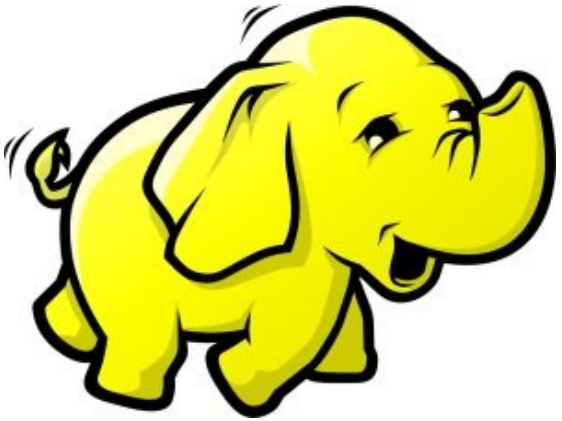
淺談海量資料的趨勢、挑戰與因應對策

Big Data : the Trends, Challenges and Solutions

Jazz Wang

Yao-Tsung Wang

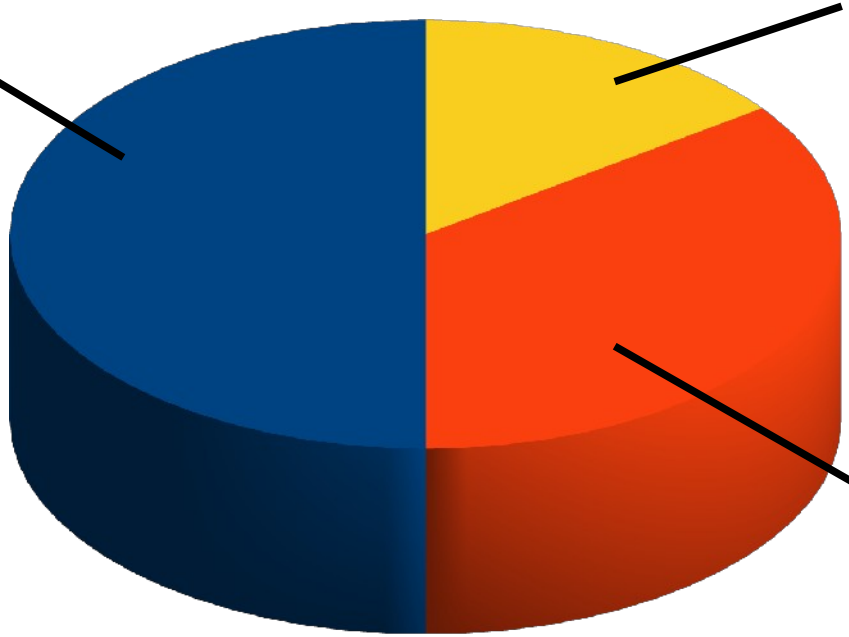
jazz@nchc.org.tw



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

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

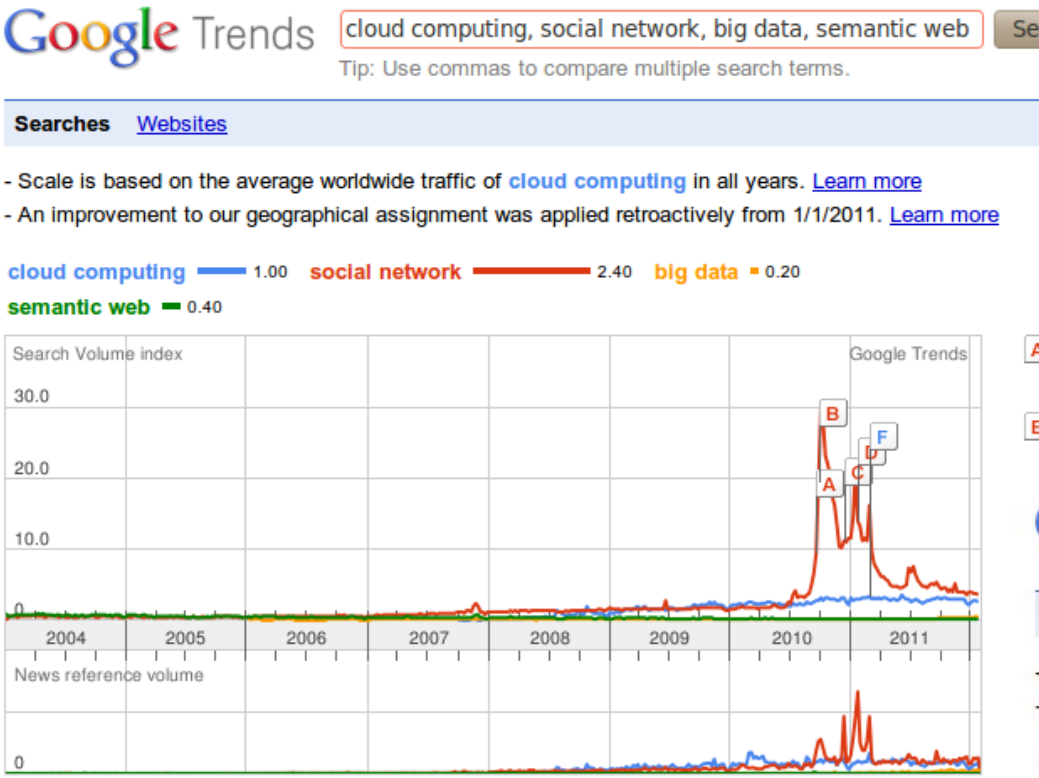
趨勢
Trends

定義
Definitions

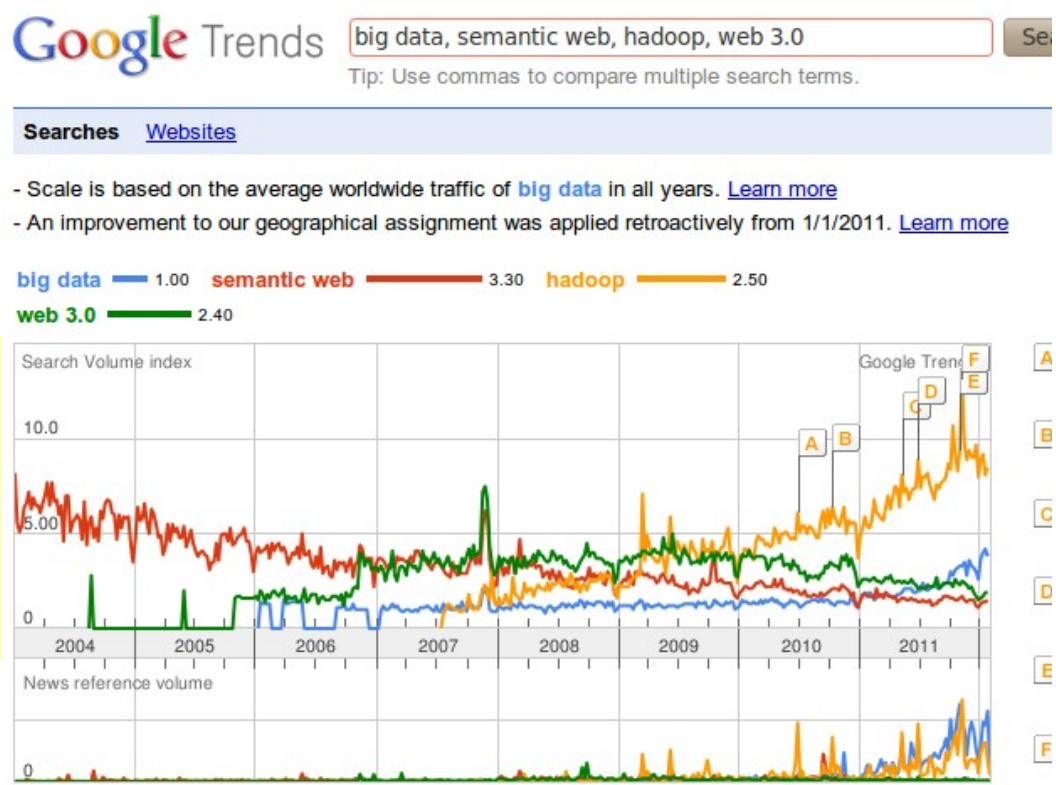
挑戰：管理維度
The Six Dimensions

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, ….

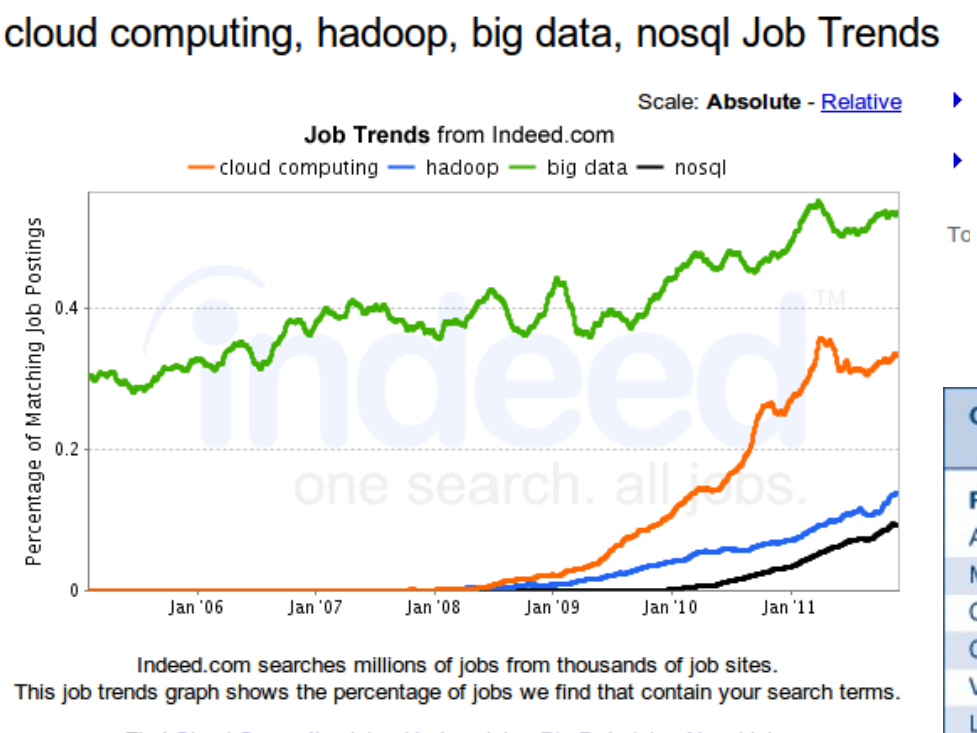


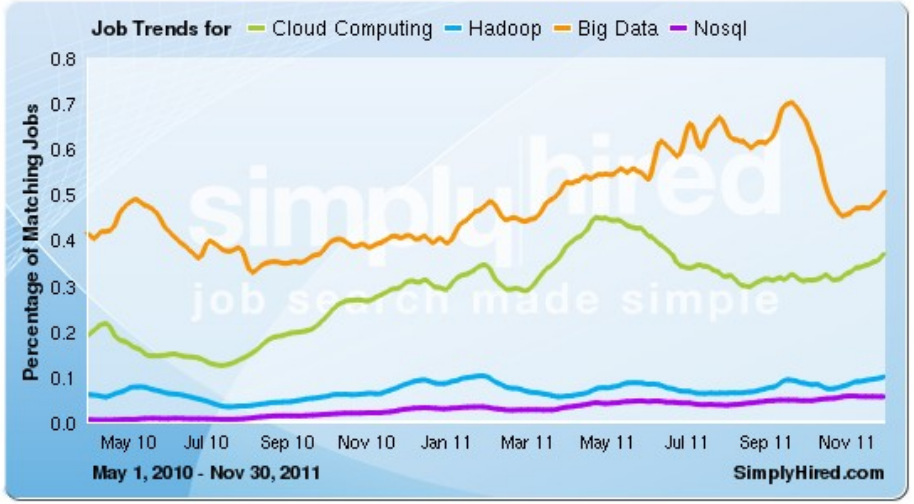
語意網（Semantic Web）從 2001 年開始制定標準後，逐漸下滑。而同義詞 Web 3.0 也呈現相似趨勢。海量資料（Big Data）與其關鍵技術 Hadoop，則仍在上揚中。



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

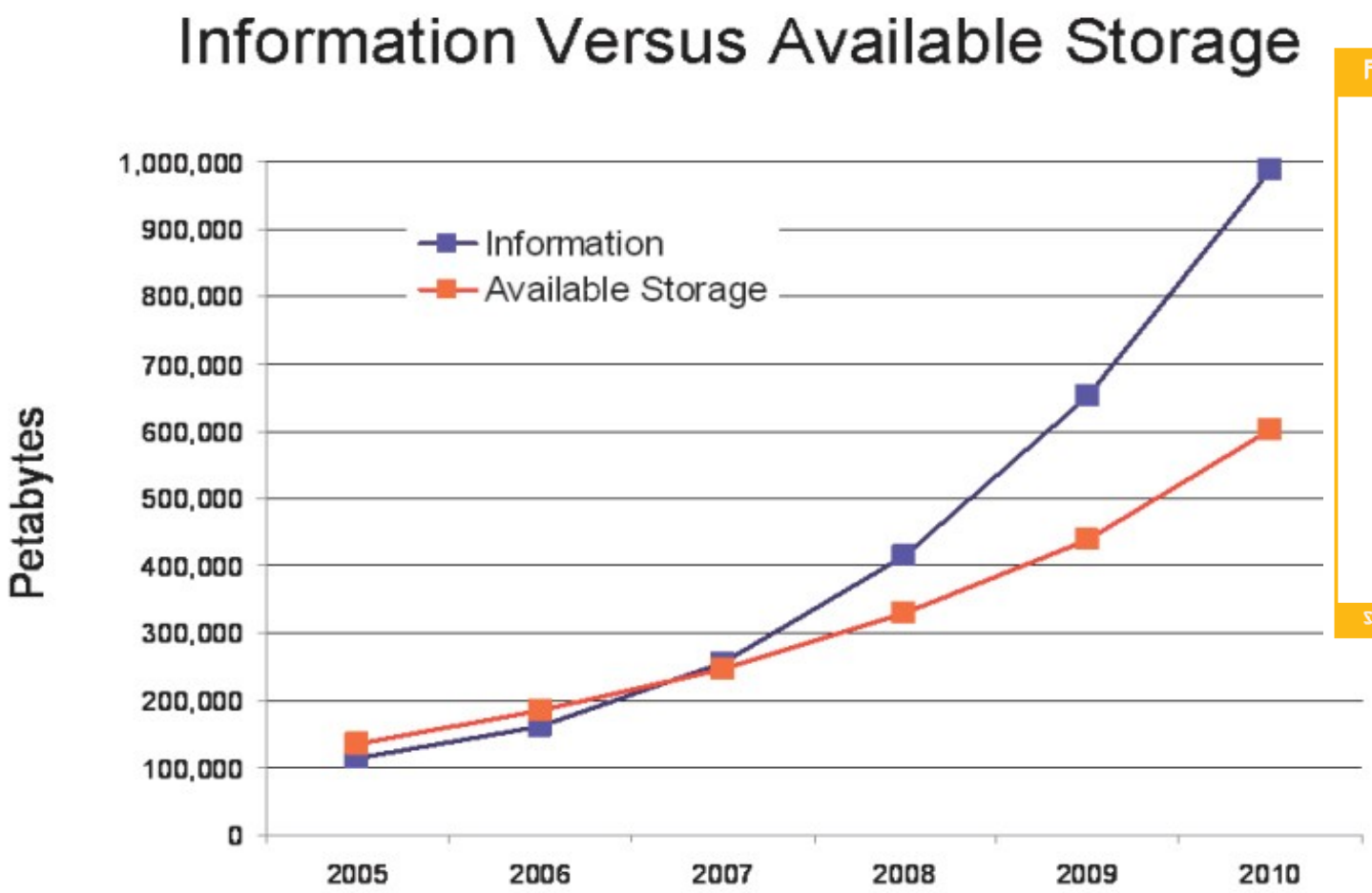
[3] Cloud Computing

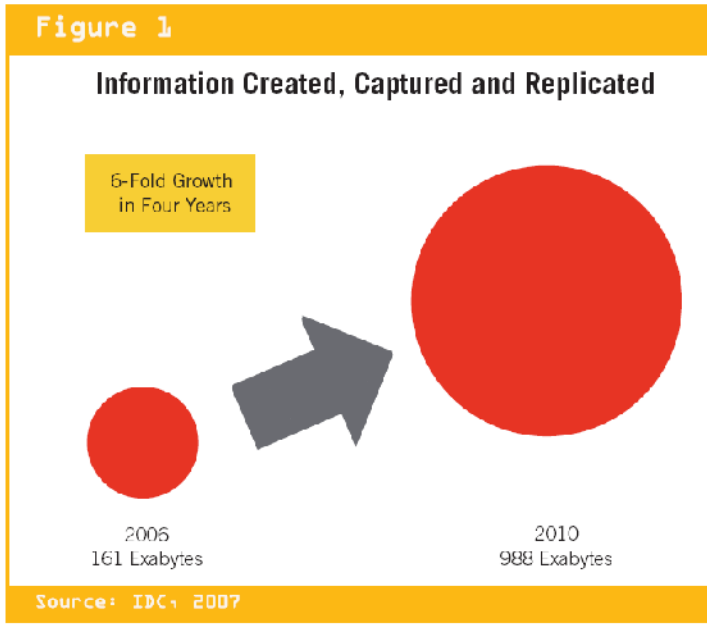
參考來源：http://www.gartner.com/DisplayDocument?id=1524714

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





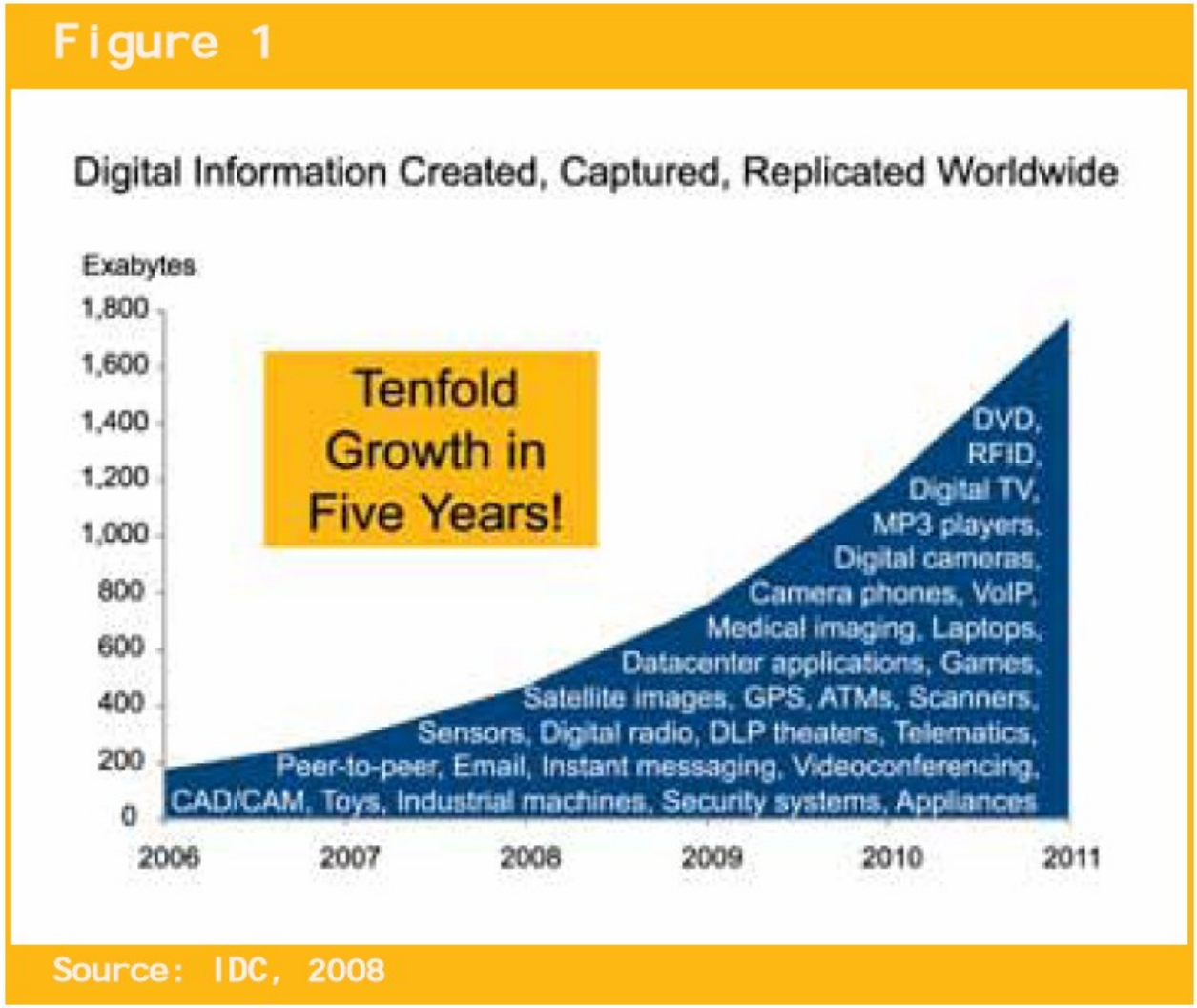
2007 年， IDC 預估
2010 年會成長六倍！
（相較 2006 年）

Source: IDC, 2007

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

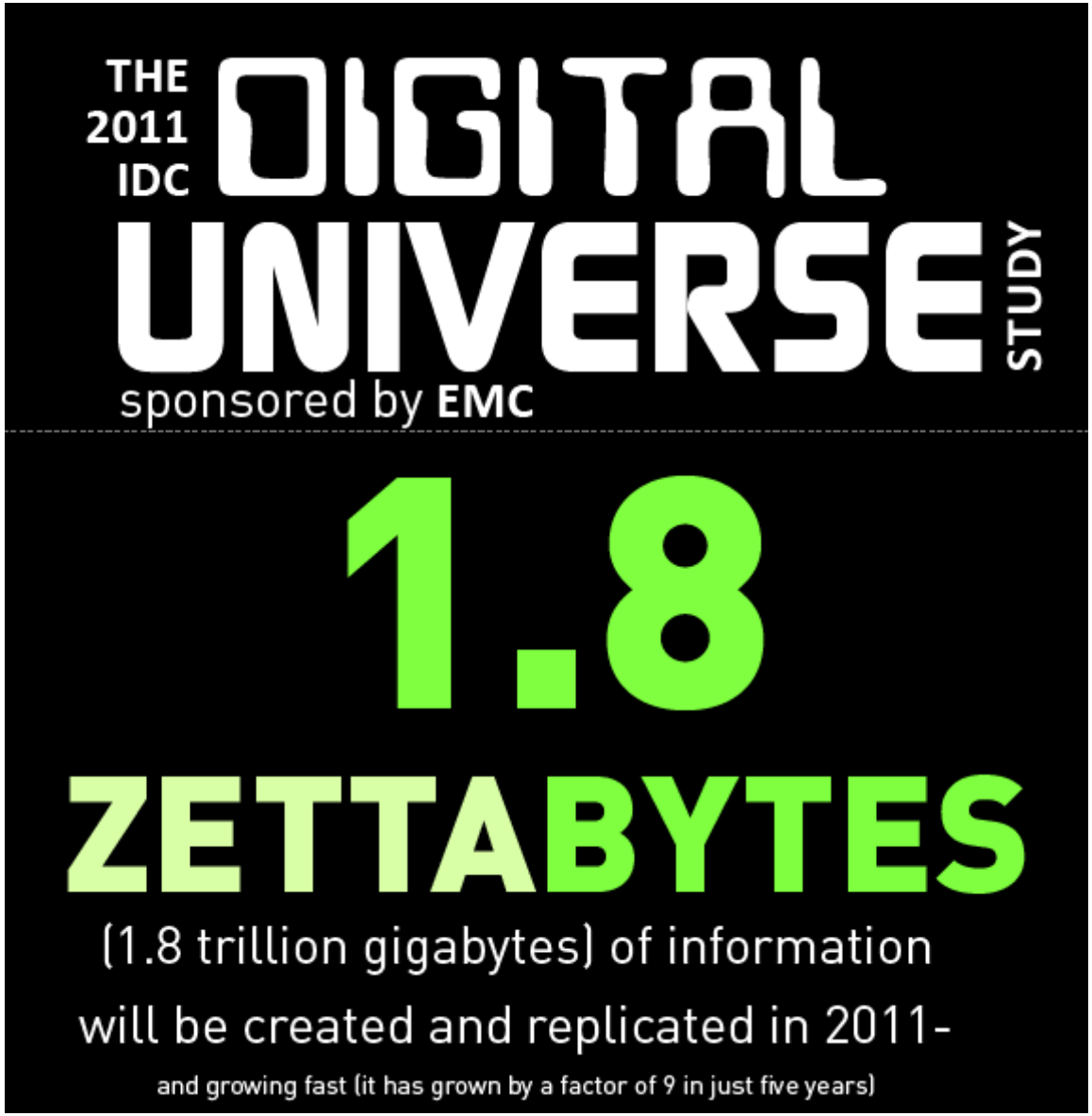


2009 年， IDC 預估
2011 年會成長十倍！
（相較 2006 年）

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.

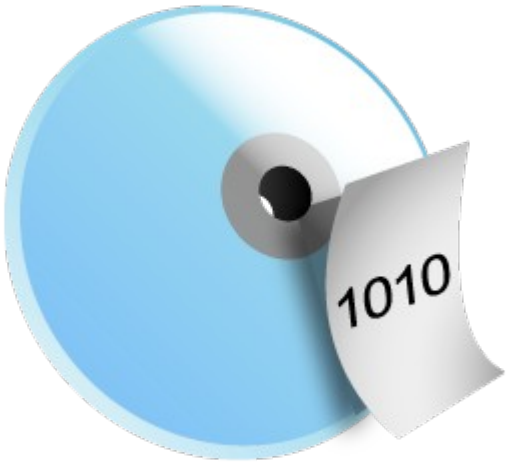
Definition [edit]

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]

多個檔案，容量 10TB

一個資料庫，容量 10TB

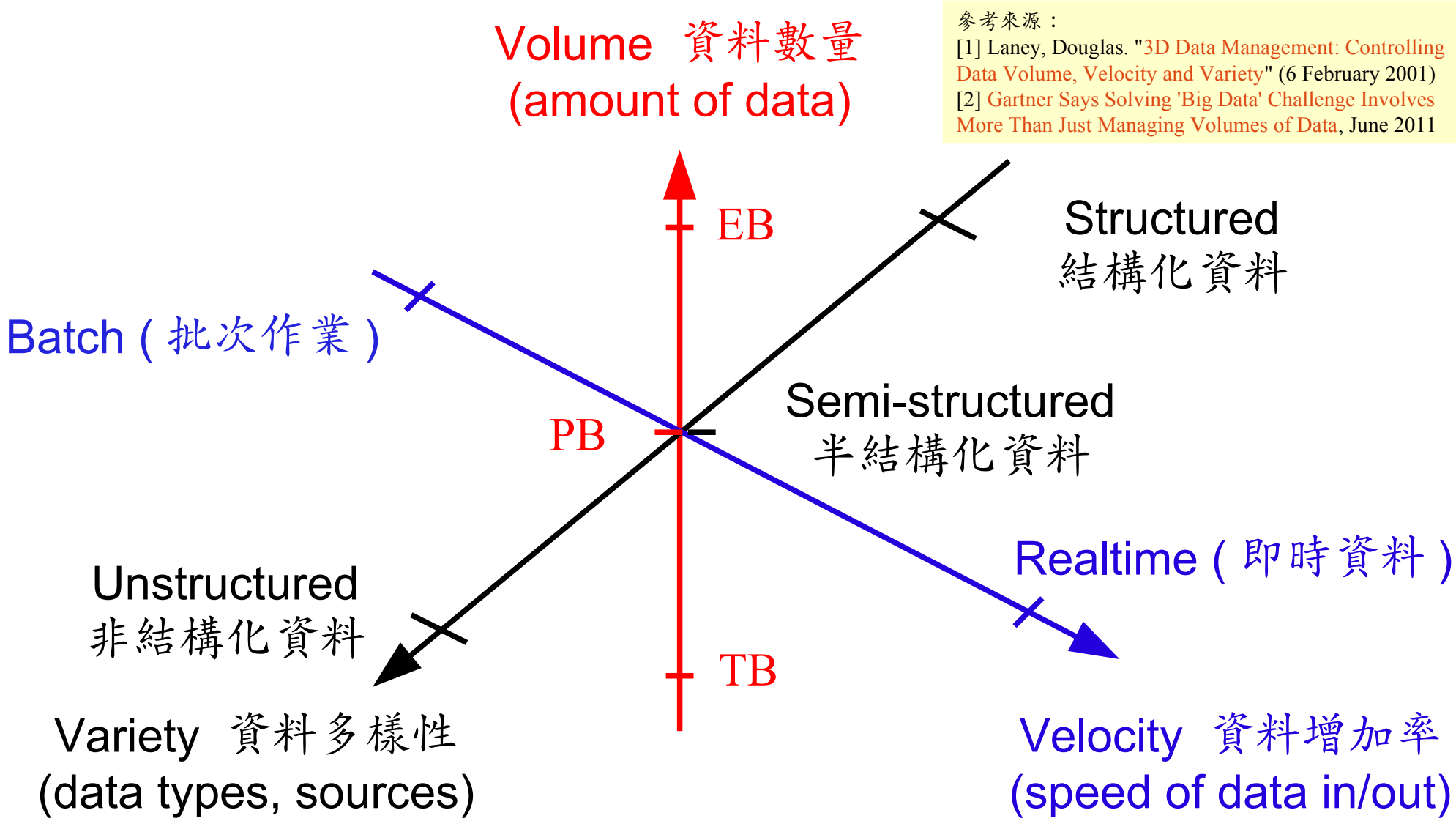


一個檔案，容量 10TB

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

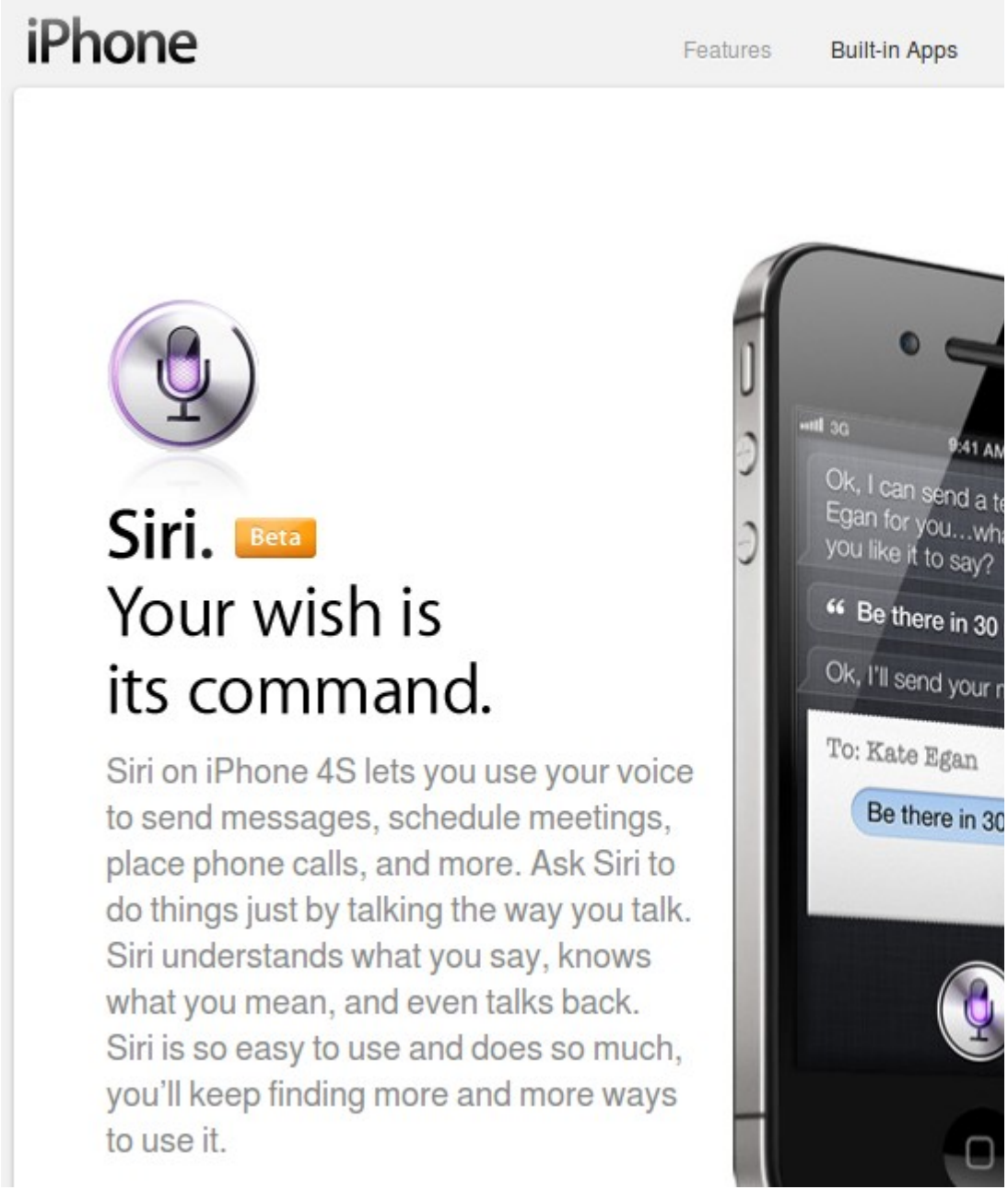
Agenda 演講大綱

What is Big Data ? 何謂海量資料

Why should we care? 為何需要關切

資料 Data → 知識 Knowledge → 智慧 Wisdom

Can Machine understand You? 讓機器更懂你？



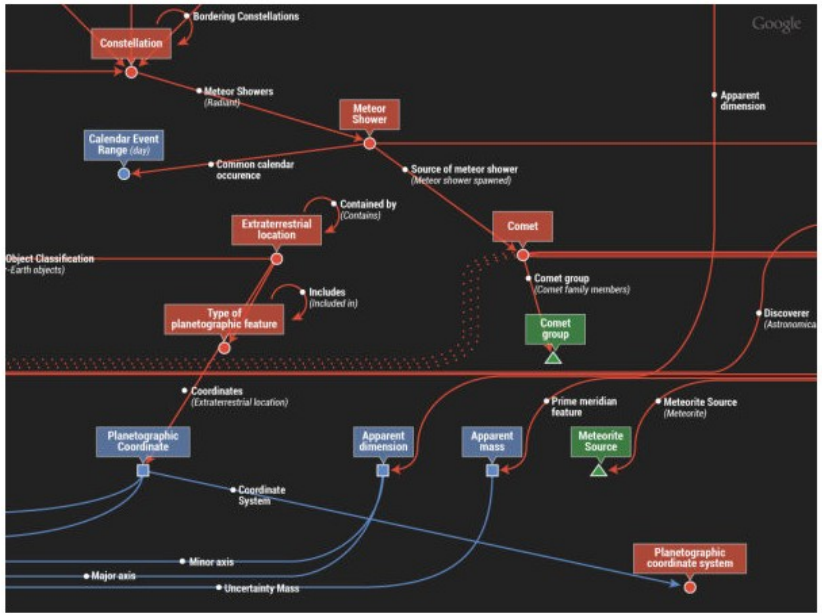
Google將發展「人工智慧」 永久改變搜尋引擎

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』

e-Taiwan 成果系列報導 → 行動台灣 應用推動計畫 → 發展優質網路社會 (U-Taiwan) 計畫 → 智慧台灣 領航系列報導

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

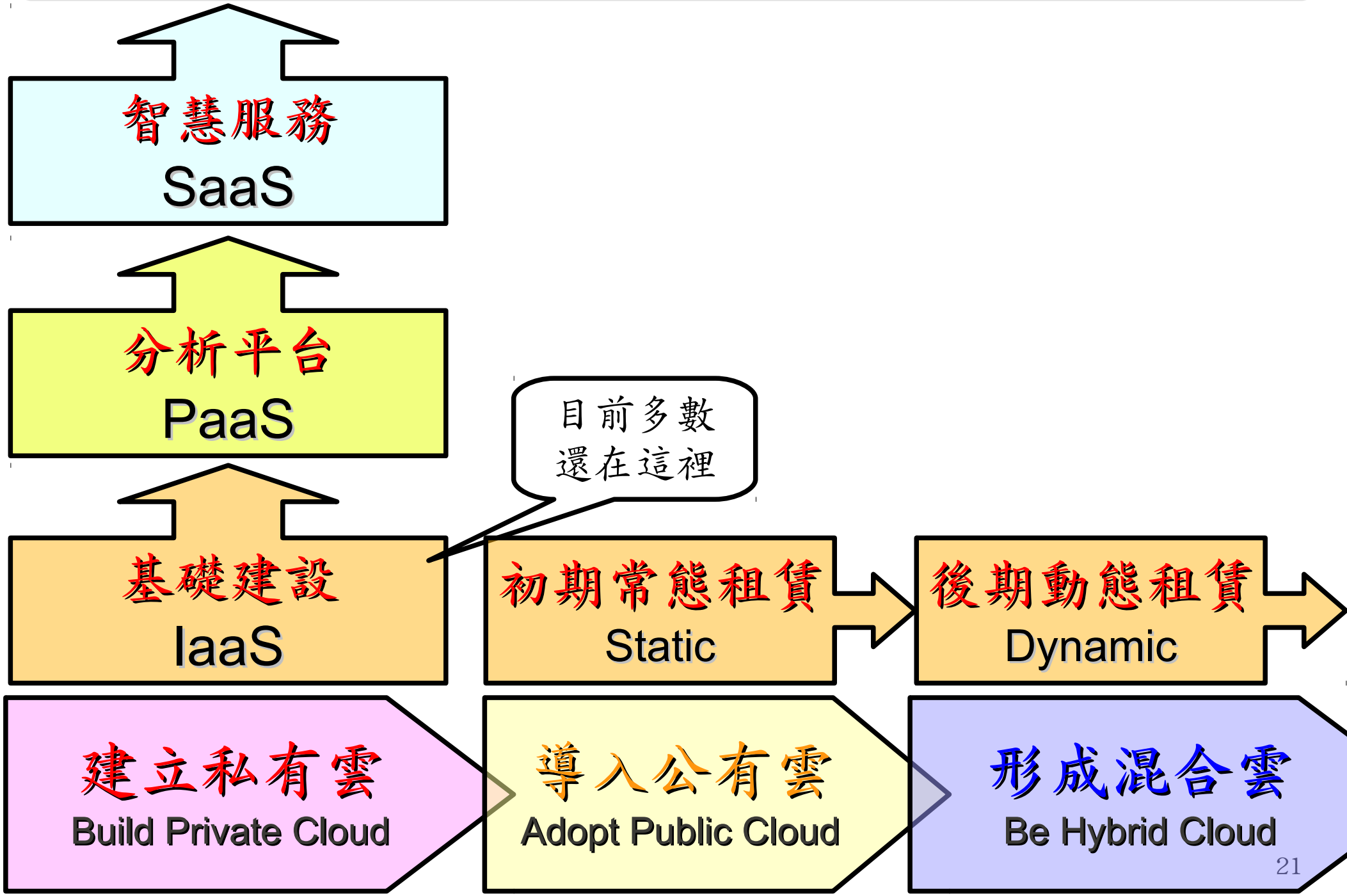
What is Big Data ? 何謂海量資料

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

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 ??

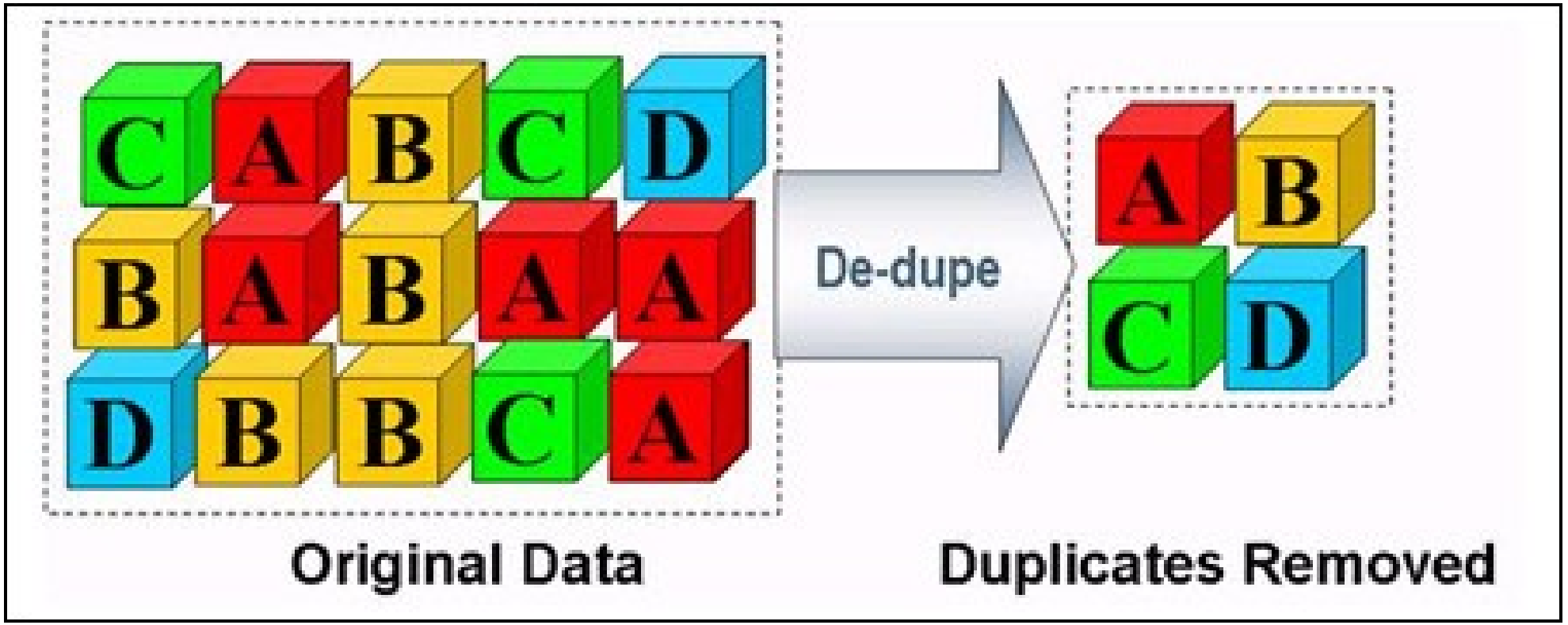
Application Virtualization	應用程式虛擬化
Desktop Virtualization Client Virtualization	桌面虛擬化
Presentation Virtualization	顯示虛擬化
OS-level Virtualization	作業系統虛擬化
Network Virtualization	網路虛擬化
Storage Virtualization	儲存虛擬化

Database Virtualization 資料庫虛擬化

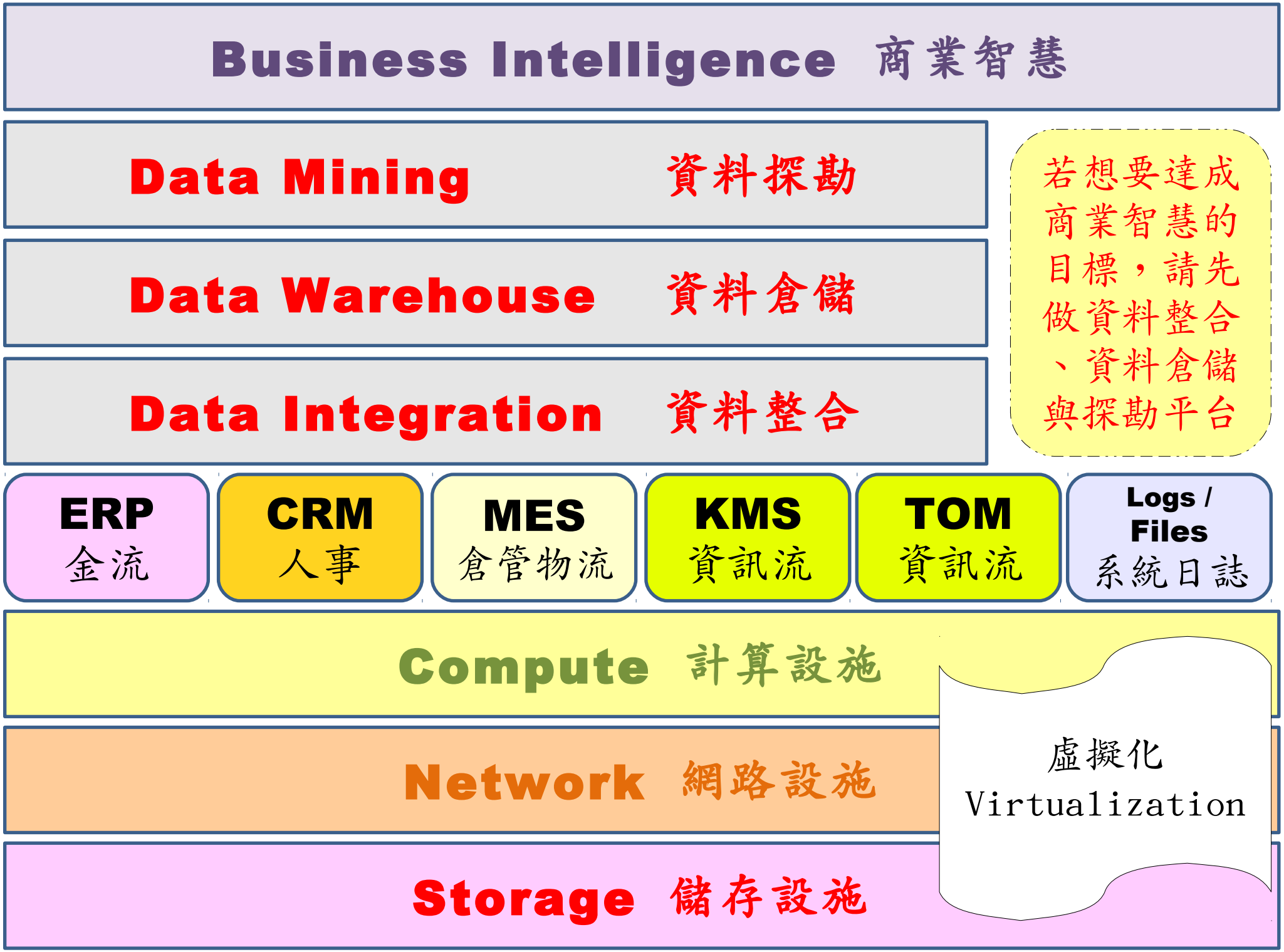
Data Virtualization 資料虛擬化

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

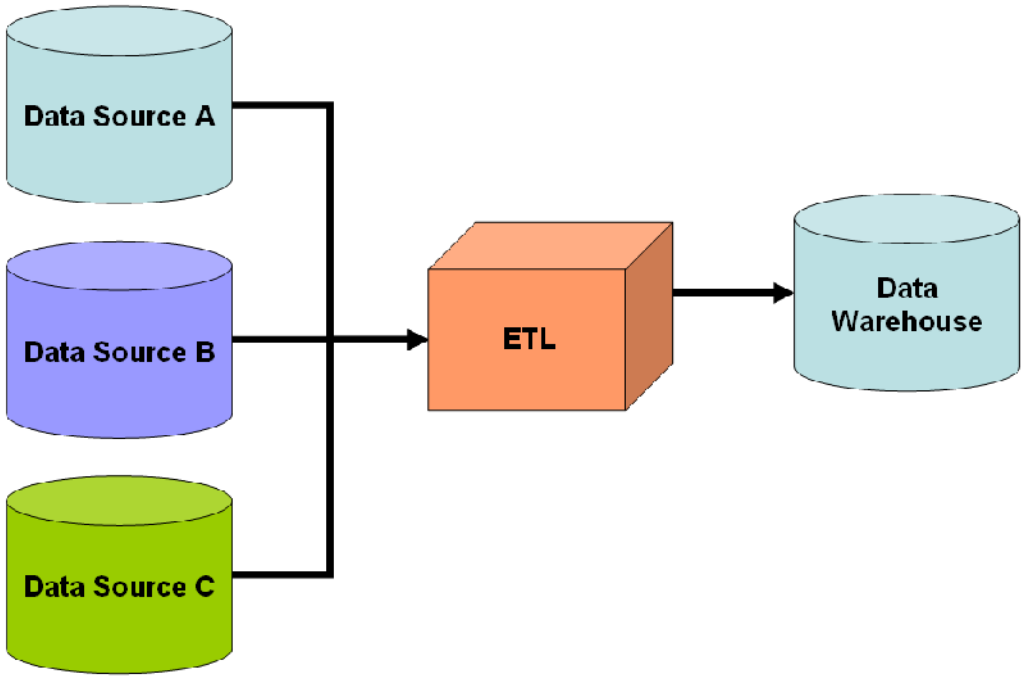


Figure 1: Simple schematic for a data warehouse. The ETL process extracts information from the source databases, transforms it and then loads it into the data warehouse.

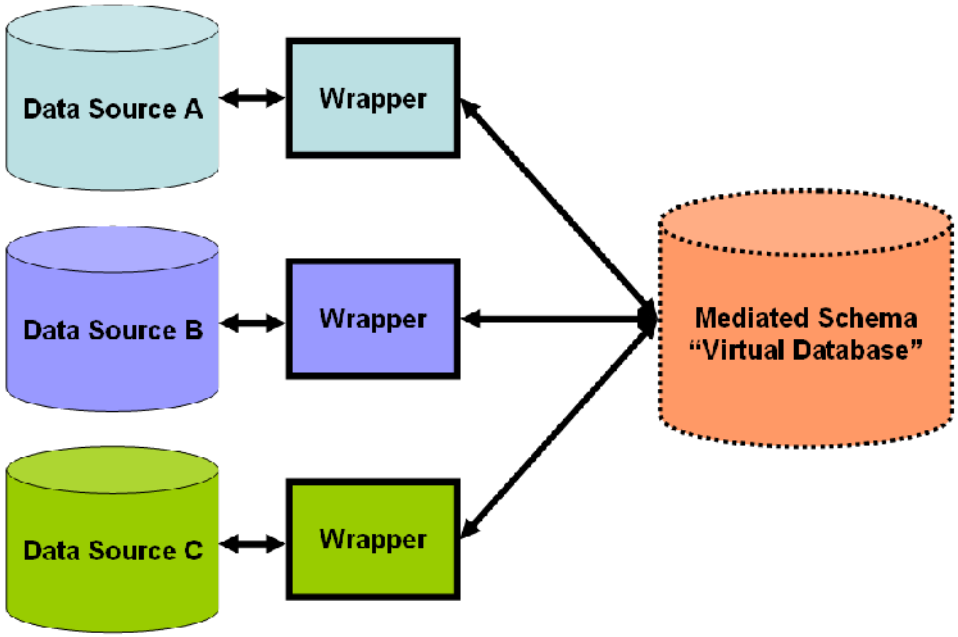
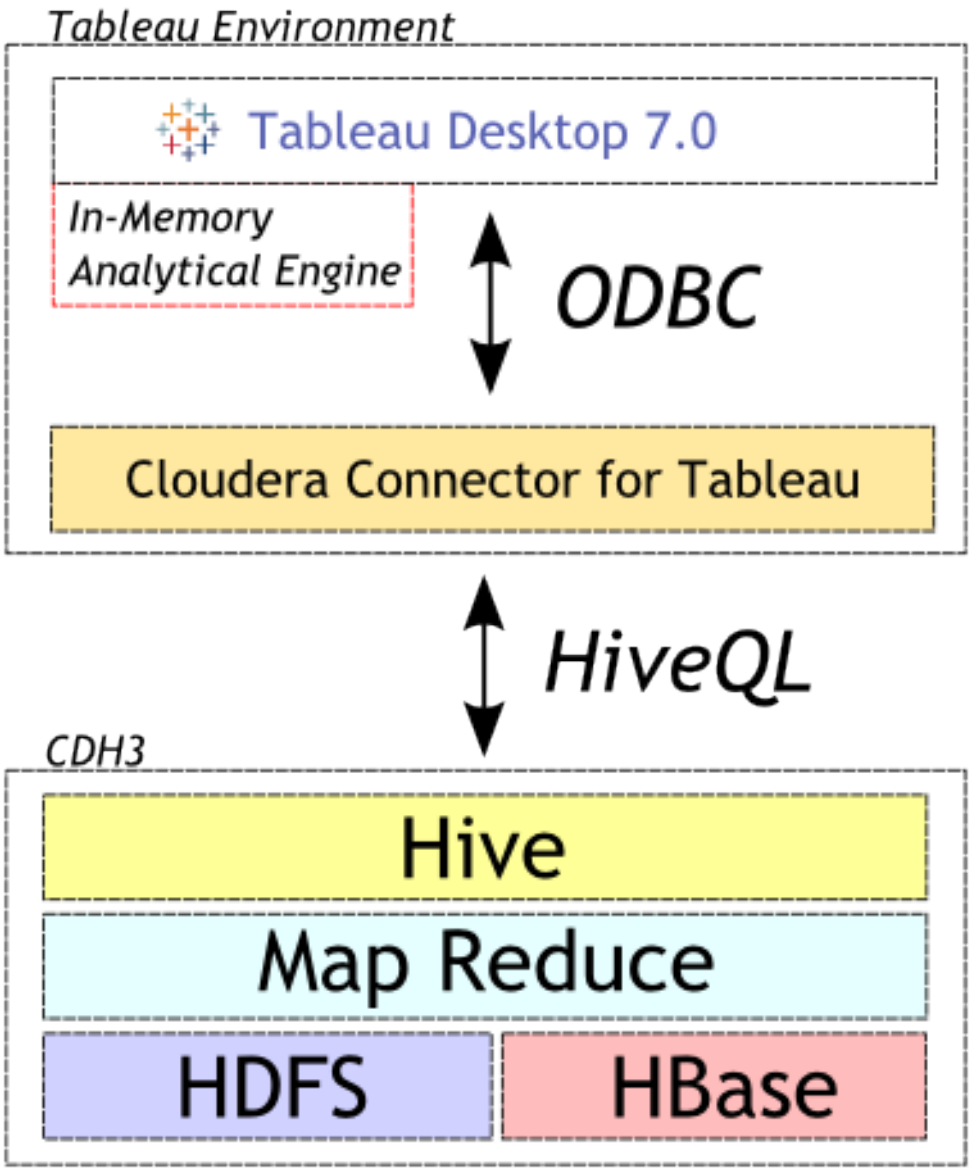
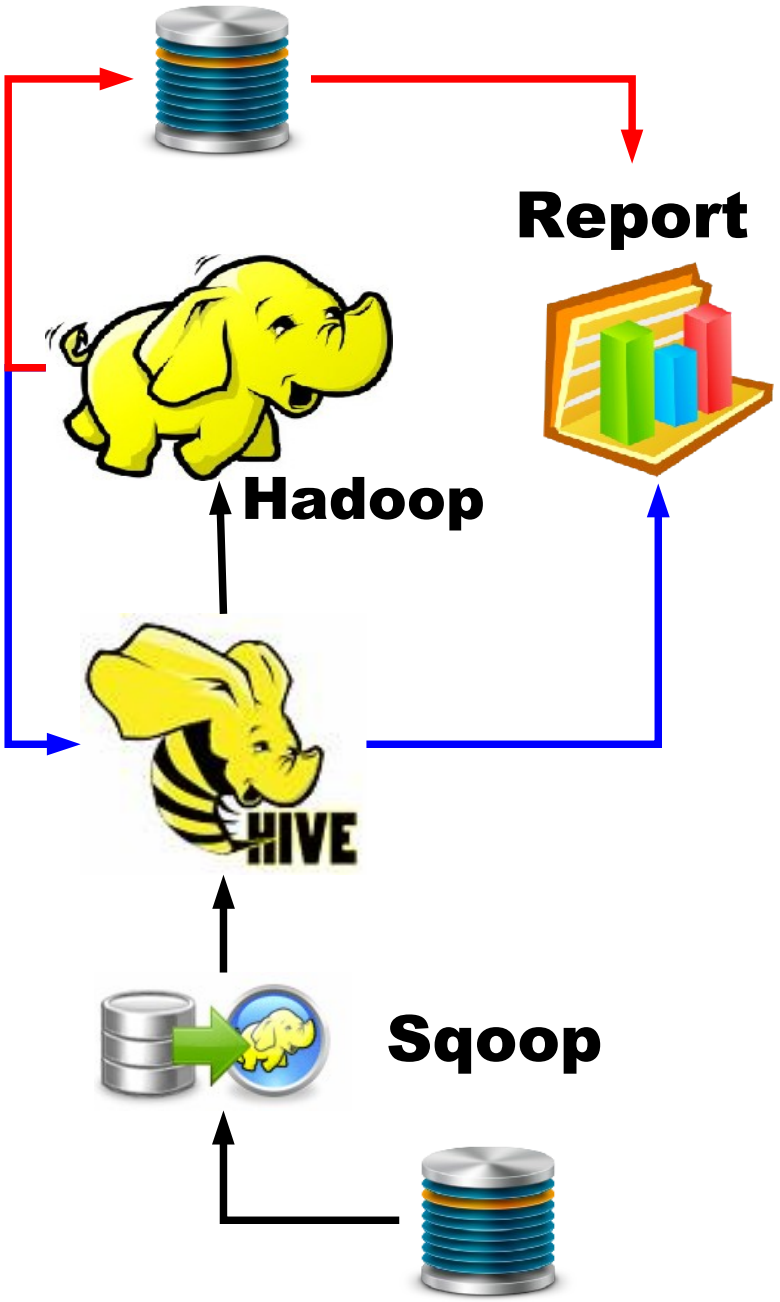


Figure 2: Simple schematic for a data-integration solution. A system designer constructs a mediated schema against which users can run queries. The virtual database interfaces with the source databases via wrapper code if required.

Data Mining & Visualization 資料探勘與視覺化





Agenda 演講大綱

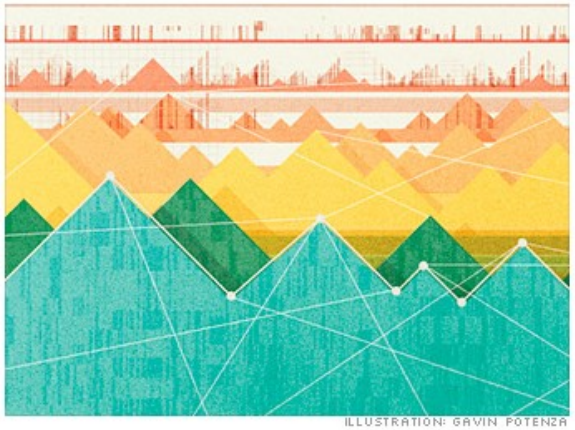
What is Big Data ?	何謂海量資料
Why should we care?	爲何需要關切
When to deploy it ?	何時導入技術
How to handle it ?	三大因應策略
Who is key player ?	誰是成功關鍵

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.



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/
http://visualoop.tumblr.com/post/4052912103/the-role-of-the-data-scientist

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

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 .....

Questions?

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

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw