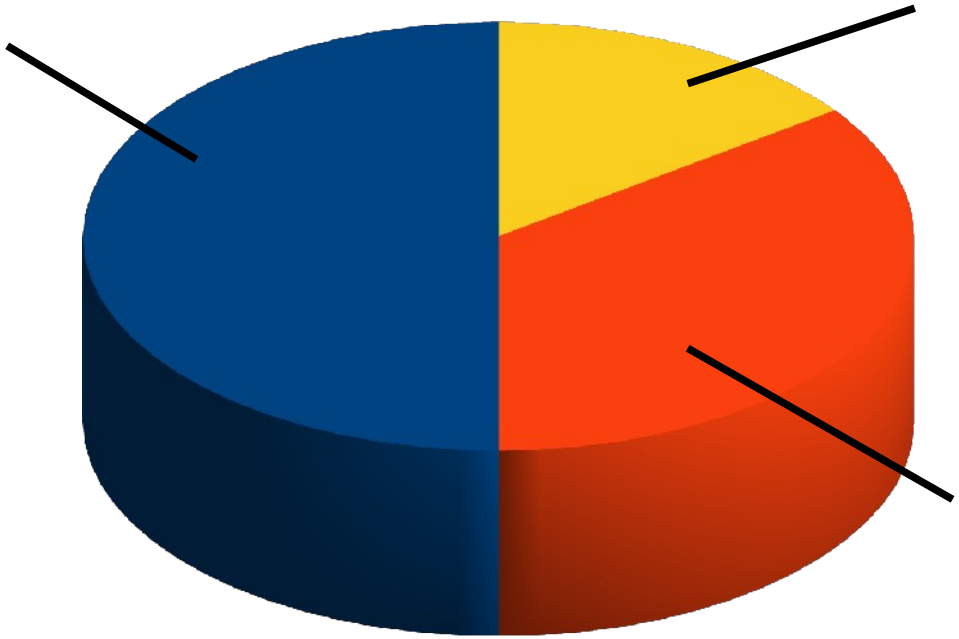
WHO AM I？這傢伙是誰啊？JAZZ？

- 講者介紹：
  - 國網中心　王耀聰　副研究員 / 交大電控八九級碩士
  - jazz@nchc.org.tw
- 所有投影片、參考資料與操作步驟均在網路上
  - http://trac.nchc.org.tw/cloud
  - 由於雲端資訊變動太快，愛護地球，請減少不必要之列印。

FOSS 使用者
Debian/Ubutnu
Access Grid
Motion/VLC
Red5
Debian Router
DRBL/Clonezilla
Hadoop



行動力薄弱的開發者
TRTC WSU/
Haduzilla /
Hadop4Win / Ezilla

推廣者
DRBL/Clonezilla
Partclone/Tuxboot
Hadoop Ecosystem

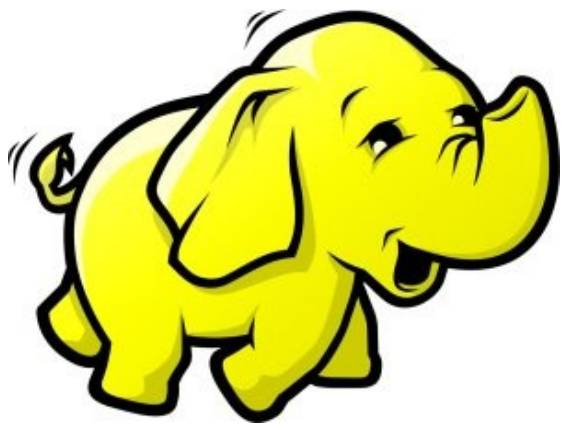
淺談海量資料的趨勢、挑戰與因應對策

Big Data : the Trends, Challenges and Solutions

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



Agenda 演講大綱

What is Big Data ? 何謂海量資料

Why should we care? 爲何需要關切

When to deploy it ? 何時導入技術

How to handle it ? 三大因應策略

Who is key player ? 誰是成功關鍵

WHAT

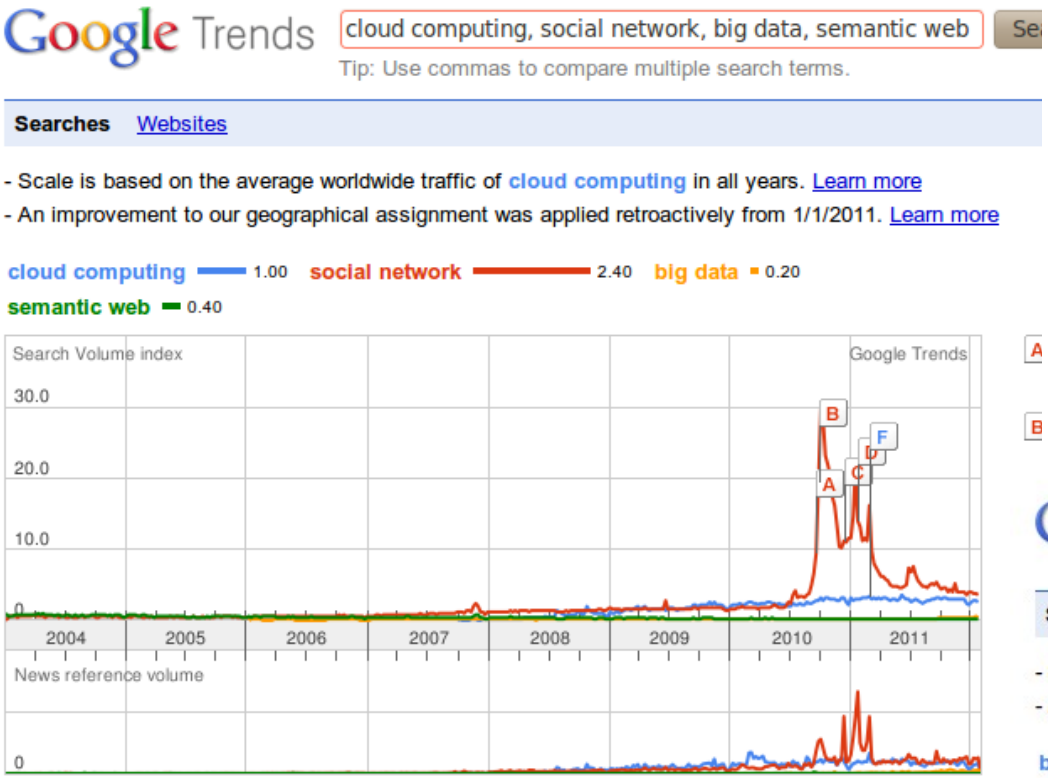
What is Big Data ?　何謂海量資料

趨勢 Trends → 定義 Definitions → 挑戰：管理維度 The Six Dimensions

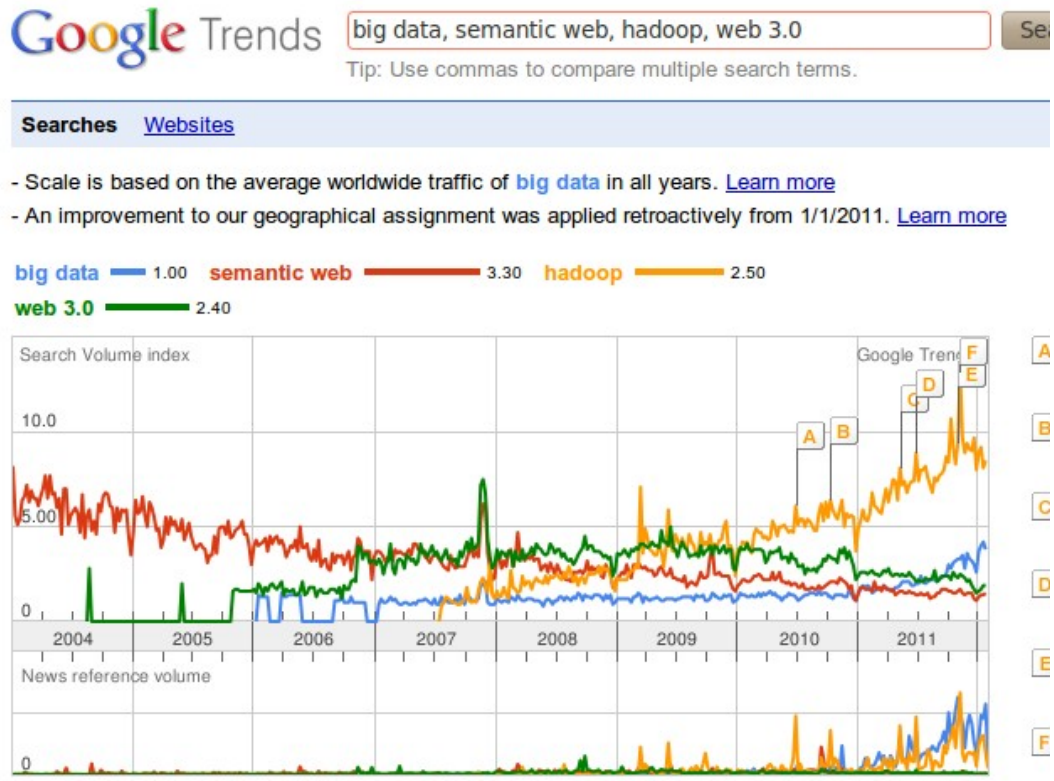
Source: http://www.2010taipeiexpo.tw/ct.asp?xItem=17186&CtNode=5952&mp=3

Trends …. It's all about Buzzwords ….. 「趨勢」亦或「流行語」？

Web 3.0, Cloud Computing, Social Network, Big Data, ….

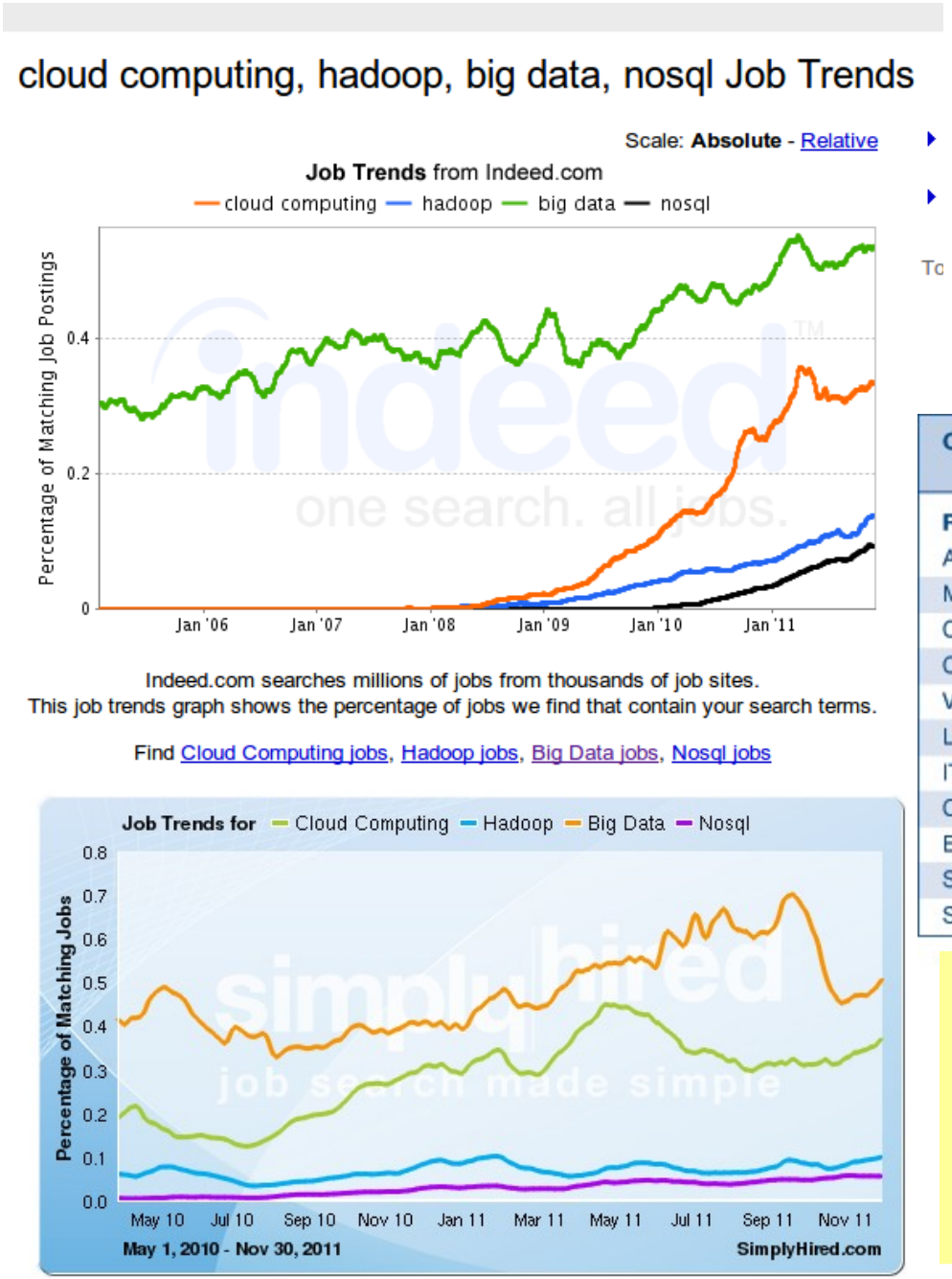


語意網（Semantic Web）從2001年開始制定標準後，逐漸下滑。而同義詞Web 3.0也呈現相似趨勢。海量資料（Big Data）與其關鍵技術Hadoop，則仍在上揚中。



整體而言，雲端運算（Cloud Computing）與社交網路（Social Network）呈現上揚。且社交網路比雲端運算還引人注目。

Trends of Market Needs 市場需求趨勢



美國軟體就業市場分析，根據 indeed 與 simply hired 兩間公司的趨勢觀察，都得到一樣的結果：

Big Data > Cloud Computing > Hadoop > NoSQL

CIO technologies	Ranking of technologies CIOs selected as one of their top 3 priorities in 2012			
Ranking	**2012**	**2011**	**2010**	**2009**
Analytics and business intelligence	1	5	5	1
Mobile technologies	2	3	6	12
Cloud computing (SaaS, IaaS, PaaS)	3	1	2	16
Collaboration technologies (workflow)	4	8	11	5
Virtualization	5	2	1	3
Legacy modernization	6	7	15	4
IT management	7	4	10	*
Customer relationship management	8	18	*	*
ERP applications	9	13	14	2
Security	10	12	9	8
Social media/Web 2.0	11	10	3	15

Gartner CIO Agenda 2012 前三名：

[1] Business Intelligence (Big Data)

[2] Mobile technology

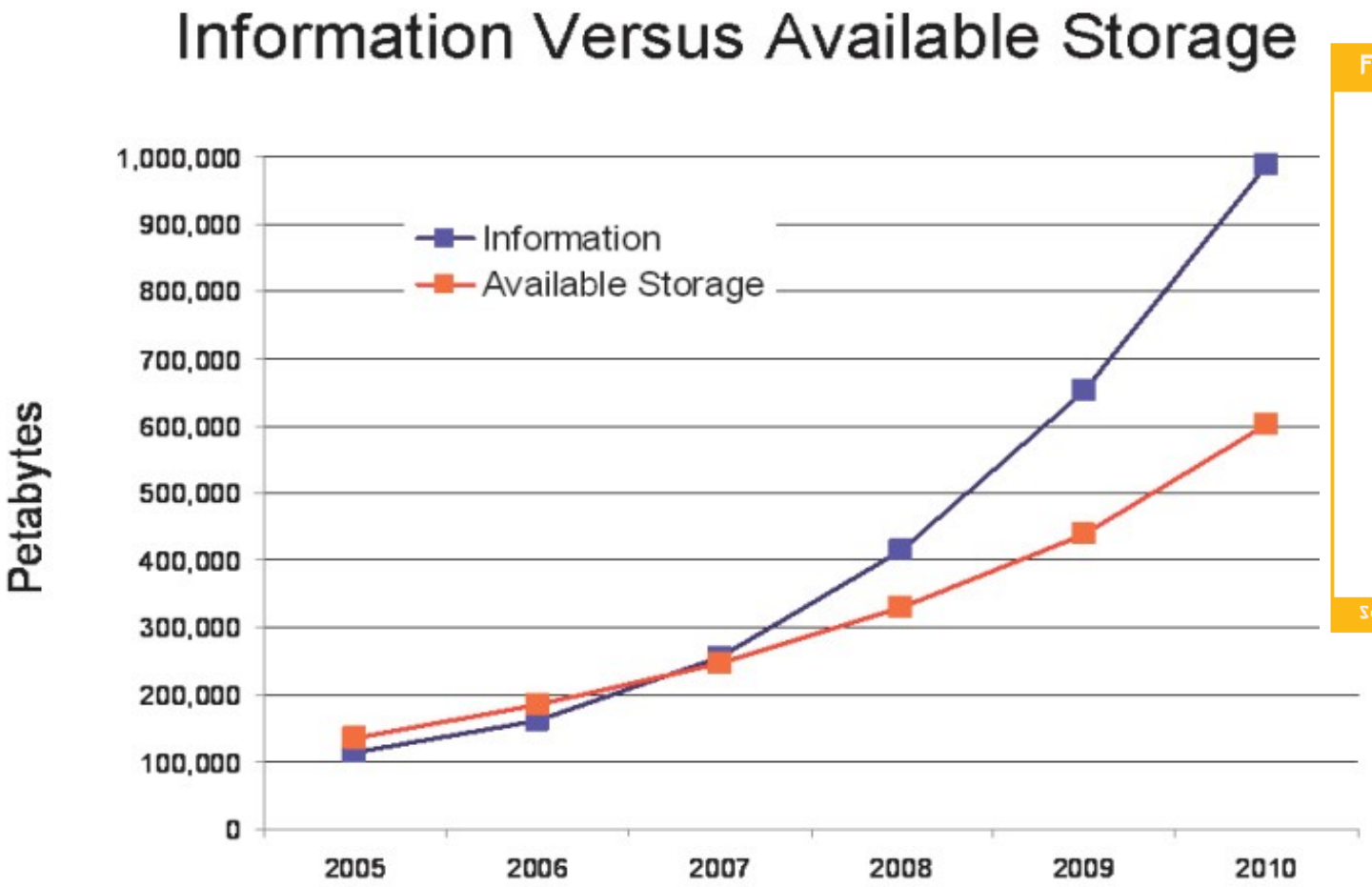
[3] Cloud Computing

參考來源：http://www.gartner.com/DisplayDocument?id=1524714

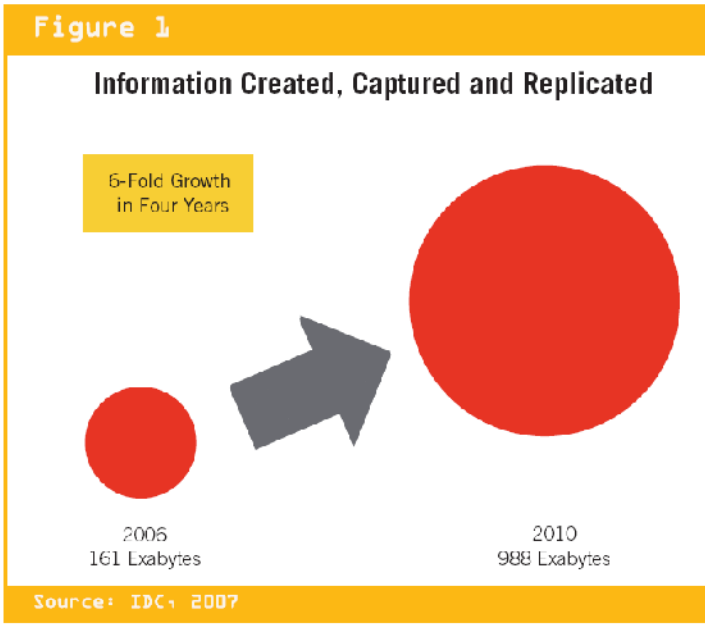
How BIG? 讓我們先來認識一下容量單位

Bit (b)	1 or 0
Byte (B)	8 bits
Kilobyte (KB)	1,000 bytes
Megabyte (MB)	1,000 KB
Gigabyte (GB)	1,000 MB
Terabyte (TB)	1,000, GB
Petabyte (PB)	1,000 TB
Exabyte (EB)	1,000 PB
Zettabyte (ZB)	1,000 EB

Data Explosion!! 始於 2007 的「資料大爆炸」時代



Source: IDC, 2007

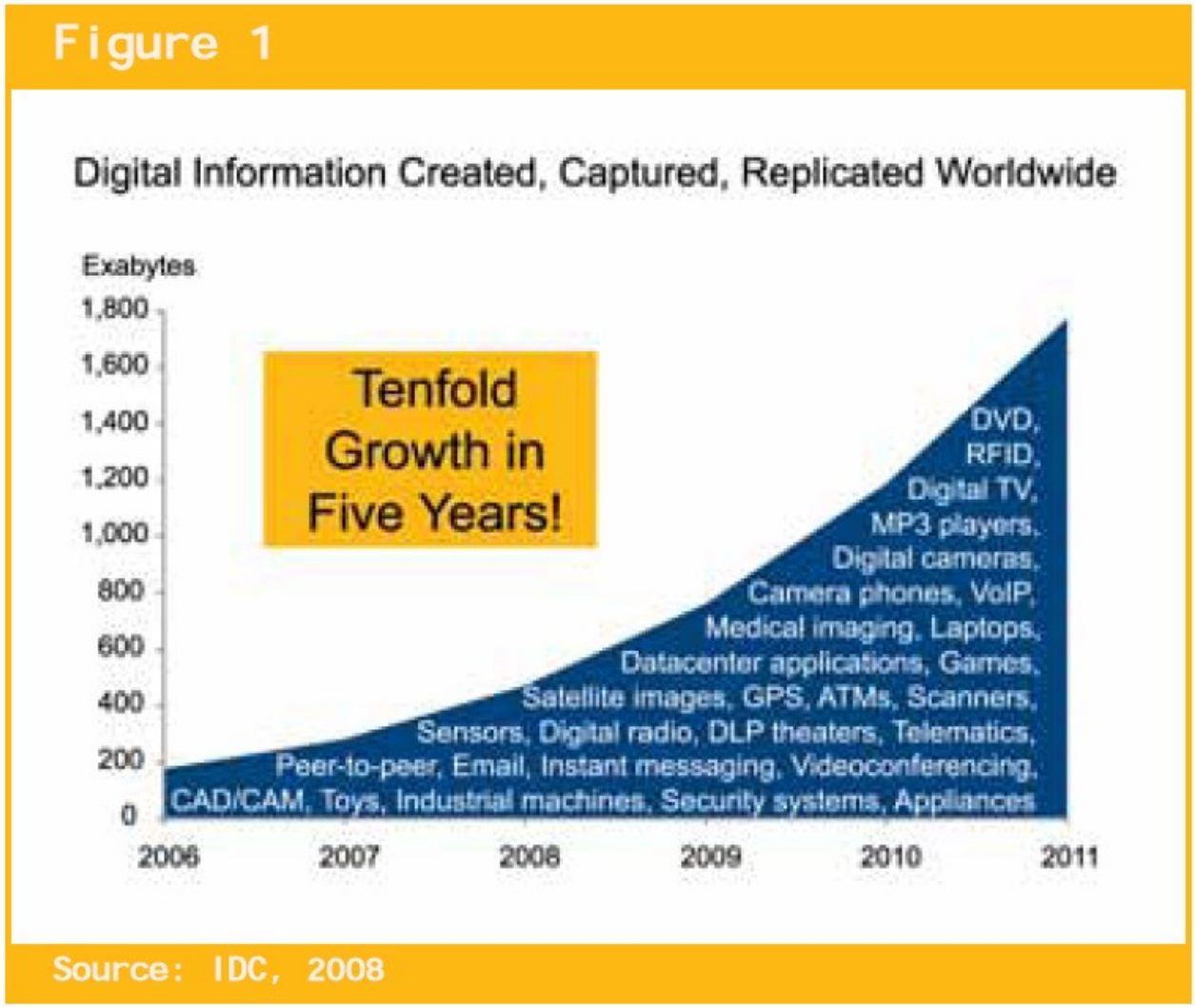


2007 年， IDC 預估
2010 年會成長六倍！
（相較 2006 年）

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 Explosion!! 始於 2007 的「資料大爆炸」時代

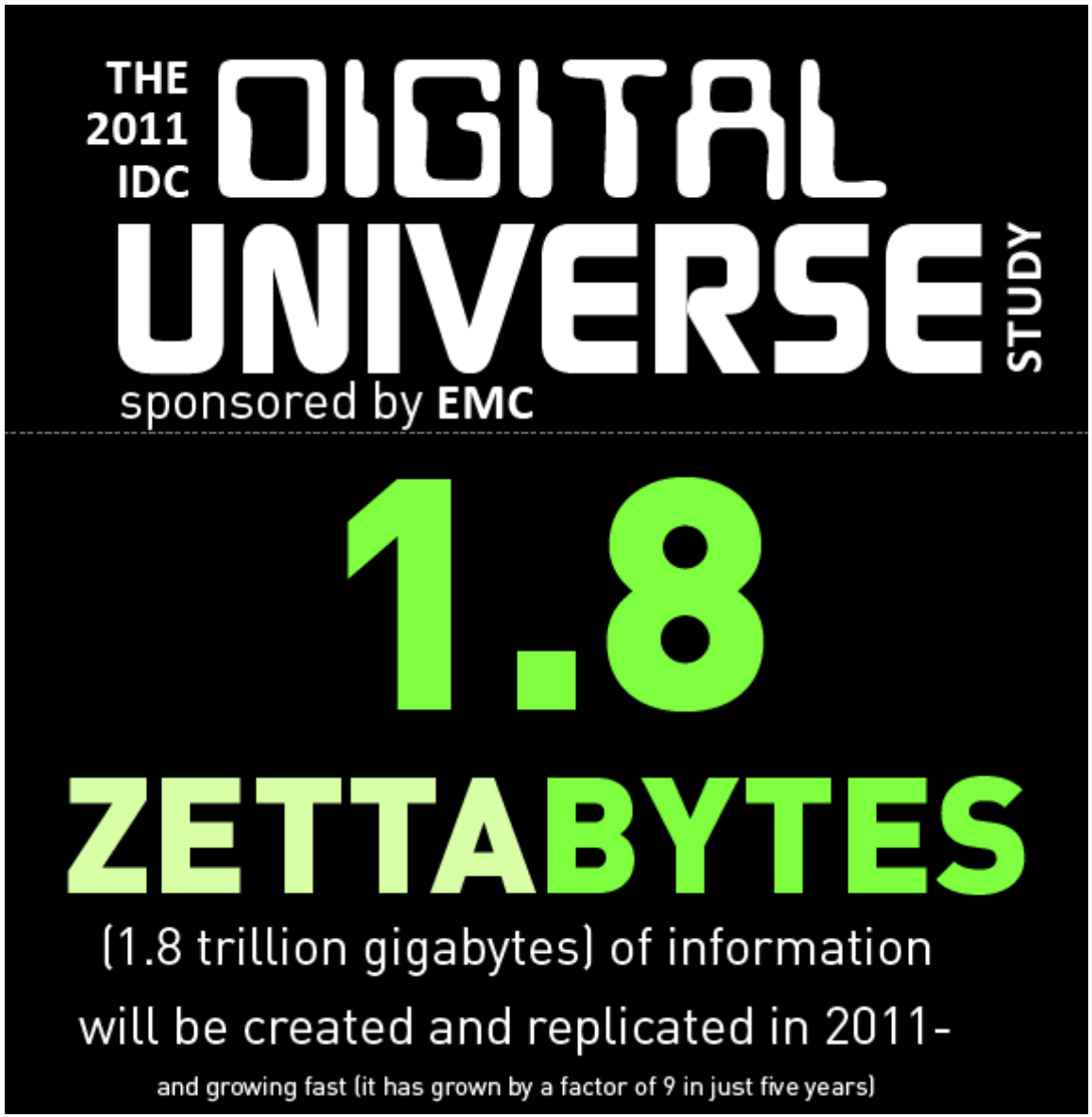


2009 年， IDC 預估
2011 年會成長十倍！
（相較 2006 年）

2006　161 EB
2007　281 EB
2010　988 EB（預測）
2011　1773 EB（預測）

出處：The Diverse and Exploding Digital Universe,
An Updated Forecast of Worldwide Information Growth Through 2011
March 2008, An IDC White Paper - sponsored by EMC
http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

Data expanded 2x each year !! 每年約略兩倍



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]

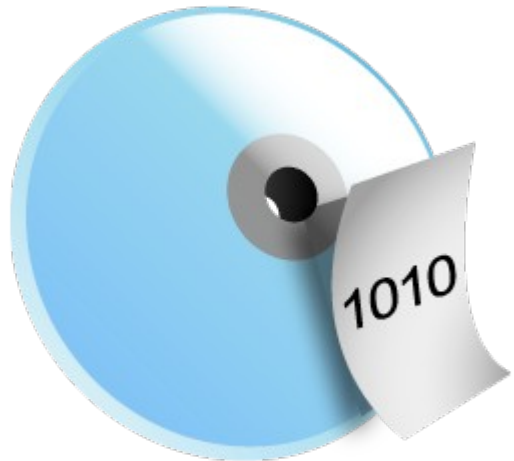
Big data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a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

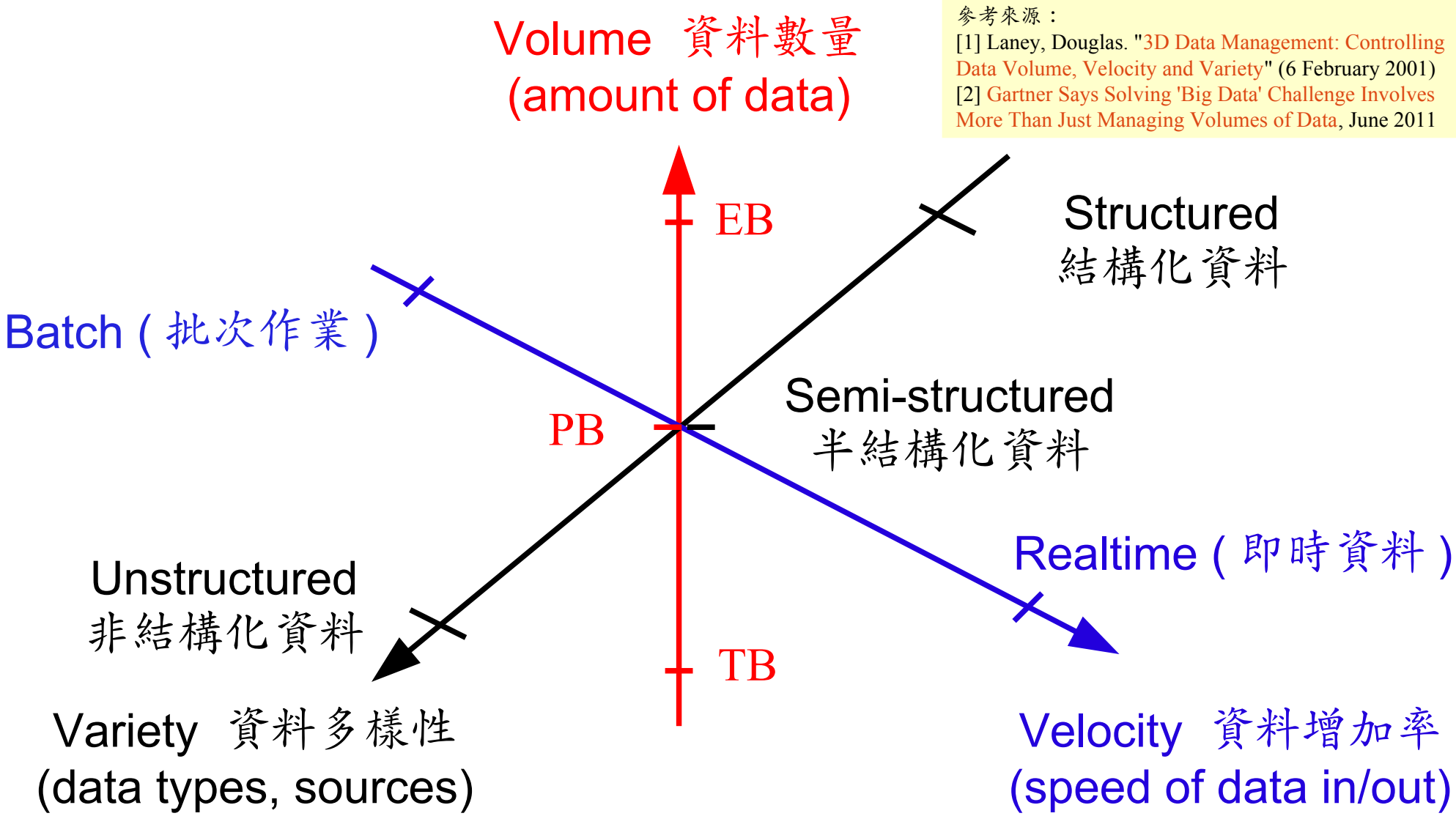


一個檔案，容量 100TB

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」

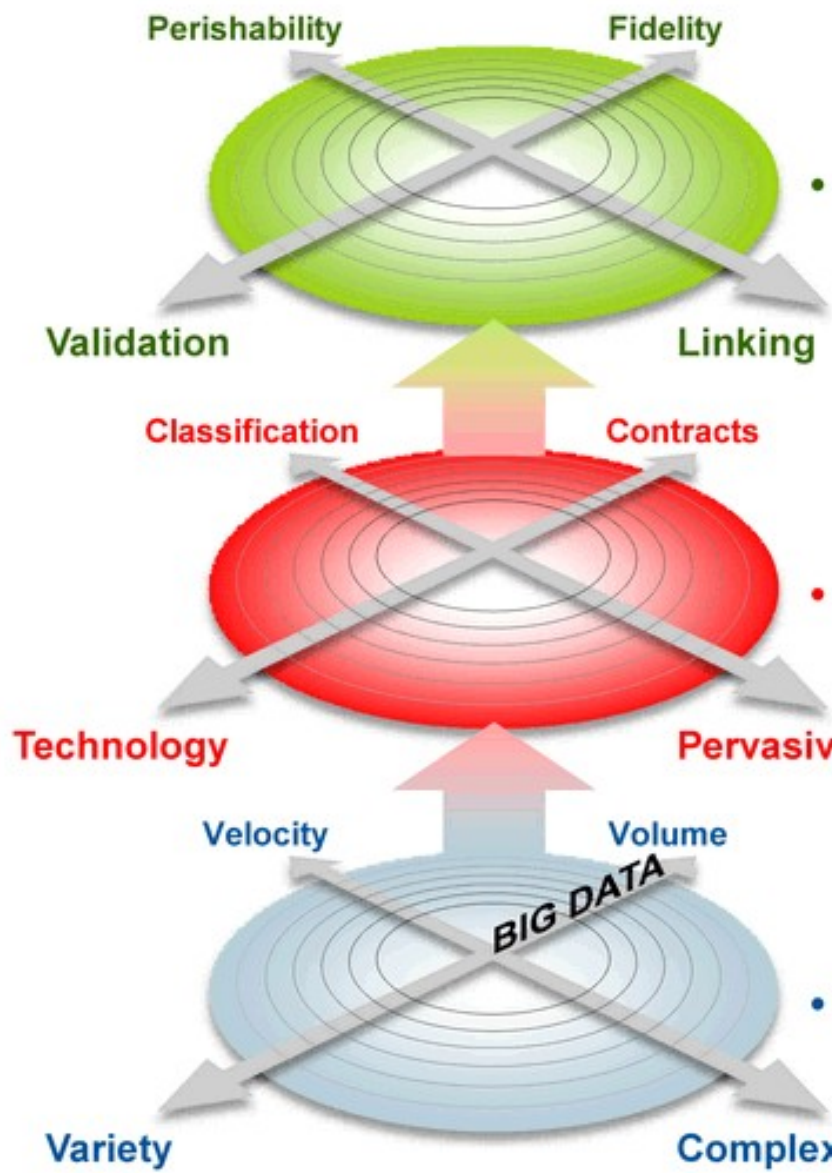
參考來源：
[1] Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety" (6 February 2001)
[2] Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data, June 2011



Six Dimensions of Big Data? 六個維度？

Source: Big Data, not Big Problems, http://www.talend.com/products-big-data/

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

Agenda 演講大綱

What is Big Data ? 何謂海量資料

Why should we care? 爲何需要關切

資料 Data → 知識 Knowledge → 智慧 Wisdom

Why we call it " SMART " !!
智慧打哪兒來？！

Smart Phone
智慧手機

Smart Car
智慧車輛

Smart Grid
智慧電網

SMART
哪裡長
智慧了？

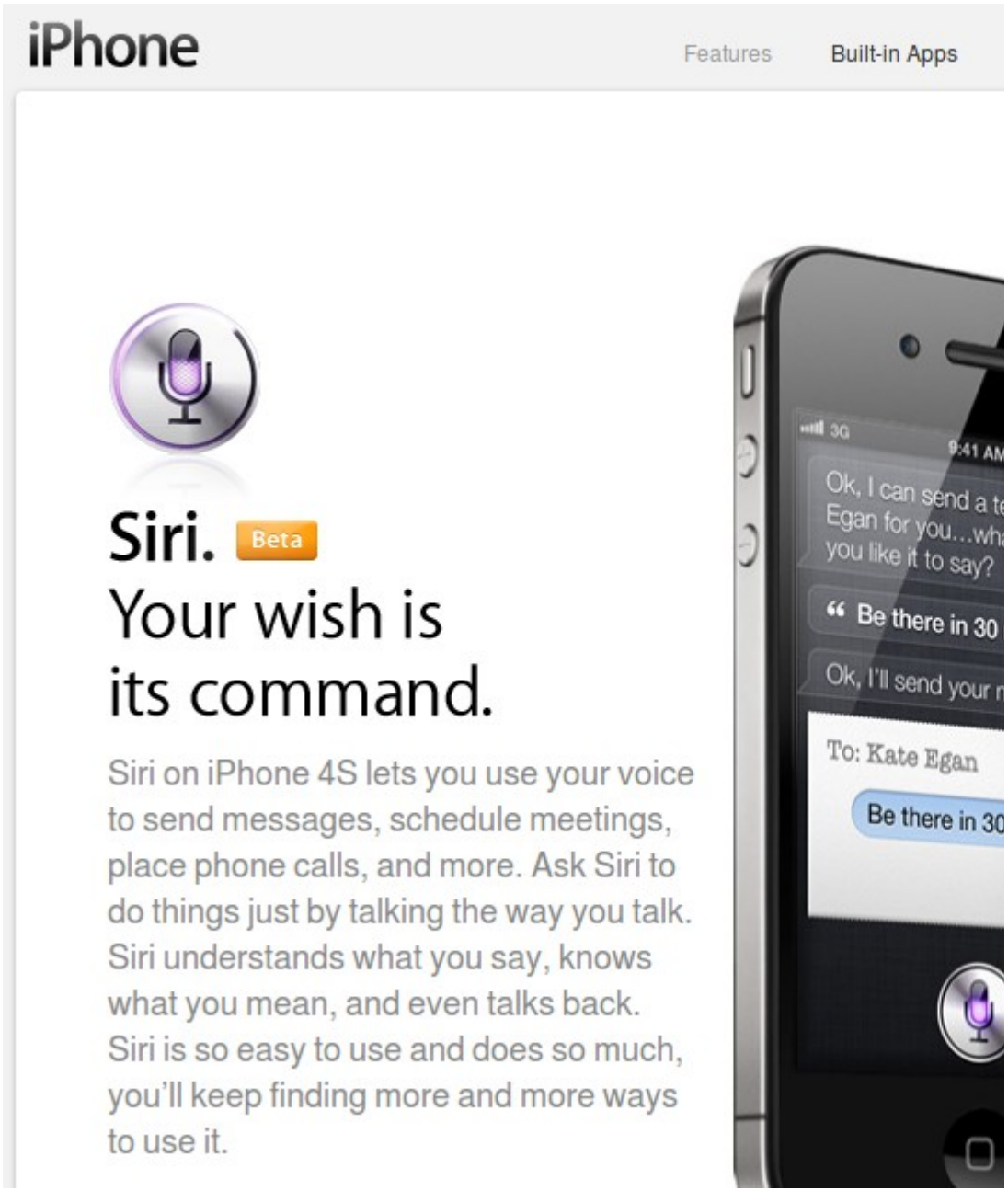
Smart City
智慧城市

Smart Home
智慧家庭

Smart Meter
智慧電錶

資料 Data → 知識 Knowledge → 智慧 Wisdom

Can Machine understand You? 讓機器更懂你？



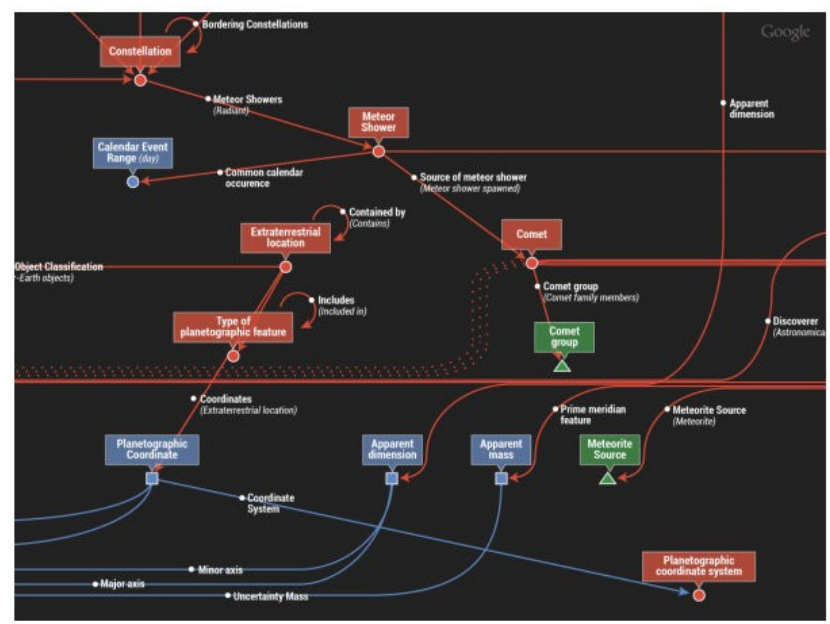
Google將發展「人工智慧」　永久改變搜尋引擎

2012年02月15日 00:11

點評：*超級阿斯拉，衝啊！（阿斯拉：好的，隼人！）*

記者黃郁棋／綜合報導

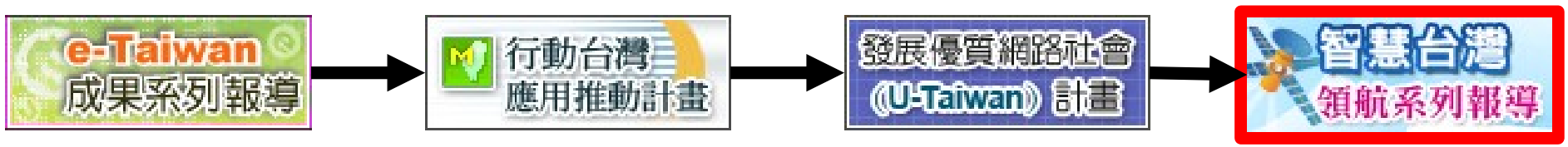
每個人都在猜，下一波網路革命是什麼？每個人都在猜，未來的世界會如何運作？Google的資深副總Amit Singhai透露了一點訊息。「Google正努力從『單字』層面進展到『意義』層面，未來搜尋引擎提供的不只是關鍵字搜尋，搜尋引擎甚至會『明白』你到底要什麼。」



▲Google未來將會朝「人工智慧」前進。(圖／取自mashable.com)

http://www.ettoday.net/news/20120215/25085.htm

Evolution of Software / Service
軟體演化勢必走向『智能化』



實體 Physical	單機版 個人使用 Personal Software	網路版 多人共享 Share Service Software	行動版 隨時存取 Mobile Cloud Service
信箱 Mailbox	E-Mail	Web Mail	Mobile Mail
電視 TV	電視盒 Setop Box	Web TV Ex. Youtube	Mobile TV
打字機 Typer Writer	Office	Google Docs	M-Office
電話 Telephone	數位電話 PBX	Skype	Flash Wengo
佈告欄 Bullet Borad	電子佈告欄 BBS	部落格 Blog	微網誌 Twitter

The wisdom of Clouds (Crowds)
雲端序曲：雲端的智慧始終來自於群眾的智慧

2006 年 8 月 9 日

Google 執行長施密特（Eric Schmidt）於SES '06會議中首次使用「雲端運算（Cloud Computing）」來形容無所不在的網路服務

2006 年 8 月 24 日

Amazon 以 Elastic Compute Cloud 命名其虛擬運算資源服務

Source: http://www.cnet.co.uk/i/c/blg/cat/software/cloudcomputing/clouds1.jpg

Data is the source of Wisdom !!
用雲掌握資料，加以分析，形成智能給端用

雲

資料中心
提供服務

雲端設計新思維：端的智能來自於雲的服務

Devices share the wisdom of Cloud

端

各類裝置
存取服務

Agenda 演講大綱

- What is Big Data ? 何謂海量資料
- Why should we care? 爲何需要關切
- **When** to deploy it ? 何時導入技術

基礎建設 IaaS → 分析平台 PaaS → 智慧服務 SaaS

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

Target Market is S.M.B.
主要客户爲中小企業

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

Hybrid Cloud

以大型企業爲主要客户
Enterprise is key market

私有雲端
Private Cloud

Community Cloud
社群雲端

Academia 學術爲主

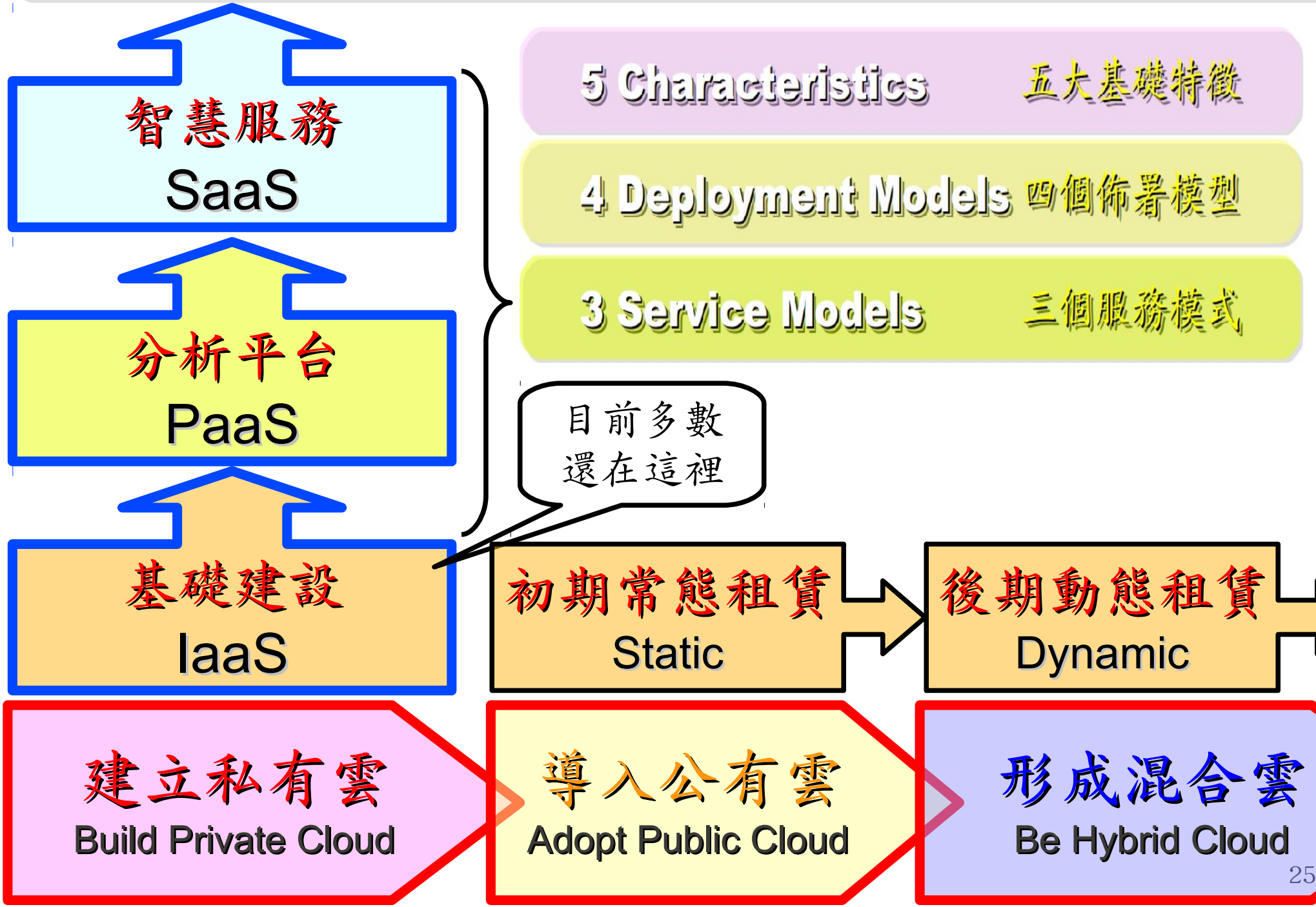
3 Service Models of Cloud Computing
雲端運算的三種服務模式（市場區隔）

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

Roadmap to build Your Enterprise Cloud !!
佈建企業雲端的時程規劃



Agenda 演講大綱

What is Big Data ? 何謂海量資料

Why should we care? 爲何需要關切

When to deploy it ? 何時導入技術

How to handle it ? 三大因應策略

儲存虛擬化 Dedup. → 資料安全 Security → 智慧服務 SaaS

Three Solutions !! 三種服務模式 vs. 三類因應對策

SaaS Software as a Service 軟體即服務	**Web 2.0** 網頁服務	(A) 提供 API 介面 (B) 分散式資料庫
PaaS Platform as a Service 平台即服務	**Data Analysis** 資料分析	(A) 資料整合 (B) 資料探勘
IaaS Infrastructure as a Service 架構即服務	**Virtualization** 虛擬化技術	(A) 儲存虛擬化 (B) 備援與加密

What is Virtualization ??
虛擬化技術有哪些呢 ??

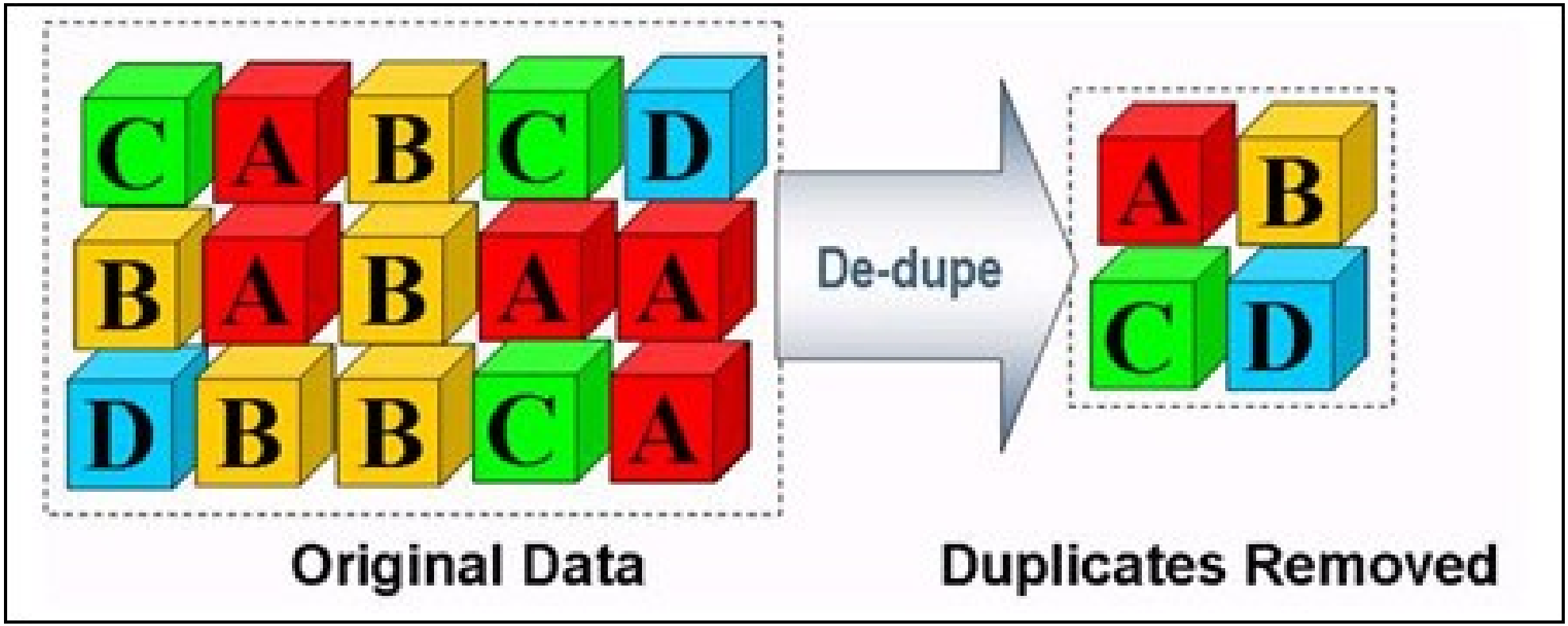
- Application Virtualization 應用程式虛擬化
- Desktop Virtualization / Client Virtualization 桌面虛擬化
- Presentation Virtualization 顯示虛擬化
- OS-level Virtualization 作業系統虛擬化
- Network Virtualization 網路虛擬化
- Storage Virtualization 儲存虛擬化

Database Virtualization 資料庫虛擬化

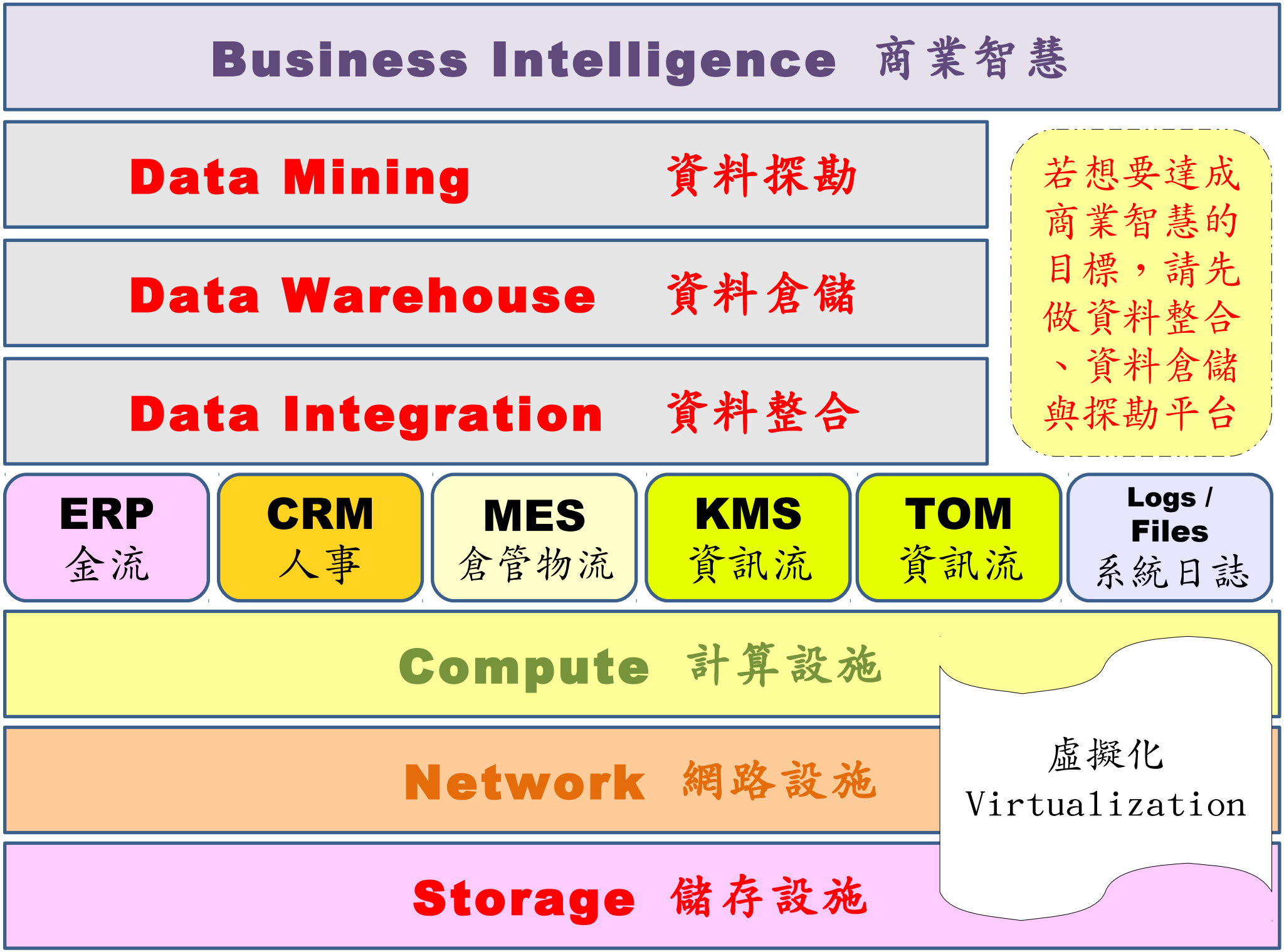
Data Virtualization 資料虛擬化

Source: http://en.wikipedia.org/wiki/Virtualization

Deduplication? 去除重複儲存的資料？



- 資料整合爲跨單位整合的第一步 !!
- 商業硬體方案： EMC、 NetApp
- 自由軟體方案：
  - ZFS、 Lessfs、 SDFS...


Business Intelligence 商業智慧
Data Mining 資料探勘
Data Warehouse 資料倉儲
Data Integration 資料整合
若想要達成商業智慧的目標，請先做資料整合、資料倉儲與探勘平台
ERP 金流
CRM 人事
MES 倉管物流
KMS 資訊流
TOM 資訊流
Logs / Files 系統日誌
Compute 計算設施
Network 網路設施
Storage 儲存設施
虛擬化 Virtualization

Data Integration ? 怎麼做資料整合？

Source : http://en.wikipedia.org/wiki/Data_integration

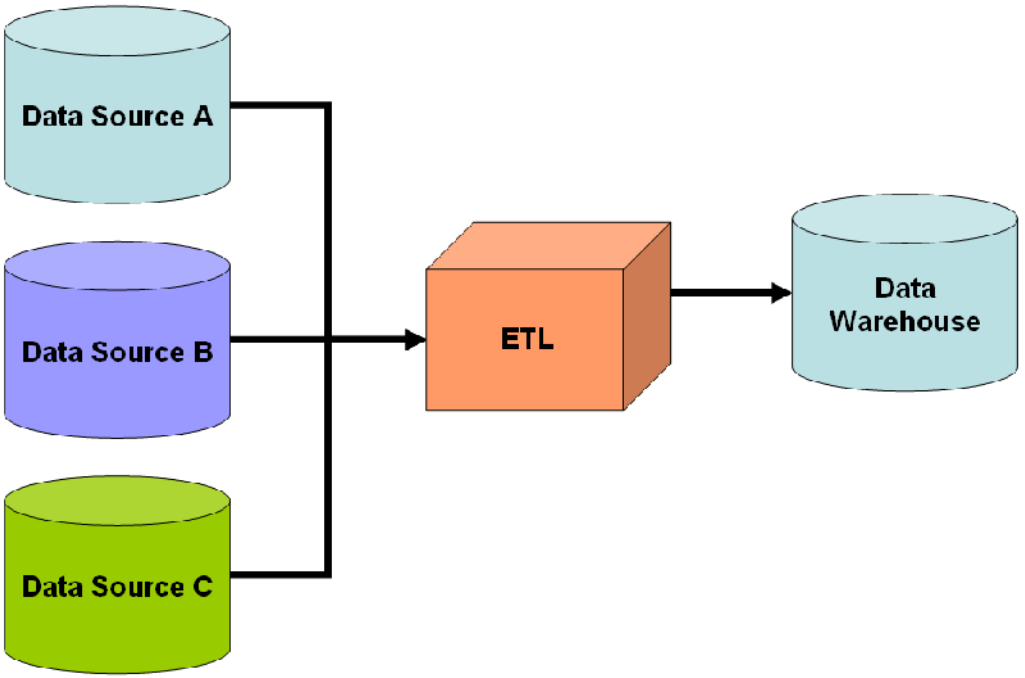


Figure 1: Simple schematic for a data warehouse. The ETL process extracts information from the source databases, transforms it and then loads it into the data warehouse.

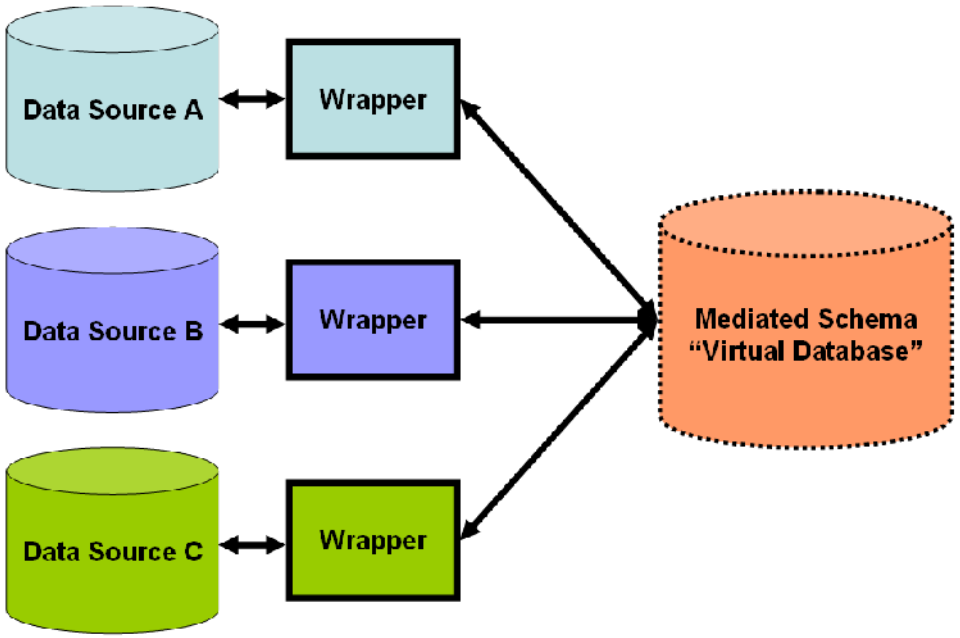


Figure 2: Simple schematic for a data-integration solution. A system designer constructs a mediated schema against which users can run queries. The virtual database interfaces with the source databases via wrapper code if required.

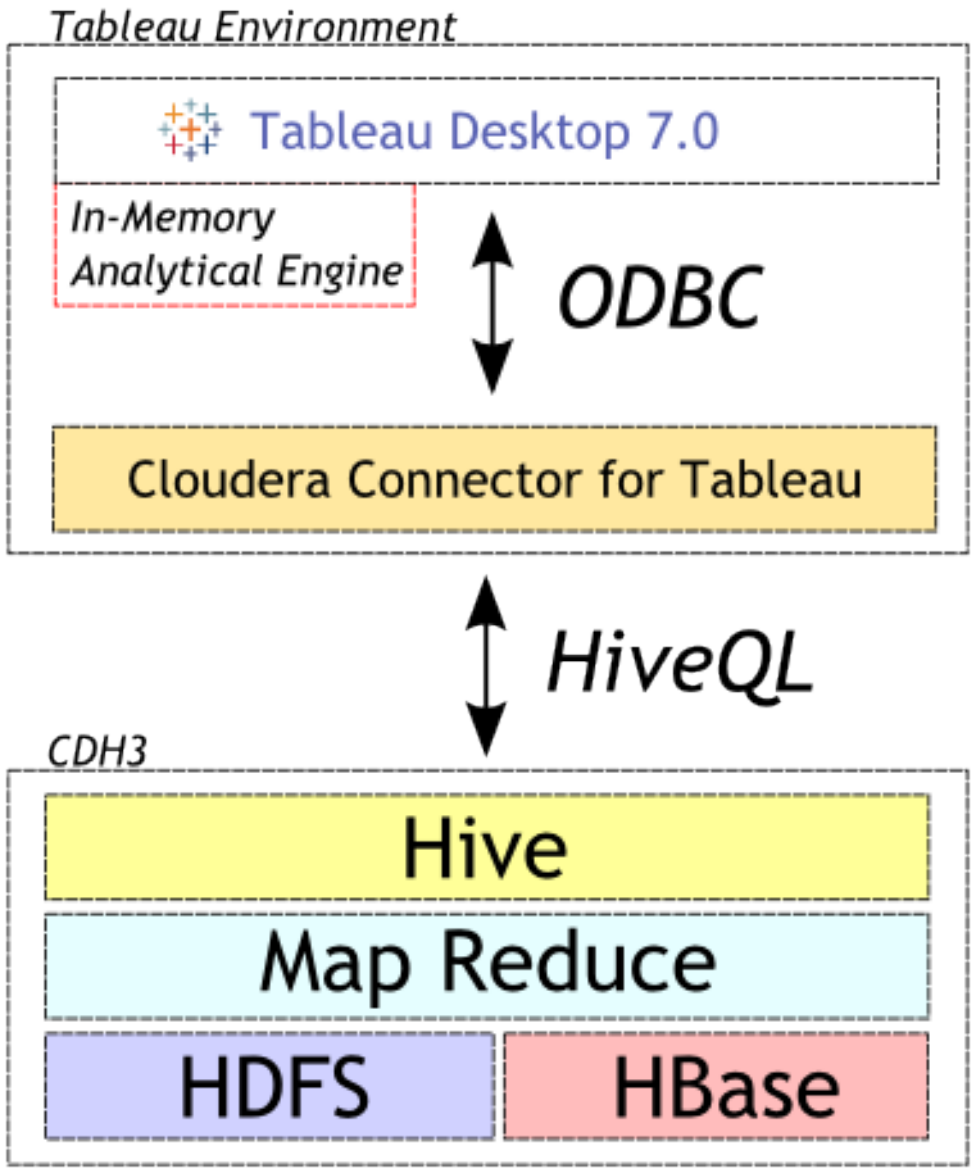
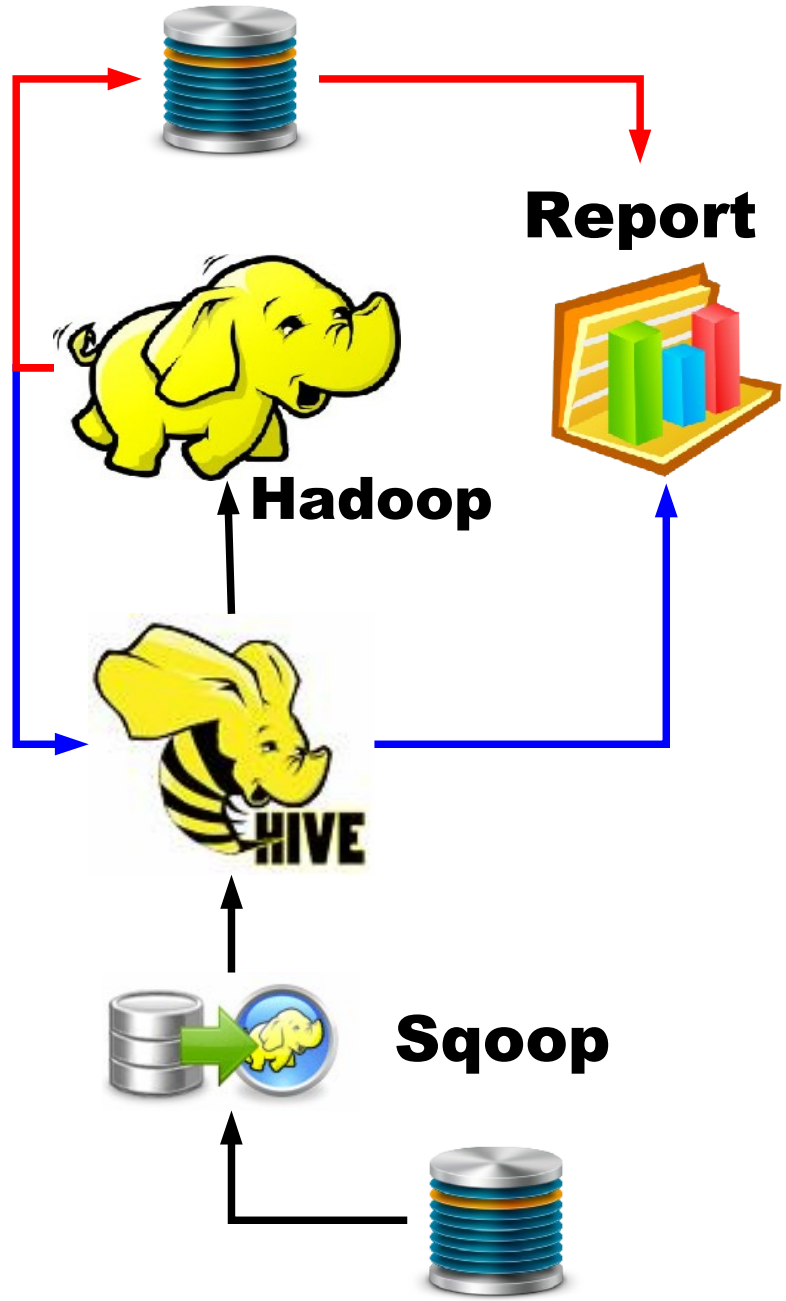


Tableau Environment
Tableau Desktop 7.0
In-Memory Analytical Engine
ODBC
Cloudera Connector for Tableau
HiveQL
CDH3
Hive
Map Reduce
HDFS
HBase


Report
Hadoop
HIVE
Sqoop

Agenda 演講大綱

- What is Big Data ? 何謂海量資料
- Why should we care? 爲何需要關切
- When to deploy it ? 何時導入技術
- How to handle it ? 三大因應策略
- **Who** is key player ? 誰是成功關鍵

Data Scientist !! 資料科學家！！

Data scientist: The hot new gig in tech

By Michal Lev-Ram, writer September 6, 2011: 5:00 AM ET

Companies that want to make sense of all their bits and bytes are hiring so-called data scientists - if they can find any.

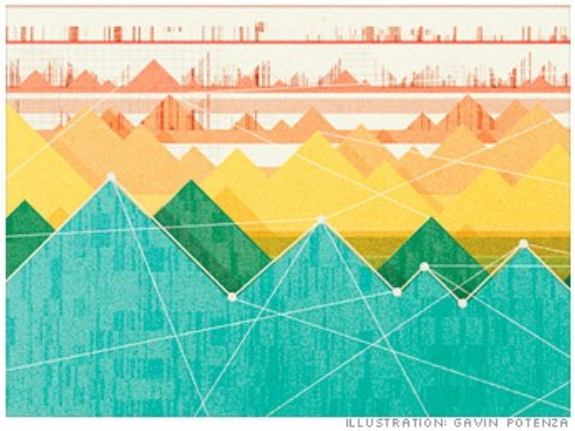


ILLUSTRATION: GAVIN POTENZA

FORTUNE -- The unemployment rate in the U.S. continues to be abysmal (9.1% in July), but the tech world has spawned a new kind of highly skilled, nerdy-cool job that companies are scrambling to fill: data scientist.

會「統計」的人照過來！

財星雜誌(FORTUNE)等均報導今年最熱門的職缺是「資料科學家」！

Source :
http://tech.fortune.cnn.com/2011/09/06/data-scientist-the-hot-new-gig-in-tech/

The way toward Business Inteligence
通往商業智慧的漫長道路

Business Inteligence	商業智慧
Data Mining	資料探勘
Data Warehouse	資料倉儲
Data Integration	資料整合
OS-level Virtualization	作業系統虛擬化
Network Virtualization	網路虛擬化
Storage Virtualization	儲存虛擬化

pentaho
POWERFUL ANALYTICS MADE EASY™

cloudera

HITACHI
Inspire the Next

talend*
open integration solutions

vmware

CISCO

EMC²

NetApp

What we learn today ?

WHAT	海量資料泛指介於TB到PB之間的資料集!! few dozen TeraBytes to PetaBytes in single data set !!
WHY	透過統計分析人類的資料，讓機器更有智慧~ Make Machine Smart !
WHEN	先建私有雲的虛擬化架構，然後才建分析平台 Build Private IaaS first, then PaaS !!
HOW	儲存虛擬化、資料備援與加密、分析平台 Deduplication , Data Recovery / Encryption, Data Analysis
WHO	資料科學家！接下來的講者都是佼佼者！ Data Scientist ! Next Speaker are all Key Players ....

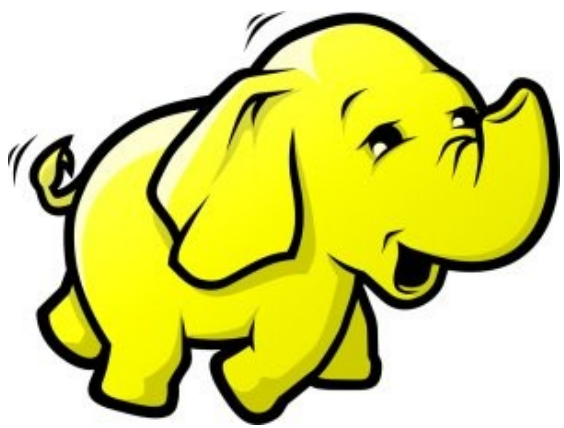
處理海量資料的資訊架構與關鍵技術

Technologies to build IT Stack for Big Data

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



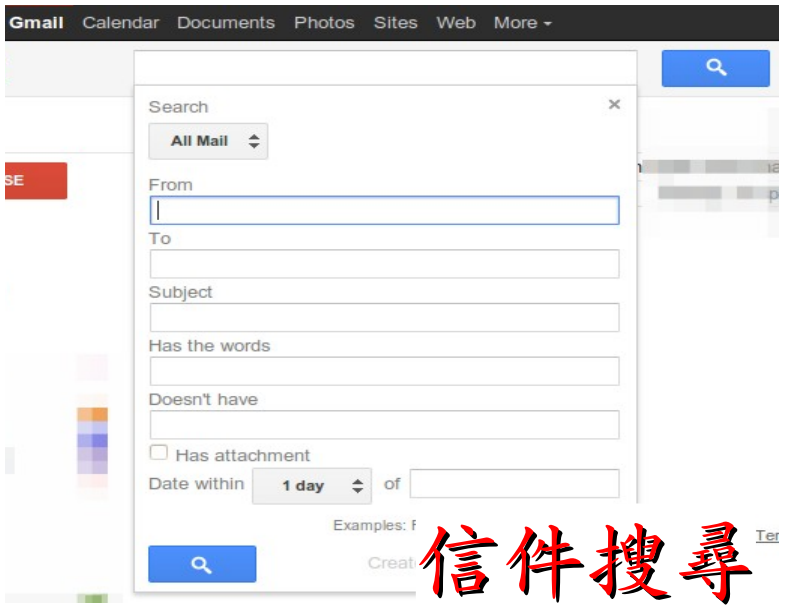
Hot Jobs in Big Data
從海量資料的熱門工作談起

Data Mining	資料探勘
Data Visualization	資料視覺化
Data Analysis	資料分析
Data Manipulation	資料操控
Data Discovery	資料鑑識

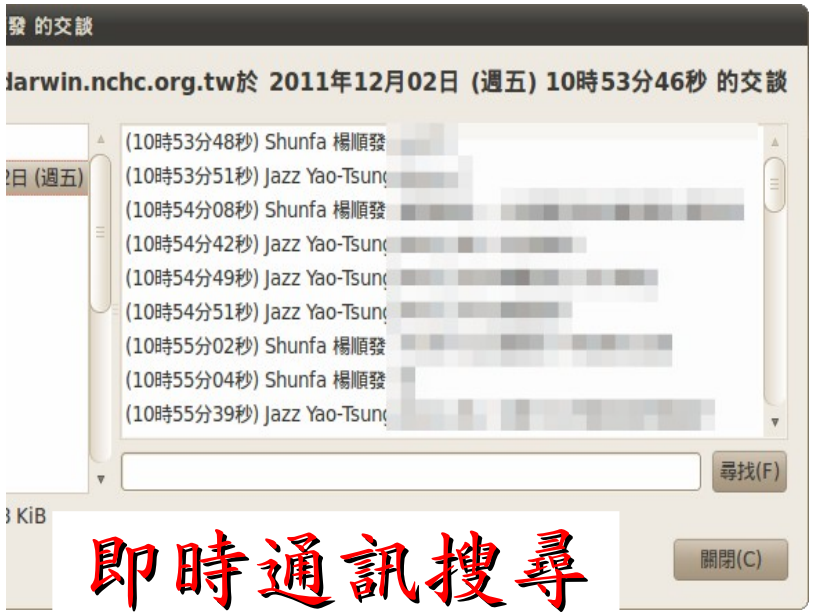
How to Get a Hot Job in Big Data, Dan Tynan, InfoWorld, March 19, 2012
出處：http://www.cio.com/article/print/702388

Applications of Data Mining
資料探勘的應用 - 搜尋引擎

檔案搜尋



信件搜尋

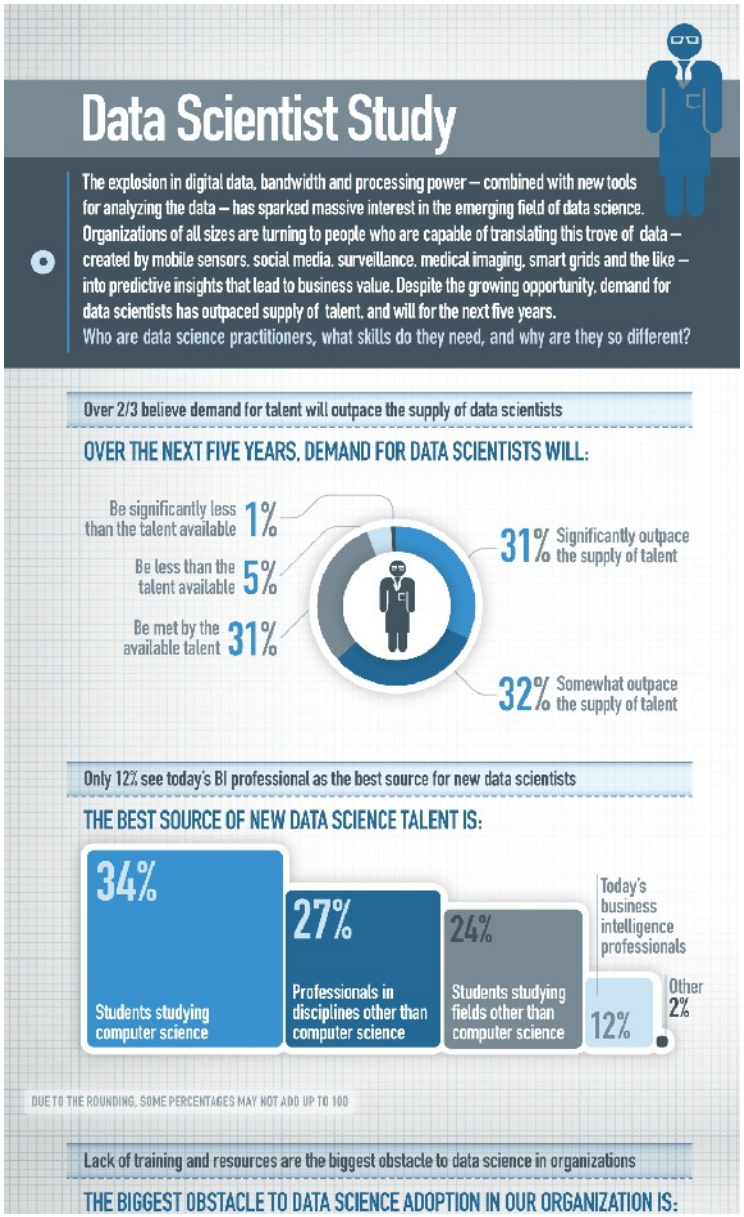


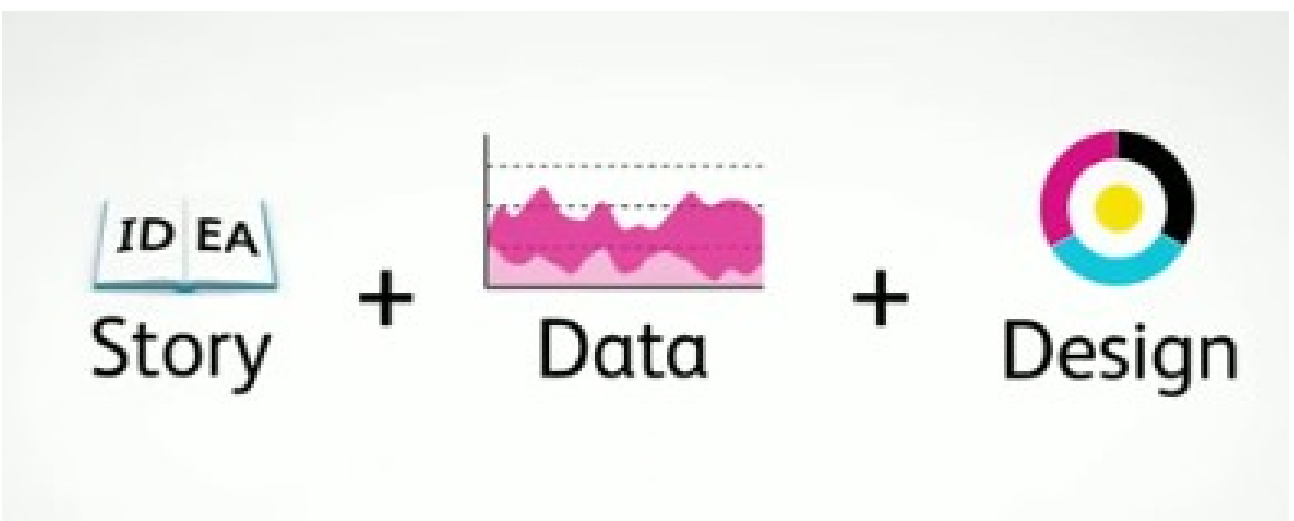
即時通訊搜尋

資料庫搜尋

「網頁搜尋」

Applications of Data Visualization
資料視覺化的應用 -Infographics





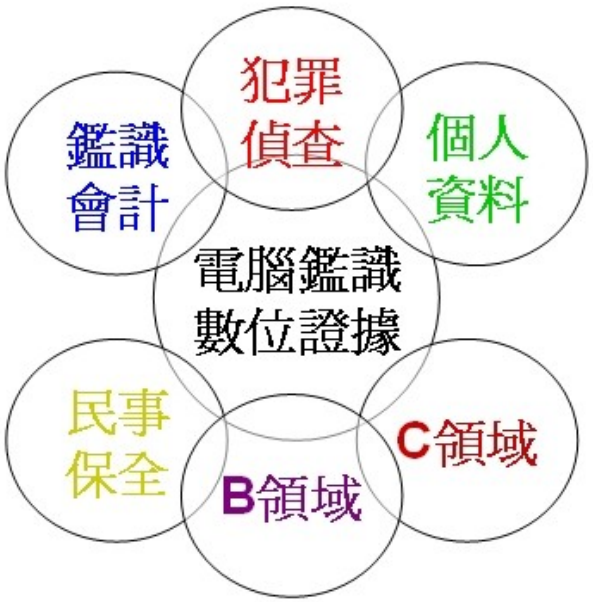
參考來源：未來「夯」職業：資料科學家 http://www.bnext.com.tw/print/article/id/21740
淺談超吸睛的資訊圖表 http://www.inside.com.tw/2011/04/13/infographics

Applications of Data Analysis
資料分析的應用 - 商業智慧 (BI)

圖片來源： http://bizcubic.com/wp-content/uploads/2010/01/business-intelligence-dashboard.jpg

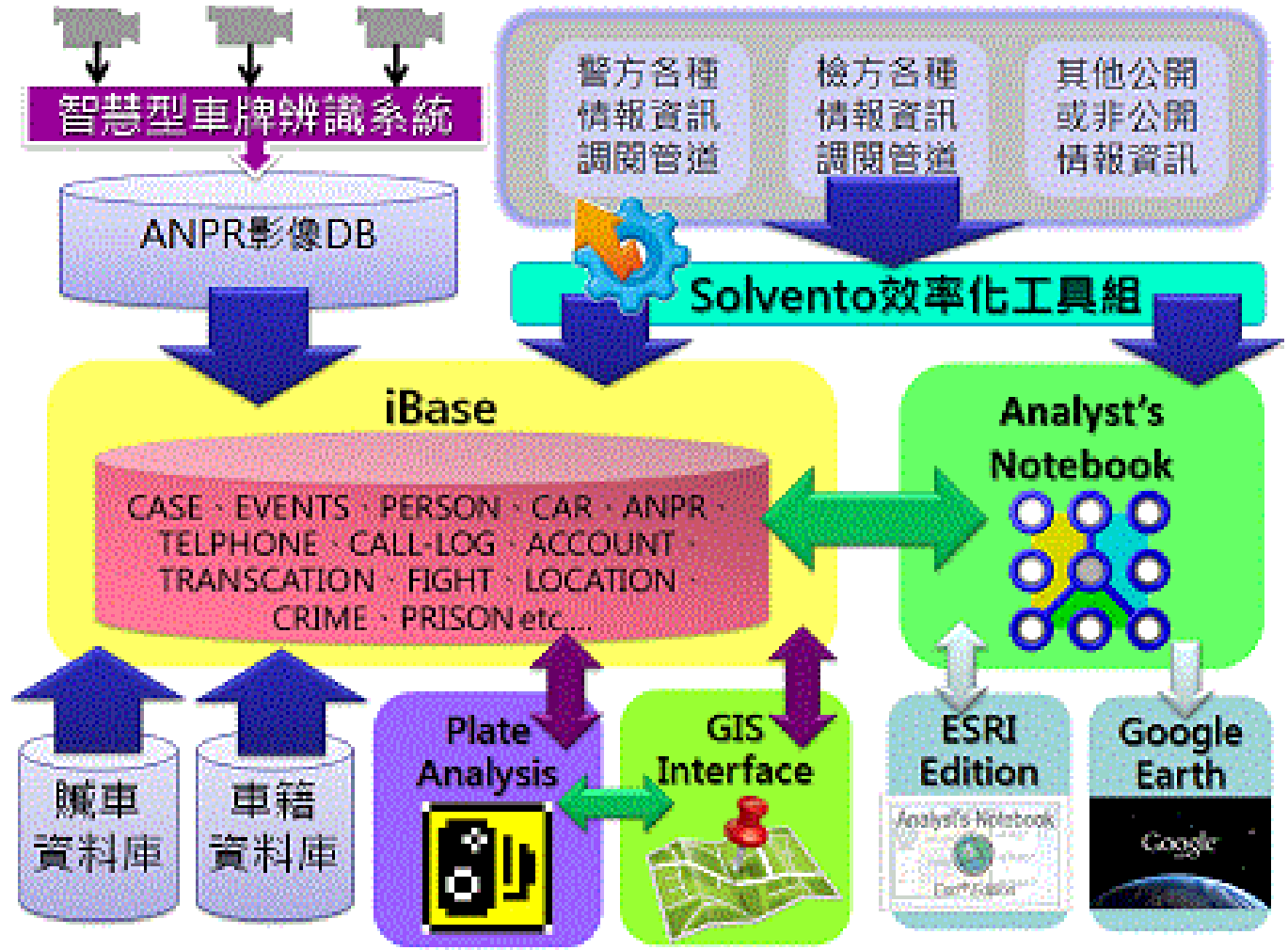
Applications of Data Discovery
數位鑑識 - 資訊與法律的結合

電腦鑑識&會計鑑識

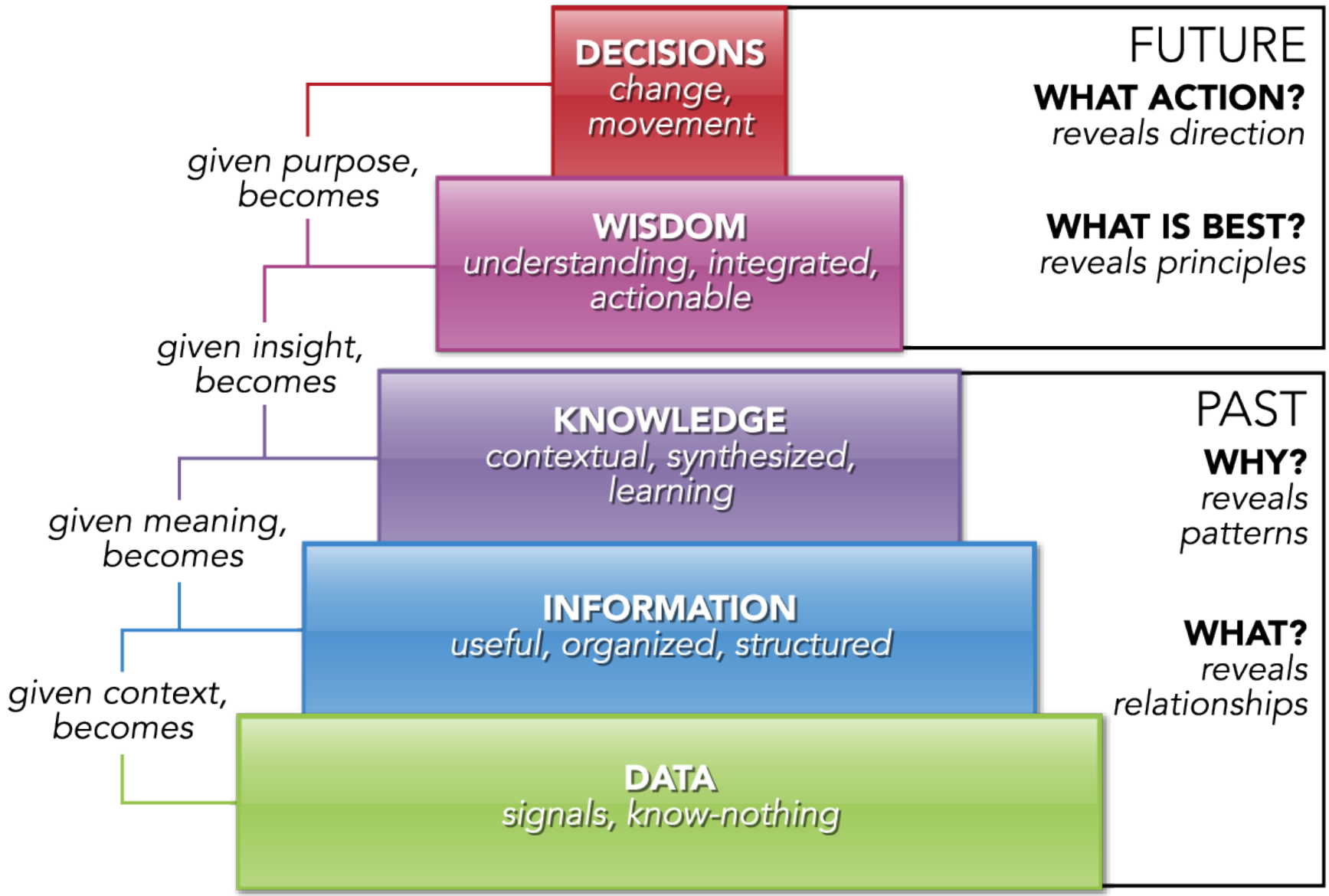


http://blog.udn.com/kf0630/6018593

http://www.solventosoft.com/upload/ANPR_02s.gif



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

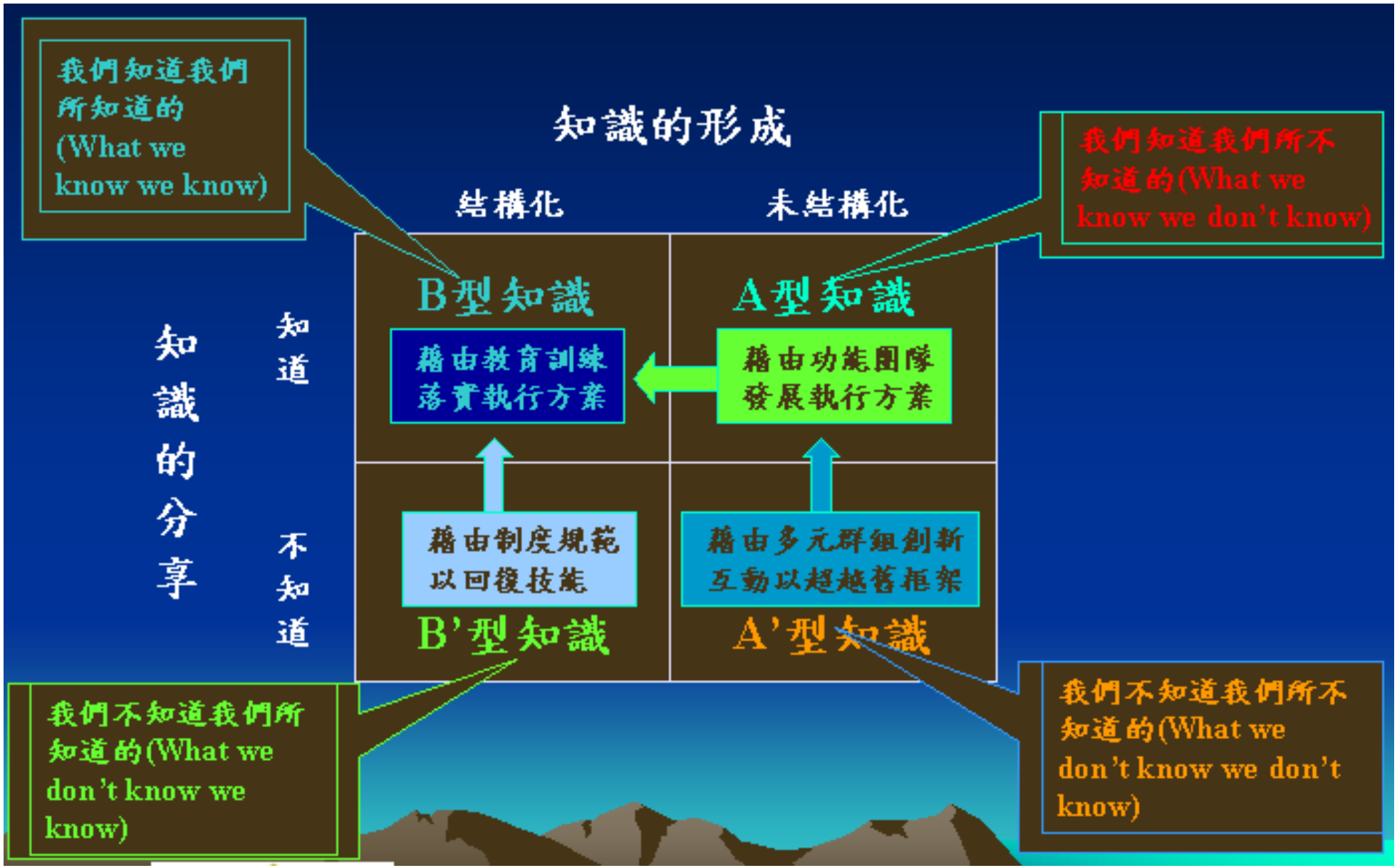


http://www.pursuantgroup.com/blog/tag/dikw-model/

Knowledge Spiral Model
知識螺旋模型（竹内弘高，1996）

出處： http://ecollege.nacs.gov.tw/nacs_file/upload/layoutDownload/130b575bacb00000b978.html

Type of Knowledge
知識的種類

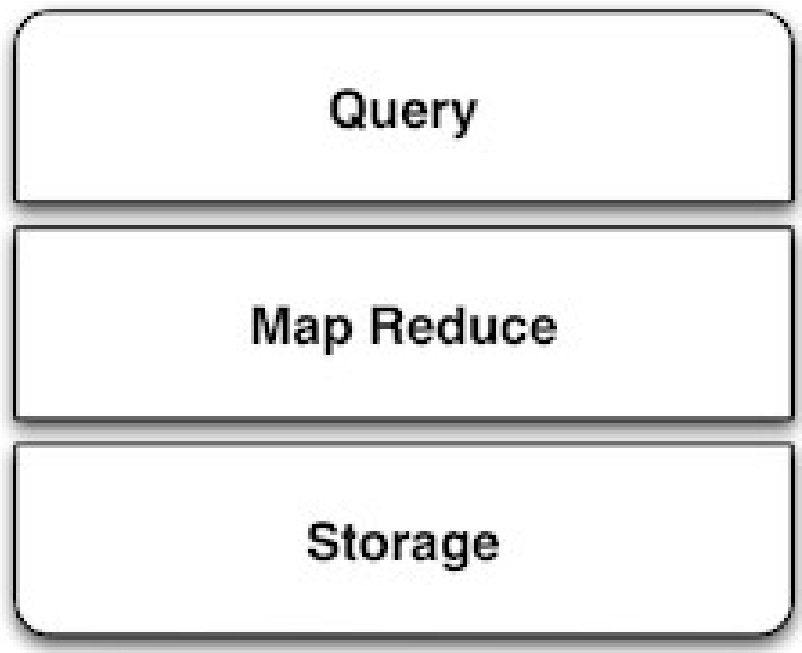


出處： http://www.goldenlamp.idv.tw/wiki/images/e/ef/Skm4.png

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

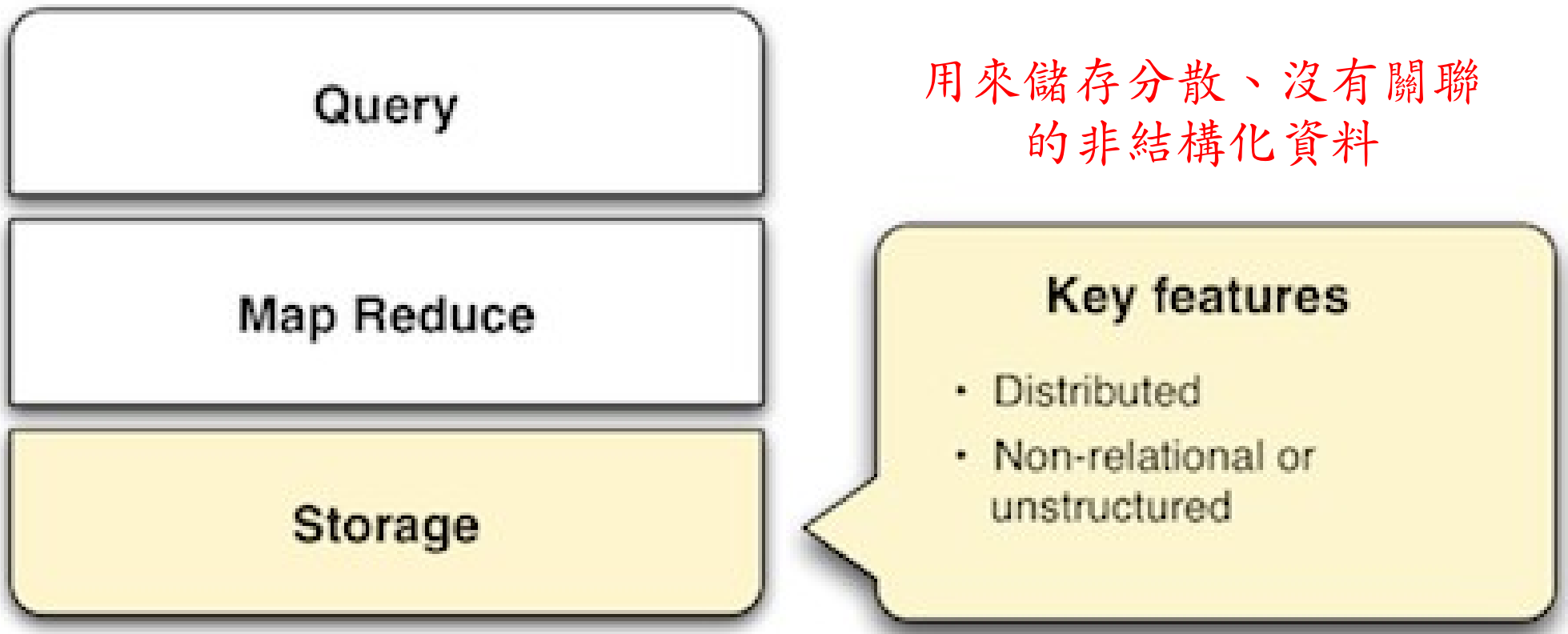
做網頁相關的人可能聽過 LAMP

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

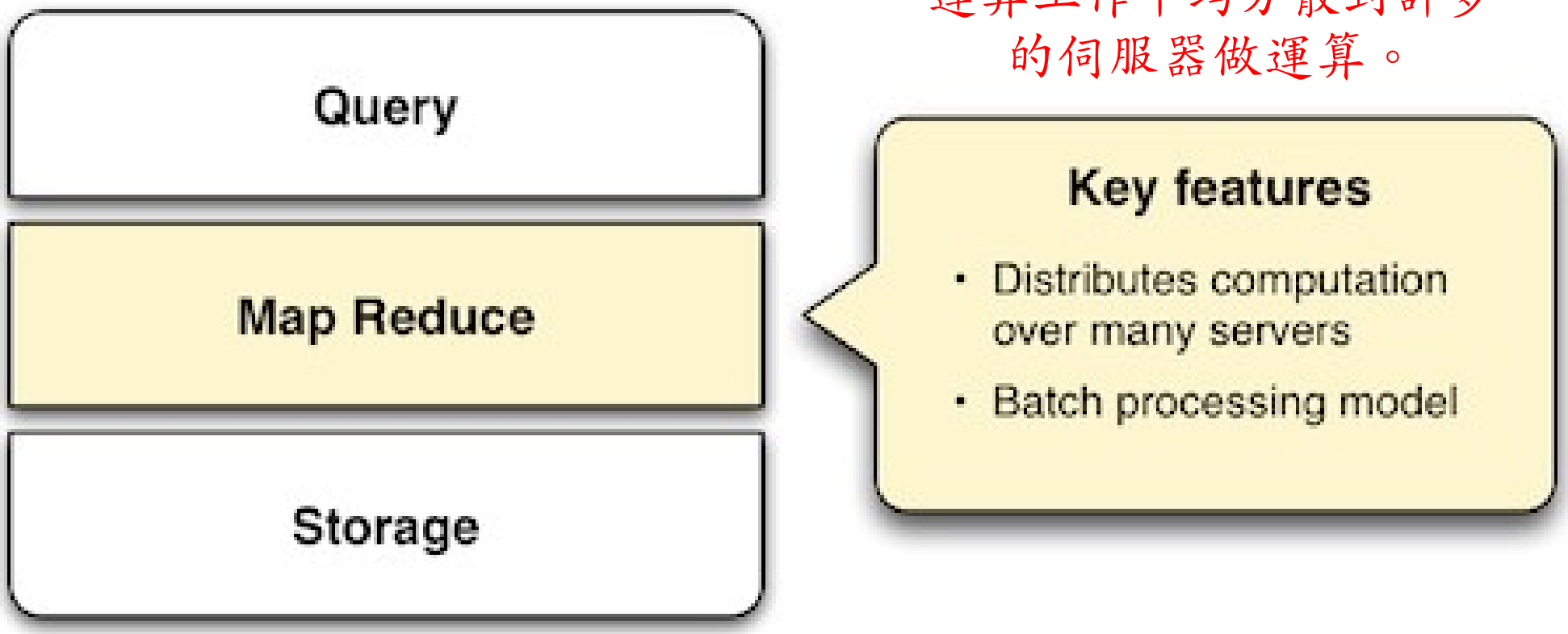


用來儲存分散、沒有關聯的非結構化資料

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

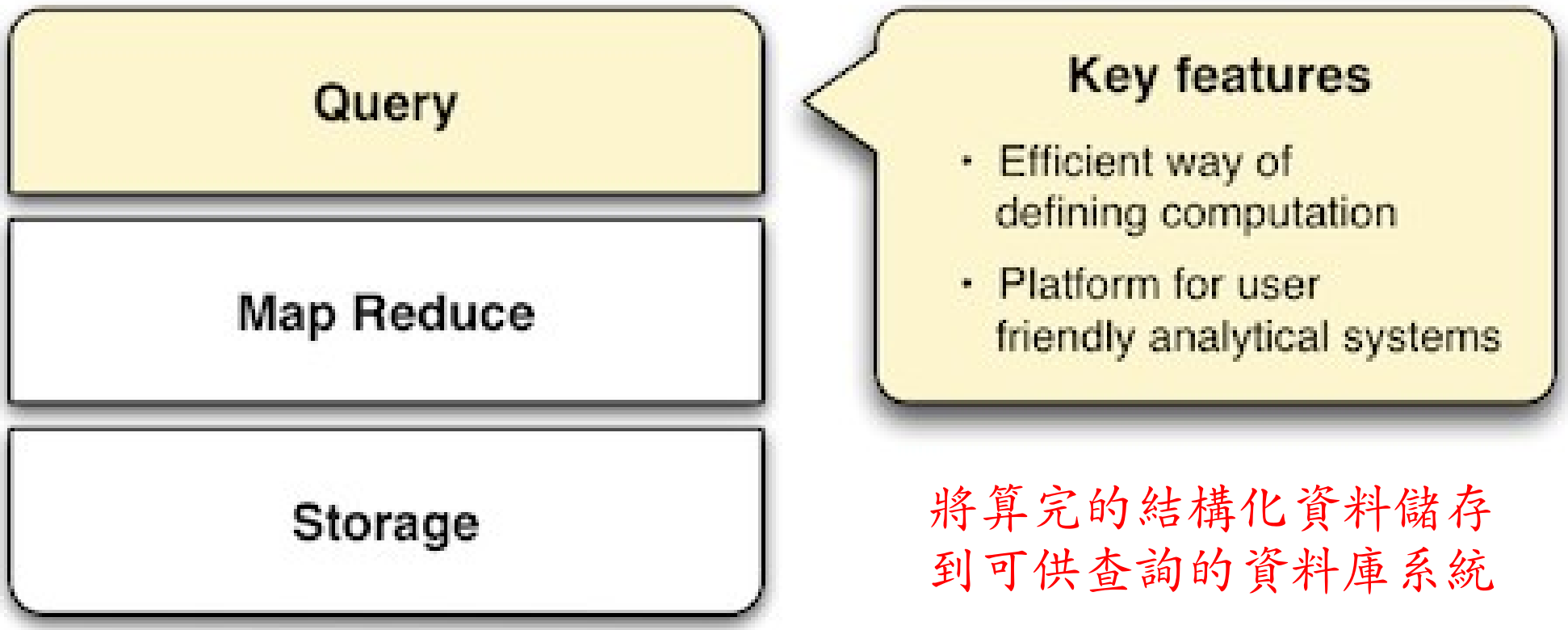
The SMAQ stack for big data
海量資料處理的資訊架構

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

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

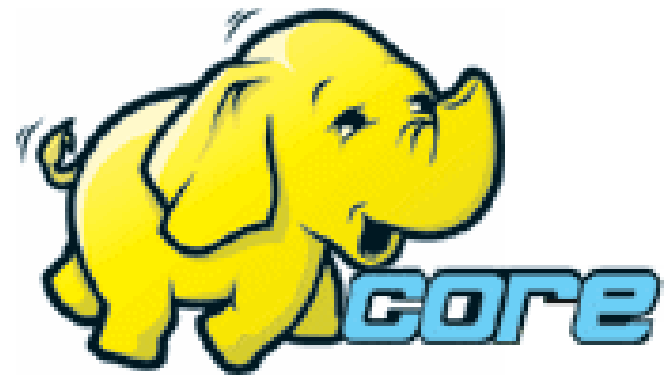
更多不同語言的 MapReduce API 實作：
http://trac.nchc.org.tw/grid/intertrac/wiki%3Ajazz/09-04-14%23MapReduce

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/

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

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**

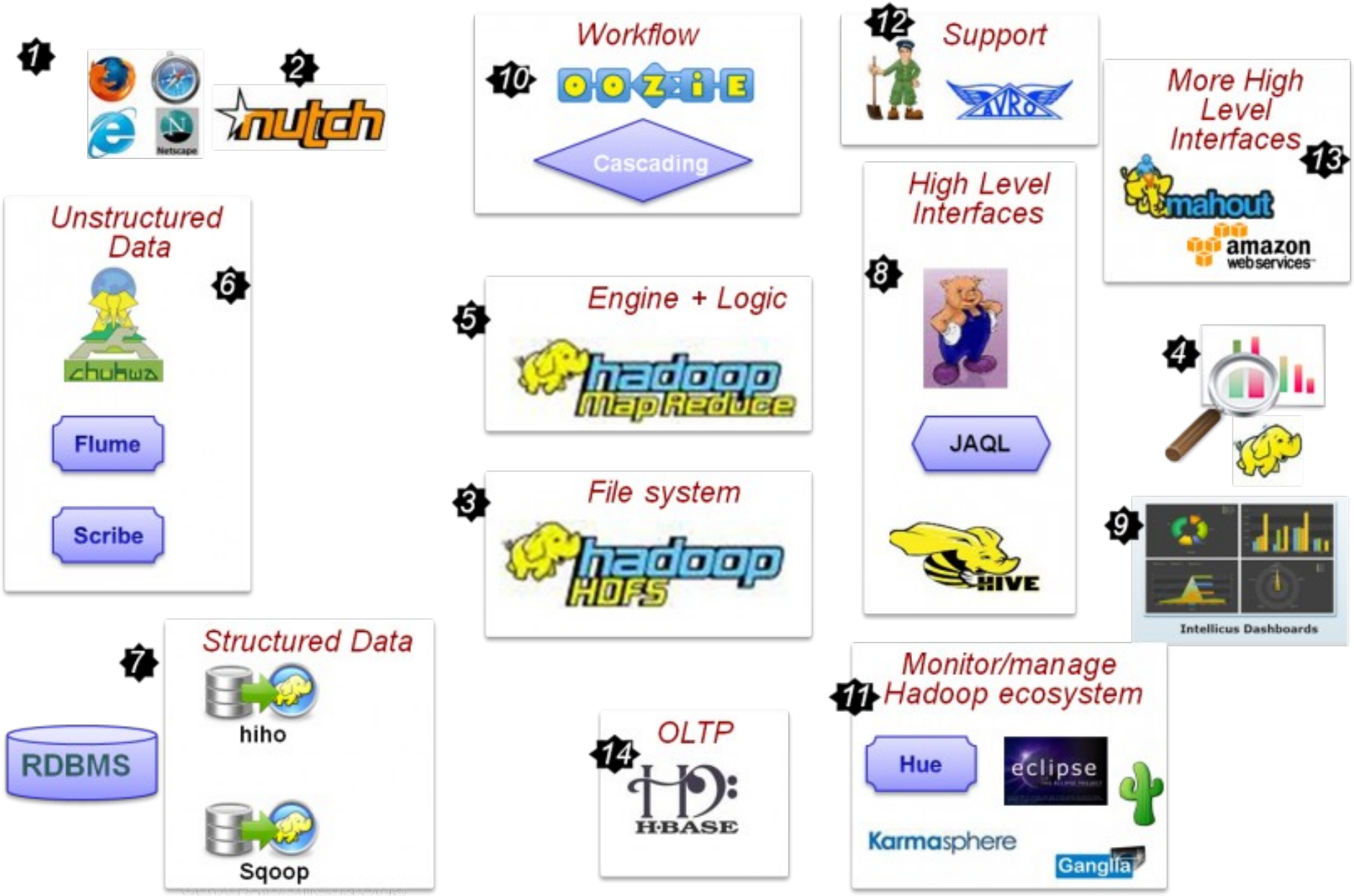
Sector-Sphere

National Center for Data Mining
University of Illinois at Chicago
UIC

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

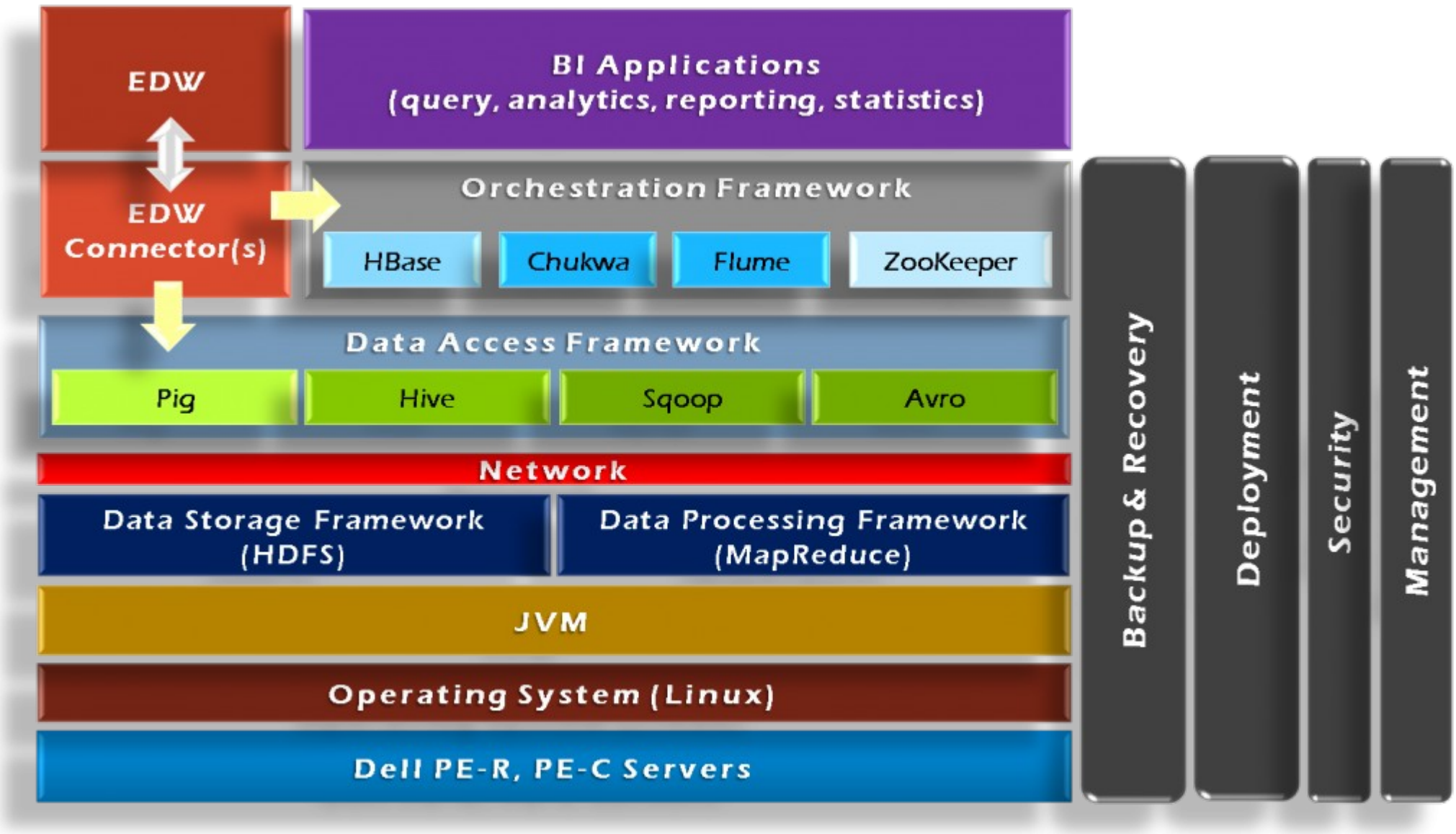
Why we choice Hadoop? Good Ecosystem!

豐富的生態系建構出處理海量資料的工具庫



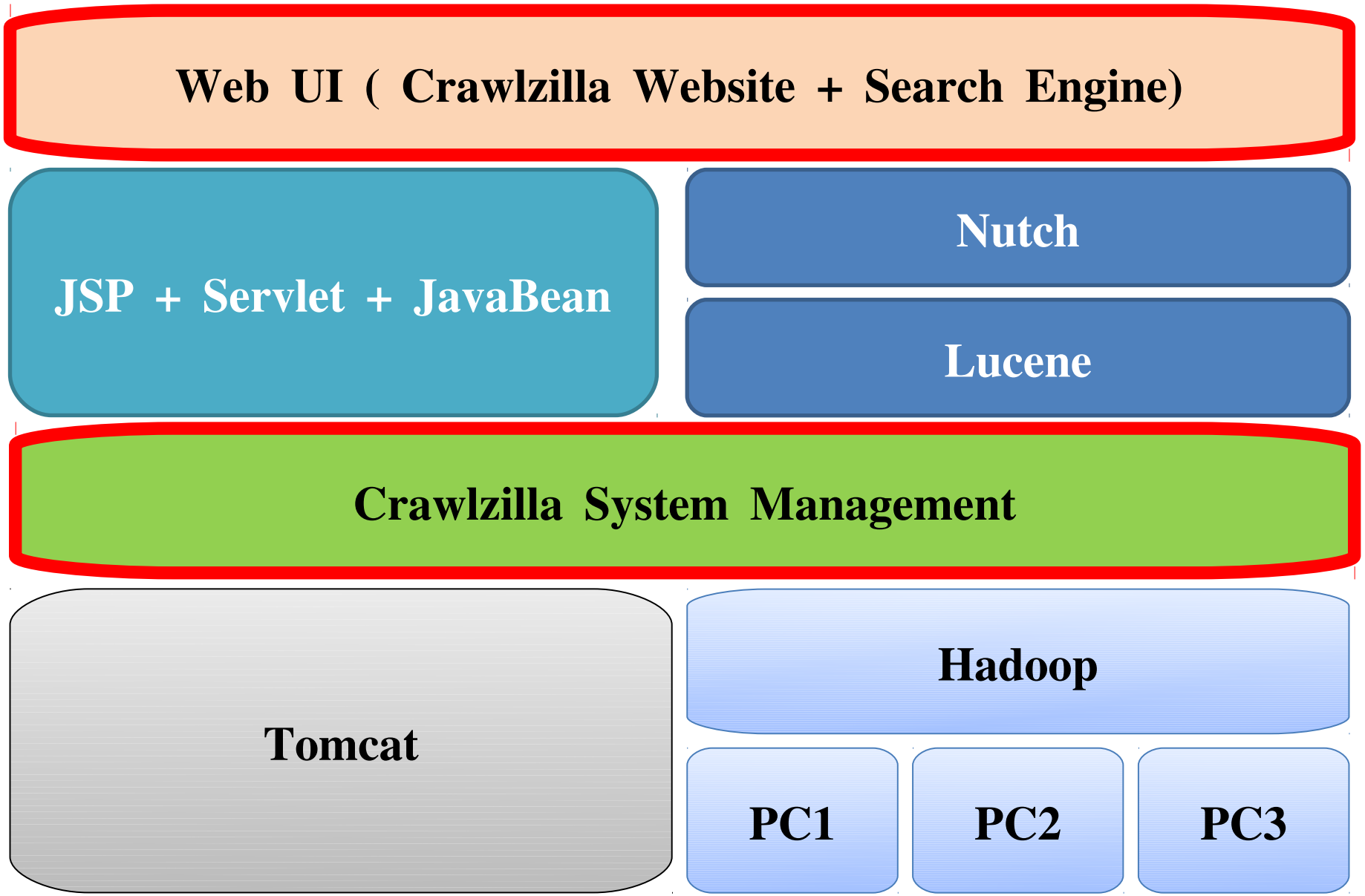
http://rationalintelligence.com/wp_log/?p=104

BI and EDW build on Hadoop Ecosystem
運用 Hadoop 生態系搭建資料倉儲與商業智慧分析

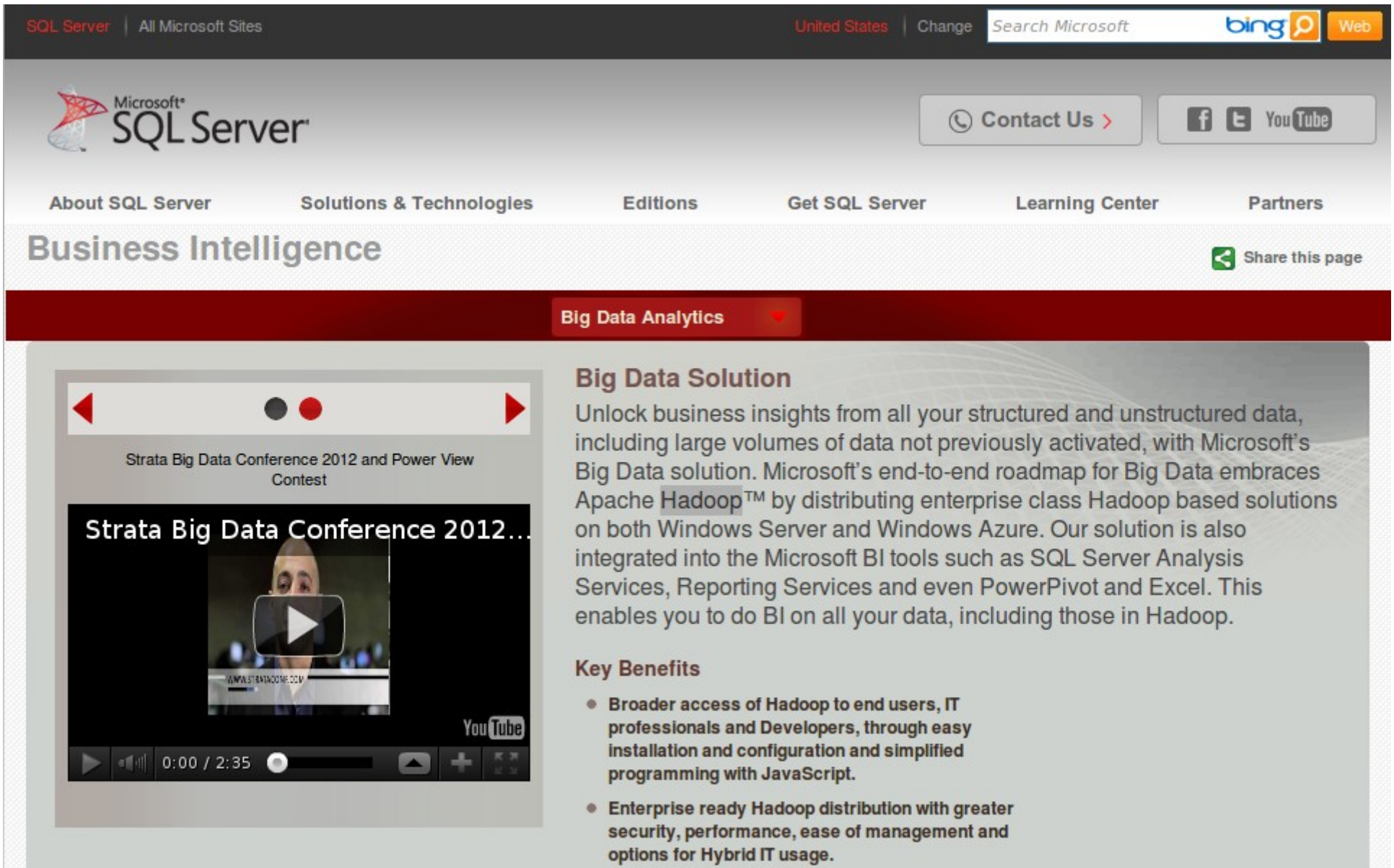


http://rationalintelligence.com/wp_log/?p=104

Build your own search engine, too
您也能用 Hadoop 搭建自己的搜尋引擎



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：分析駭客行爲，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

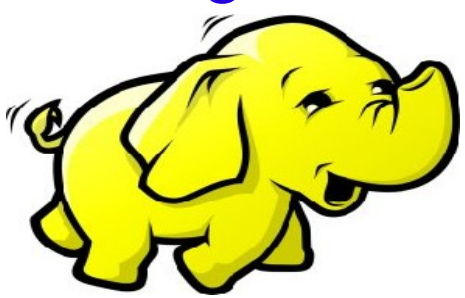
Hadoop 這套軟體的歷史源起 ... 2005~2006

- 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

Lucene - Java / LUCENE-1

lock files don't work in JDK 1.1

Log In

Views

Details

Type:	Bug	Status:	Closed
Priority:	Major	Resolution:	Fixed
Affects Version/s:	1.2	Fix Version/s:	None
Component/s:	core/store		
Labels:	None		
Environment:	Operating System: All		

People

Assignee: Lucene Developers

Reporter: Doug Cutting

Vote (0) Watch (0)

Dates

Created: 09/Oct/01 16:19

Nutch / NUTCH-1

test

Log In

Views

Details

Type:	Bug	Status:	Closed
Priority:	Trivial	Resolution:	Fixed
Affects Version/s:	None	Fix Version/s:	None
Component/s:	None		
Labels:	None		

People

Assignee: Unassigned

Reporter: Doug Cutting

Vote (0) Watch (0)

Dates

Created: 11/Feb/05 04:23

Hadoop Common / HADOOP-1

initial import of code from Nutch

Log In

Views

Details

Type:	Task	Status:	Closed
Priority:	Major	Resolution:	Fixed
Affects Version/s:	None	Fix Version/s:	0.1.0
Component/s:	None		
Labels:	None		

People

Assignee: Doug Cutting

Reporter: Doug Cutting

Vote (0) Watch (0)

Dates

Created: 01/Feb/06 02:54

There are serveral Hadoop subprojects

- Hadoop Common: The common utilities that support the other Hadoop subprojects.
- HDFS: A distributed file system that provides high throughput access to application data.
- MapReduce: A software framework for distributed processing of large data sets on compute clusters.

Other Hadoop related projects

- Chukwa: A data collection system for managing large distributed systems.
- HBase: A scalable, distributed database that supports structured data storage for large tables.
- Hive: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- Pig: A high-level data-flow language and execution framework for parallel computation.
- ZooKeeper: A high-performance coordination service for distributed applications.

Avro

- Avro is a data serialization system.
- It provides:
  - *Rich data structures.*
  - *A compact, fast, binary data format.*
  - *A container file, to store persistent data.*
  - *Remote procedure call (RPC).*
  - *Simple integration with dynamic languages.*
- Code generation is not required to read or write data files nor to use or implement RPC protocols. Code generation as an optional optimization, only worth implementing for statically typed languages.
- For more detail, please check the official document: http://avro.apache.org/docs/current/

Zoo Keeper

- http://hadoop.apache.org/zookeeper/
- ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications.
- *Each time they are implemented there is a lot of work that goes into fixing the bugs and race conditions that are inevitable. Because of the difficulty of implementing these kinds of services, applications initially usually skimp on them ,which make them brittle in the presence of change and difficult to manage. Even when done correctly, different implementations of these services lead to management complexity when the applications are deployed.*

Pig

- http://hadoop.apache.org/pig/
- Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.
- Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs
- Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:
  - Ease of programming
  - Optimization opportunities
  - Extensibility

Hive

- http://hadoop.apache.org/hive/
- Hive is a data warehouse infrastructure built on top of Hadoop that provides tools to enable easy data summarization, adhoc querying and analysis of large datasets data stored in Hadoop files.
- Hive QL is based on SQL and enables users familiar with SQL to query this data.

Chukwa

- http://hadoop.apache.org/chukwa/
- Chukwa is an open source data collection system for monitoring large distributed systems.
- built on top of HDFS and Map/Reduce framework
- includes a flexible and powerful toolkit for displaying, monitoring and analyzing results to make the best use of the collected data.

Mahout

- http://mahout.apache.org/
- Mahout is a scalable machine learning libraries.
- implemented on top of Apache Hadoop using the map/reduce paradigm.
- Mahout currently has
  - Collaborative Filtering
  - User and Item based recommenders
  - K-Means, Fuzzy K-Means clustering
  - Mean Shift clustering
  - More ...

Questions?

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

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw