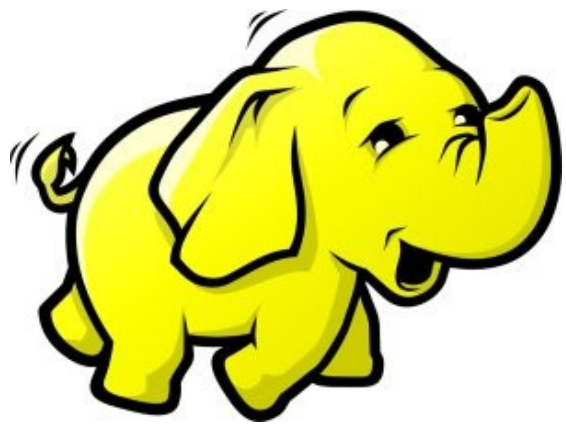
海量資料處理平台 Hadoop 與抓抓龍

Introduction to Hadoop and Crawlzilla

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



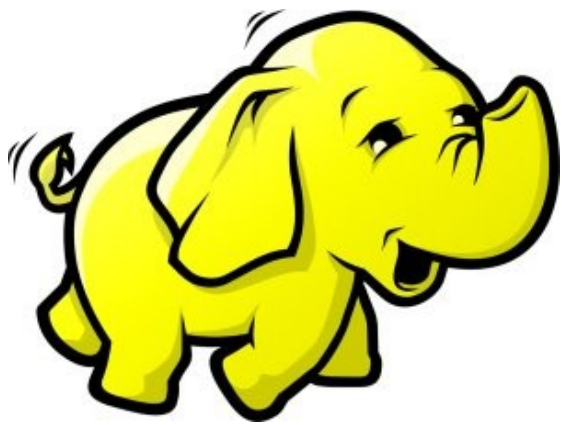
海量資料的趨勢與挑戰

Trends and Challenges of Big Data

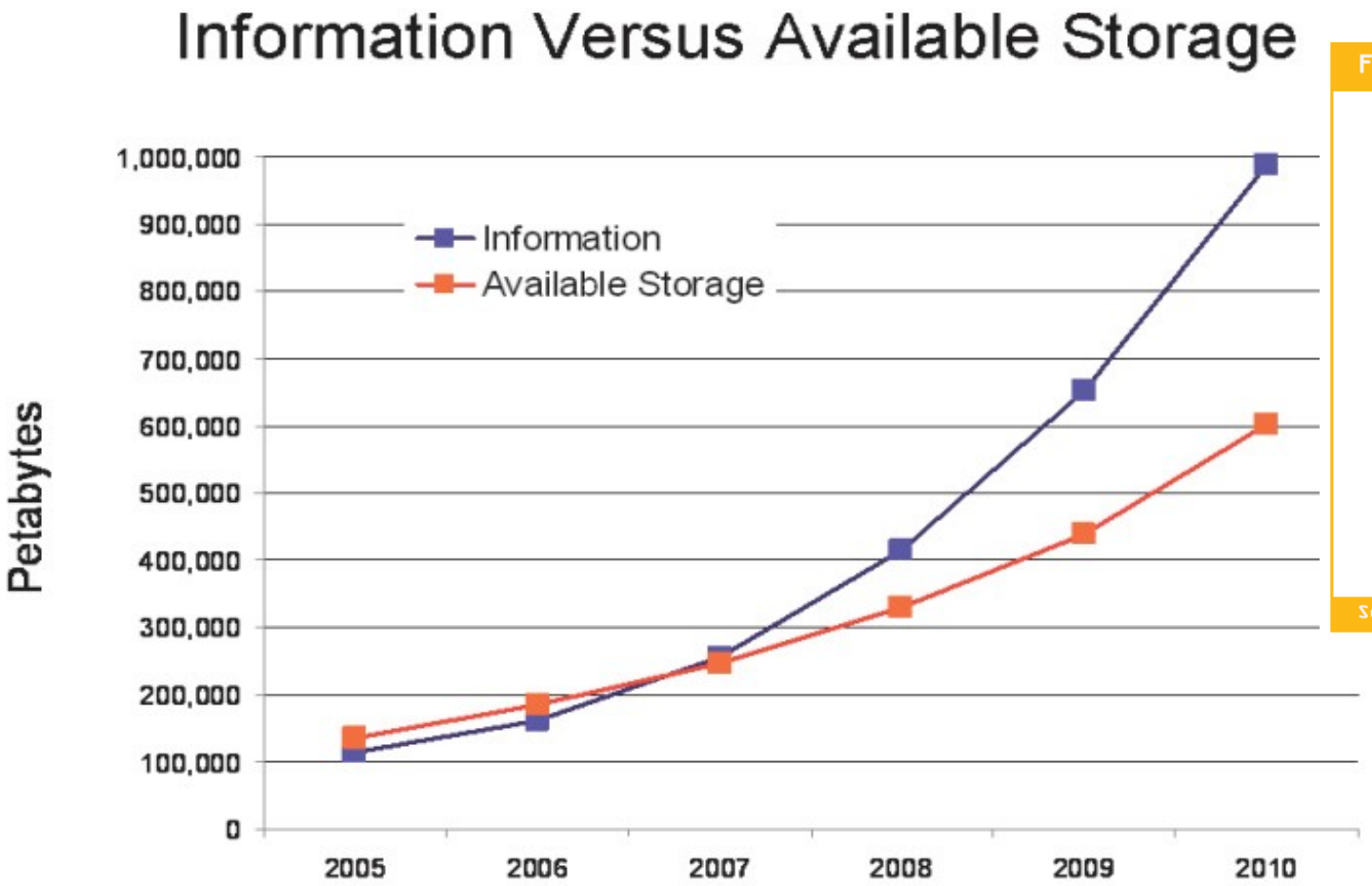
Jazz Wang

Yao-Tsung Wang

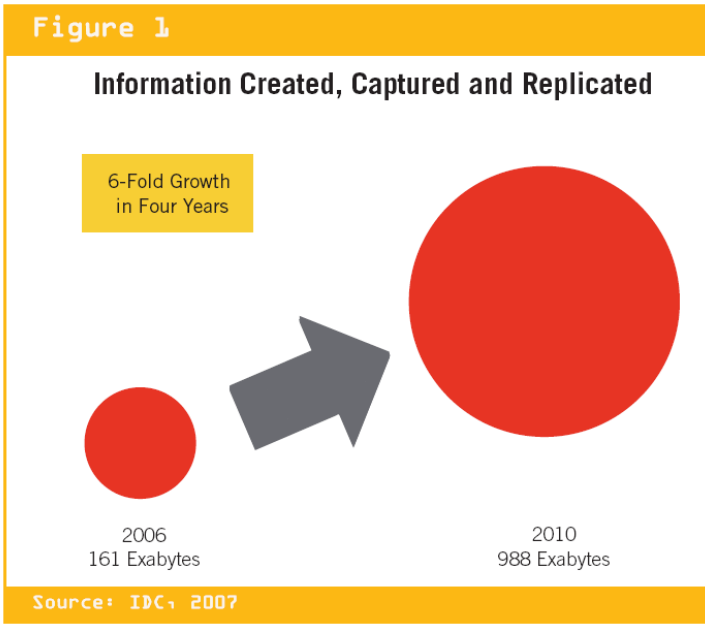
jazz@nchc.org.tw



Data Explosion!! 始於 2007 的「資料大爆炸」時代



Source: IDC, 2007

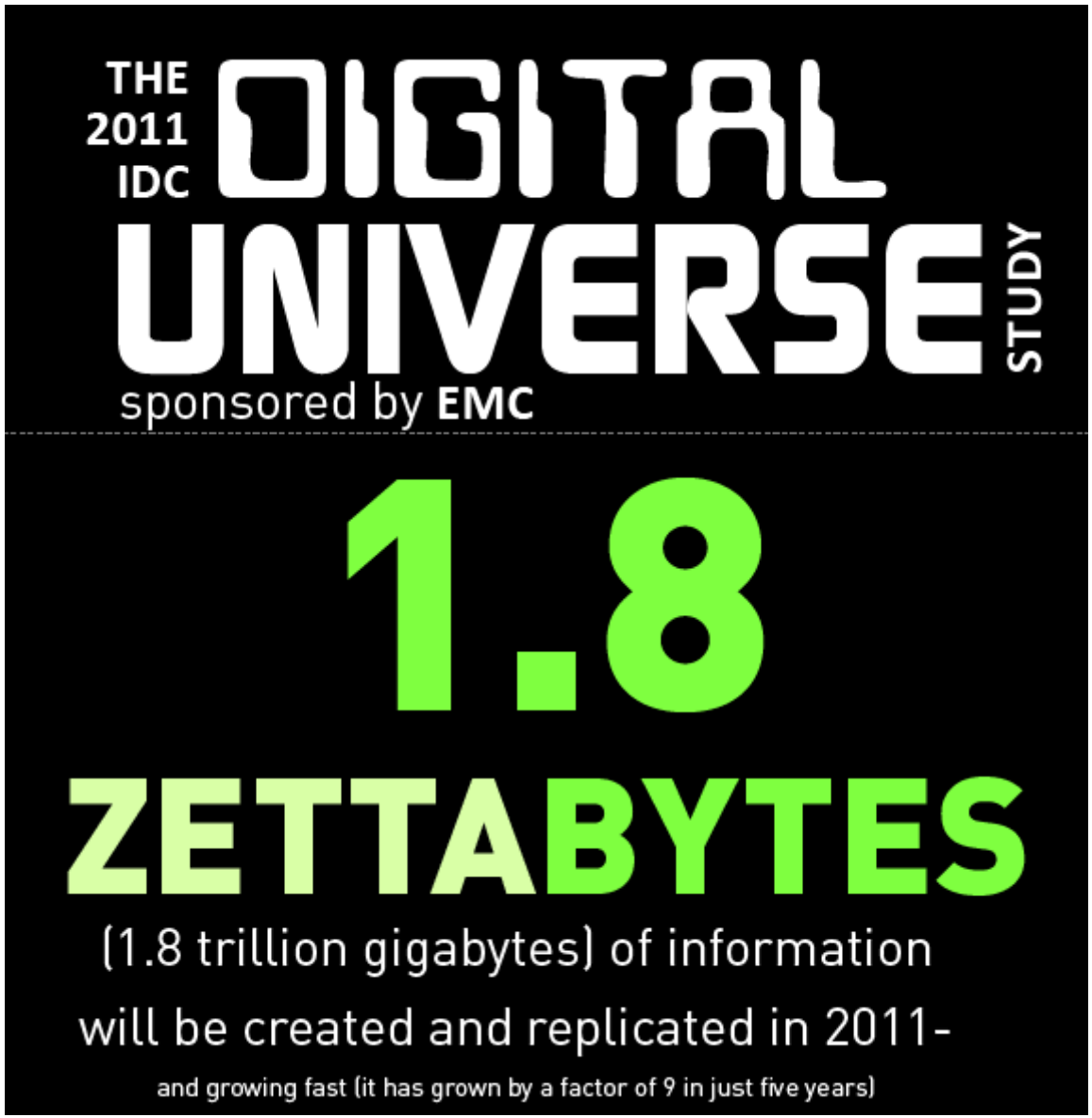


2007 年， IDC 預估
2010 年會成長六倍！
（相較 2006 年）

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]

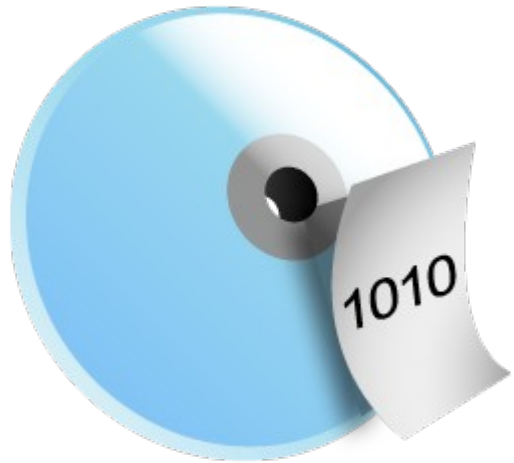
Big data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a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

In a 2001 research report[14] 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.[15]

出處：http://en.wikipedia.org/wiki/Big_data

多個檔案，容量 100TB

一個資料庫，容量 100TB

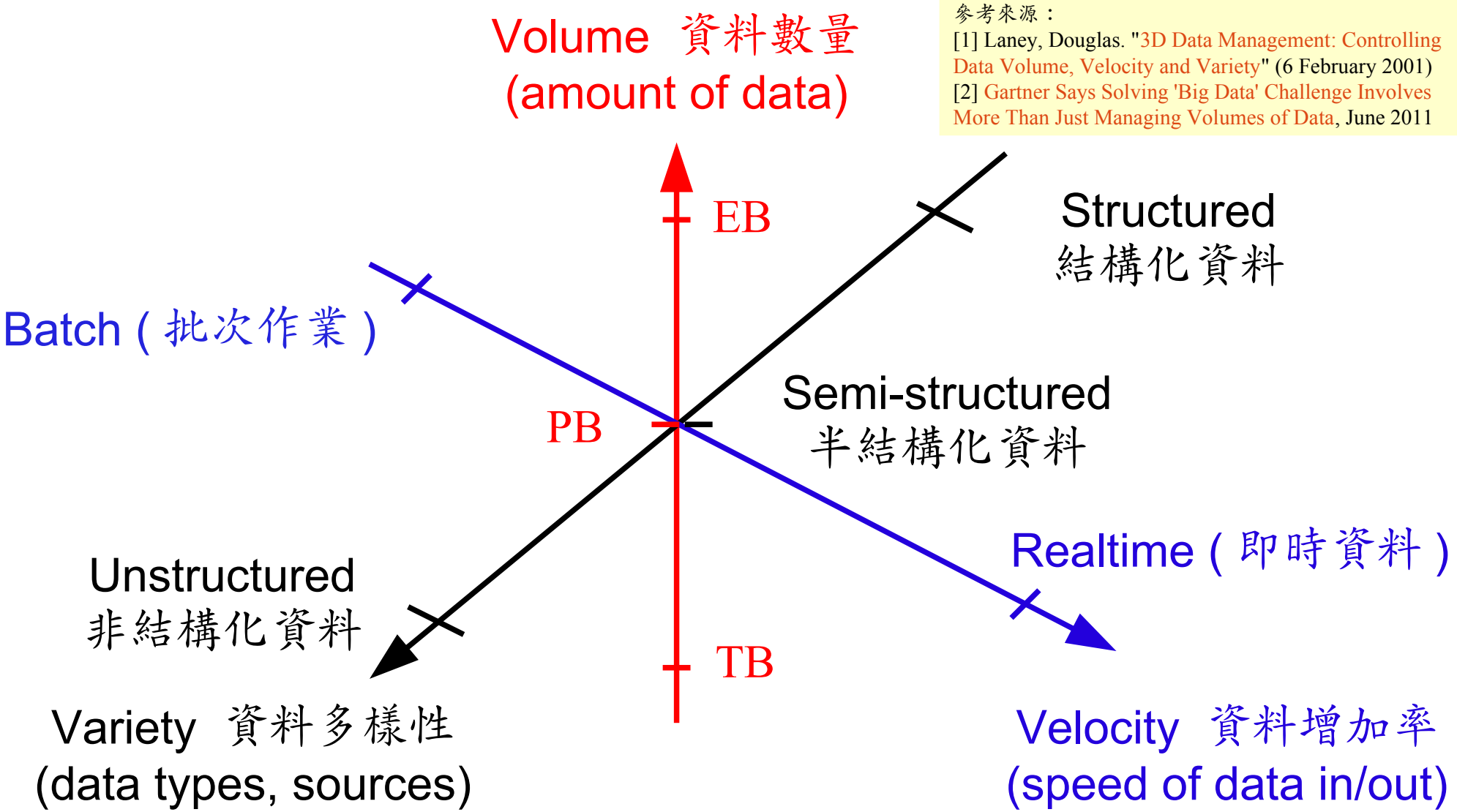


一個檔案，容量 100TB

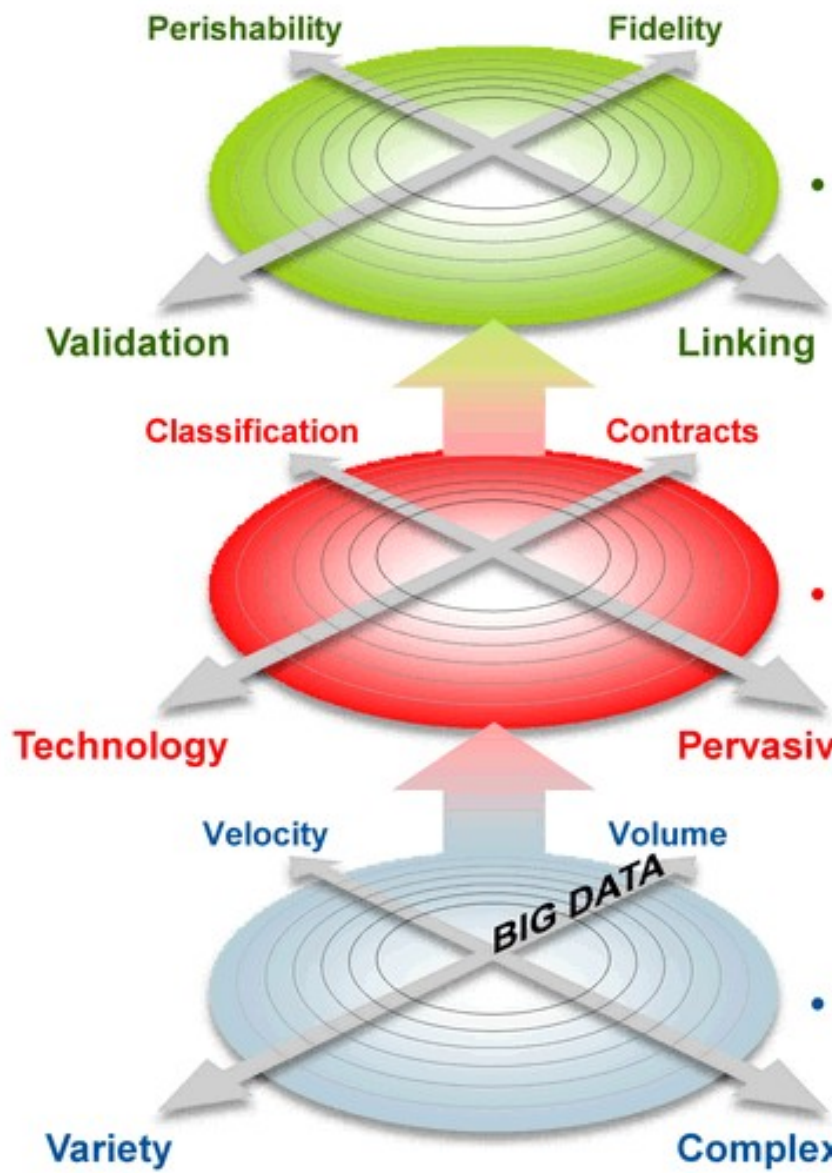
Gartner Big Data Model？ 海量資料的模型？

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？



品質管控

- Qualification and Assurance

權限管控

- Access Enablement and Control

數量管控

- Quantification

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

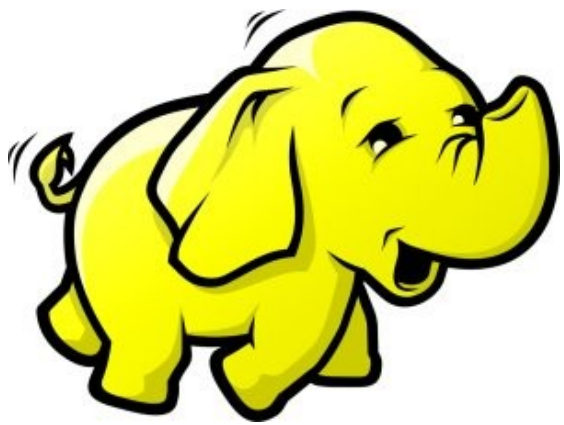
處理海量資料的資訊架構與關鍵技術

Technologies to build IT Stack for Big Data

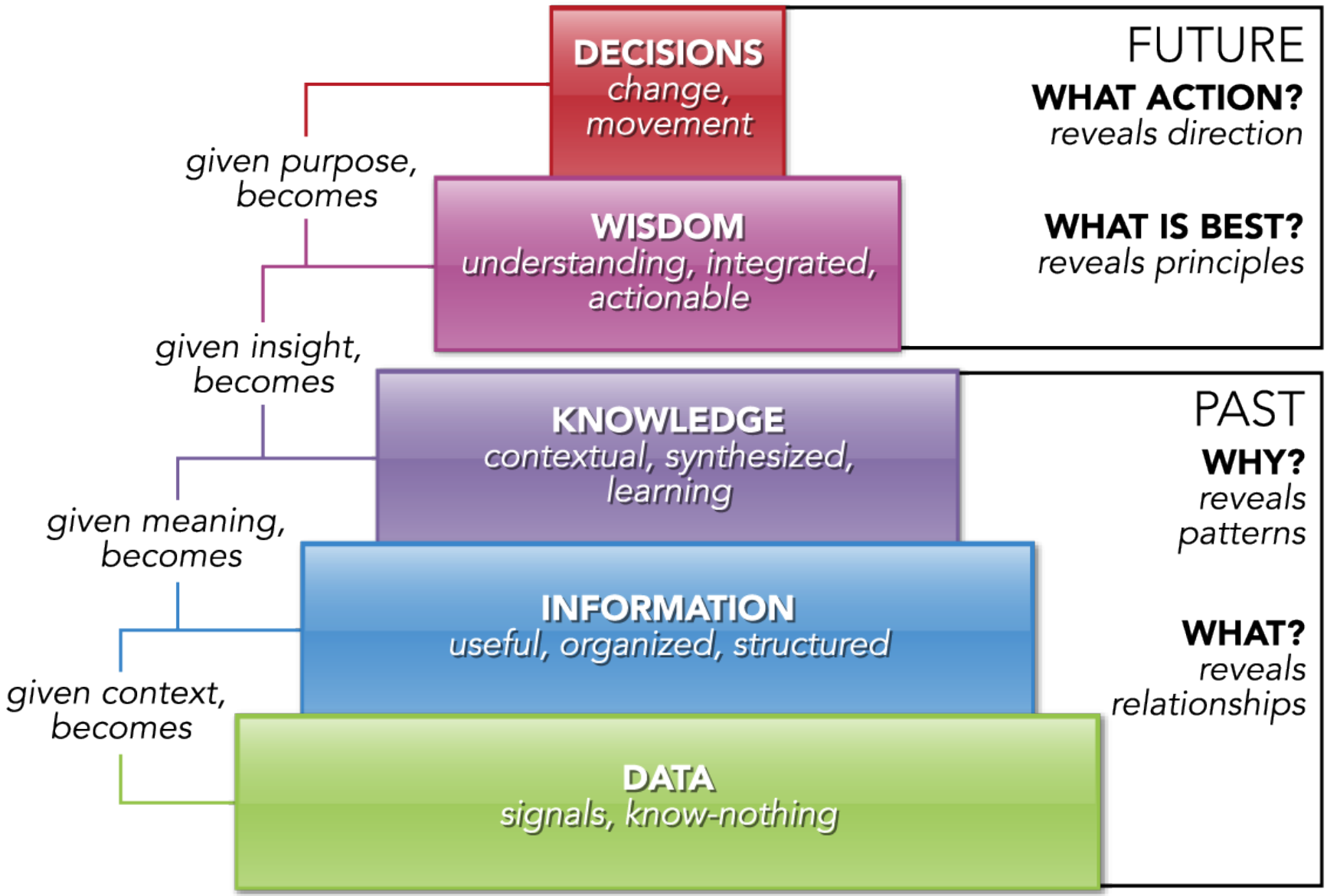
Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



Data, Information, Knowledge, Wisdom
知識管理模型：資料、資訊、知識與智慧

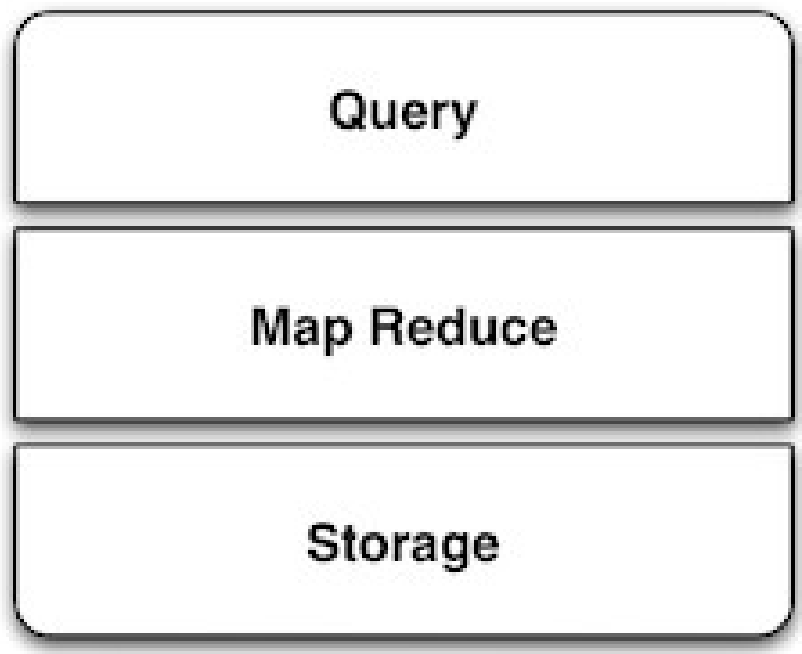


http://www.pursuantgroup.com/blog/tag/dikw-model/

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

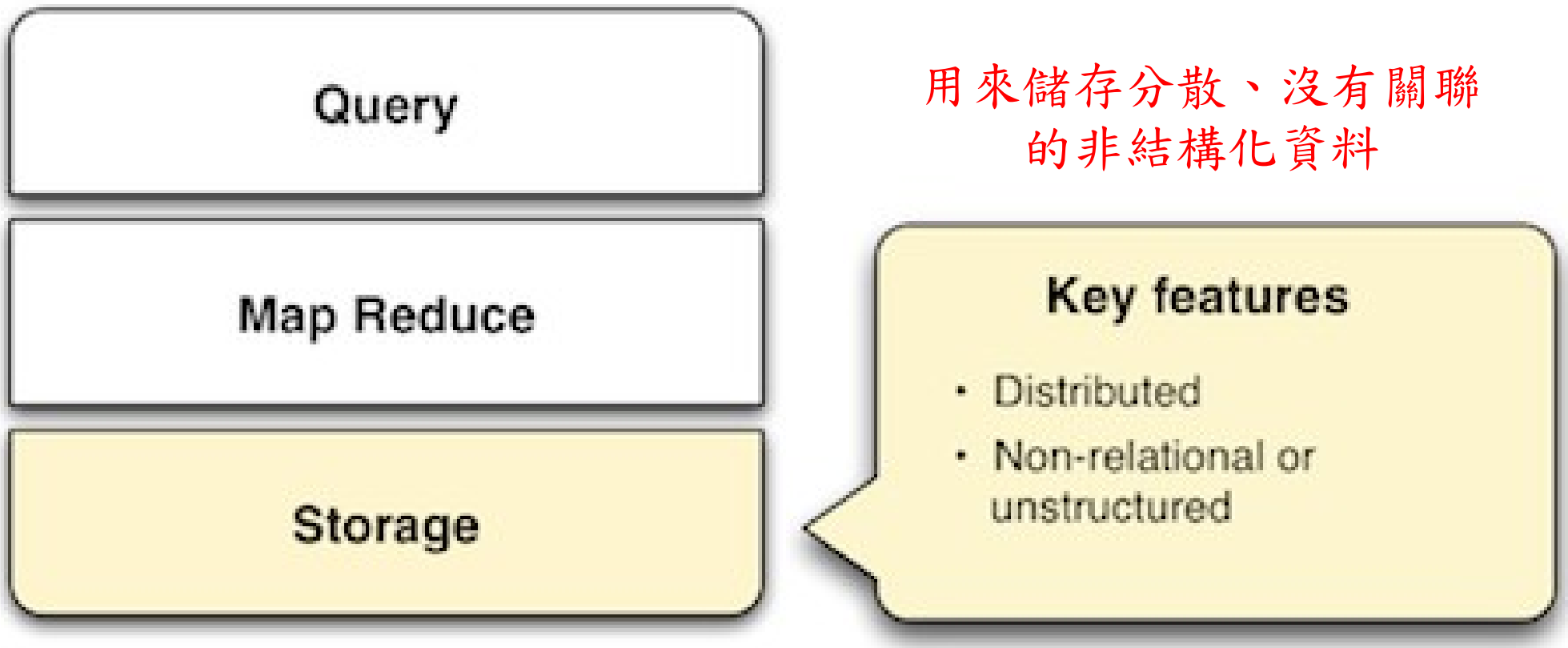
未來處理海量資料的人必需知道 SMAQ（Storage, MapReduce and Query）



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

The SMAQ stack for big data
海量資料處理的資訊架構

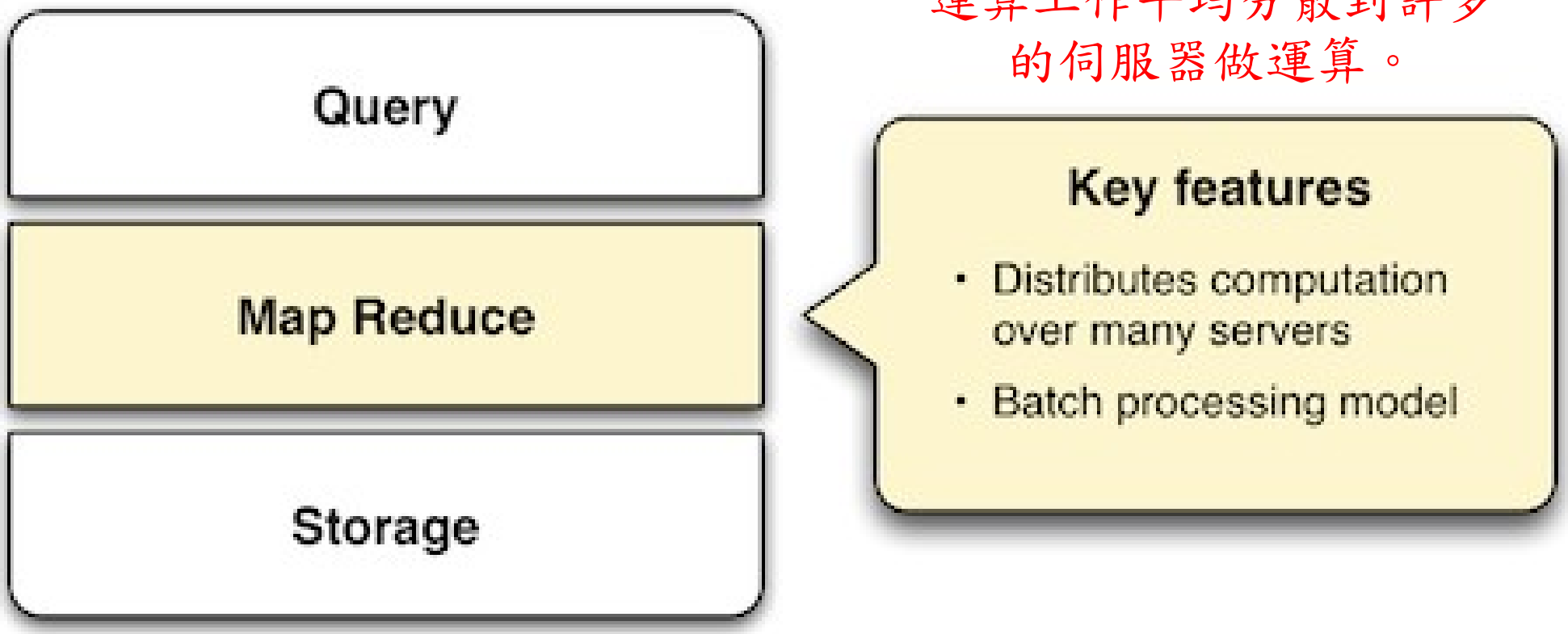


用來儲存分散、沒有關聯的非結構化資料

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

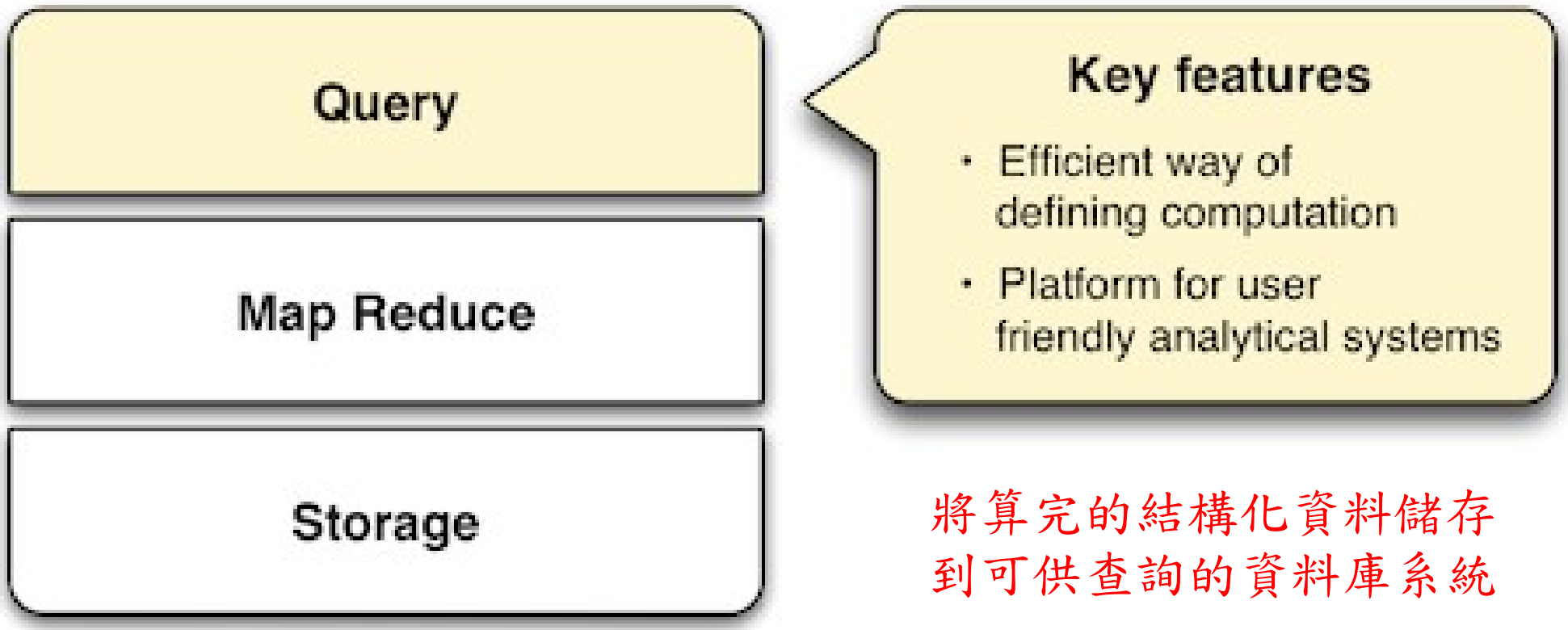
The SMAQ stack for big data
海量資料處理的資訊架構

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。



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

The SMAQ stack for big data
海量資料處理的資訊架構



將算完的結構化資料儲存到可供查詢的資料庫系統

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

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 : Simplifed 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

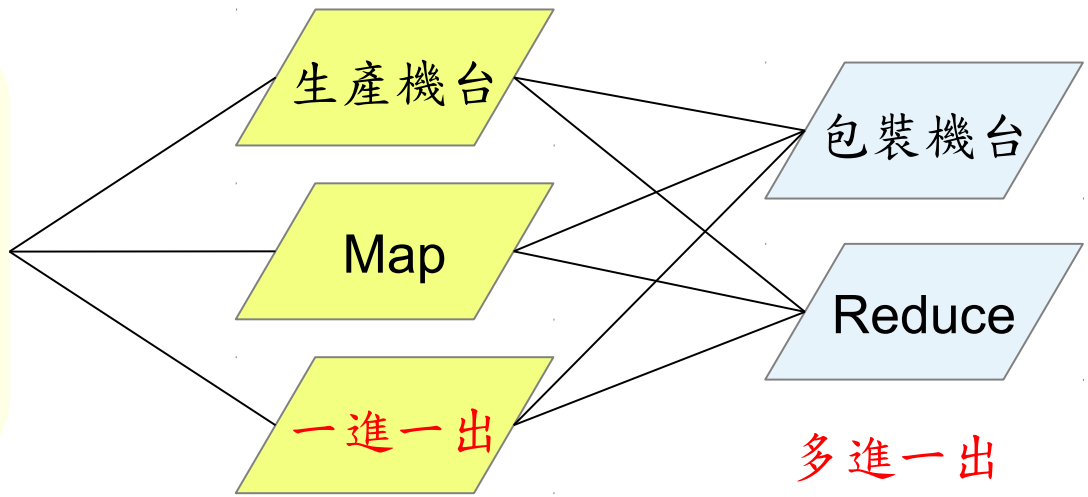
Hadoop 簡介

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

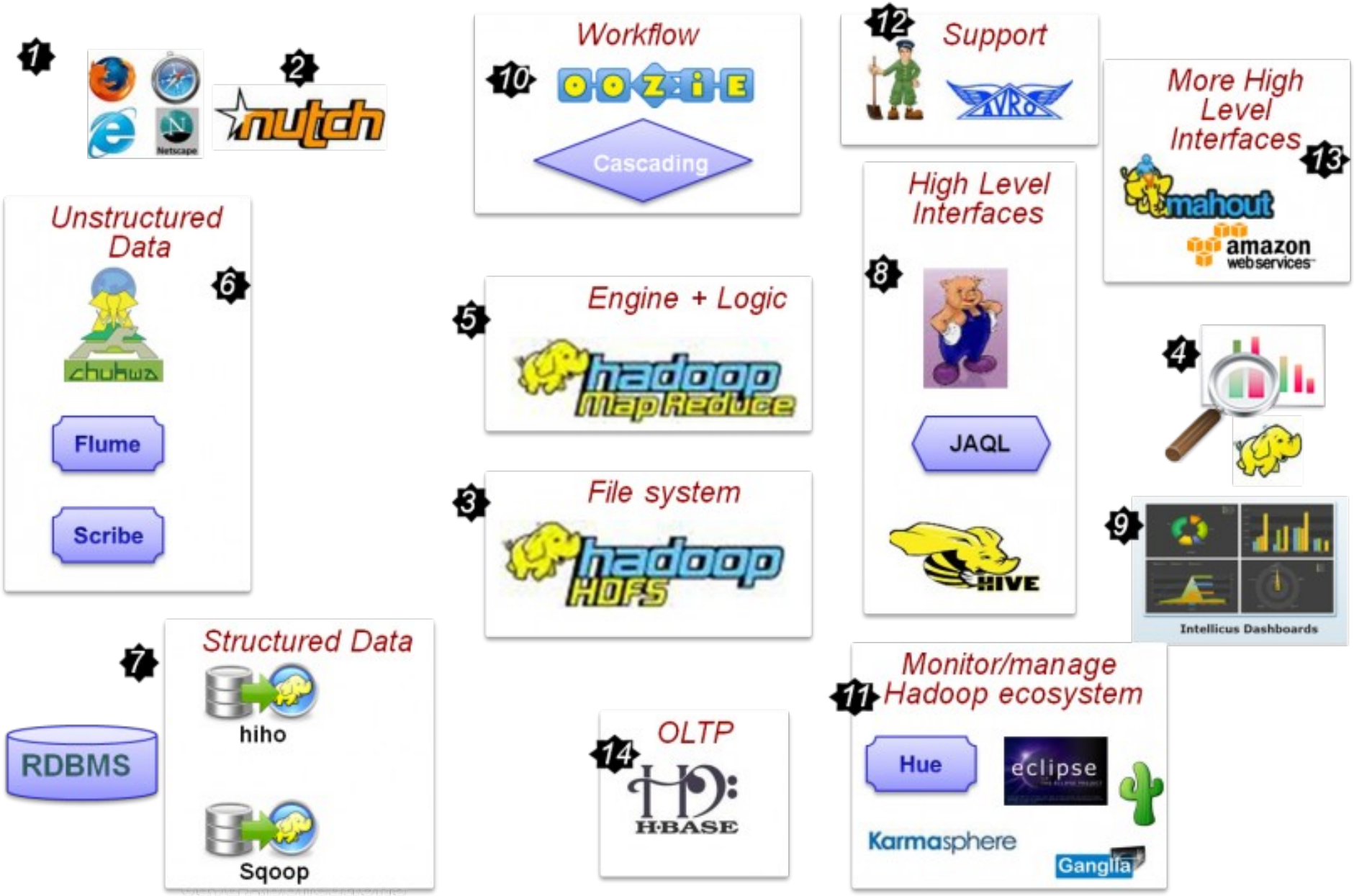
Sector-Sphere

National Center for Data Mining
University of Illinois at Chicago
UIC

open data
Open Data Group
http://www.opendatagroup.com/

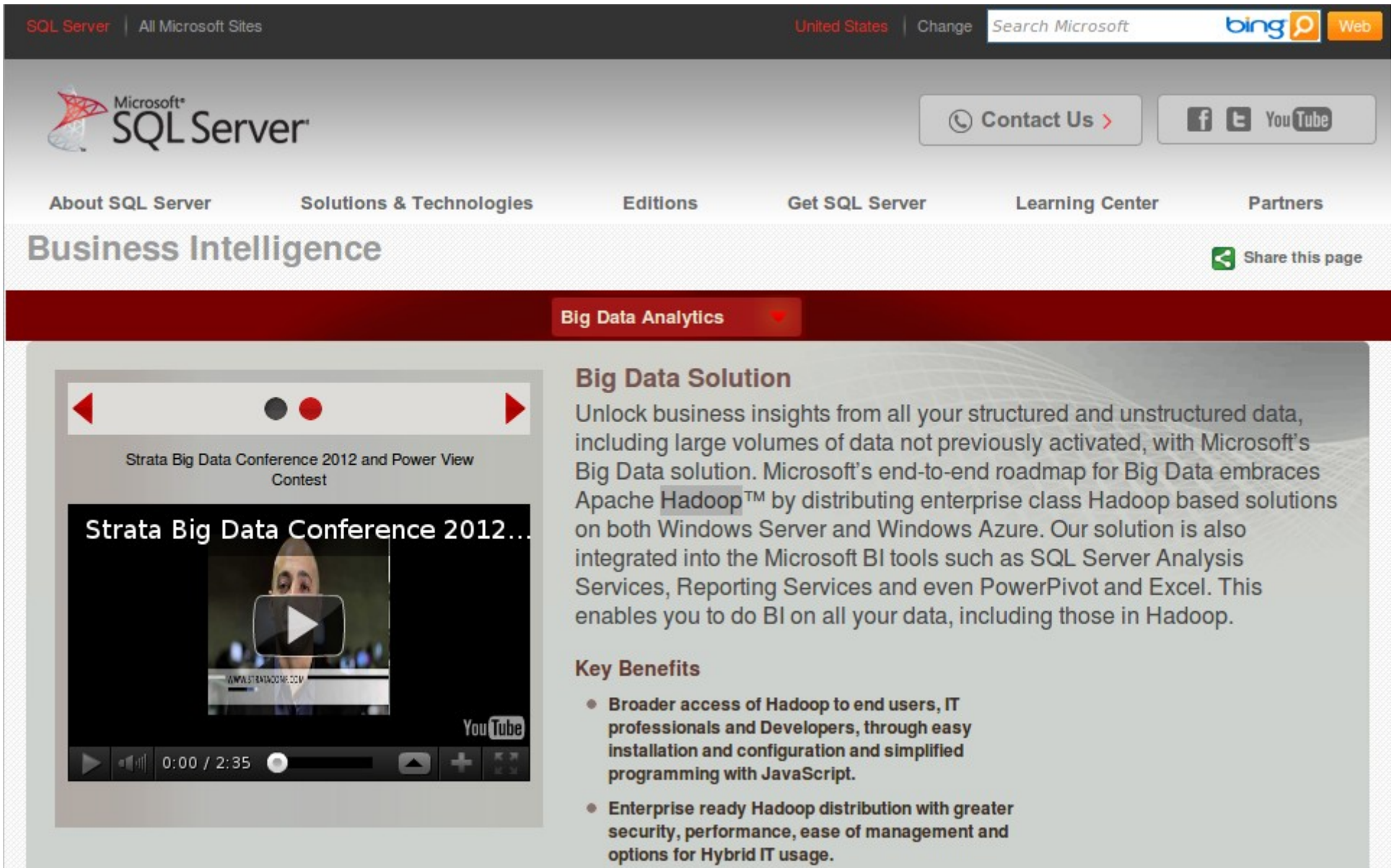
Why we choice Hadoop? Good Ecosystem!

豐富的生態系建構出處理海量資料的工具庫



http://rationalintelligence.com/wp_log/?p=104

Microsoft love Hadoop, too
微軟幫 Azure 還有 SQL Server 都接上 Hadoop



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.

參考來源：Cloudera teams up to connect Oracle and Hadoop
http://news.cnet.com/8301-13846_3-20008242-62.html

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]

全台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、行動加值服務，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

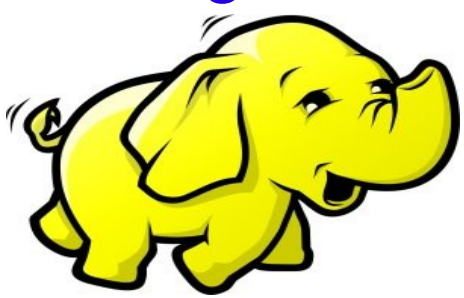
Hadoop 這套軟體的歷史源起 ... 2005~2006

- 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.)
- Doung 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

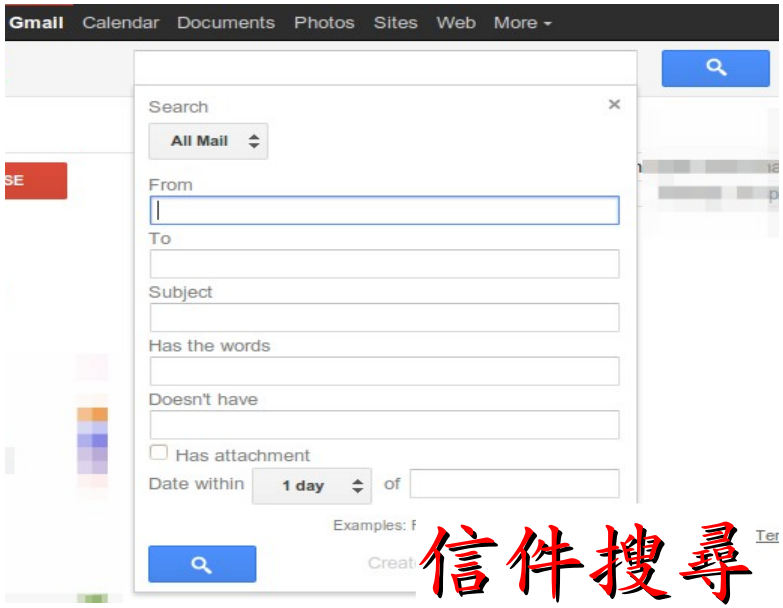
Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

Search is everywhere in our daily life !!

「搜尋」已經成為我們生活中的一部分



信件搜尋

即時通訊搜尋

資料庫搜尋

今天要談的是「網頁搜尋」

To speed up search, We need "Index"

爲了加速搜尋的效率，我們需要「索引」

出現頁碼

關鍵字

Index

Symbols

A

Do you like to write notes?

你有寫筆記的習慣嘛？

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/

December 2011

Su	Mo	Tu	We	Th	Fr	Sa
				1	2	
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

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		
chinese-utf8.pl	**Chinese Translations**	http://www.oddmuse.org/cgi-bin/wiki/Chinese
download.pl	**Download Extension**	http://www.oddmuse.org/cgi-bin/wiki/Download_Extension
headers.pl	**Header Markup Extension**	http://www.oddmuse.org/cgi-bin/wiki/Header_Markup_Extension
tables-long.pl	**Long Table Markup Extension**	http://www.oddmuse.org/cgi-bin/wiki/Long_Table_Markup_Extension
tables.pl	**Table Markup Extension**	http://www.oddmuse.org/cgi-bin/wiki/Table_Markup_Extension
usemod.pl	**Usemod Markup Extension**	http://www.oddmuse.org/cgi-bin/wiki/Usemod_Markup_Extension
2005-02-16		
calendar.pl	**Calendar Extension**	http://www.oddmuse.org/cgi-bin/wiki/Calendar_Extension

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/

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/

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/

trac
Integrated SCM & Project Management

Search

Login | Preferences | Help/Guide | About Trac

Wiki | Timeline | Roadmap | Browse Source | View Tickets | Search

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 Realtion
【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

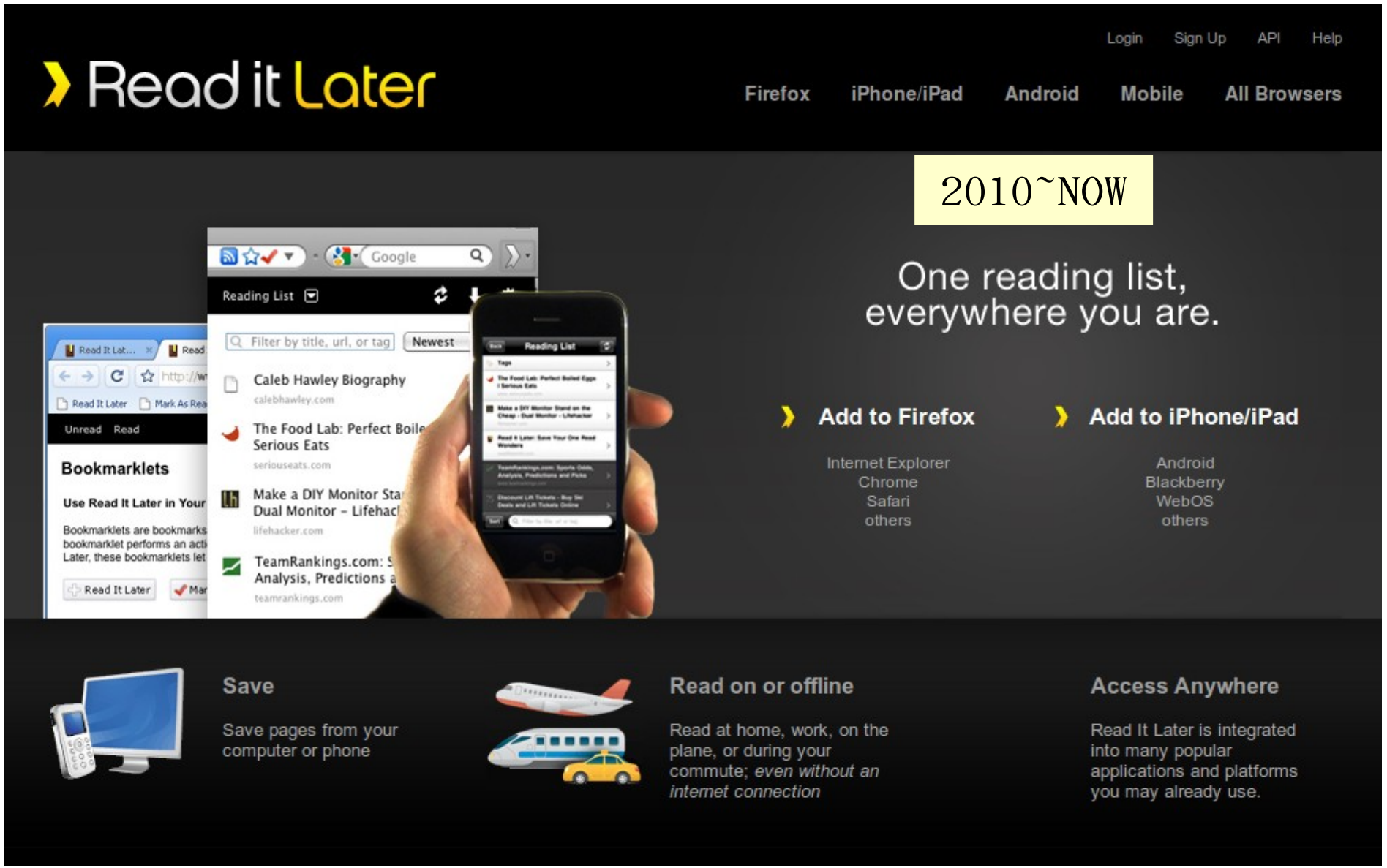
- 2008-05-27: How to Setup BOINC example project
- 2008-02-13: How to Build BONIC Server

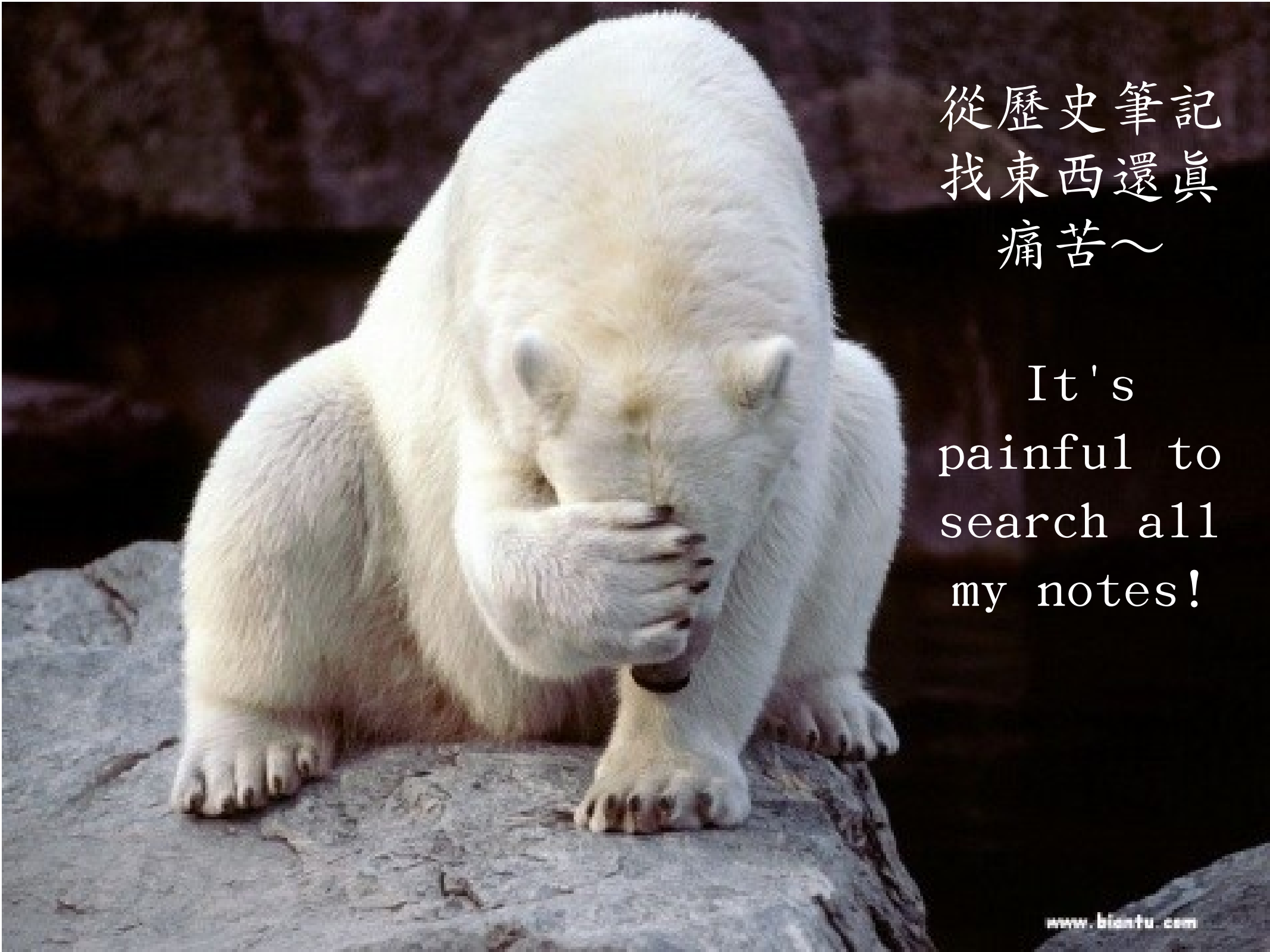
2006~NOW

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/




從歷史筆記
找東西還眞
痛苦～
It's
painful to
search all
my notes!
www.biantu.com

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？

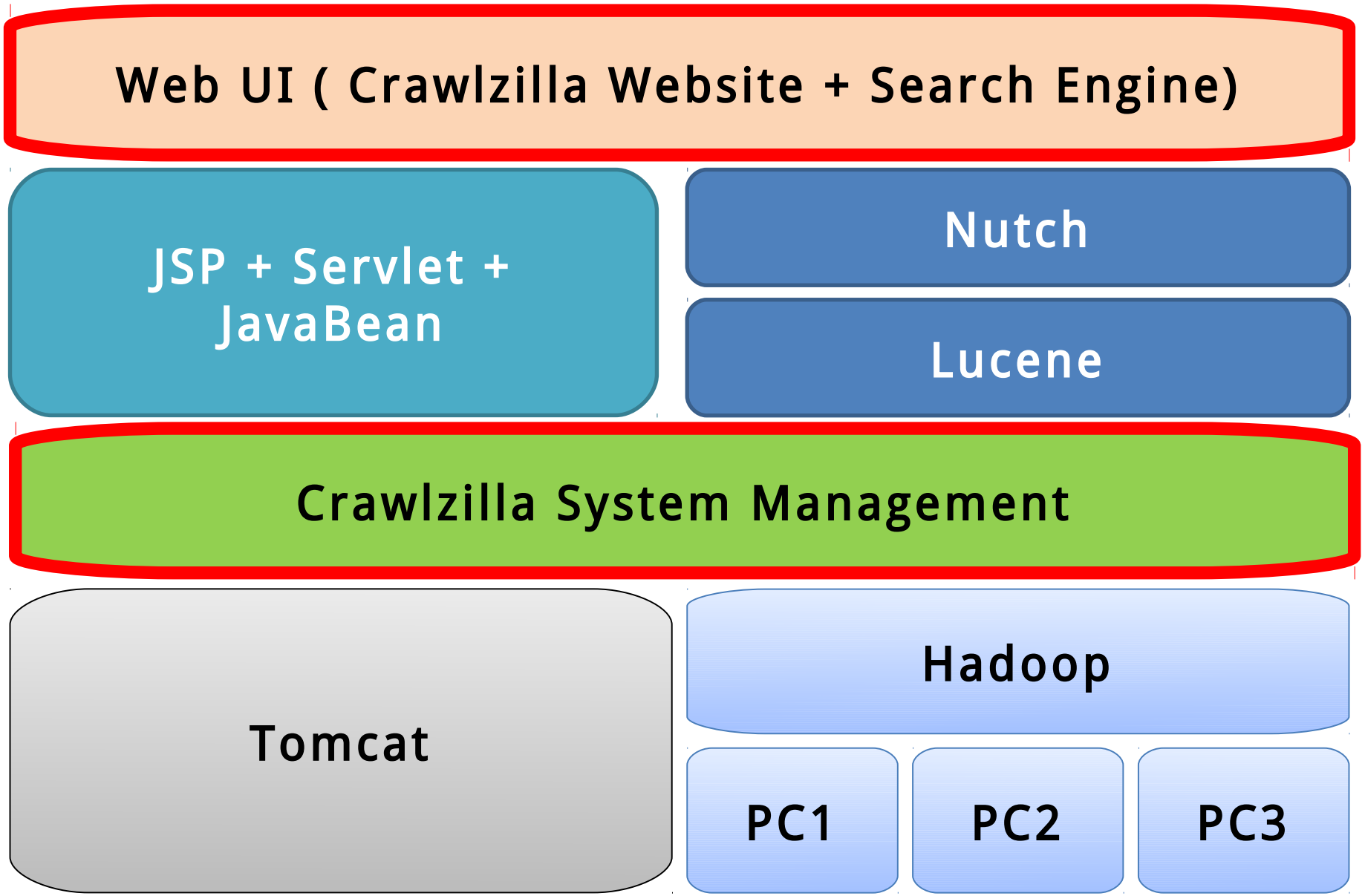
Crawlzilla 系統功能
Feature of Crawlzilla

- 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

System Architecture of Crawlzilla

抓抓龍系統架構



Comparison with other projects

抓抓龍與其他搜尋引擎專案的比較

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

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)

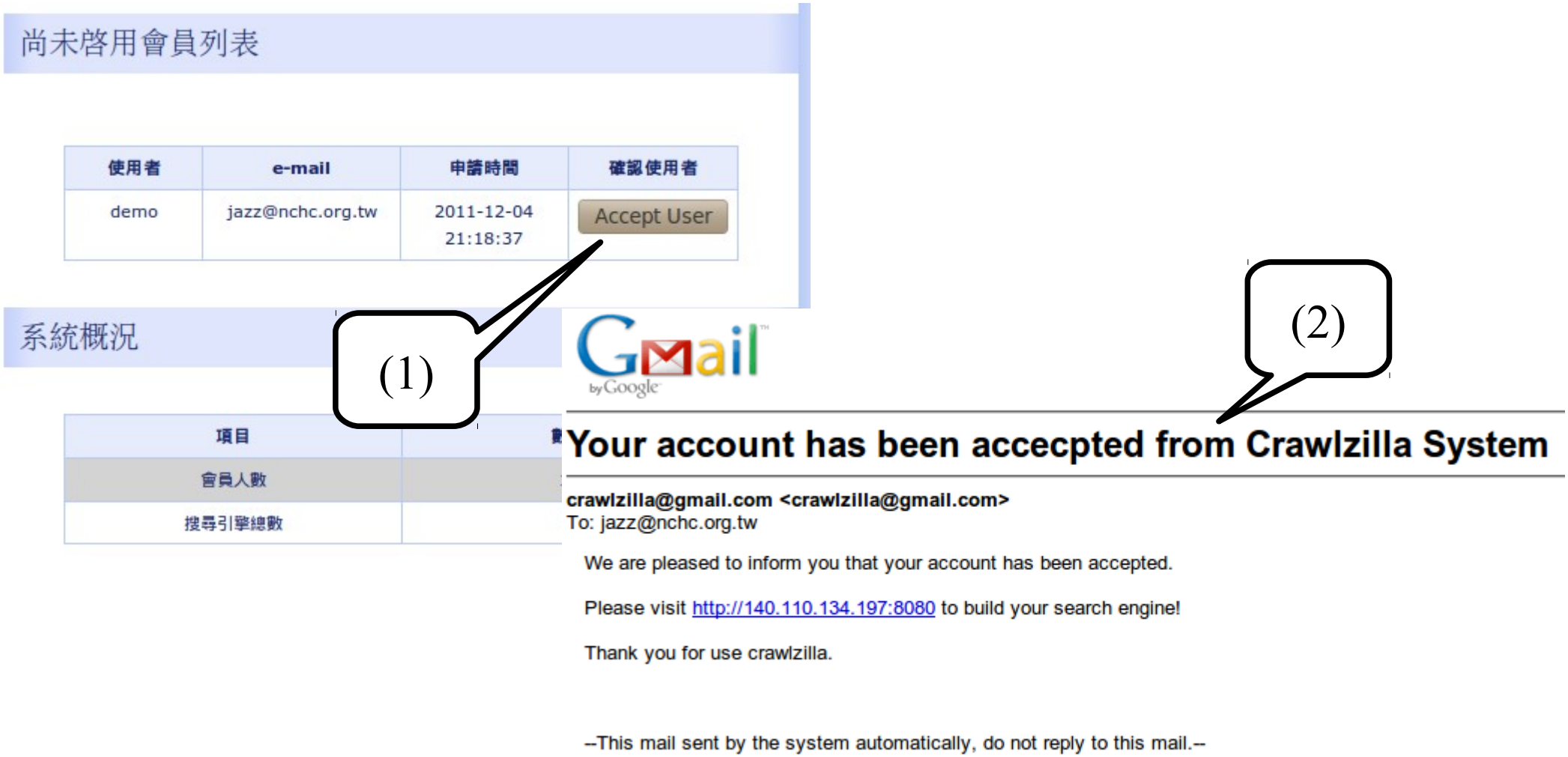
首先連線到 http://demo.crawlzilla.info

▲ 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！



▲ 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)

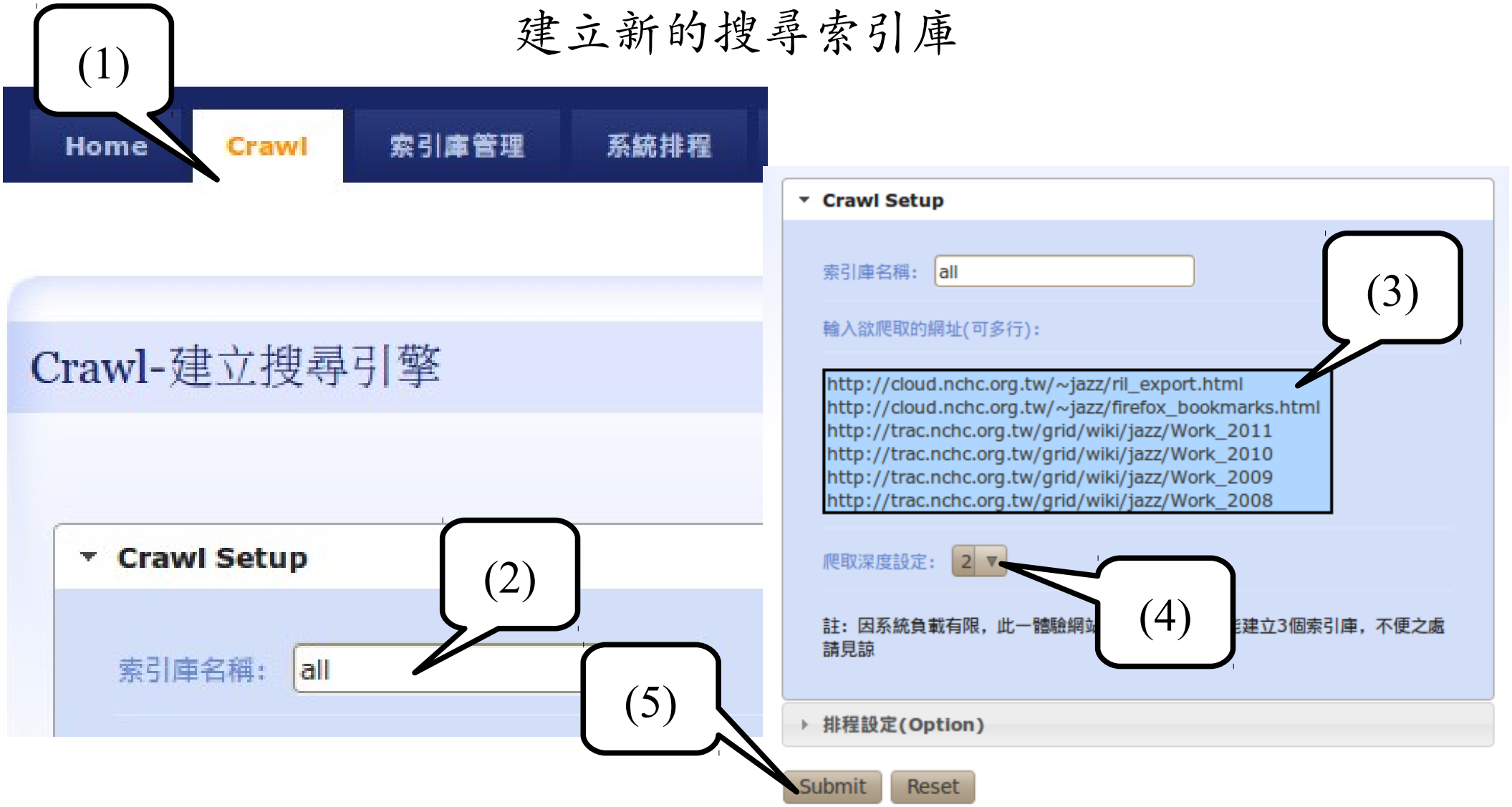
重新連線到 http://demo.crawlzilla.info

▲ 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



▲ Step 4 ：輸入索引庫名稱、起始網址與搜尋深度

Multi-user Web Search Cloud Service : Crawlzilla 1.0

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

▲ Step 5 ：等待抓抓龍幫您建立專屬的搜尋索引庫

Multi-user Web Search Cloud Service : Crawlzilla 1.0

Crawlzilla 1.0 多人版雲端服務 (6)

索引庫管理

▲ 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

▲ 您可以在索引庫管理進行手動重新爬取或刪除索引庫

Multi-user Web Search Cloud Service : Crawlzilla 1.0

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 ）

▲ 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

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

▲ 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

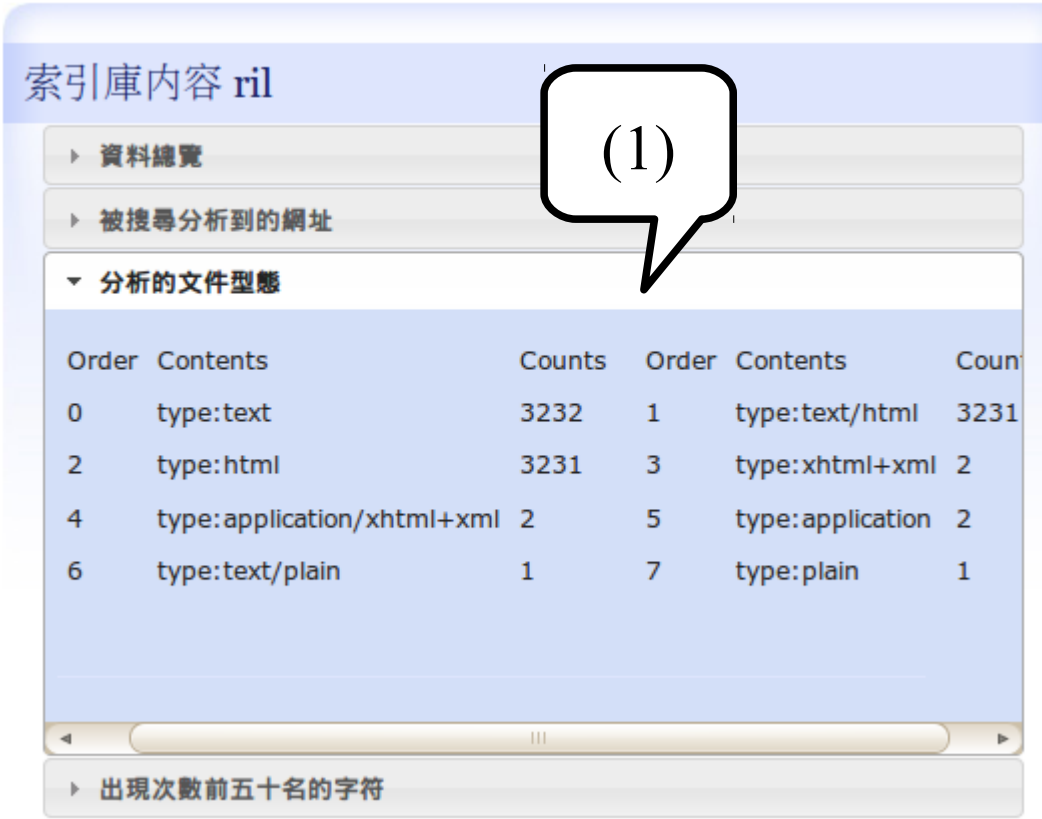
Order	Contents	Counts	Order	Contents
0	site:www.digitimes.com.tw	204	1	site:www.bnext.c
2	site:groups.google.com	74	3	site:highscalabilit
4	site:www.theregister.co.uk	49	5	site:www.ithome
6	site:www.cloudera.com	47	7	site:gigaom.com
8	site:en.wikipedia.org	39	9	site:www.networ
10	site:wiki.apache.org	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

▲ 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



Order	Contents	Counts	Order	Contents	Coun
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

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

▲ 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 (Toturial 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

waue0920@gmail.com

郭文傑 (中)

rock@nchc.org.tw

goldjay1231@gmail.com

楊順發 (右)

shunfa@nchc.org.tw

shunfa@gmail.com

Questions?

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

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw