巨量資料的趨勢、挑戰與因應對策

Big Data : Trends, Challenges and Solutions

Who AM I

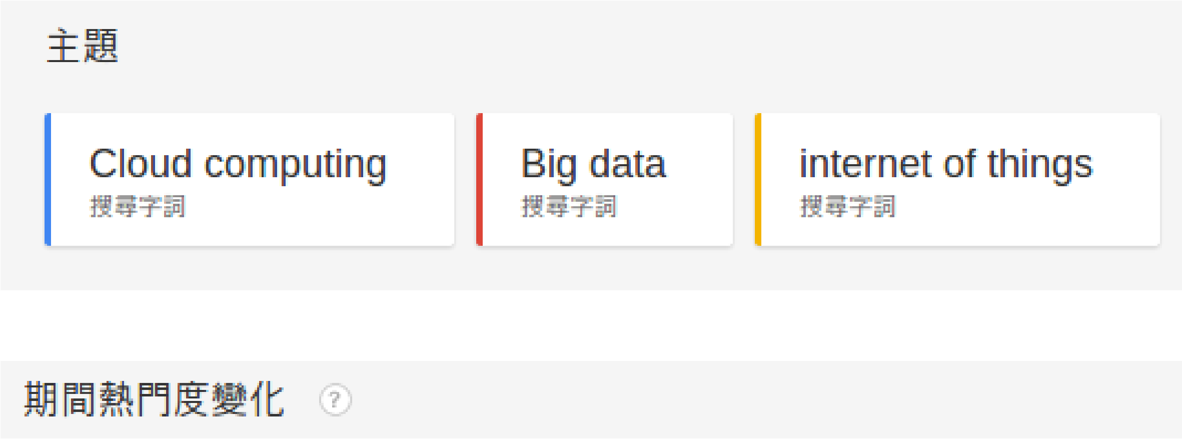
- 王耀聰 Jazz Yao-Tsung Wang
- Hadoop.TW 共同創辦人
- Hadoop The Definitive Guide 譯者
- Hadoop Operations 譯者
- 自由軟體愛好者 / 推廣者 / 開發者
- http://about.me/jazzwang - slideshare, github, etc.
- http://trac.3du.me/cloud - 原 http://trac.nchc.org.tw/cloud

趨勢

http://www.cns11643.gov.tw/AIDB/fm_view.do?font=cns_0000&id=77771
http://www.cns11643.gov.tw/AIDB/fm_view.do?font=cns_0000&id=76913

鄉野調查(1)

巨量資料?!



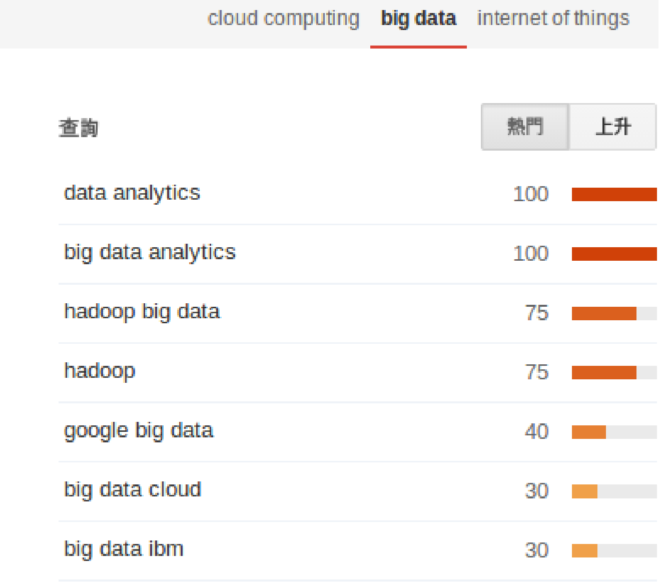
主題

Cloud computing
搜尋字詞

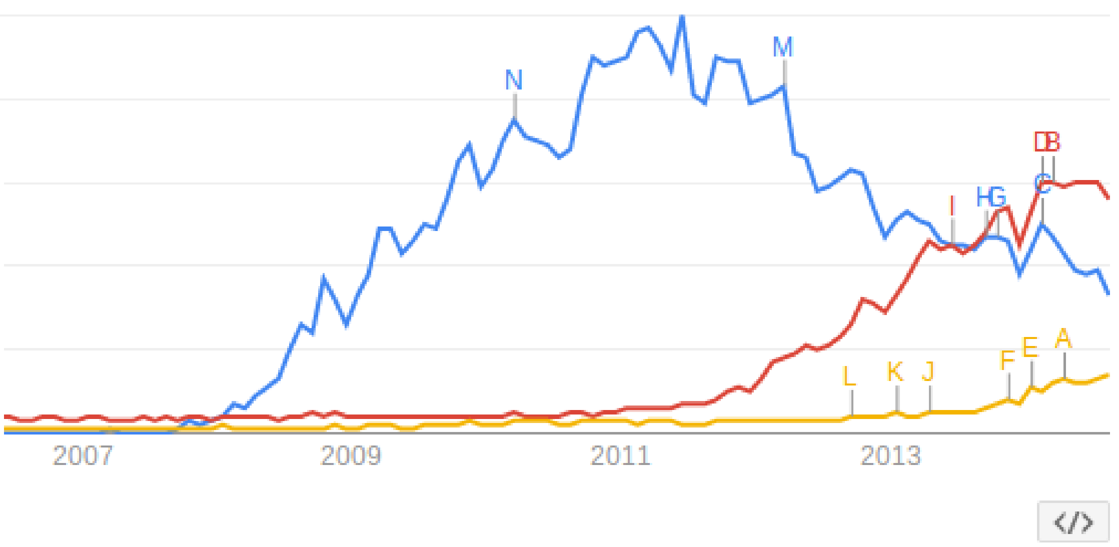
Big data
搜尋字詞

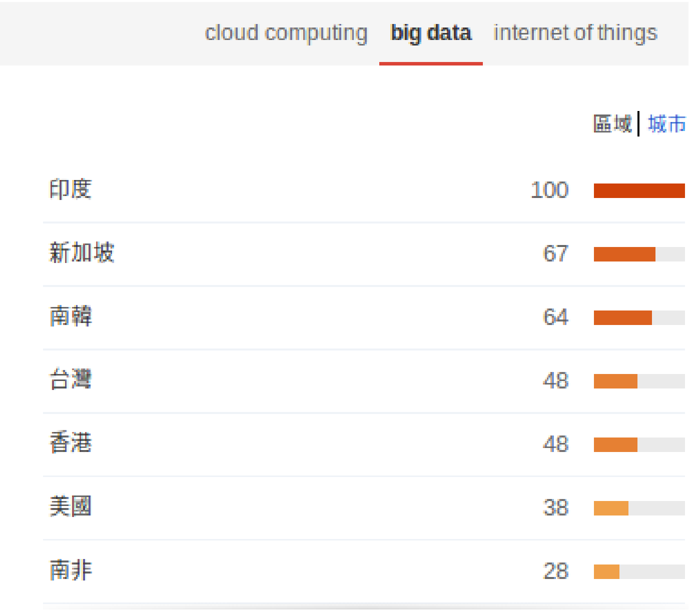
internet of things
搜尋字詞

期間熱門度變化



查詢	
data analytics	100
big data analytics	100
hadoop big data	75
hadoop	75
google big data	40
big data cloud	30
big data ibm	30





印度	100
新加坡	67
南韓	64
台灣	48
香港	48
美國	38
南非	28

3 Buzzwords in 2013
三大年度熱門關鍵字

物聯網

Internet of Things

雲端運算

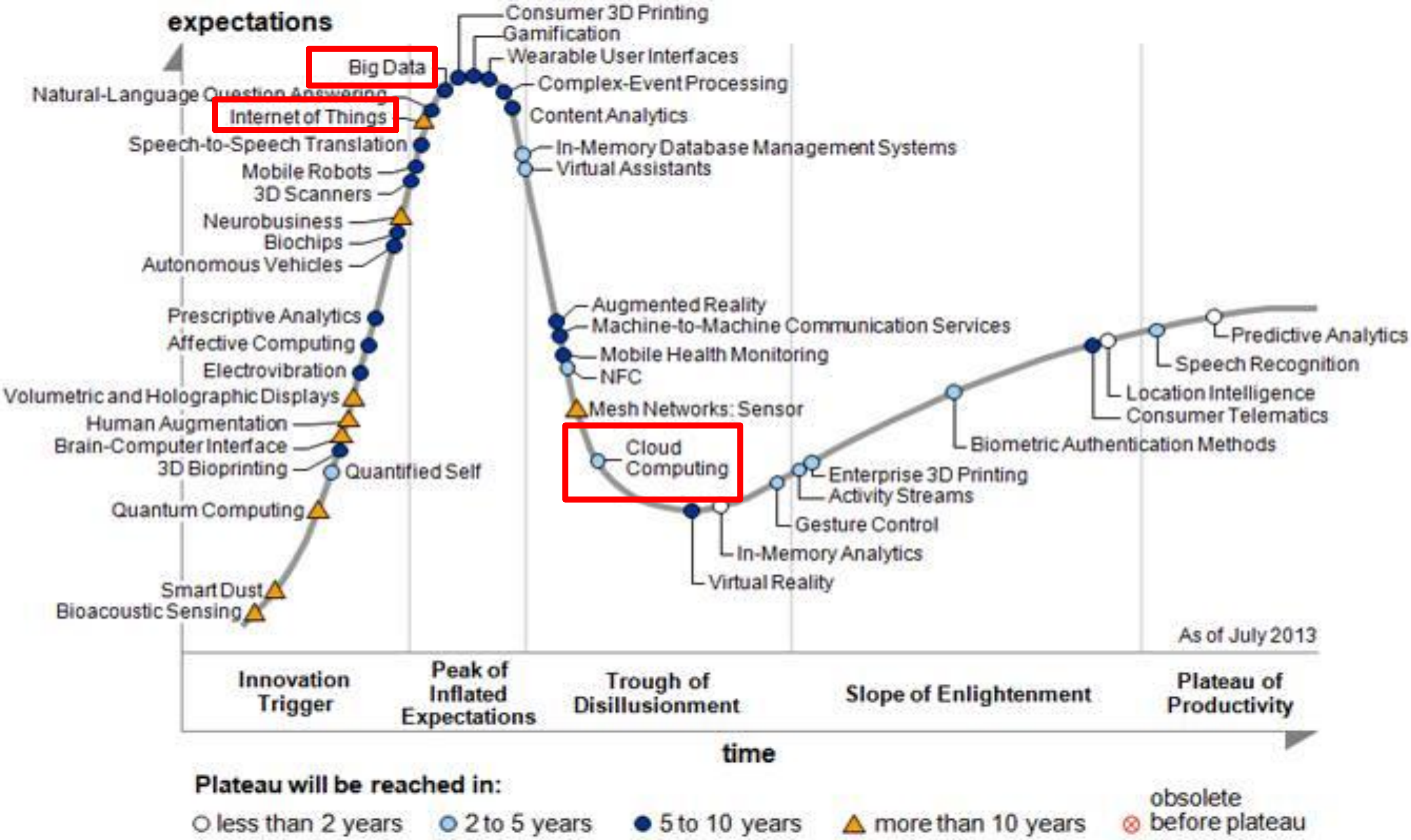
Cloud Computing

巨量資料

Big Data

市場現況：Gartner Hype Cycle 2013

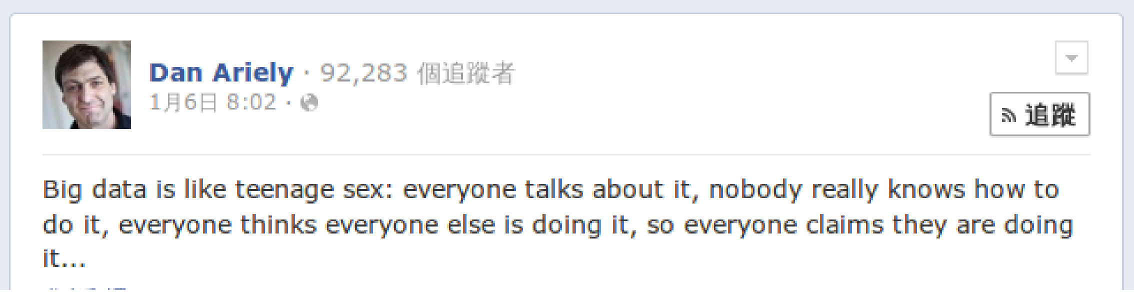
萌芽期　萝幻期　幻滅期　平原期　高原期



巨量資料的現況 ...
Current Status of Big Data .....

" **Big data** is like **teenage sex**:
everyone **talks about** it,
nobody really knows **how to do** it,
everyone **thinks** everyone else is doing it,
so everyone **claims** they are doing it .. "

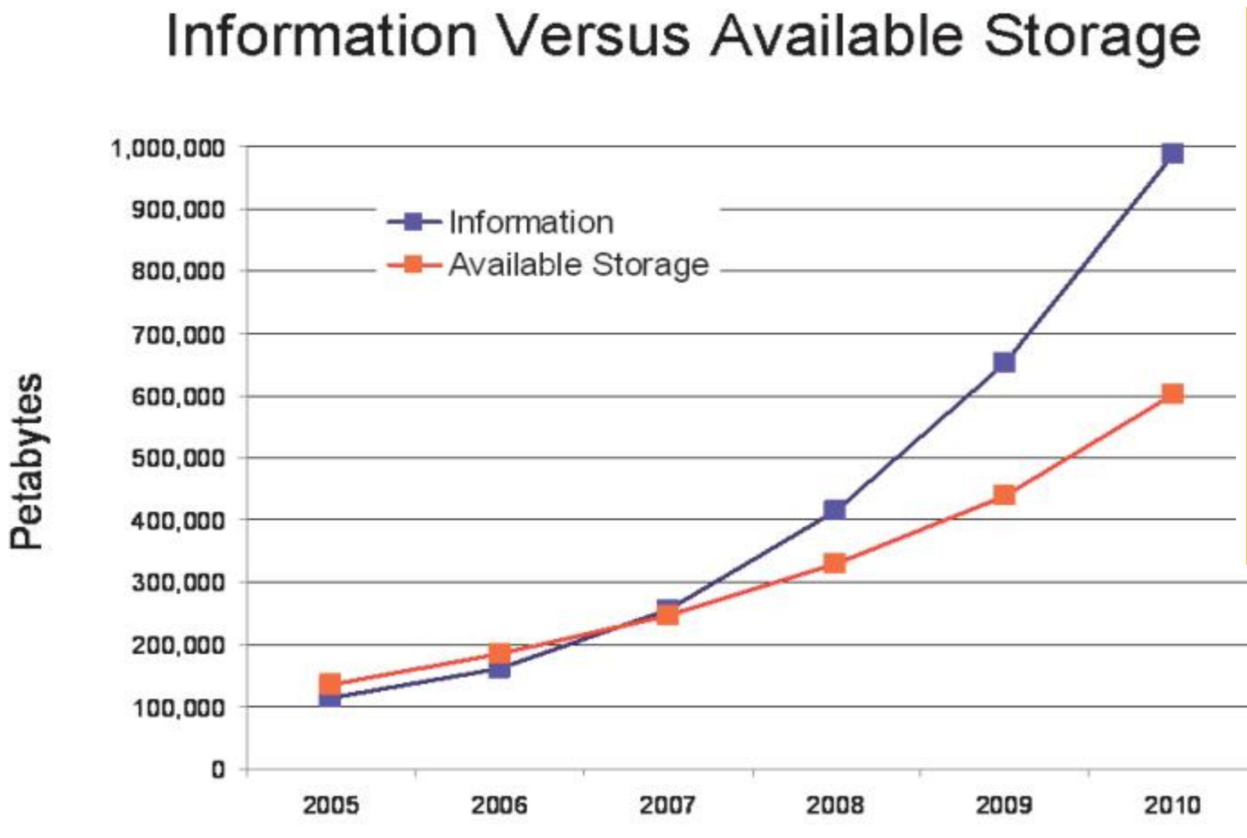
– Dan Ariely, Professor at Duke University
and Professor at Center for Advanced Hindsight



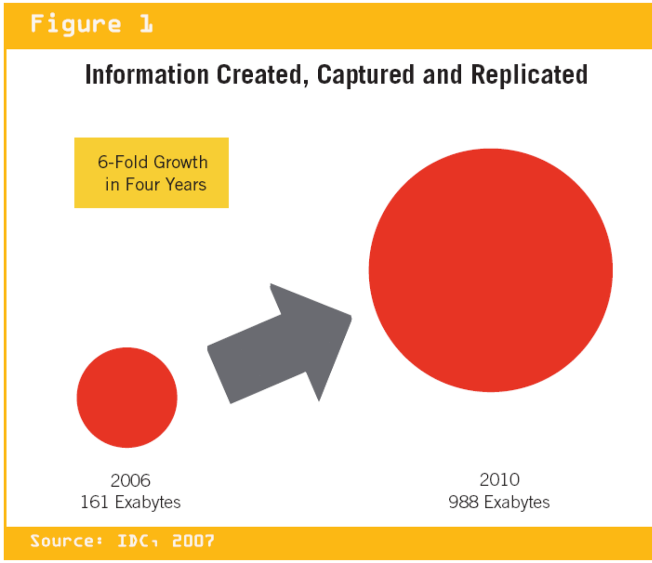
Dan Ariely · 92,283 個追蹤者
1月6日 8:02 ·

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

始於2007的「資料大爆炸」時代 Data Explosion!!



Source: IDC, 2007



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

數位宇宙以每年 1.5 倍速度成長

APRIL 2014
EMC²
EMC DIGITAL UNIVERSE
With Research & Analysis by IDC
The DIGITAL UNIVERSE of OPPORTUNITIES
RICH DATA & the Increasing Value of the INTERNET OF THINGS
4.4 ZETTABYTES
44 ZETTABYTES
The Digital Universe Is Huge– And Growing Exponentially
If the Digital Universe were represented by the memory in a stack of tablets, in 2013 it would have stretched two-thirds the way to the Moon
By 2020, there would be 6.6 stacks from the Earth to the Moon
2013
2020
Total Number of Connectable Things
BILLION
(2013)
(2020)
187

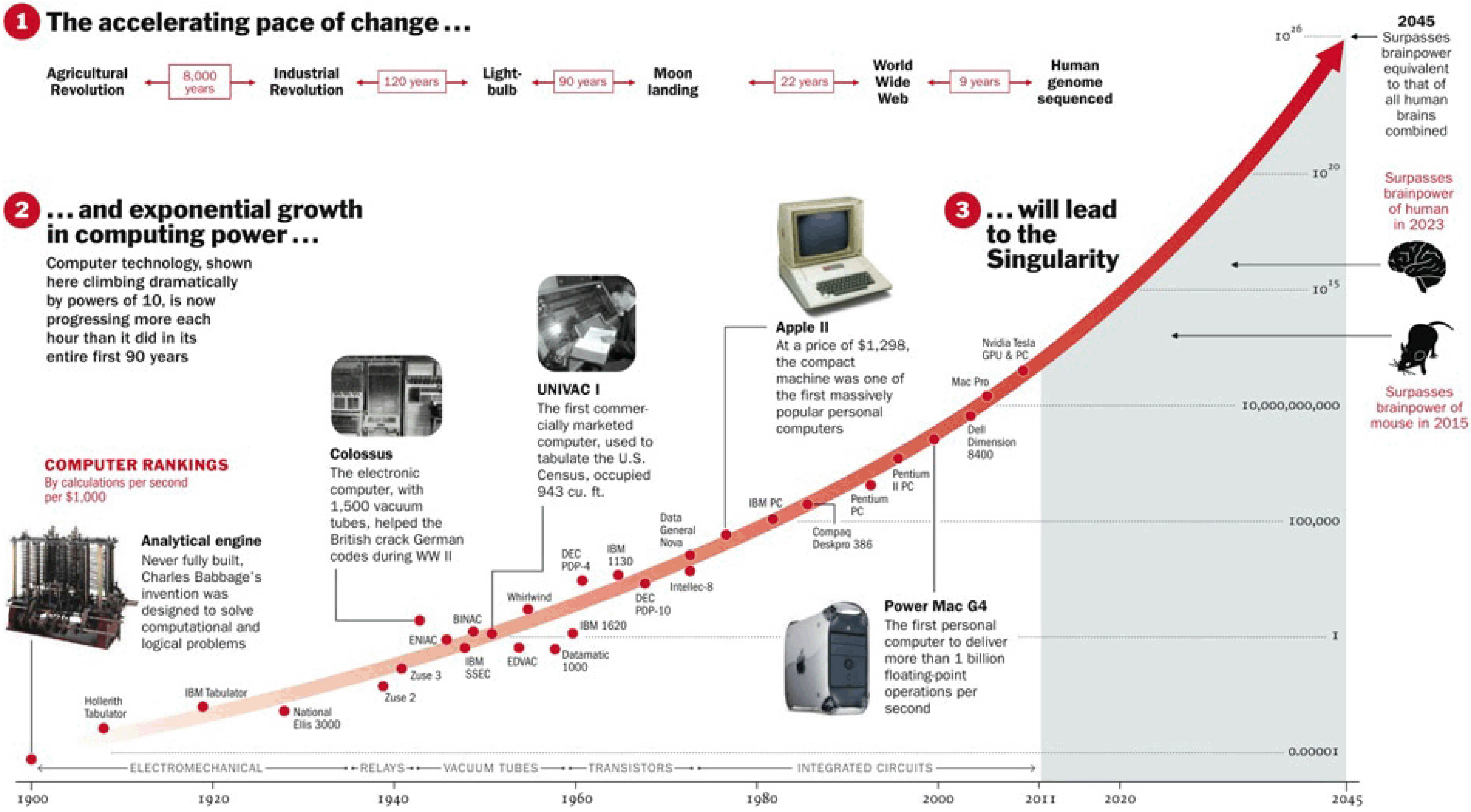
追蹤歷年的IDC數據：

2006	161 EB
2007	281 EB
2008	487 EB
2009	800 EB (0.8 ZB)
2010	988 EB (預測)
2010	1227 EB (1.2 ZB)
2011	1773 EB (預測)
2011	1800 EB (1.8 ZB)
2012	2837 EB (2.8 ZB)
2013	4400 EB (4.4 ZB)

景氣差而成長趨緩？
或受新技術抑制？

出處：The Digital Universe of Opportunities
http://www.emc.com/infographics/digital-universe-2014.htm

典範轉移的時間間距愈來愈短



1 The accelerating pace of change …

2 … and exponential growth in computing power …

Computer technology, shown here climbing dramatically by powers of 10, is now progressing more each hour than it did in its entire first 90 years

3 … will lead to the Singularity

COMPUTER RANKINGS
By calculations per second per $1,000

Analytical engine
Never fully built, Charles Babbage's invention was designed to solve computational and logical problems

Colossus
The electronic computer, with 1,500 vacuum tubes, helped the British crack German codes during WW II

UNIVAC I
The first commercially marketed computer, used to tabulate the U.S. Census, occupied 943 cu. ft.

Apple II
At a price of $1,298, the compact machine was one of the first massively popular personal computers

Power Mac G4
The first personal computer to deliver more than 1 billion floating-point operations per second

2045
Surpasses brainpower equivalent to that of all human brains combined

Surpasses brainpower of human in 2023

Surpasses brainpower of mouse in 2015

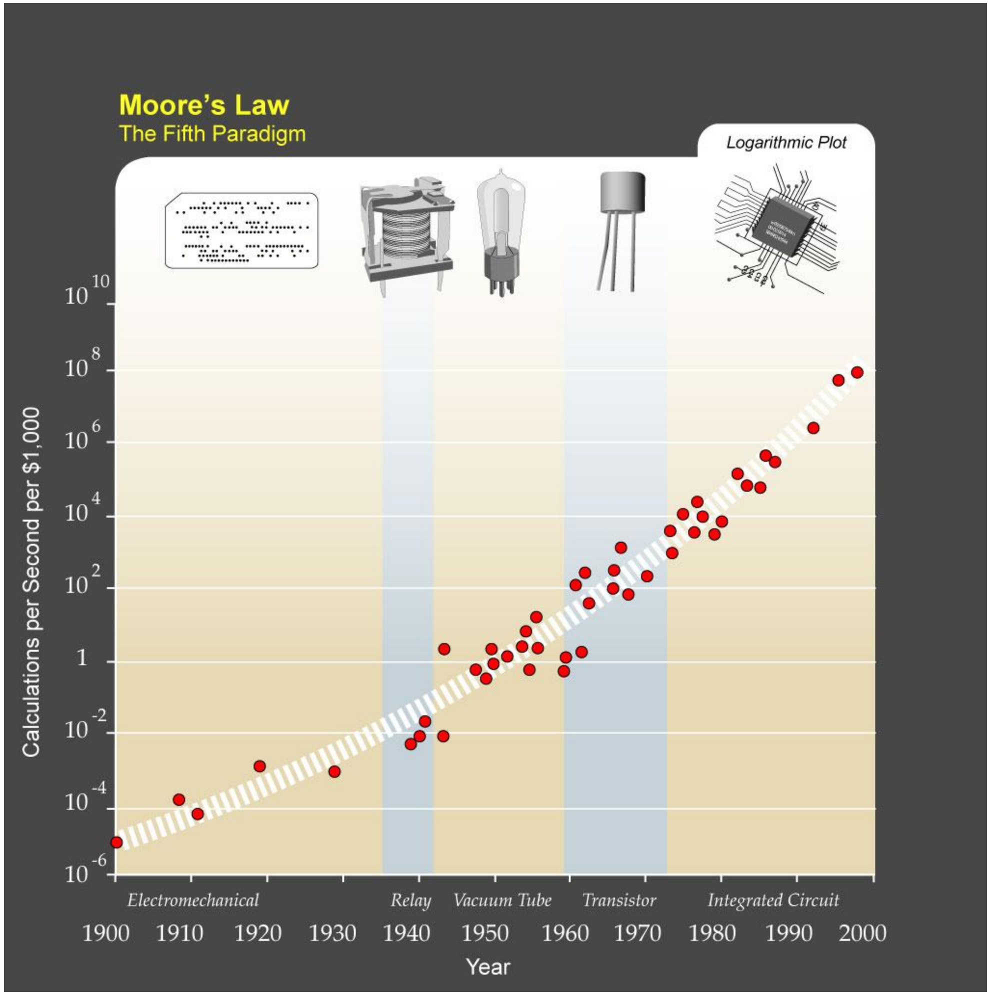
Source : TIME Magazine, "2045: The Year Man Becomes Immortal", Feb. 10, 2011
http://content.time.com/time/magazine/article/0,9171,2048299,00.html
Image Source : http://trickvilla.com/wp-content/uploads/Moores-law-graph.gif

Trend of Computing
– Moore's Law

摩爾定律是1965年由英特爾創始人之一戈登·摩爾提出來的。

在積體電路上可容納的電晶體數目，約每隔24個月便會增加一倍。

英特爾執行長David House所說：每隔18個月晶片的效能提高一倍。

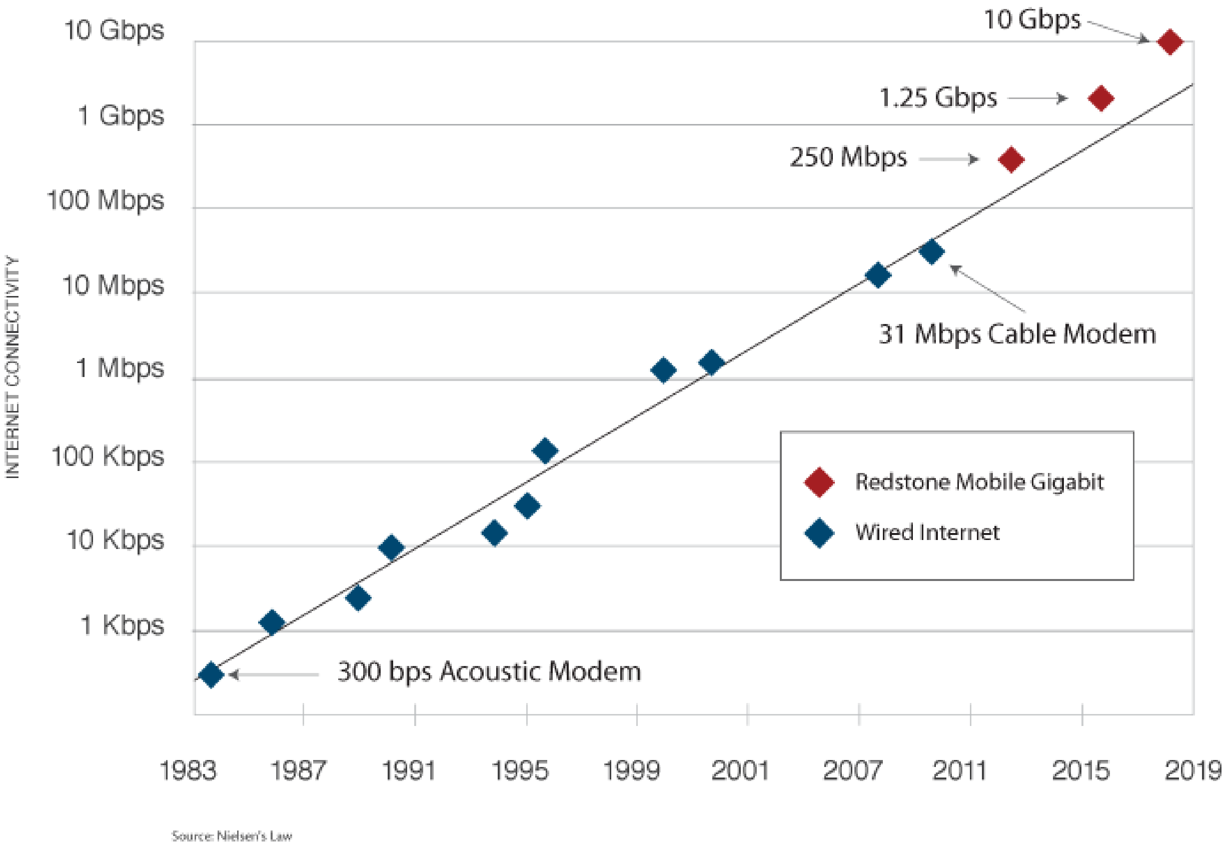


Source: Moore's Law, Wikipedia
http://upload.wikimedia.org/wikipedia/commons/c/c5/PPTMooresLawai.jpg

Trend of Network
– Nielsen's Law of Internet Bandwidth

尼爾森定律是1998年由Jakob Nielsen提出。

每隔20個月，網際網路頻寬會增加一倍。



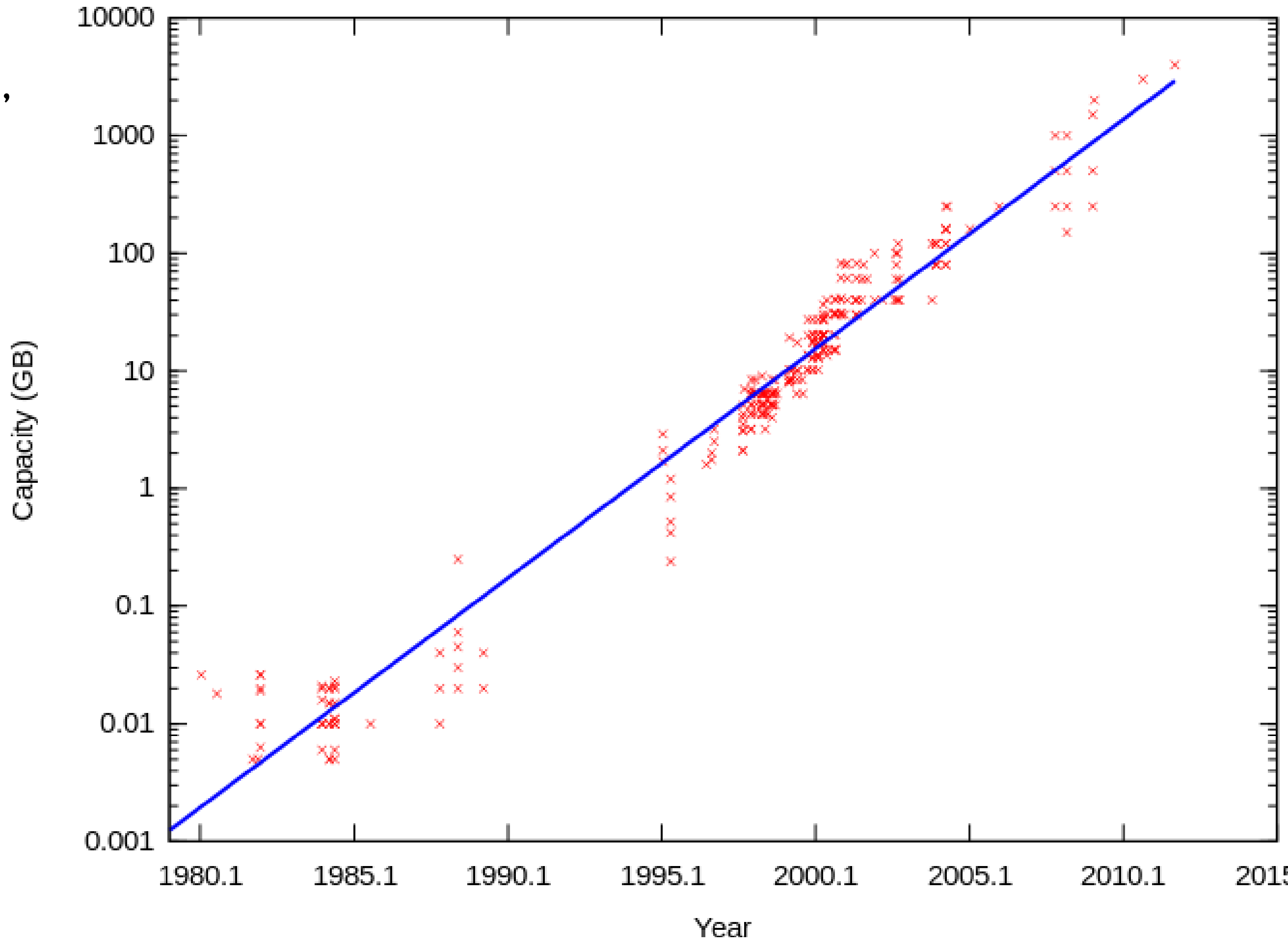
Source: Nielsen's Law

Source: “Nielsen's Law of Internet Bandwidth”, April 5, 1998
http://www.nngroup.com/articles/law-of-bandwidth/
Image Source: “Nielsen's Law”, May 31, 2013
http://redstone.us.com/2013/05/31/subject-nielsens-law/

Trend of Storage
– Kryder's Law

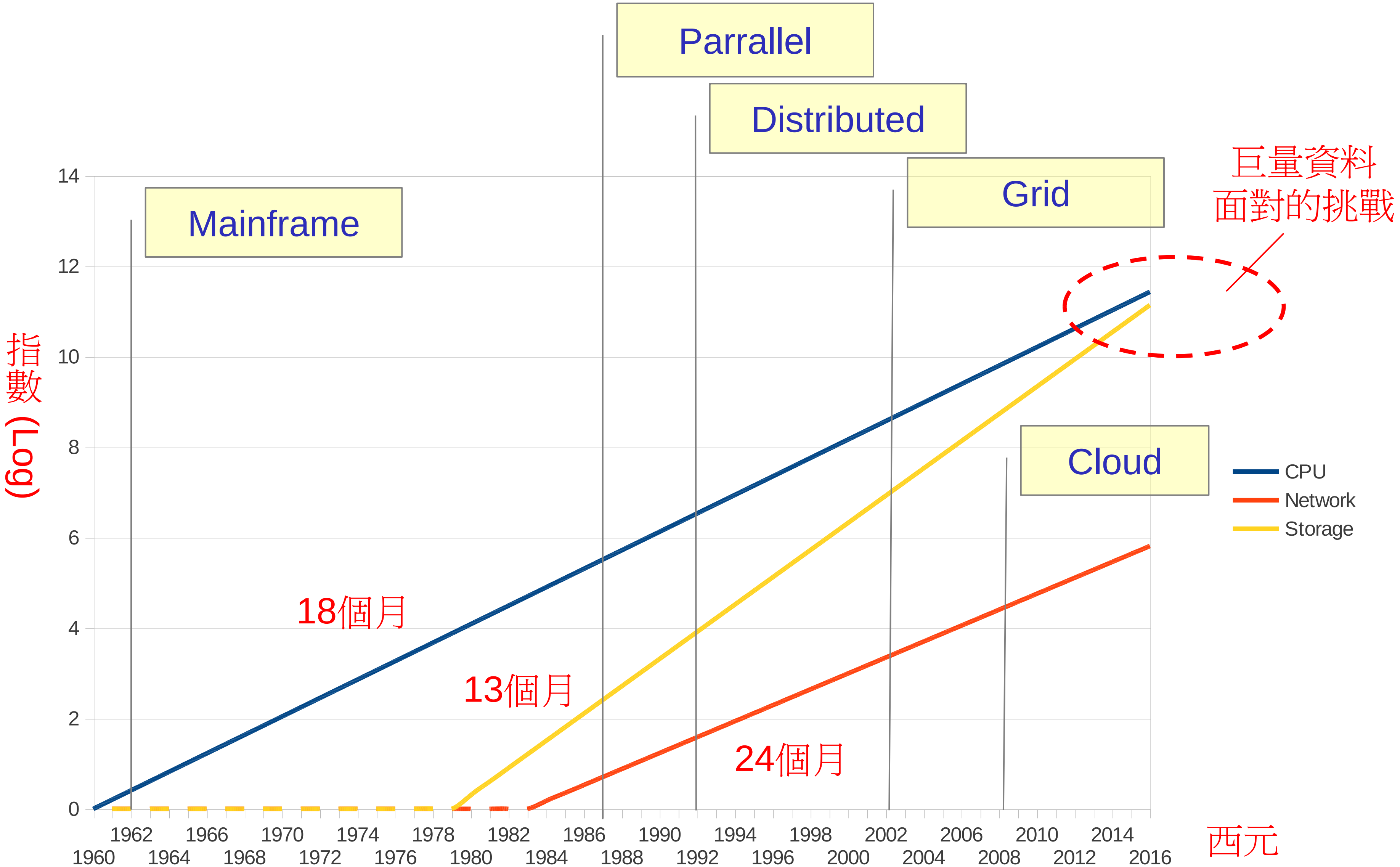
奎德定律是2005年，由希捷資深研發副總馬克奎德提出的。

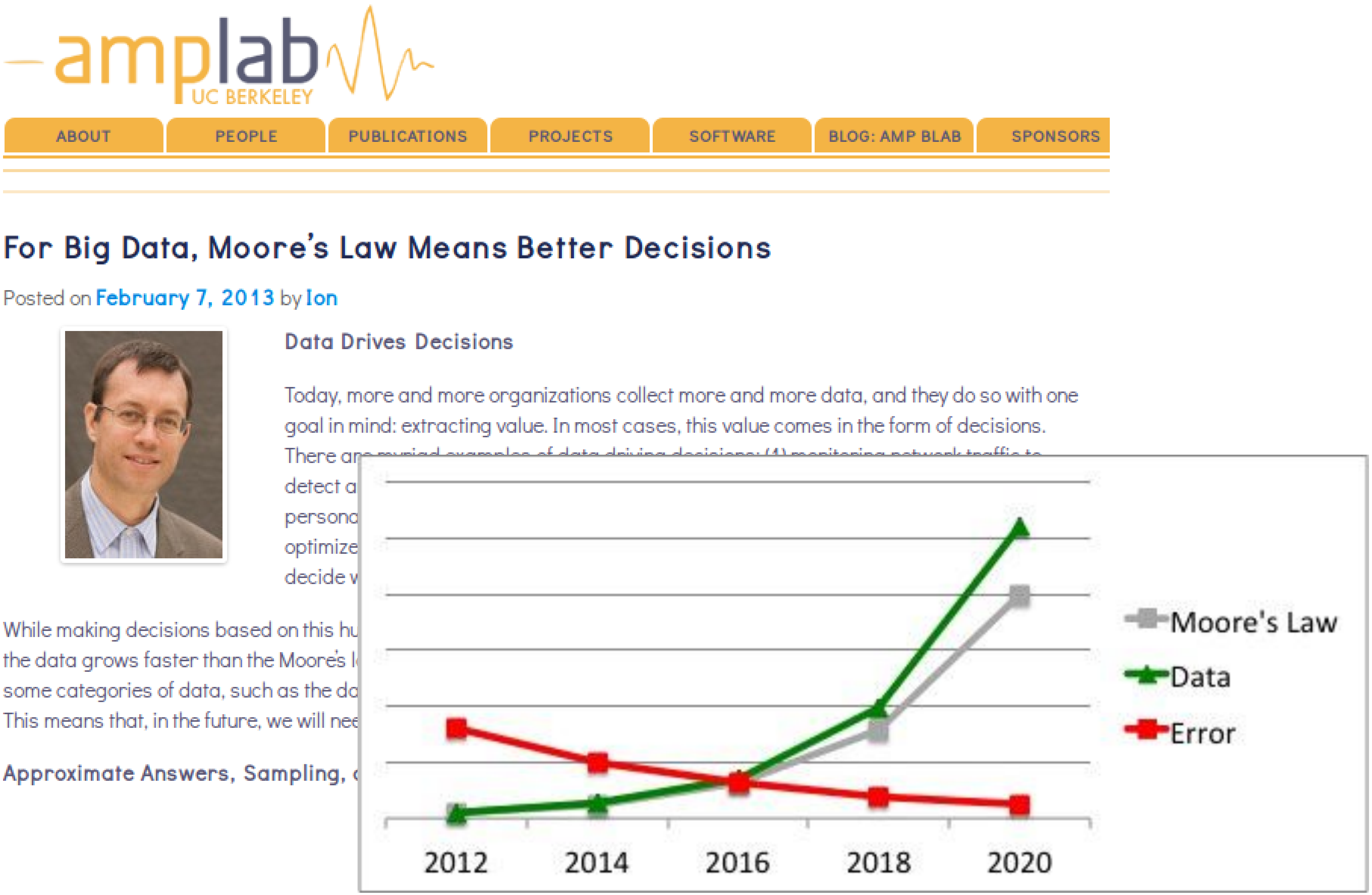
每隔13個月相同價格的儲存容量就會增加一倍。



Source: 奎德定律，科學人雜誌，2005年9月號，第43期
http://203.68.243.199/saweb/read.asp?docsn=2005092489
Image Source: “Hard drive capacity over time following Kryder's Law (1980-2011)”, Wikepedia
http://en.wikipedia.org/wiki/File:Hard_drive_capacity_over_time.svg

Moore's Law , Nielsen's Law , Kryder's Law





amplab UC BERKELEY

ABOUT PEOPLE PUBLICATIONS PROJECTS SOFTWARE BLOG: AMP BLAB SPONSORS

For Big Data, Moore's Law Means Better Decisions

Posted on February 7, 2013 by Ion

Data Drives Decisions

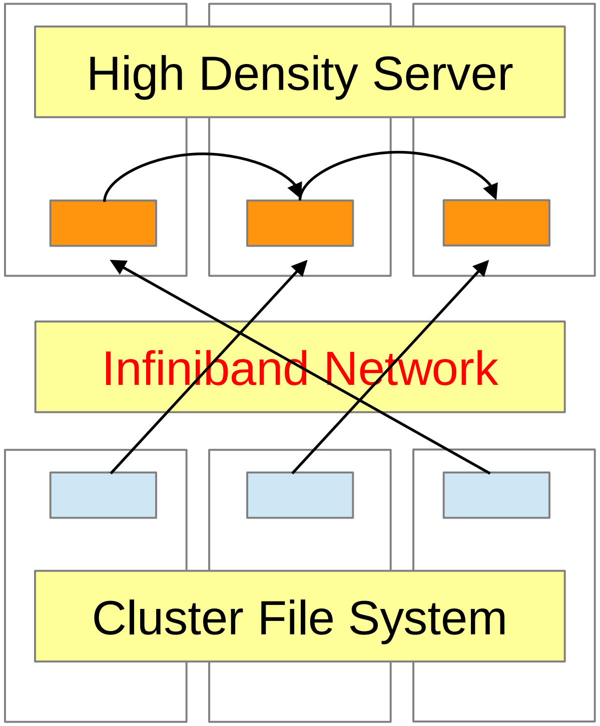
Today, more and more organizations collect more and more data, and they do so with one goal in mind: extracting value. In most cases, this value comes in the form of decisions. There are myriad examples of data driving decisions: (1) monitoring network traffic to
detect a
persona
optimize
decide w

While making decisions based on this hu
the data grows faster than the Moore's l
some categories of data, such as the da
This means that, in the future, we will nee

Approximate Answers, Sampling, c

https://amplab.cs.berkeley.edu/2013/02/07/for-big-data-moores-law-means-better-decisions/

Paradigm Shift in Architecture
from Computing Center to Data Center



High Density Server

Infiniband Network

Cluster File System

Computing Center

Move Data

To Compute

Message Passing

減少資料搬運

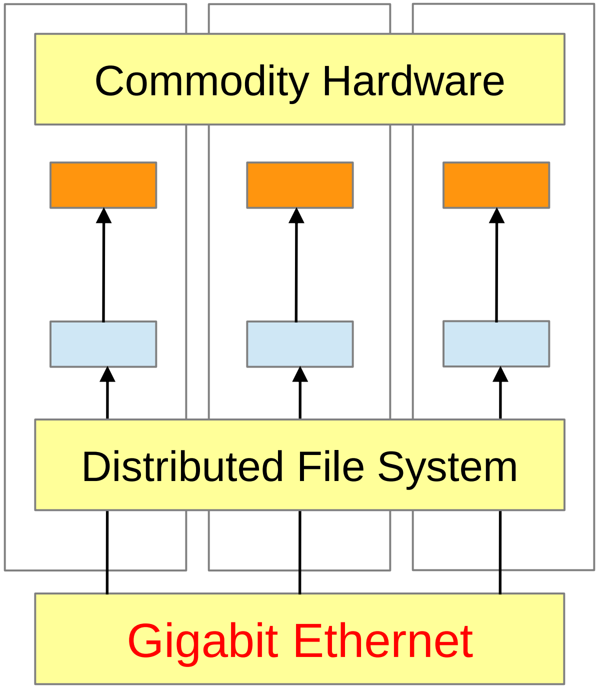
Reduce
Data Transfer

強調能源效率

Energy-
Efficiency

易於橫向擴充

High-
Scalability



Commodity Hardware

Distributed File System

Gigabit Ethernet

Data Center

Move Compute

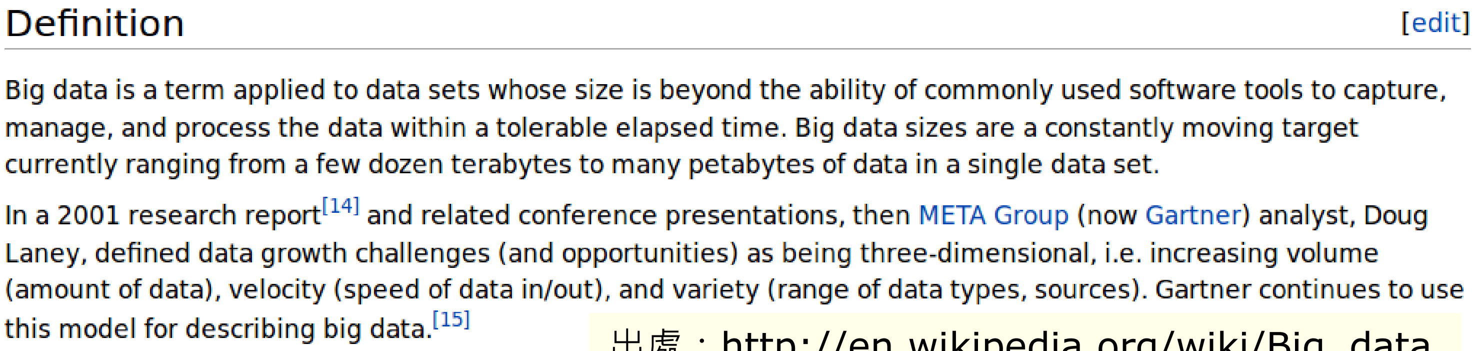
To Data

Share Noting

巨量資料的標準定義 What is Big Data?!

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.

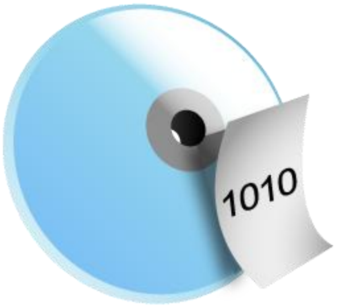
In a 2001 research report[14] and related conference presentations, then META Group (now Gartner) analyst, Doug Laney, defined data growth challenges (and opportunities) as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in/out), and variety (range of data types, 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

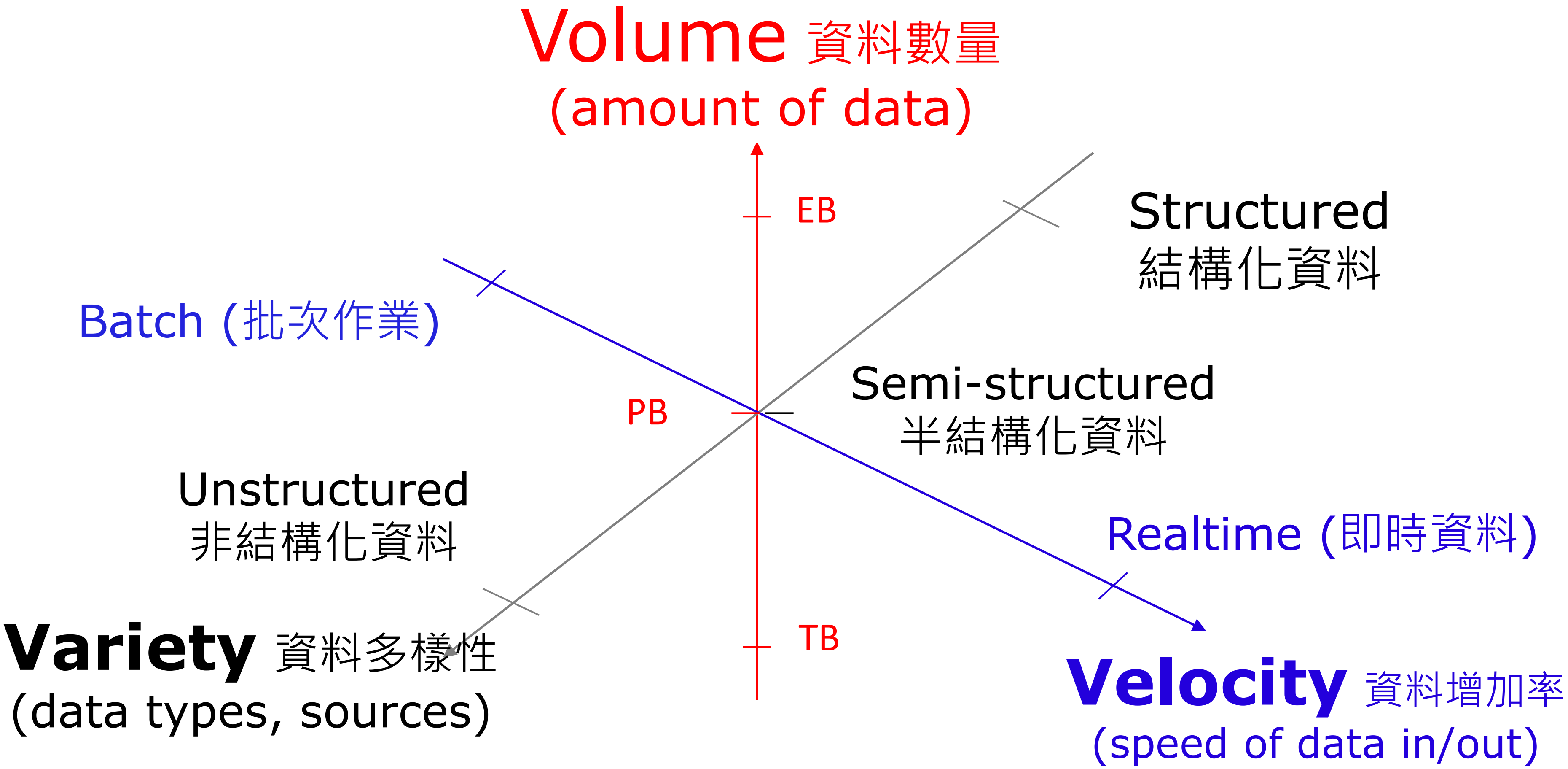
一個檔案，容量100TB



巨量資料的三大挑戰 Challenges - 3 Vs of Big Data

巨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



Volume 資料數量
(amount of data)

EB

PB

TB

Structured
結構化資料

Semi-structured
半結構化資料

Unstructured
非結構化資料

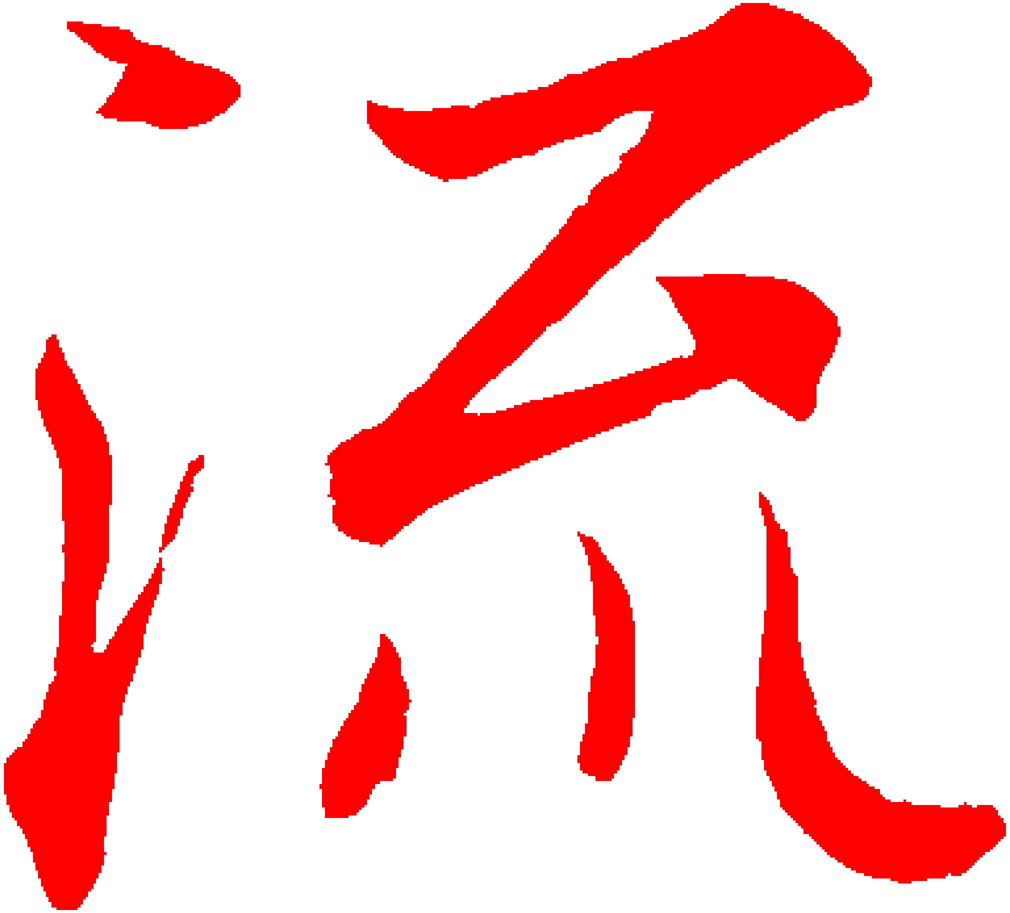
Batch (批次作業)

Realtime (即時資料)

Variety 資料多樣性
(data types, sources)

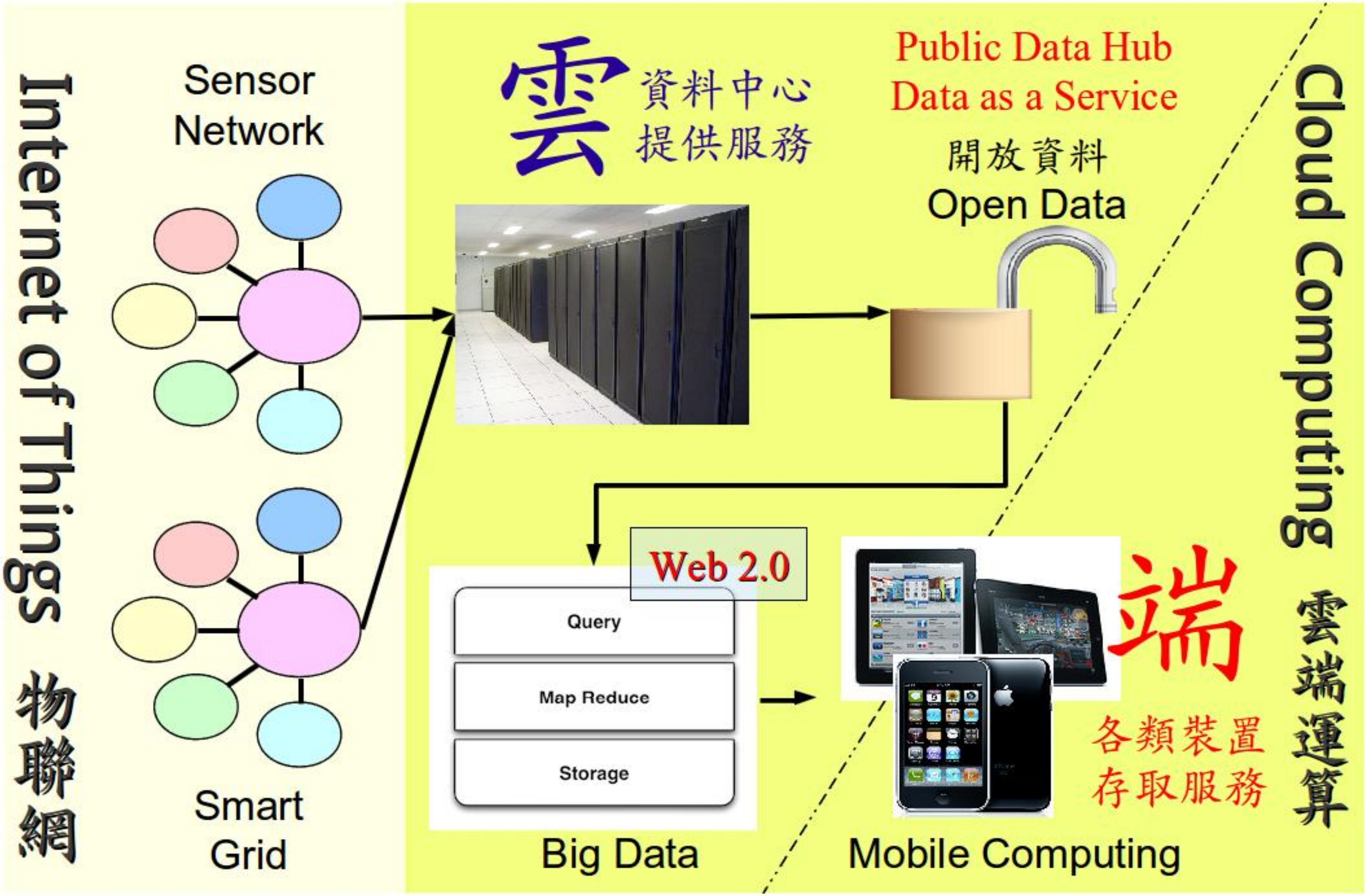
Velocity 資料增加率
(speed of data in/out)

觀點：巨量資料應用的本質



出處碑帖：王羲之 / 臨鍾繇千字文
http://www.cns11643.gov.tw/AIDB/fm_view.do?font=cns_0000&id=38552

巨量資料的奇幻漂流　Life of Big Data



Internet of Things

物聯網

Sensor Network

Smart Grid

雲 資料中心提供服務

Public Data Hub

Data as a Service

開放資料

Open Data

Web 2.0

Query

Map Reduce

Storage

Big Data

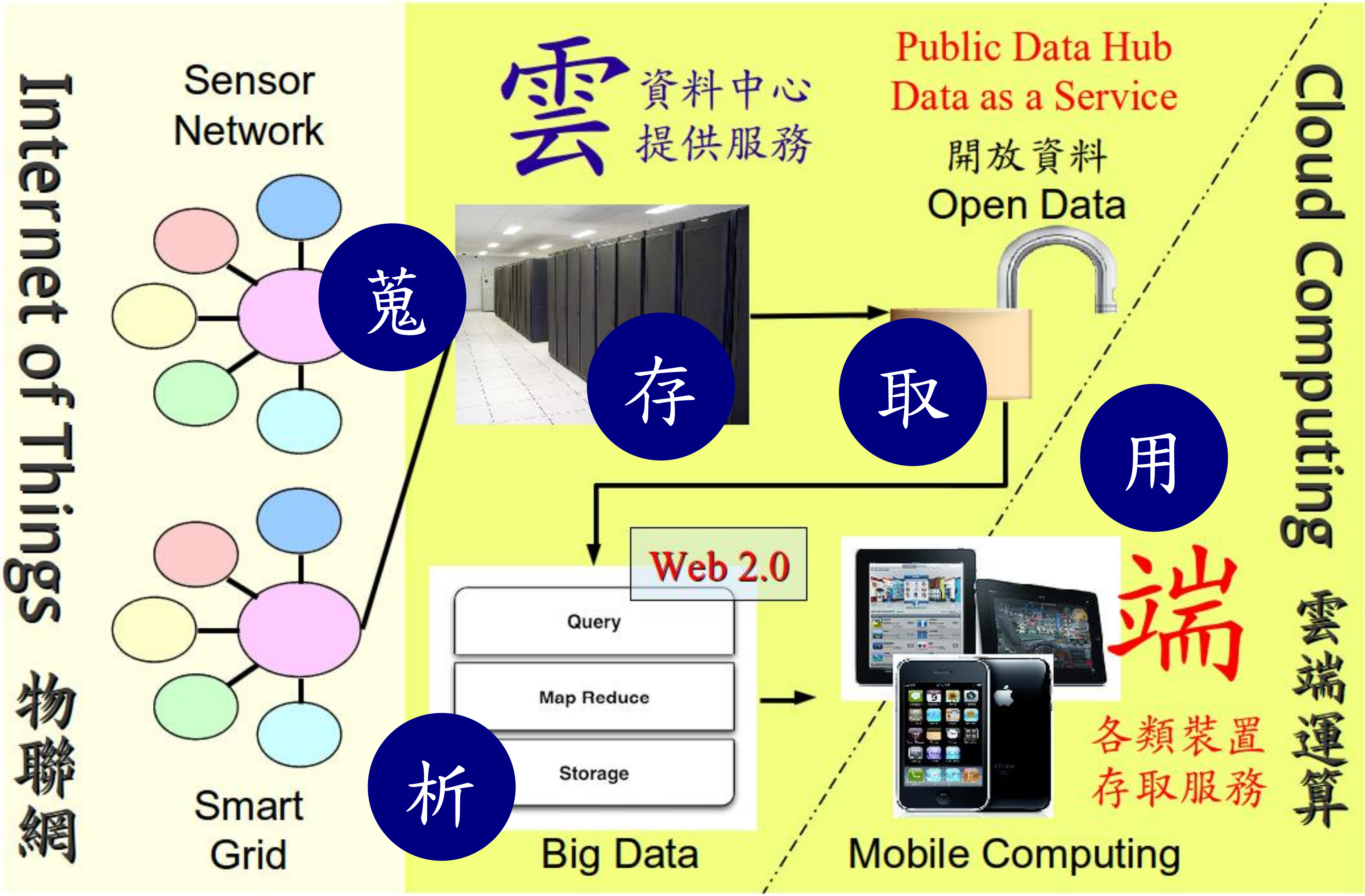
端 各類裝置存取服務

Mobile Computing

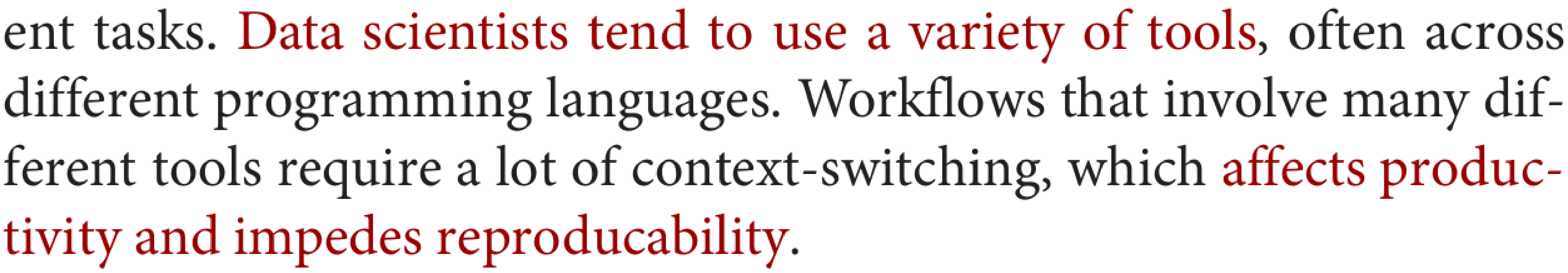
Cloud Computing

雲端運算

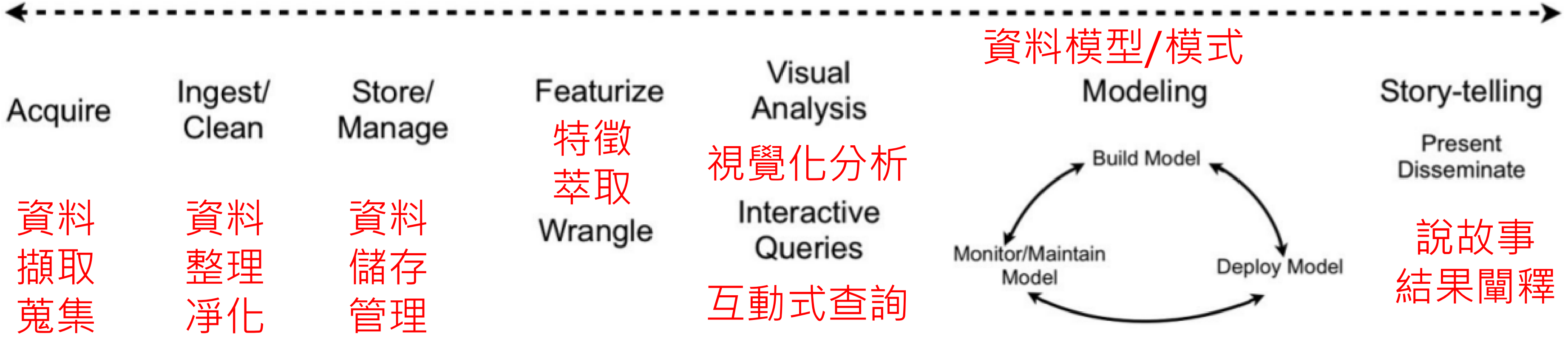
巨量資料的五大生命週期 5 Stage of Big Data Life Cycle



資料科學工作流程 Data Science Workflow



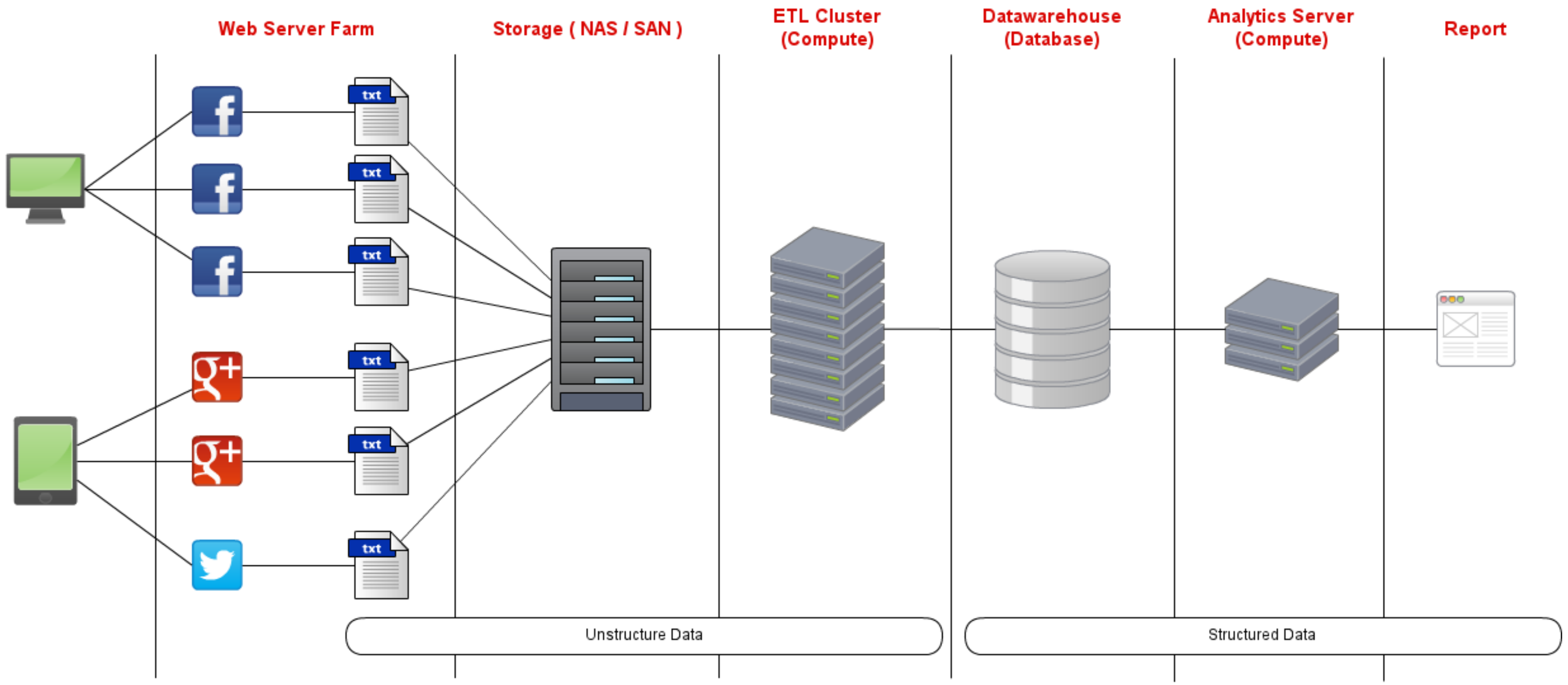
ent tasks. Data scientists tend to use a variety of tools, often across different programming languages. Workflows that involve many different tools require a lot of context-switching, which affects productivity and impedes reproducability.



"Data Analysis: Just One Component of the Data Science Workflow",
By Ben Lorica, 'Big Data Now 2013', OR'eilly
http://www.oreilly.com/data/free/files/bigdatanow2013.pdf

Ex. Data Flow of Log Analysis

生 流 存 算 用 看



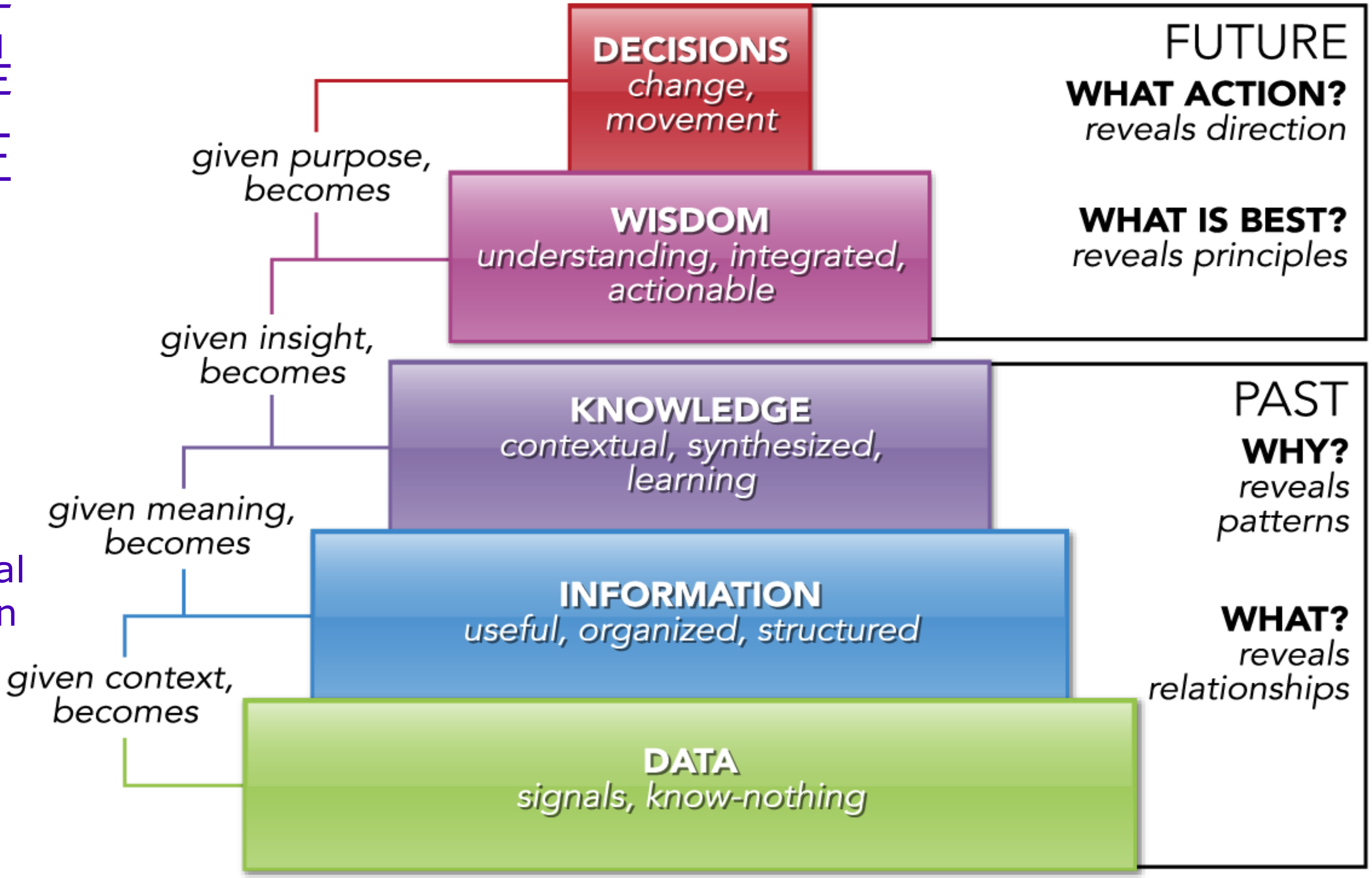
企業導入巨量資料技術的挑戰

知識源自彙整過去，智慧在能預測未來

Knowledge is from the PAST, Wisdom is for the FUTURE.

資料多寡不是重點，重點是我們想要產生什麼價值呢？時效合理嘛？成本合理嘛？

It does not matter how big is your data. The goal is to create VALUE within reasonable time period and total cost of ownership.

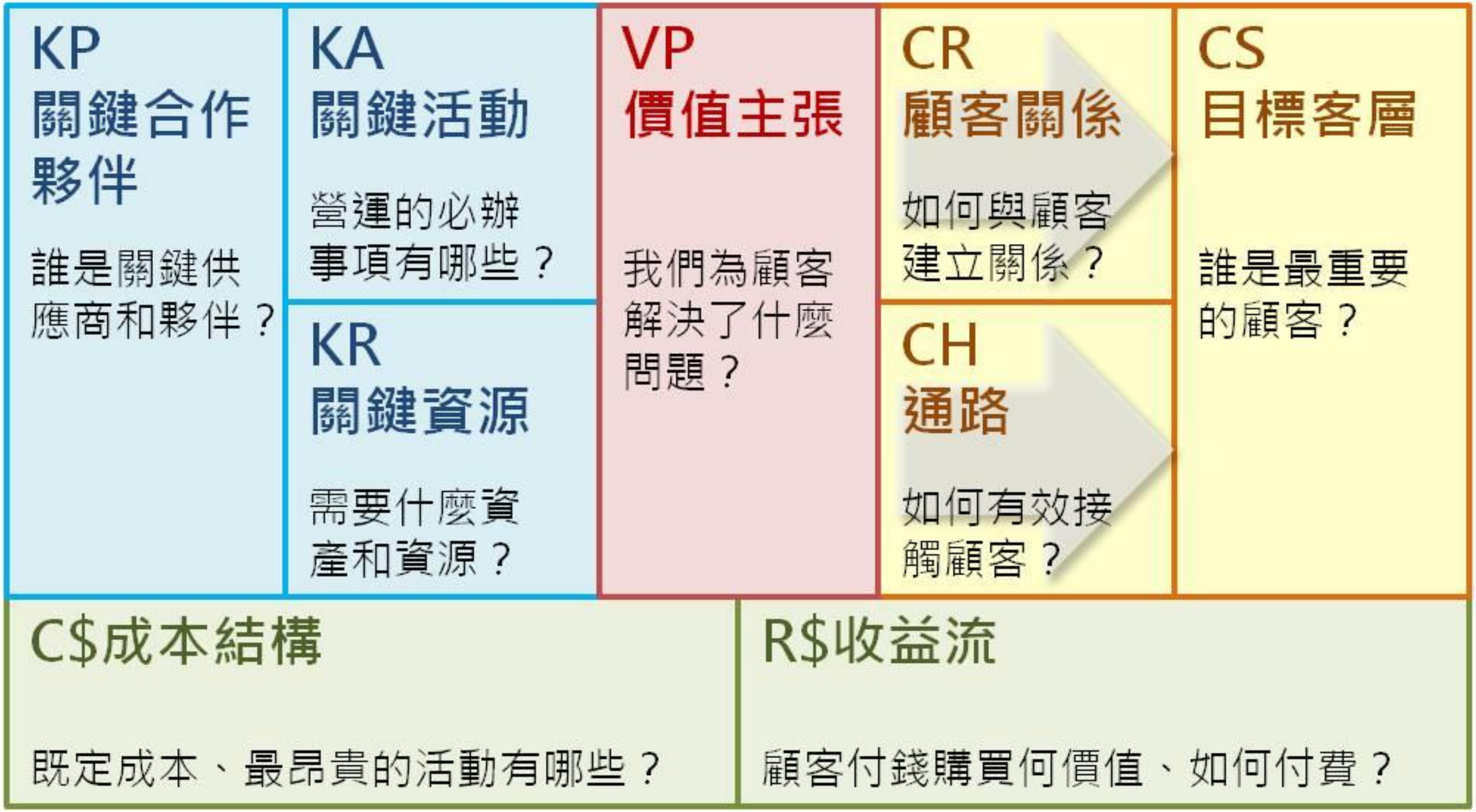


http://www.pursuantgroup.com/blog/tag/dikw-model/

大家都說「資料是金礦」，
那就讓我們拿採礦當類比吧！

國際金價	**提供給客戶的價值**	產品通路
開採成本	總擁有成本	軟硬體投資
提煉廠	分析平台與工具軟體	SMAQ
含金度	資料鑑價？	**商業模式**
開採權	分析資料的合法性	個資法
金礦	資料集	Open Data

從創新到創業，最難的是『創造價值』！



KP
關鍵合作夥伴

誰是關鍵供應商和夥伴？

KA
關鍵活動

營運的必辦事項有哪些？

KR
關鍵資源

需要什麼資產和資源？

VP
價值主張

我們為顧客解決了什麼問題？

CR
顧客關係

如何與顧客建立關係？

CH
通路

如何有效接觸顧客？

CS
目標客層

誰是最重要的顧客？

C$成本結構

既定成本、最昂貴的活動有哪些？

R$收益流

顧客付錢購買何價值、如何付費？

獲利世代：自己動手，畫出你的商業模式
Business Model Generation
http://www.books.com.tw/products/0010567254

雲端運算的商業模式
Business Model of Cloud Computing

規模經濟 (Economies of Scale)

眾人共用資料中心的軟硬體資源，降低總持有成本

資料即服務 (Data as a Service)

綁架你的資料，當資料越來越大，網路傳不動，你就付錢吧！

網路即通路 (Network as a Channel)

一雲多螢，雲端的成功關鍵在於網路頻寬普及率

運算即價值 (Compute as a Value)

當資料集中，連結愈多，愈能透過運算的手段，
找出群眾的智慧，就是提供給客戶最好的價值！