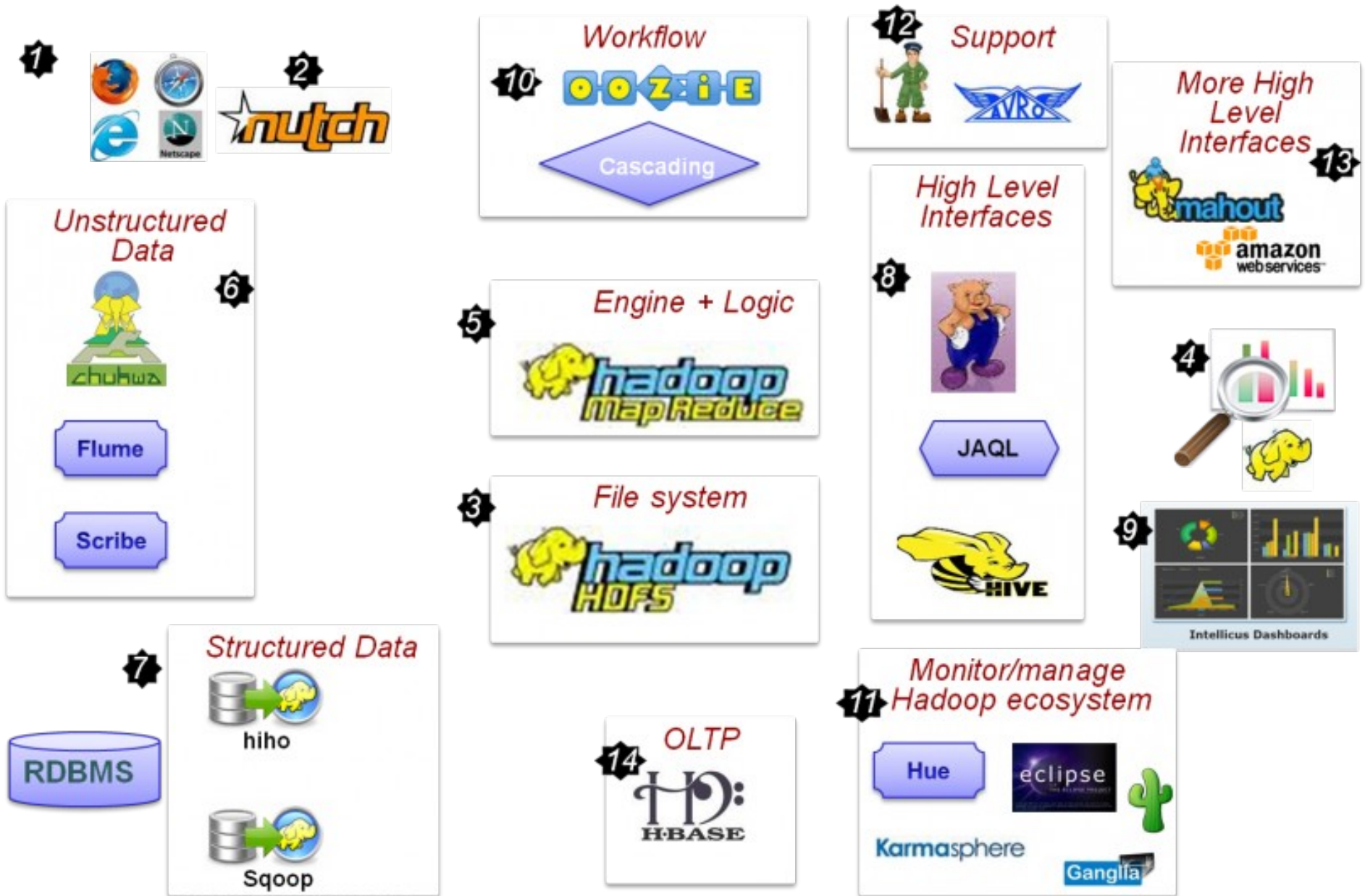


Open Data Group

<http://www.opendatagroup.com/>

# Why we choice Hadoop? Good Ecosystem!

豐富的生態系建構出處理海量資料的工具庫





# Microsoft love Hadoop, too

## 微軟幫 Azure 還有 SQL Server 都接上 Hadoop



The screenshot shows the Microsoft SQL Server website. At the top, there's a navigation bar with "SQL Server" and "All Microsoft Sites" on the left, "United States" and "Change" in the middle, and a search bar with "Search Microsoft" and "bing" logo on the right. Below the navigation bar is the Microsoft SQL Server logo. To the right of the logo are "Contact Us" and social media icons for Facebook, Twitter, and YouTube. Below the logo are navigation links: "About SQL Server", "Solutions & Technologies", "Editions", "Get SQL Server", "Learning Center", and "Partners". The main content area is titled "Business Intelligence" and has a "Share this page" button. A red banner highlights "Big Data Analytics". Below this is a video player for "Strata Big Data Conference 2012 and Power View Contest" with a play button and a progress bar. To the right of the video is the "Big Data Solution" section, which includes a paragraph of text and a "Key Benefits" list.

SQL Server | All Microsoft Sites | United States | Change | Search Microsoft | bing | Web

Microsoft SQL Server

Contact Us > | Facebook | Twitter | YouTube

About SQL Server | Solutions & Technologies | Editions | Get SQL Server | Learning Center | Partners

Business Intelligence | Share this page

Big Data Analytics

Strata Big Data Conference 2012 and Power View Contest

Strata Big Data Conference 2012...

YouTube

### Big Data Solution

Unlock business insights from all your structured and unstructured data, including large volumes of data not previously activated, with Microsoft's Big Data solution. Microsoft's end-to-end roadmap for Big Data embraces Apache Hadoop™ by distributing enterprise class Hadoop based solutions on both Windows Server and Windows Azure. Our solution is also integrated into the Microsoft BI tools such as SQL Server Analysis Services, Reporting Services and even PowerPivot and Excel. This enables you to do BI on all your data, including those in Hadoop.

#### Key Benefits

- Broader access of Hadoop to end users, IT professionals and Developers, through easy installation and configuration and simplified programming with JavaScript.
- Enterprise ready Hadoop distribution with greater security, performance, ease of management and options for Hybrid IT usage.

參考來源：Big Data Solution | Microsoft SQL Server 2008 R2

<http://www.microsoft.com/sqlserver/en/us/solutions-technologies/business-intelligence/big-data-solution.aspx>



# Oracle love Hadoop, too

## Oracle 也接上 Hadoop



CNET > News > Software, Interrupted

## Cloudera teams up to connect Oracle and Hadoop

Cloudera and Quest software are partnering to provide connectivity between Oracle and Hadoop.



by [Dave Rosenberg](#) | June 21, 2010 5:30 AM PDT

[Follow](#)

This week [Cloudera](#), a provider of software and services for the Apache Hadoop project, is set to announce a partnership with [Quest Software](#) to develop, support, and distribute an Oracle connector for Hadoop.



# Hinet Application of Big Data

## 中華電信已經在做的海量資料應用

Business Next 數位時代

### 中華電信：分析駭客行為，拓展對外新服務

撰文者：趙郁竹

發表日期：2012-03-06



[214期雜誌精選]

全球最大的中華電信提供行動電話、市話、寬頻固網、MOD……，各種業務服務，加起來的用戶數就有3000萬，比全台灣人口還多，光是單月帳務數量就高達100億筆資料。除了電信、寬頻服務，還有日益增加的數位服務、行動增值服務，從服務內容到客戶端，累積出的資料相當驚人。

「資料量越來越大，日常分析工作需要很多時間，但新的運算技術有效解決了這個問題，」中華電信資訊處處長陳明仕說。2010年開始，因為中華電信本身的資料運算需求，採用分散式運算架構Hadoop技術，打造出大資料運算平台，不但解決了自身的資料問題，還能對外提供資料運算應用。

以MOD為例，一天有幾千萬筆資料，如何找出使用者在什麼時段做了什麼事？廣告效益又如何？「用傳統的方法，需要400分鐘才能分析完；用Hadoop大資料平台，13分鐘就能解決，節省非常多時間，」他說。

#### 追蹤再拆解

大資料運算技術除了節省時間，還能防止駭客入侵。「駭客的攻擊行為都有模式可循，」陳明仕解釋，就像球賽一樣，了解進攻模式就能防守。用戶的資料保護是第一要務，因此透過行為模式分析，能有效保護企業資訊安全，也保障客戶的個資安全。

參考來源：中華電信：分析駭客行為，拓展對外新服務，發表日期：2012-03-06

<http://www.bnext.com.tw/print/article/id/22333>



# Hinet Application of Big Data

## 中華電信已經在做的海量資料應用

IT ithome.com.tw

### 中華電信用Hadoop技術分析通話明細

READ LATER

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。

中華電信用來分析的資料模型最早於10多年前已有雛形，但當初主要用於行動語音分析。一直到2009年，他們完整導入Teradata的電信業邏輯資料模型cLDM 9.0版，整合更多電信服務的用戶資料。楊秀一表示，當初導入該模型的目的主要是為了整合行動語音、固網、數據的資料，進行以人為中心的分析模式。在導入之前，中華電信的資料模型是以設備為中心，因為不同設備的記錄資料儲存在不同的資料庫，無法進行整合性的分析。

參考來源：中華電信用 Hadoop 技術分析通話明細，發表日期：2011-06-12  
<http://www.ithome.com.tw/itadm/article.php?c=68023>

# History of Hadoop ... 2001~2005

## Hadoop 這套軟體的歷史源起 ... 2001~2005



- 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 ... 2005~2006

## Hadoop 這套軟體的歷史源起 ... 2005~2006

- 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

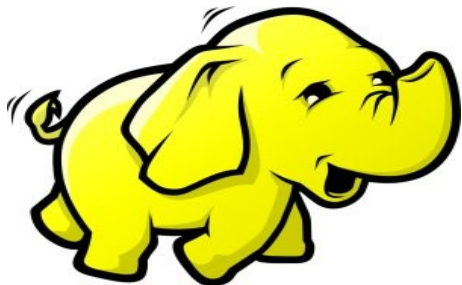
- It builds on **Lucene and Solr**, adding web-specifics, such as a **crawler**, a **link-graph database**, parsers for HTML and other document formats, etc.



# History of Hadoop ... 2006 ~ Now

## *Hadoop* 這套軟體的歷史源起 ... 2006 ~ 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**





# 運用抓抓龍製作個人化書籤搜尋引擎

Build Your Personal Bookmark Search Engine using Crawlzilla

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



Powered by DRBL

# Search is everywhere in our daily life !!

## 「搜尋」已經成爲我們生活中的一部分

搜尋結果

### 檔案搜尋

網址(D) 搜尋結果


搜尋小幫手

您想要搜尋什麼?

- 圖片、音樂，或視訊(P)
- 文件(文字處理、試算表，等等)(O)
- 所有檔案和資料夾(L)
- 電腦或人員(C)
- 說明和支援中心裡的資訊(I)

您也可能想要...

- 搜尋網際網路(S)
- 變更喜好(G)



0 個物件

Gmail Calendar Documents Photos Sites Web More -

Search

All Mail

From

To

Subject

Has the words

Doesn't have

Has attachment

Date within 1 day of

Examples: f

Search

### 信件搜尋

發 的交談

larwin.nchc.org.tw 於 2011年12月02日 (週五) 10時53分46秒 的交談

10時53分48秒) Shunfa 楊順發

10時53分51秒) Jazz Yao-Tsung

10時54分08秒) Shunfa 楊順發

10時54分42秒) Jazz Yao-Tsung

10時54分49秒) Jazz Yao-Tsung

10時54分51秒) Jazz Yao-Tsung

10時55分02秒) Shunfa 楊順發

10時55分04秒) Shunfa 楊順發

10時55分39秒) Jazz Yao-Tsung

尋找(F)

關閉(C)

### 即時通訊搜尋

IEEE Xplore DIGITAL LIBRARY

IEEE

BROWSE

- Journals & Magazines
- Conference Proceedings
- Standards
- Books
- Educational Courses

Search 3,076,887 documents

SEARCH

Advanced Search | Preferences | Search Tips

### 資料庫搜尋

## 今天要談的是「網頁搜尋」

設Yahoo!奇摩為首頁 資訊展PK線上搶先

# YAHOO! 奇摩

網頁 | 知識+ | 圖片 | 影片 | 部落格 | 字典 | 新聞 | 購物 BETA

網頁搜尋

熱門: 第一美腿 12歲父親 嫩模女神 幼稚病 51區 花心星座 解夢 知識: 傷口癢竟是 電鍋料理

2011 資訊月 ONLINE 3G特展搶先看!!

# To speed up search, We need “Index”

爲了加速搜尋的效率，我們需要「索引」

## Index

出現頁碼

關鍵字

### Symbols

! (exclamation mark) command prefix, 368

### A

ack queue in HDFS, 66

ACLs (access control lists)  
for Hadoop services, 283  
ZooKeeper, 446, 456

ActiveKeyValueStore class (example), 464

ad hoc analysis and product feedback  
(hypothetical use case), 511

adjacency list, 560

installation, 565

prerequisites, 565

TeraByte sort on, 553

Apache Hadoop project, 10, 12

Apache Lucene project, 9

Apache Nutch, 9

Apache Thrift services, 49

Apache ZooKeeper, 442

(see also ZooKeeper)

APIs in ZooKeeper, 453

archive files, copying to tasks, 253

archive tool, 72

archives, 72

# Do you like to write notes?

## 你有寫筆記的習慣嘛？



想知道更多 | 升級至專業版 | 檔案下載 | 部落格 | 百寶箱

• 登入  
• 立即註冊

### 什麼都記得住



#### 擷取每件事。

儲存點子、喜愛事物以及所見所聞。

#### 隨處存取。

Evernote 幾乎能在所有電腦、手機以及其他行動裝置上使用。

#### 快速尋找事物。

根據標題、標籤或甚至是影像內的印刷與手寫文字進行搜尋。

取得 Evernote

立即下載

免費使用。

## 大腦記憶力有限，只好靠筆記啦！



# Tools that I used to write notes

## 我作筆記的工具 (1) 維基 Oddmuse Wiki

軟體下載：<http://www.oddmuse.org/>

[關於本站](#) [更新紀錄](#) [[年曆\(c\)](#)] [[登入\(l\)](#)] [[上一頁](#)] [[首頁\(h\)](#)] [[12-02\(9\)](#)] [[12-03\(0\)](#)] [[12-04\(1\)](#)] [[12-05\(2\)](#)] [[12-06\(3\)](#)]

December 2011						
Su	Mo	Tu	We	Th	Fr	S
				1	2	
4	5	6	7	8	9	1
11	12	13	14	15	16	1
18	19	20	21	22	23	2
25	26	27	28	29	30	3

## 關於本站

架站日期：2005/01/01

基本架構：<http://www.oddmuse.org/> Oddmuse Wiki

## 更新紀錄

2005-01-01		
<a href="#">chinese-utf8.pl</a>	<b>Chinese Translations</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Chinese">http://www.oddmuse.org/cgi-bin/wiki/Chinese</a>
<a href="#">download.pl</a>	<b>Download Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Download_Extension">http://www.oddmuse.org/cgi-bin/wiki/Download_Extension</a>
<a href="#">headers.pl</a>	<b>Header Markup Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Header_Markup_Extension">http://www.oddmuse.org/cgi-bin/wiki/Header_Markup_Extension</a>
<a href="#">tables-long.pl</a>	<b>Long Table Markup Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Long_Table_Markup_Extension">http://www.oddmuse.org/cgi-bin/wiki/Long_Table_Markup_Extension</a>
<a href="#">tables.pl</a>	<b>Table Markup Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Table_Markup_Extension">http://www.oddmuse.org/cgi-bin/wiki/Table_Markup_Extension</a>
<a href="#">usemod.pl</a>	<b>Usemod Markup Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Usemod_Markup_Extension">http://www.oddmuse.org/cgi-bin/wiki/Usemod_Markup_Extension</a>
2005-02-16		
<a href="#">calendar.pl</a>	<b>Calendar Extension</b>	<a href="http://www.oddmuse.org/cgi-bin/wiki/Calendar_Extension">http://www.oddmuse.org/cgi-bin/wiki/Calendar_Extension</a>

2005~2008

# Tools that I used to write notes

## 我作筆記的工具 (2) 維基 PmWiki

軟體下載：<http://www.pmwiki.org/>

### 導覽列 (編輯)

首頁  
民視專訪  
福山參訪  
超級視訊  
阿聰的首頁  
穎燦的首頁  
測試沙箱  
近期更新

### PmWiki 官方網站

程式安裝  
常見問題  
外掛程式  
布景主題  
臭蟲管理

### PmWiki 使用手冊

頁面編輯入門  
頁面編輯技巧  
建立新的頁面

Main » [Home Page](#)

[本文](#) · [編輯](#) · [附檔](#) · [列印](#) · [歷史](#)

Search

縮小 正常 放大

### 歡迎使用 PmWiki:

以下是一些 [PmWiki](#) 安裝後的預設頁面，你可以從這些頁面開始瞭解 PmWiki:

- [Pm Wiki](#) 的 [相關文件目錄](#)。
- 什麼是 [WikiWikiWeb](#) ?
- [頁面編輯技巧](#) 和 [寫作語法](#) 這兩份完文件描述如何建立一個 Wiki 頁面。
- 你可以用 [測試沙箱](#) 來練習 Wiki 的語法。

請定期拜訪 [PmWiki 官方網站](#) 以便獲得最新的資訊。

« [October 2011 period](#) · [February 2012 period](#) »

### December 2011

Mon	Tue	Wed	Thu	Fri	Sat	Sun
			<a href="#">01</a>	<a href="#">02</a>	<a href="#">03</a>	<a href="#">04</a>
<a href="#">05</a>	<a href="#">06</a>	<a href="#">07</a>	<a href="#">08</a>	<a href="#">09</a>	<a href="#">10</a>	<a href="#">11</a>
<a href="#">12</a>	<a href="#">13</a>	<a href="#">14</a>	<a href="#">15</a>	<a href="#">16</a>	<a href="#">17</a>	<a href="#">18</a>
<a href="#">19</a>	<a href="#">20</a>	<a href="#">21</a>	<a href="#">22</a>	<a href="#">23</a>	<a href="#">24</a>	<a href="#">25</a>
<a href="#">26</a>	<a href="#">27</a>	<a href="#">28</a>	<a href="#">29</a>	<a href="#">30</a>	<a href="#">31</a>	

### January 2012

Mon	Tue	Wed	Thu	Fri	Sat	Sun
						<a href="#">01</a>
<a href="#">02</a>	<a href="#">03</a>	<a href="#">04</a>	<a href="#">05</a>	<a href="#">06</a>	<a href="#">07</a>	<a href="#">08</a>
<a href="#">09</a>	<a href="#">10</a>	<a href="#">11</a>	<a href="#">12</a>	<a href="#">13</a>	<a href="#">14</a>	<a href="#">15</a>
<a href="#">16</a>	<a href="#">17</a>	<a href="#">18</a>	<a href="#">19</a>	<a href="#">20</a>	<a href="#">21</a>	<a href="#">22</a>
<a href="#">23</a>	<a href="#">24</a>	<a href="#">25</a>	<a href="#">26</a>	<a href="#">27</a>	<a href="#">28</a>	<a href="#">29</a>
<a href="#">30</a>	<a href="#">31</a>					

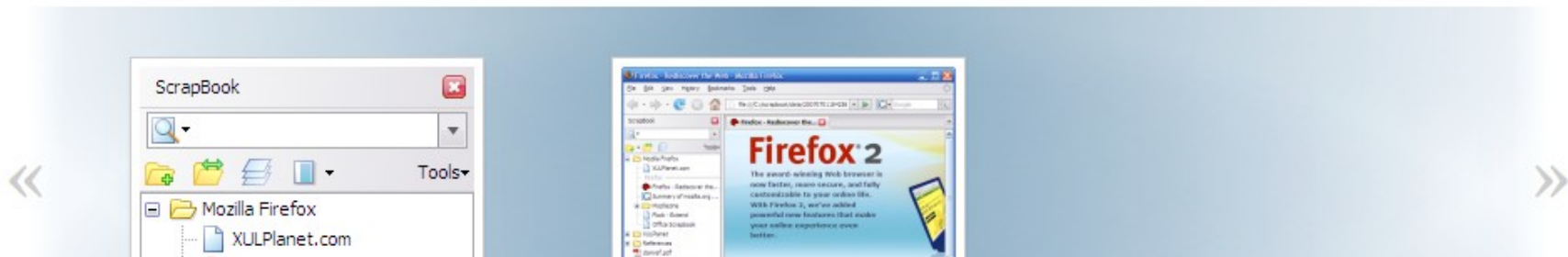
# Tools that I used to write notes

## 我作筆記的工具 (3) 離線網頁 ScrapBook

安裝：<https://addons.mozilla.org/zh-TW/firefox/addon/scrapbook/>



The screenshot shows the Mozilla Add-ons website for the ScrapBook extension. At the top, there's a navigation bar with "註冊 或 登入" and "其他應用程式" next to the Mozilla logo. Below that, the page title "附加元件" (Add-ons) is displayed with a search bar and a "搜尋附加元件" button. The main content area features the ScrapBook 1.4.8 extension card, which includes the author's name "Gomita", a description "協助您擷取網頁，方便整理擷取後的資料。", and a green "新增至 Firefox" button. To the right of the card, there's a star rating of 4.5, "564 位使用者評論", and "448,868 位使用者". Below the card, there's a "捐款" button with a price of "建議 US\$2.99". A yellow box on the right side of the page contains the text "2005~NOW".





# Tools that I used to write notes

## 我作筆記的工具 (4) 維基 + 版本控制 Trac

軟體下載：<http://trac.edgewall.org/>

[Login](#) | [Preferences](#) | [Help/Guide](#) | [About Trac](#)

	<a href="#">Wiki</a>	<a href="#">Timeline</a>	<a href="#">Roadmap</a>	<a href="#">Browse Source</a>	<a href="#">View Tickets</a>	<a href="#">Search</a>
--	----------------------	--------------------------	-------------------------	-------------------------------	------------------------------	------------------------

[Start Page](#) | [Index](#) | [History](#) | [Last Change](#)

~ Welcome to NCHC Grid Architecture Research Group ~

### 【Project News】

- 2011-01-28: [2011 Project Deliverable Results](#) had updated~
- 2010-07-01: [2010 Project Deliverable Results](#) had updated~
- 2009-03-18: [2009 Project Deliverable Results](#) had updated~
- 2008-10-08: [2008 Project Deliverable Results](#) had updated~
- [雲端平台維護日誌](#)
- [Paper Reading Schedule \(讀書會行程\)](#)
- [Submission Deadline List \(投稿截止日期列表\)](#)
- [2008 Related NEWS](#)

### 【Project News】

#### 【1: Distributed / Parallel Computing】

- 1.1: [Kerrighed](#)
- 1.2: [BOINC](#)
- 1.3: [WebOS](#)
- 1.4: [Parallel Computing](#)

#### 【2: File System & Data Grid】

- 2.1 [Distributed & Parallel File System](#)
- 2.2 [FS Realition](#)

#### 【3: Virtualization】

#### 【Projects】

#### 【Related Conference Paper / Poster / Demo Submission】

#### 【Possible Technical Whitepaper Submission】

#### 【Technical Tutorials】

[Visitor](#)

### 【1: Distributed / Parallel Computing】

#### 1.1: [Kerrighed](#)

#### 1.2: [BOINC](#)

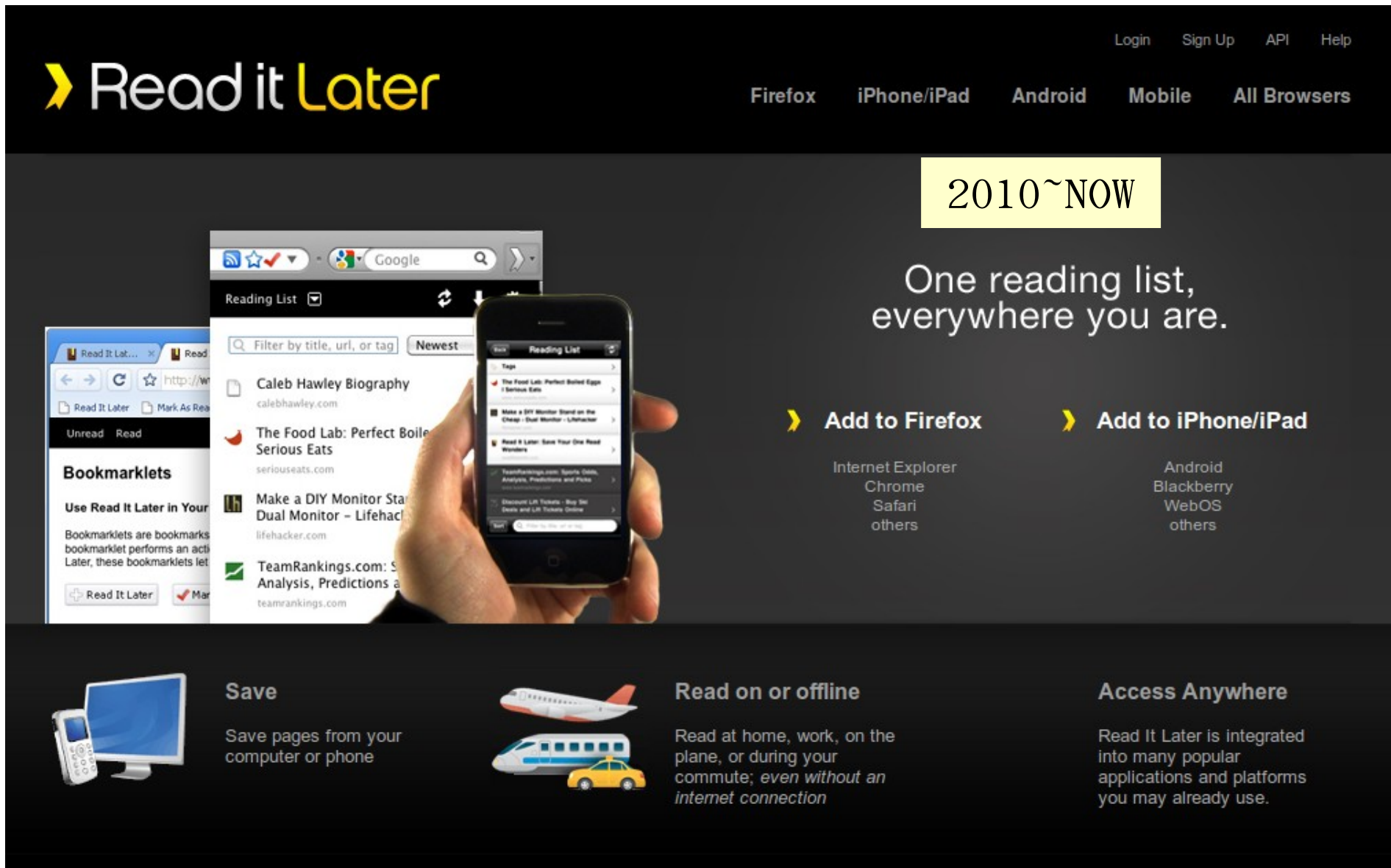
- 2008-05-27: [How to Setup BOINC example project](#)
- 2008-02-13: [How to Build BONIC Server](#)

2006~NOW

# Tools that I used to write notes

## 我作筆記的工具 (5) 線上書籤 **ReadItLater**

安裝：<http://readitlaterlist.com/>



The image shows a screenshot of the Read It Later website. At the top, there is a navigation bar with links for 'Login', 'Sign Up', 'API', and 'Help'. Below this, the 'Read it Later' logo is displayed on the left, and a list of supported browsers and devices is on the right: 'Firefox', 'iPhone/iPad', 'Android', 'Mobile', and 'All Browsers'. A yellow box highlights the text '2010~NOW'. The main content area features the slogan 'One reading list, everywhere you are.' and two call-to-action buttons: 'Add to Firefox' and 'Add to iPhone/iPad'. Below these buttons, there are lists of supported browsers (Internet Explorer, Chrome, Safari, others) and devices (Android, BlackBerry, WebOS, others). A central image shows a hand holding a smartphone displaying the Read It Later app interface, with a desktop browser window and a tablet interface overlaid behind it. The bottom section of the page is divided into three columns: 'Save' (with an icon of a computer and phone), 'Read on or offline' (with icons of an airplane, train, and car), and 'Access Anywhere' (with an icon of a globe).

2010~NOW

One reading list, everywhere you are.

**Add to Firefox**

Internet Explorer  
Chrome  
Safari  
others

**Add to iPhone/iPad**

Android  
Blackberry  
WebOS  
others

**Save**

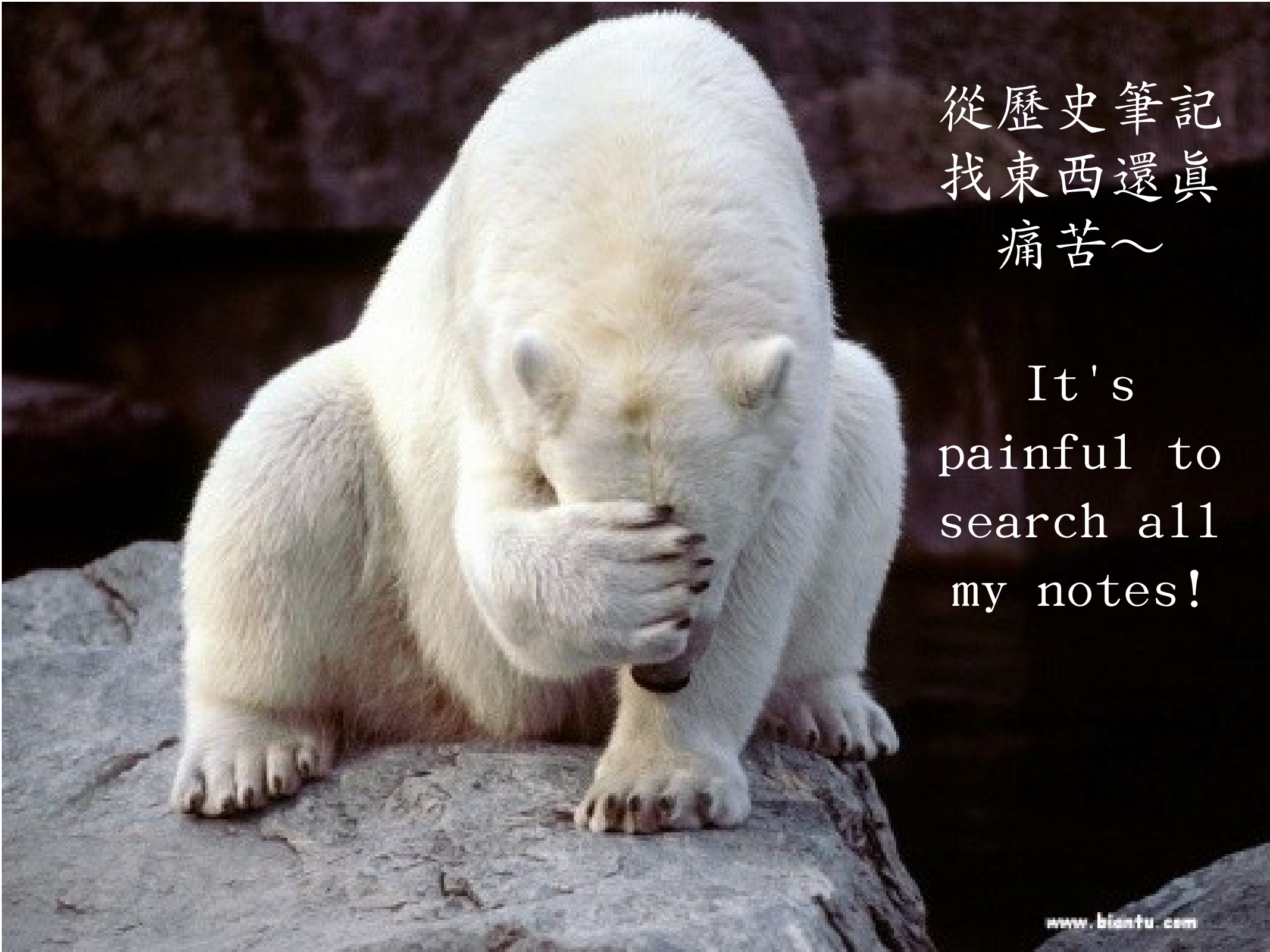
Save pages from your computer or phone

**Read on or offline**

Read at home, work, on the plane, or during your commute; *even without an internet connection*

**Access Anywhere**

Read It Later is integrated into many popular applications and platforms you may already use.



從歷史筆記  
找東西還真  
痛苦～

It's  
painful to  
search all  
my notes!



既然我有許多筆記放在網頁上，  
何不試試自家研發的抓抓龍呢？

# Crawlzilla 系統功能

## Feature of Crawlzilla

- 支援**叢集運算**及顧全  
安全性
- 支援**中文分詞**功能
- 支援多工網頁爬取
- 支援**多重搜尋引擎**
- **即時瀏覽資料庫資訊**
- 解決中文亂碼及中文  
支援
- 支援多國語言
- 網頁管理

如果您還不認識抓抓龍，

不妨看一下 2010 年 Hadoop 使用者會議的錄影

[http://cloud.nchc.org.tw/20101202/slides/01\\_Crawlzilla.wmv](http://cloud.nchc.org.tw/20101202/slides/01_Crawlzilla.wmv)



# System Architecture of Crawlzilla

## 抓抓龍系統架構

Web UI ( Crawlzilla Website + Search Engine)

JSP + Servlet +  
JavaBean

Nutch

Lucene

Crawlzilla System Management

Tomcat

Hadoop

PC1

PC2

PC3

# Comparison with other projects

## 抓抓龍與其他搜尋引擎專案的比較

	<b>Spidr</b>	<b>Larbin</b>	<b>Jcraw</b>	<b>Nutch</b>	<b>Crawlzilla</b>
Install	Rube Package Install	Gmake Compiler and Install	Java Compiler and Install	Deploy Configure Files	<b>Provide Auto Installation</b>
Crawl website pages	O	O	O	O	<b>O</b>
Parser Content	X	X	X	O	<b>O</b>
Cluster Computing	X	X	X	O	<b>O</b>
Interface	Command	Command	Command	Command	<b>Web-UI</b>
Support Chinese Segmentation	X	X	X	X	<b>O</b>

# New Feature of Crawlzilla 1.0

## 抓抓龍 1.0 的新功能

- 支援多重使用者
- 採用 jQuery Ui 打造的新網頁管理介面
- 支援重新爬取 ( Re-Crawl )
- 支援排程爬取 ( Schedule / Crontab )
- 支援雲端服務：
  - 懶得自己建？沒關係！這裡可以試用！
  - <http://demo.crawlzilla.info>

# Multi-user Web Search Cloud Service : Crawlzilla 1.0

## Crawlzilla 1.0 多人版雲端服務 (1)

首先連線到 <http://demo.crawlzilla.info>

(1)

Home	Crawl	索引庫管理	系統排程	Slave安裝	系統設定	登入/註冊
------	-------	-------	------	---------	------	-------

使用者登入

使用者註冊

使用者帳號	<input type="text" value="demo"/>
使用者密碼	<input type="password" value="....."/>
密碼確認	<input type="password" value="....."/>
電子信箱	<input type="text" value="jazz@nchc.org.tw"/>

Submit Reset

運算節點

工作排

空間管

(2)

(3)

▲ Step 1 : 新使用者註冊頁面

# Multi-user Web Search Cloud Service : Crawlzilla 1.0

## Crawlzilla 1.0 多人版雲端服務 (2)

接著等待管理者幫您開啓帳號！

尚未啓用會員列表

使用者	e-mail	申請時間	確認使用者
demo	jazz@nchc.org.tw	2011-12-04 21:18:37	Accept User

系統概況

項目	
會員人數	
搜尋引擎總數	



**Your account has been accepted from Crawlzilla System**

crawlzilla@gmail.com <crawlzilla@gmail.com>  
To: jazz@nchc.org.tw

We are pleased to inform you that your account has been accepted.

Please visit <http://140.110.134.197:8080> to build your search engine!

Thank you for use crawlzilla.

-This mail sent by the system automatically, do not reply to this mail.-

▲ Step 2 : 等待管理者啓用，您會收到啓用通知

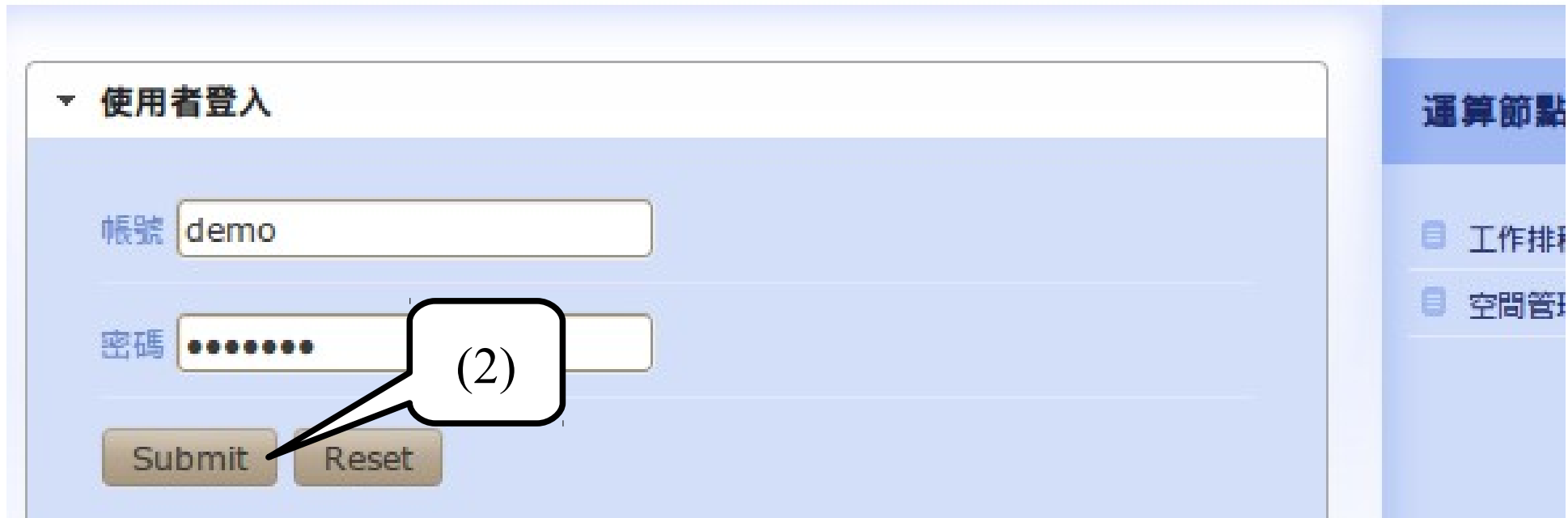


# Multi-user Web Search Cloud Service : Crawlzilla 1.0

## Crawlzilla 1.0 多人版雲端服務 (3)

重新連線到 <http://demo.crawlzilla.info>

(1)



▲ Step 3 : 登入您的個人化管理頁面

# Multi-user Web Search Cloud Service : Crawlzilla 1.0

## Crawlzilla 1.0 多人版雲端服務 (4)

建立新的搜尋索引庫

(1)

Home

Crawl

索引庫管理

系統排程

### Crawl-建立搜尋引擎

#### ▼ Crawl Setup

索引庫名稱:

(2)

(5)

#### ▼ Crawl Setup

索引庫名稱:

輸入欲爬取的網址(可多行):

```
http://cloud.nchc.org.tw/~jazz/rii_export.html
http://cloud.nchc.org.tw/~jazz/firefox_bookmarks.html
http://trac.nchc.org.tw/grid/wiki/jazz/Work_2011
http://trac.nchc.org.tw/grid/wiki/jazz/Work_2010
http://trac.nchc.org.tw/grid/wiki/jazz/Work_2009
http://trac.nchc.org.tw/qrid/wiki/jazz/Work_2008
```

(3)

爬取深度設定:

(4)

註: 因系統負載有限, 此一體驗網站僅能建立3個索引庫, 不便之處請見諒

▶ 排程設定 (Option)

Submit

Reset

▲ Step 4 : 輸入索引庫名稱、起始網址與搜尋深度

## Crawlzilla 1.0 多人版雲端服務 (5)

接著只能靜候抓抓龍幫您建立搜尋索引庫

HRER to check the status of your jobs!' and a note: 'The crawling time depends on your system performance, URLs number, and depth.' On the right side, there is a sidebar with a menu containing '搜尋引擎列表', '運算節點即時狀態', '工作排程器狀態', and '空間管理員狀態'. A speech bubble with the number '(1)' points to the '運算節點即時狀態' menu item."/>

Crawlzilla Search Engine Hands-on Display

Home Crawl 索引庫管理 系統排程 Slave安裝 系統設定 系統登出

**Setup has been submit !!! But, it need time to crawl !!!**

ex. 4URLs with 1 depth -> 10~20 minute  
4URLs with 2 depth -> 40~80 minute  
100URLs with 10 depth -> very very long

click [HRER](#) to check the status of your jobs!

**The crawling time depends on your system performance, URLs number, and depth.**

搜尋引擎列表

運算節點即時狀態 (1)

工作排程器狀態

空間管理員狀態

▲ Step 5 : 等待抓抓龍幫您建立專屬的搜尋索引庫

## Crawlzilla 1.0 多人版雲端服務 (6)

### 索引庫管理

Home Crawl **索引庫管理** 系統排程 Slave安裝 系統設定

(1)

### 索引庫管理

系統爬取狀態

索引庫名稱	爬取狀態	爬取時間	刪除狀態
all	crawling	0h:3m:31s	

(2)

- ▲ 您可以在索引庫管理看到目前爬取已使用的時間

## Crawlzilla 1.0 多人版雲端服務 (7)

搜尋索引庫建立完成後，  
可以於「索引庫管理」處進行**手動**重新爬取（re-crawl）  
或刪除索引庫的動作

索引庫管理

系統爬取狀態

索引庫名稱	爬取狀態	爬取時間	刪除狀態
-------	------	------	------

索引庫列表

索引庫名稱	建立時間	爬取時間	爬取深度	索引庫操作	執行
ril	2011-12-04 16:15:24	0:18:13	2	Select Select Re-Crawl Delete IDB	Run

告訴您  
該索引庫  
花了多久  
時間爬取

(2)

(3)

▲ 您可以在索引庫管理進行手動重新爬取或刪除索引庫



## Crawlzilla 1.0 多人版雲端服務 (8)

可以於「系統排程」處進行排程重新爬取 ( schedule )

The screenshot shows the '系統排程' (System Scheduling) page in the Crawlzilla 1.0 interface. The page is divided into two main sections: '索引庫爬取排程' (Index Library Crawl Scheduling) and '新增排程' (Add Scheduling). Callouts (1) through (4) highlight specific features:

- (1) Points to the '系統排程' (System Scheduling) menu item in the top navigation bar.
- (2) Points to the '索引庫名稱' (Index Library Name) column header in the '索引庫爬取排程' table.
- (3) Points to the '週期' (Frequency) dropdown menu in the '新增排程' form.
- (4) Points to the 'Submit' button at the bottom of the '新增排程' form.

The '索引庫爬取排程' section contains a table with the following columns: '索引庫名稱', '爬取排程時間', '週期', and '刪除排程'. The '新增排程' section includes a '索引庫選擇' dropdown menu (with 'select' and 'ril' options), a '排程日期' input field, a '排程時間' dropdown menu (with '00' and '00' options), and a '週期' dropdown menu (with '執行一次' and '說明' options). The 'Submit' and 'Reset' buttons are located at the bottom of the form.

▲ 您可以在索引庫管理看到目前爬取已使用的時間

# Multi-user Web Search Cloud Service : Crawlzilla 1.0

## Crawlzilla 1.0 多人版雲端服務 (9)

可以於「索引庫管理」處進行即時讀取索引庫資訊

索引庫列表

索引庫名稱	建立時間	爬取時間	爬取深度	索引庫操作	執行
ril	2011-12-04 16:15:24	0:18:13	2	Select ▼	Run

(1)

(2)

### 關於搜尋引擎ril

基本資訊

- 索引庫名稱 ril
- 搜尋引擎連結位置 /home/crawler/crawlzilla/user/jazzwang/IDB/ril/index
- 搜尋引擎狀態 OK
- 爬取深度 2
- 建立時間 20111204-15:57:16
- 執行時間 0:18:13
- 起始連結 [http://cloud.nchc.org.tw/~jazz/ril\\_export.html](http://cloud.nchc.org.tw/~jazz/ril_export.html)

▶ 立即搜尋及網頁嵌入語法

▲ 您可以在索引庫管理取得即時搜尋索引庫的資訊

# Multi-user Web Search Cloud Service : Crawlzilla 1.0

## Crawlzilla 1.0 多人版雲端服務 (10)

在搜尋索引庫資訊中，可以取得加入個人化搜尋引擎的語法



The screenshot displays a web page titled "關於搜尋引擎ril". It features a navigation menu with "基本資訊" and "立即搜尋及網頁嵌入語法". The "立即搜尋及網頁嵌入語法" section is expanded, showing a search engine logo for "CRAWLZILLA" and a search input field. Below this, the "語法" section displays the following HTML code:

```
  
  
<form name="search" action="http://demo.crawlzilla.info/jazzwang_ril/search.jsp" method="get">  
  
<input name="query" size=15></form>
```

Callout (1) points to the search engine logo and input field. Callout (2) points to the HTML code block.

▲ 您可以在索引庫管理取得嵌入網頁的語法

# Multi-user Web Search Cloud Service : Crawlzilla 1.0

## Crawlzilla 1.0 多人版雲端服務 (11)

索引庫內容說明了共搜尋了多少個文件（網頁），  
並且會統計最常到訪的網址排行榜

索引庫內容 ril

資料總覽

總共文字數 213155

文件檔數量 3234

索引庫更新日期 Sun Dec 04 16:15:23 CST 2011

被搜尋分析到的網址

分析的文件型態

出現次數前五十分的字符

(1)

索引庫內容 ril

資料總覽

被搜尋分析到的網址

Order	Contents	Counts	Order	Contents
0	site:www.digitimes.com.tw	204	1	site:www.bnext.c
2	site:groups.google.com	74	3	site:highscalabilit
4	site:www.theregister.co.uk	49	5	site:www.ithome
6	site:www.cloudera.com	47	7	site:gigaom.com
8	site:en.wikipedia.org	39	9	site:www.networ
10	site:wiki.apache.org	37	11	site:www.zdnet.c
12	site:www.howtoforge.com	33	13	site:www.freegro
14	site:ajaxian.com	29	15	site:news.networ
16	site:www.linuxfordevices.com	25	17	site:ieeexplore.ie
18	site:www.linux-mag.com	24	19	site:insidehpc.cor
20	site:www.readwriteweb.com	22	21	site:only-percept
22	site:www.nosqldatabases.com	19	23	site:www.openfo
24	site:www.inside.com.tw	17	25	site:www.sys-cor

(2)

▲ 索引庫內容提供了許多統計資訊

# Multi-user Web Search Cloud Service : Crawlzilla 1.0

## Crawlzilla 1.0 多人版雲端服務 (12)

此外，索引庫內容也說明了共搜尋了哪幾種文件，並且會統計最常出現的關鍵字排行榜

### 索引庫內容 ril

▶ 資料總覽

▶ 被搜尋分析到的網址

▼ 分析的文件型態

Order	Contents	Counts	Order	Contents	Counts
0	type:text	3232	1	type:text/html	3231
2	type:html	3231	3	type:xhtml+xml	2
4	type:application/xhtml+xml	2	5	type:application	2
6	type:text/plain	1	7	type:plain	1

(1)

▶ 出現次數前五十分的字符

### 索引庫內容 ril

▶ 資料總覽

▶ 被搜尋分析到的網址

▶ 分析的文件型態

▼ 出現次數前五十分的字符

Order	Contents	Counts	Order	Contents	Counts
0	content:a	2350	1	content:1	2346
2	content:2011	2240	3	content:2	2235
4	content:s	2218	5	content:3	2082
6	content:4	2030	7	content:all	1988
8	content:com	1903	9	content:about	1879
10	content:from	1861	11	content:you	1851
12	content:5	1804	13	content:10	1794
14	content:more	1786	15	content:can	1746
16	content:2010	1714	17	content:new	1653
18	content:your	1626	19	content:i	1620
20	content:use	1609	21	content:data	1602

(2)

▲ 索引庫內容提供了許多統計資訊



# 你也可以擁有自己的搜尋引擎 !!!

## Start from Here!

- **Crawlzilla** 示範多人自訂搜尋雲端服務
  - <http://demo.crawlzilla.info>
- **Crawlzilla @ Google Code Project Hosting ( 中文 )**
  - <http://code.google.com/p/crawlzilla/>
- **Crawlzilla @ Source Forge (Tutorial in English)**
  - <http://sourceforge.net/p/crawlzilla/home/>
- **Crawlzilla User Group @ Google**
  - <http://groups.google.com/group/crawlzilla-user>
- **NCHC Cloud Computing Research Group**
  - <http://trac.nchc.org.tw/cloud>

# Authors of Crawlzilla

## 抓抓龍作者群

陳威宇 (左)

[\*\*waue@nchc.org.tw\*\*](mailto:waue@nchc.org.tw)

[\*\*waue0920@gmail.com\*\*](mailto:waue0920@gmail.com)

郭文傑 (中)

[\*\*rock@nchc.org.tw\*\*](mailto:rock@nchc.org.tw)

[\*\*goldjay1231@gmail.com\*\*](mailto:goldjay1231@gmail.com)

楊順發 (右)

[\*\*shunfa@nchc.org.tw\*\*](mailto:shunfa@nchc.org.tw)

[\*\*shunfa@gmail.com\*\*](mailto:shunfa@gmail.com)





## Questions?

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

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



Powered by DRBL