當 Big Data 遇到 Open
挑戰與最佳應用分享

When Big Data Meet Open : Challenges and Best Practices

國家高速網路與計算中心
王耀聰 <jazz@narlabs.org.tw>

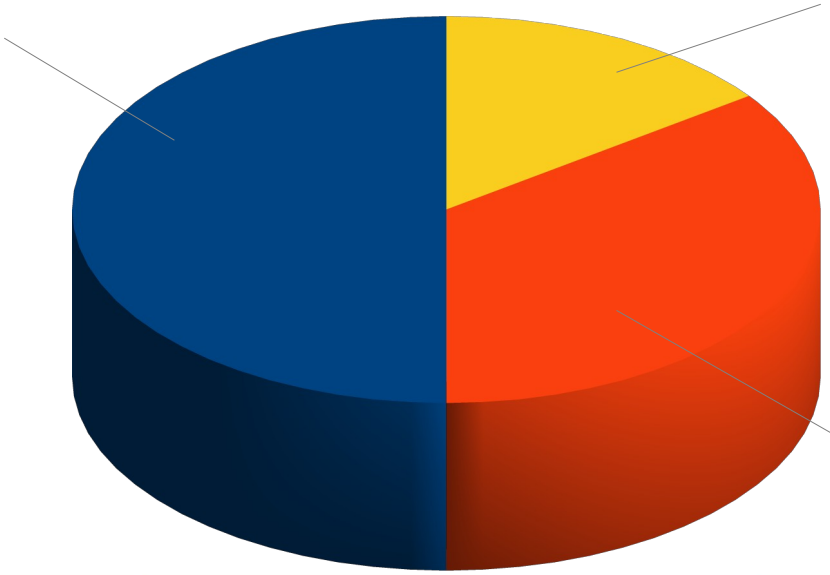
2013/11/20 - IBM 跨世代應用伺服器 趨勢論壇

WHO AM I ? 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
  - Co-Founder of Hadoop.TW 台灣 Hadoop 傳教士
  - 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

NARLabs

演講大綱 Agenda

- **Linux is everywhere** 開放無所不在
- **What is Big Data ?** 何謂巨量資料
- **Big Data in Motion !** 即時巨資應用
- **The Next Big Thing ?** 下半場的重點
- **Conclusion** 三大結論回顧

2013 三大熱門關鍵字

物聯網

Internet of Things

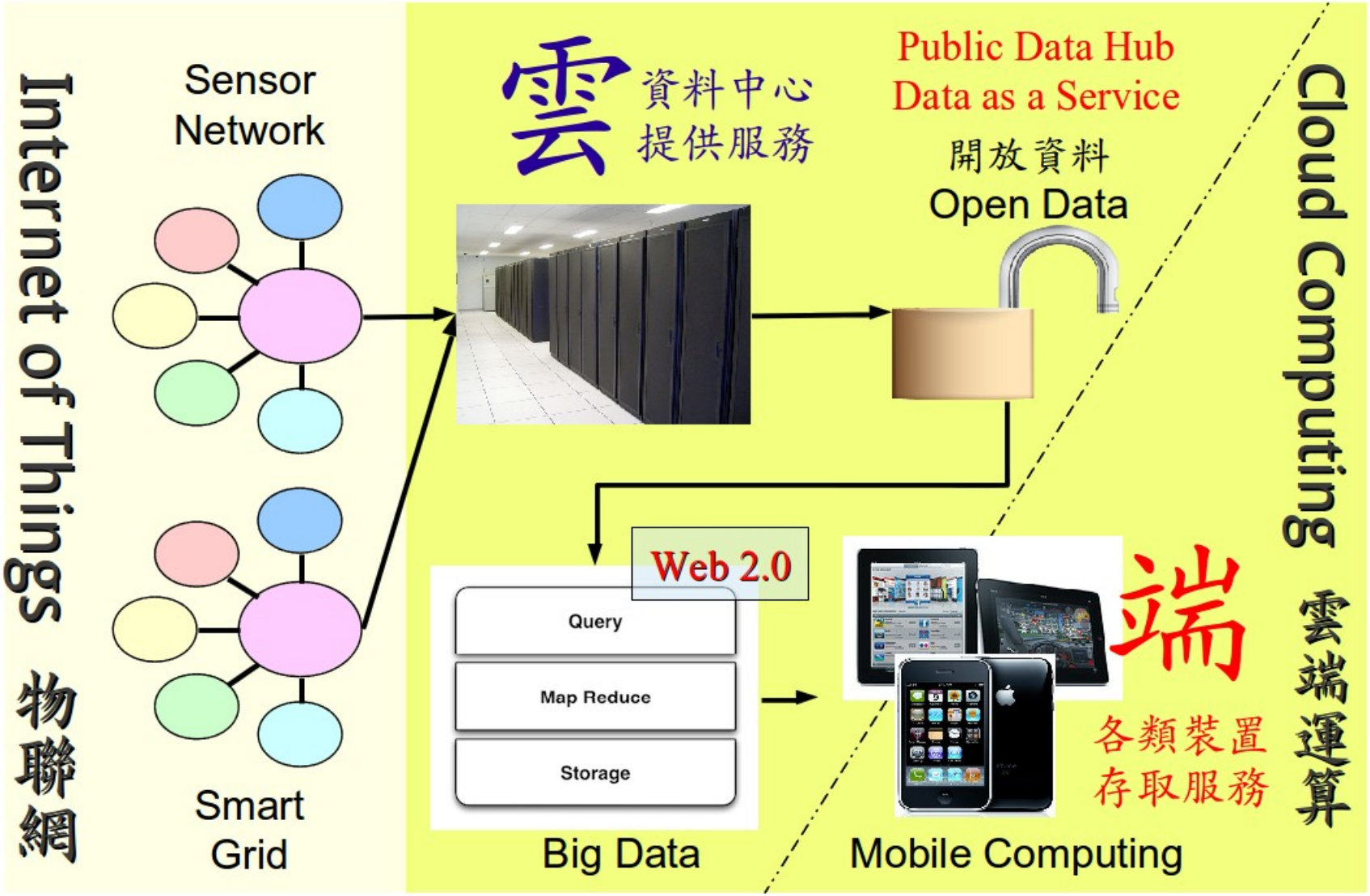
雲端運算

Cloud Computing

巨量資料

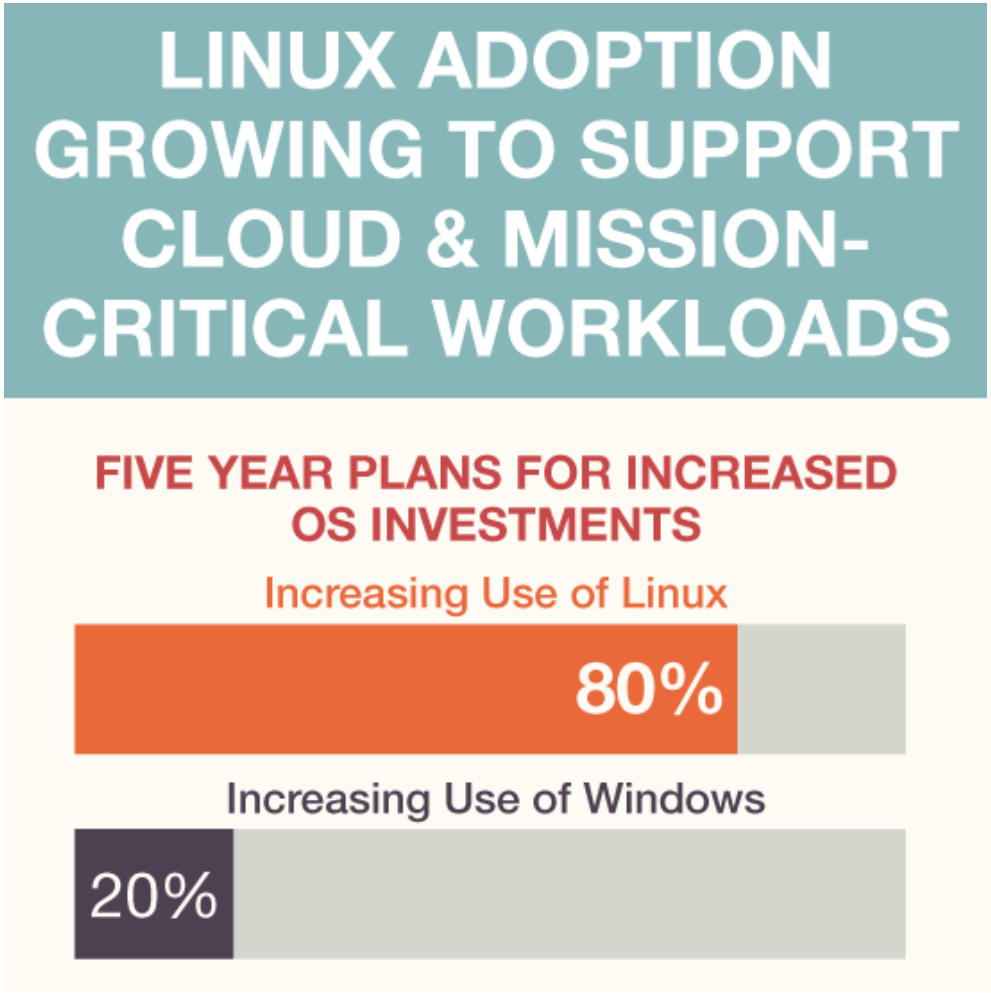
Big Data

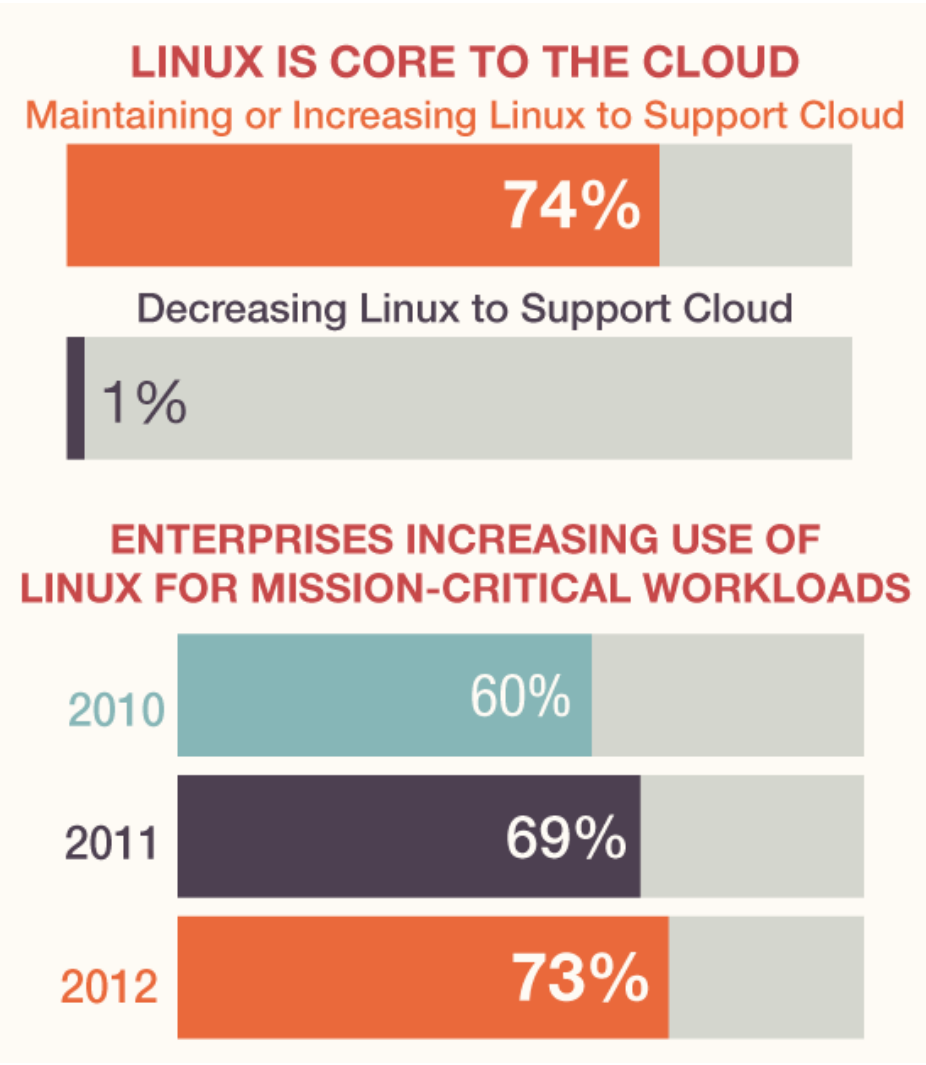
巨量資料的奇幻漂流
Life of Big Data



Linux Adoption in Enterprise increase in last 3 years

NARLabs





Source: "2013 Enterprise End User Report", Linux Foundation, March 2013

http://www.linuxfoundation.org/publications/linux-foundation/linux-adoption-trends-end-user-report-2013

NARLabs

Linux in the Cloud

← Amazon

Yahoo ↓

If you have 4000+ server, which OS will you choose?

http://www.datacenterknowledge.com/archives/2010/09/20/inside-the-yahoo-computing-coop/
http://bits.blogs.nytimes.com/2013/01/08/amazons-unknown-unknowns/?_r=0

NARLabs

Linux in the Devices !!

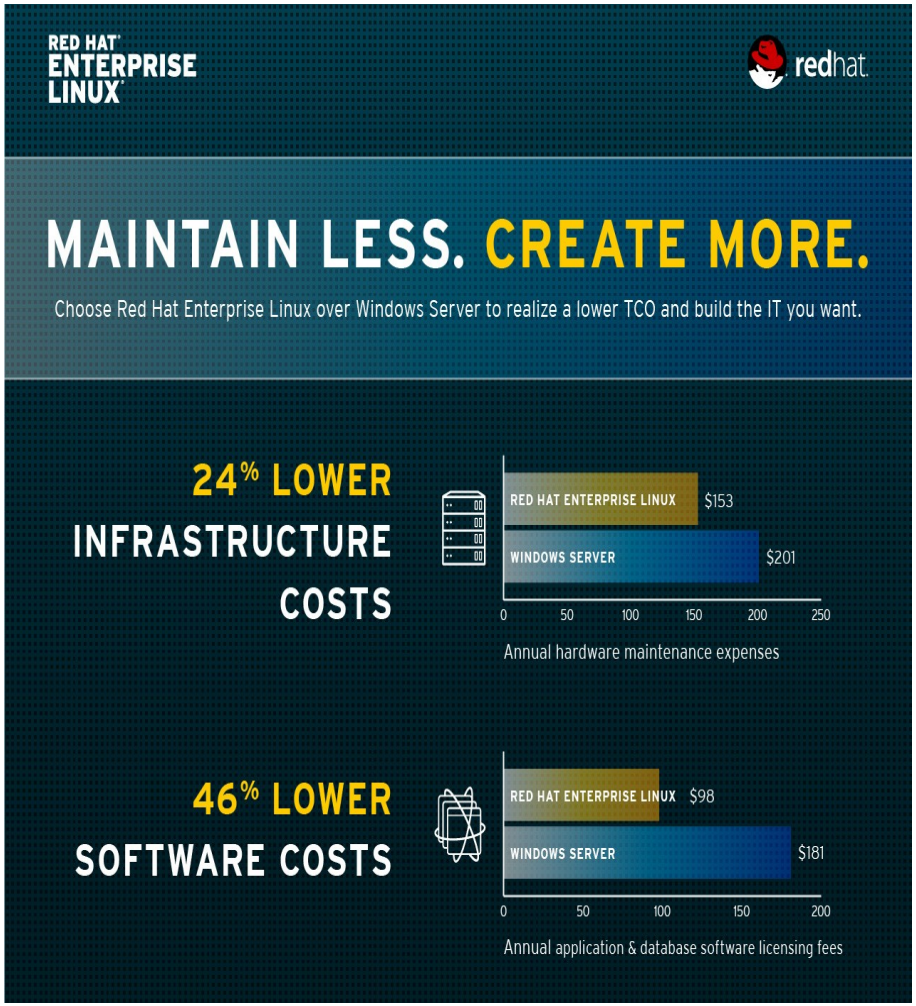
Linux have dominated Embedded, Mobile .....
maybe Internet of Things in near future ....

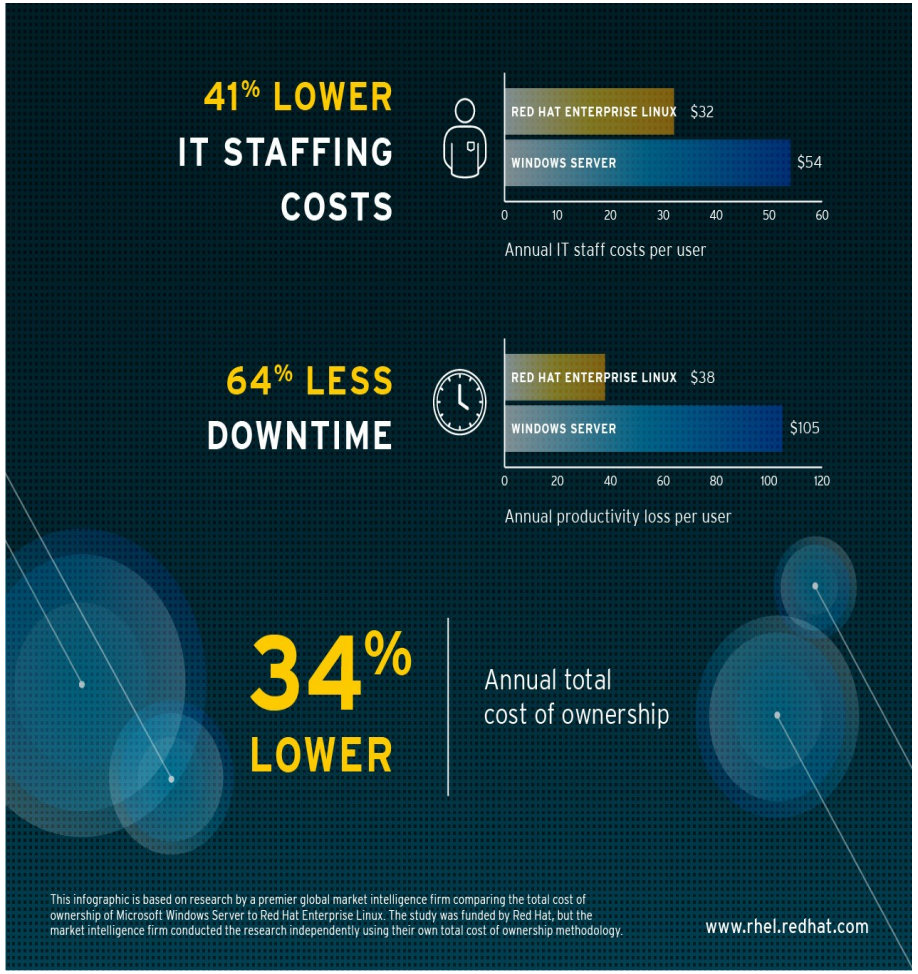
Google Chrome OS

http://crackberry.com/sites/crackberry.com/files/styles/large/public/topic_images/2013/ANDROID.png
http://boxysystems.com/myblog/wp-content/uploads/2011/01/google-chrome-OS-logo.jpg

Source: The most popular end-user Linux distributions are ...
http://www.zdnet.com/the-most-popular-end-user-linux-distributions-are-7000017223/

結論一： Why Linux ？
Total Cost of Ownership !





Source: How Red Hat Enterprise Linux trims total cost of ownership (TCO) in comparison to Windows Server, October 7, 2013
http://www.redhat.com/about/news/archive/2013/10/how-red-hat-enterprise-linux-trims-total-cost-of-ownership-in-comparison-to-windows-server

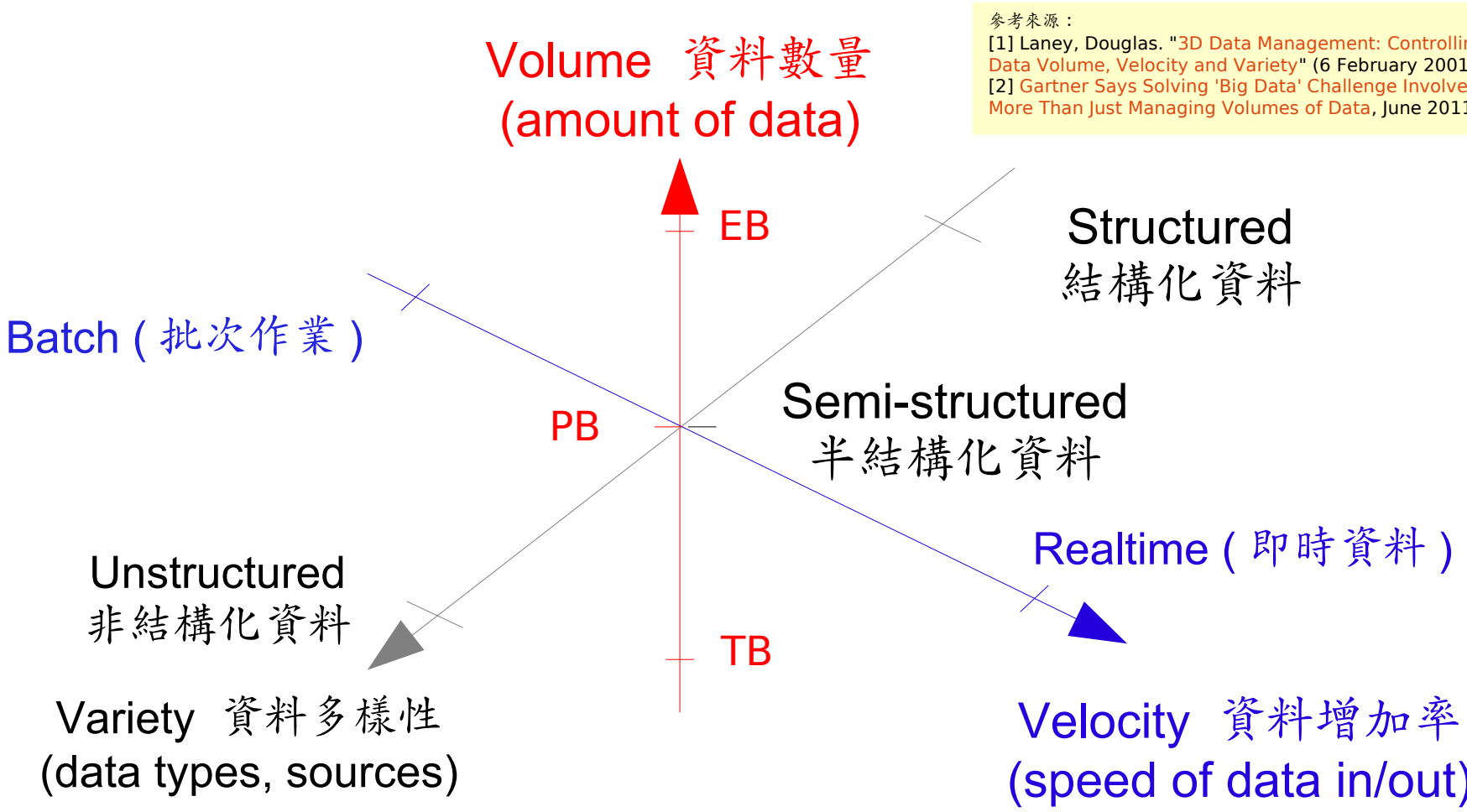
演講大綱 Agenda

- **Linux is everywhere** 開放無所不在
- **What is Big Data ?** 何謂巨量資料
- **Big Data in Motion !** 即時巨資應用
- **The Next Big Thing ?** 下半場的重點
- **Conclusion** 三大結論回顧

巨量資料的三大挑戰
3 Vs of Big Data

NARLabs

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



巨量資料的挑戰在於如何管理「數量」、「增加率」與「多樣性」

大家都說「資料是金礦」，
那就讓我們拿採礦當類比吧！

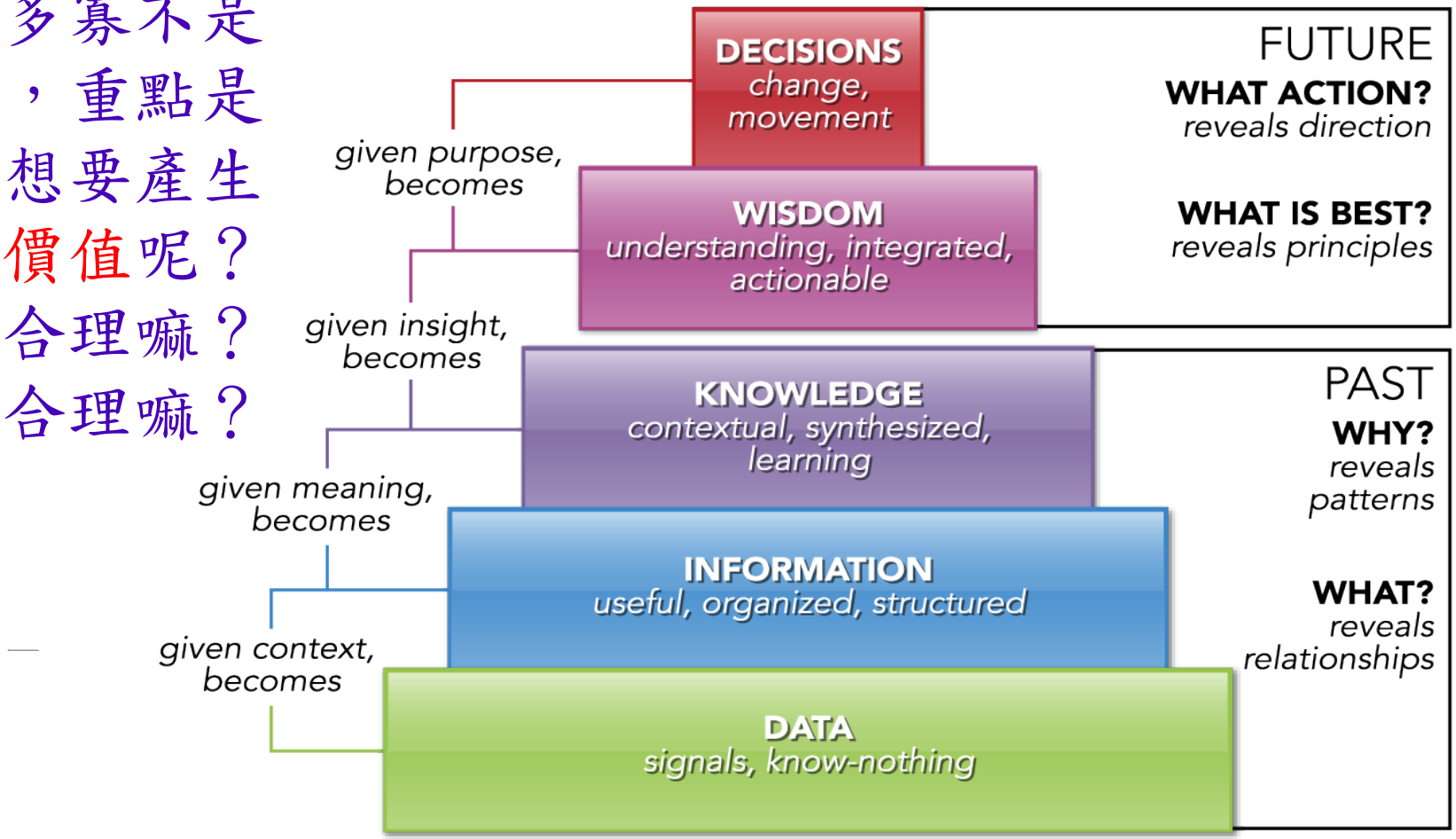
NARLabs

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**

知識源自彙整過去，
智慧在能預測未來

NARLabs

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？



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

NARLabs

PAST: Big Data at Rest

Can gigabytes predict the next Lady Gaga?

By Stacey Higginbotham

Want to know how playing on Jimmy Kimmel Live will boost the sales of an artist's album? Or how about figuring out where fans go to find artists after they hit the evening news? What about the effect Whitney Houston's death had on her YouTube and Vevo plays? They shot up 4,525 percent, by the way.

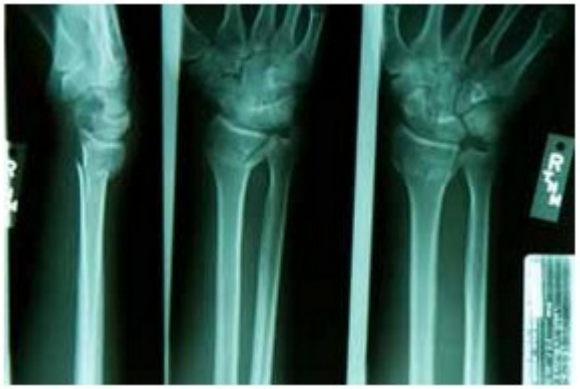
http://nextbigsound.com/

fakebook®

NETFLIX

How big data can curb the world's energy consumption

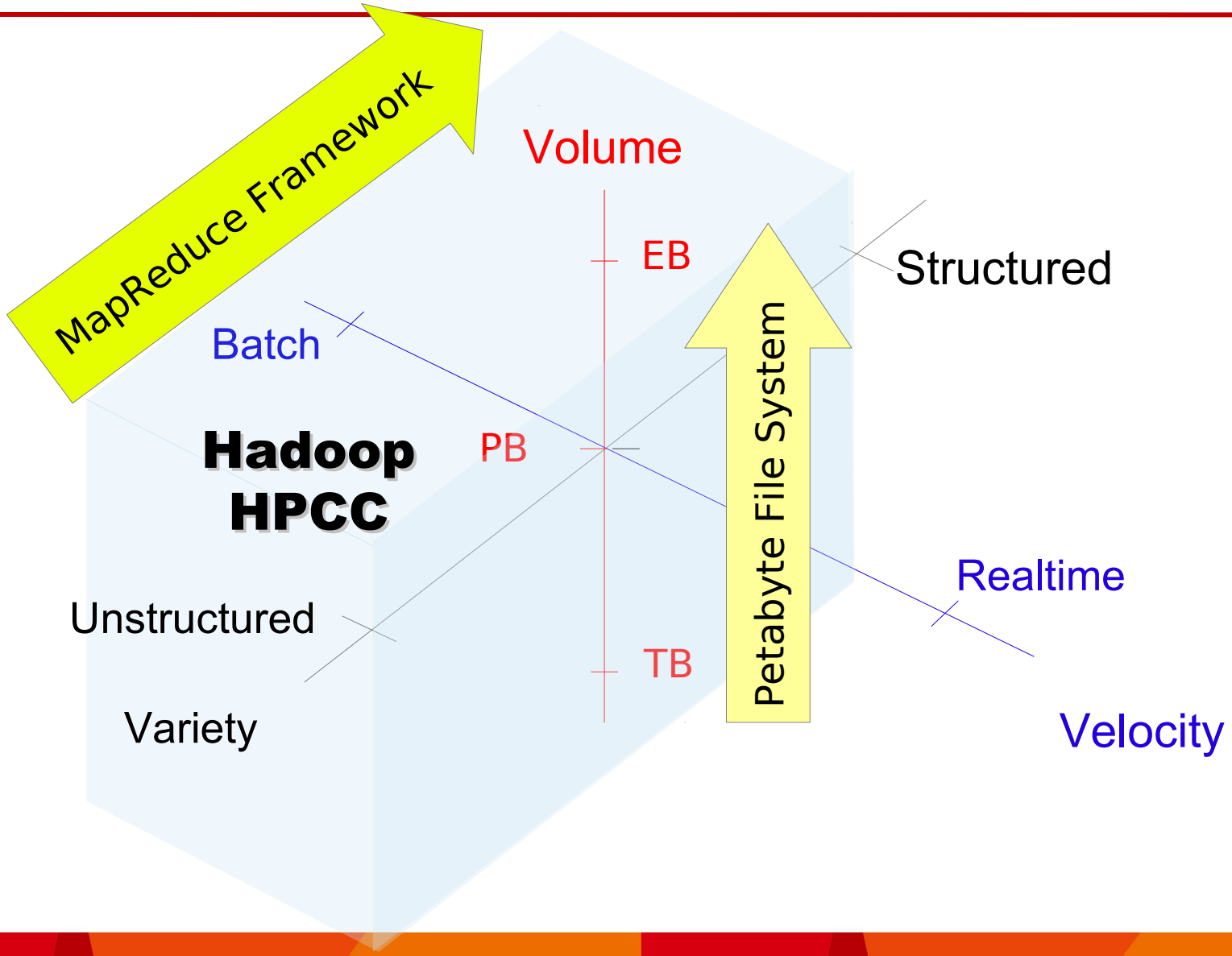
http://www.openpdc.com/



One hospital's embrace of big data

Source: 10 ways big data changes everything,
http://gigaom.com/2012/03/11/10-ways-big-data-is-changing-everything

處理巨量資料的三類技術 (1)
Data at Rest – MapReduce Framework



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

NARLabs

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

高資料通量處理平台 Hadoop
Key Concept : Data Locality

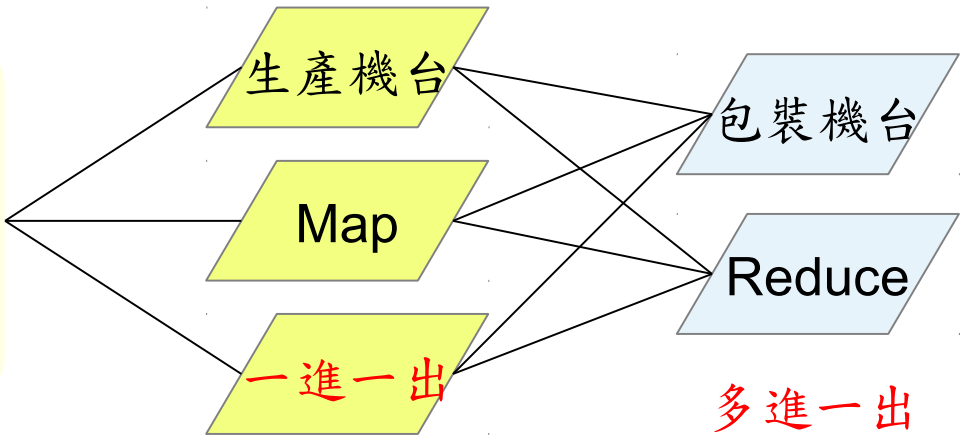
NARLabs

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。

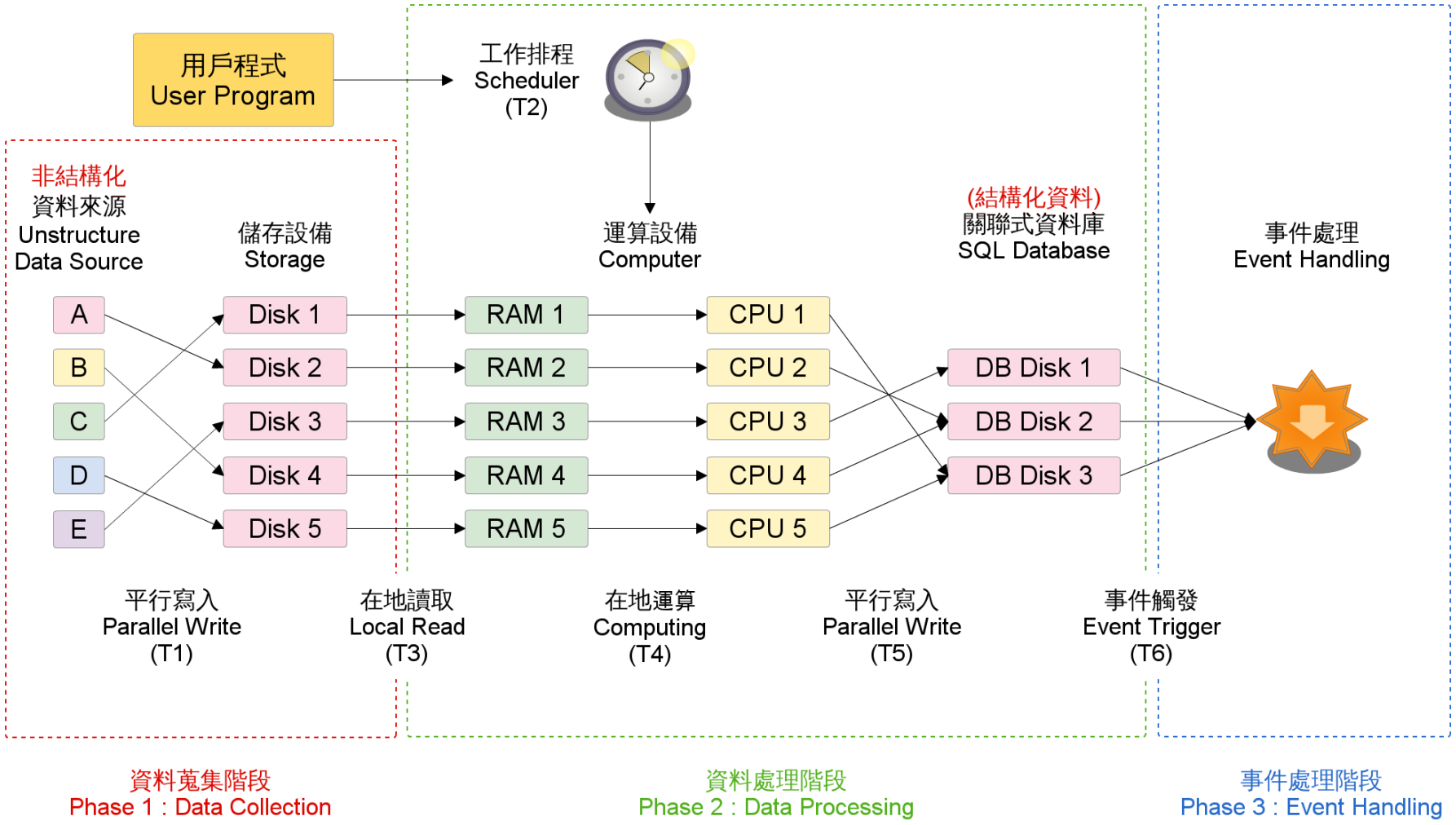
亦可以想像成一個處理海量資料的生產線，只須學會定義 **map** 跟 **reduce** 工作站該做哪些事情。

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



批次作業的運算時間
Processing Time of Batch Jobs



NARLabs

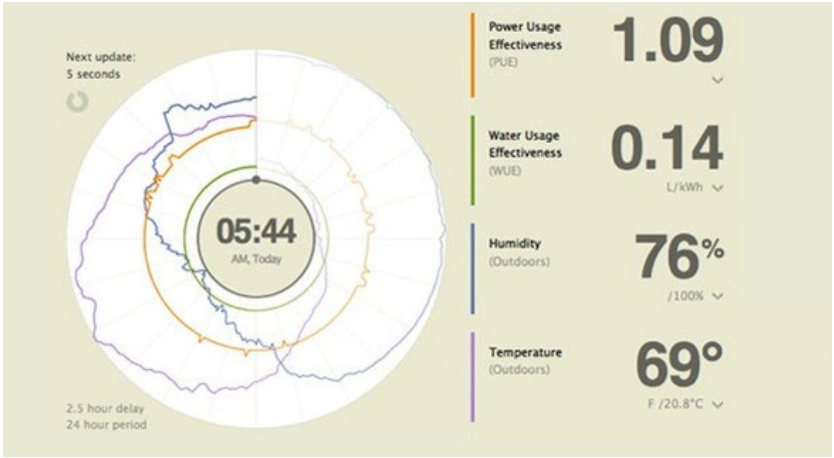
演講大綱 Agenda

Linux is everywhere	開放無所不在
What is Big Data ?	何謂巨量資料
Big Data in Motion !	即時巨資應用
The Next Big Thing ?	下半場的重點
Conclusion	三大結論回顧

NOW: Big Data in Motion

[金融] Trading Robot

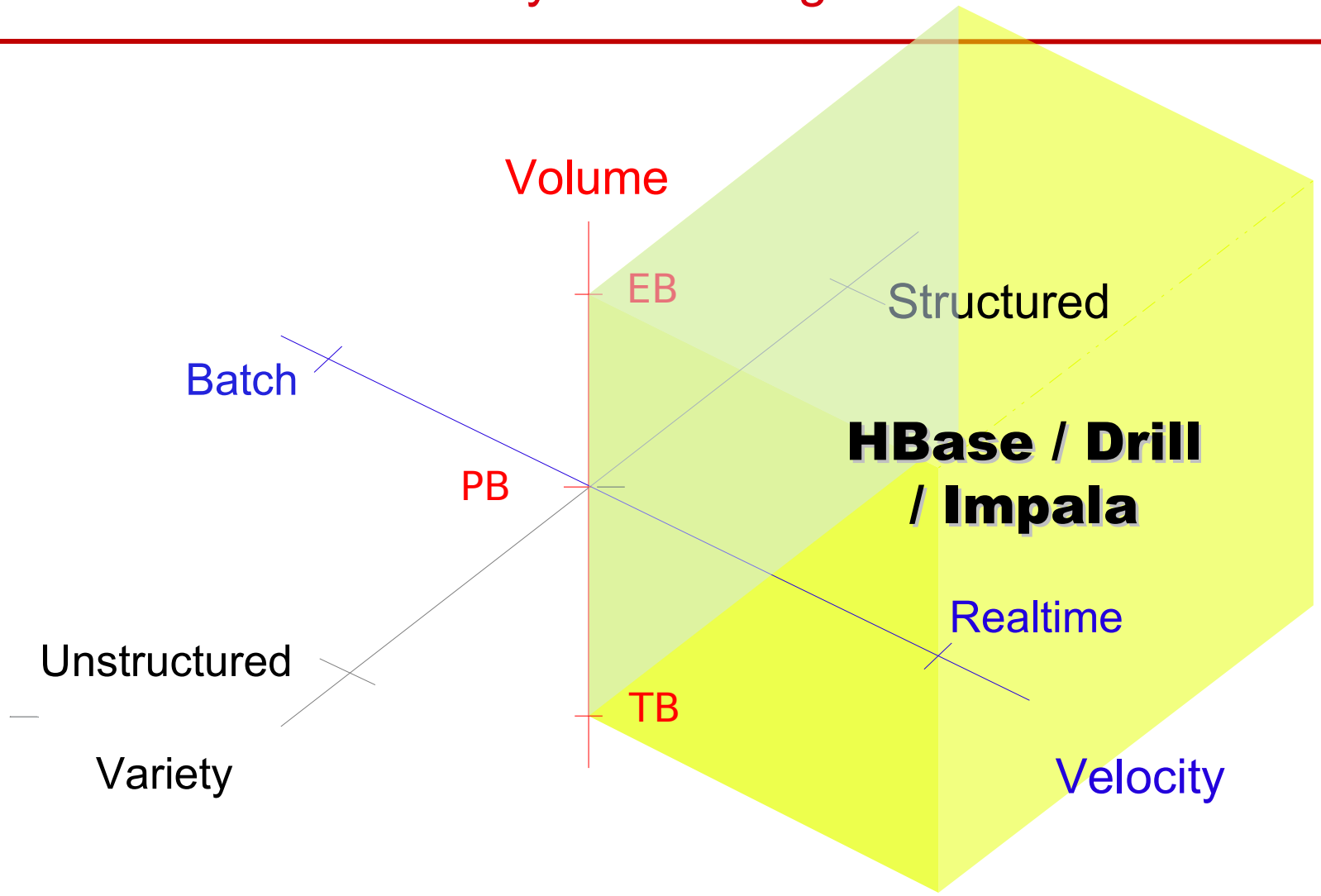
[災防] 海嘯、土石流



[資訊] 機房即時用電資訊監控、警訊

http://www.newmobilelife.com/2013/04/21/facebook-pue-real-time-charts/

處理巨量資料的三類技術 (2)
Data in Motion – In-Memory Processing



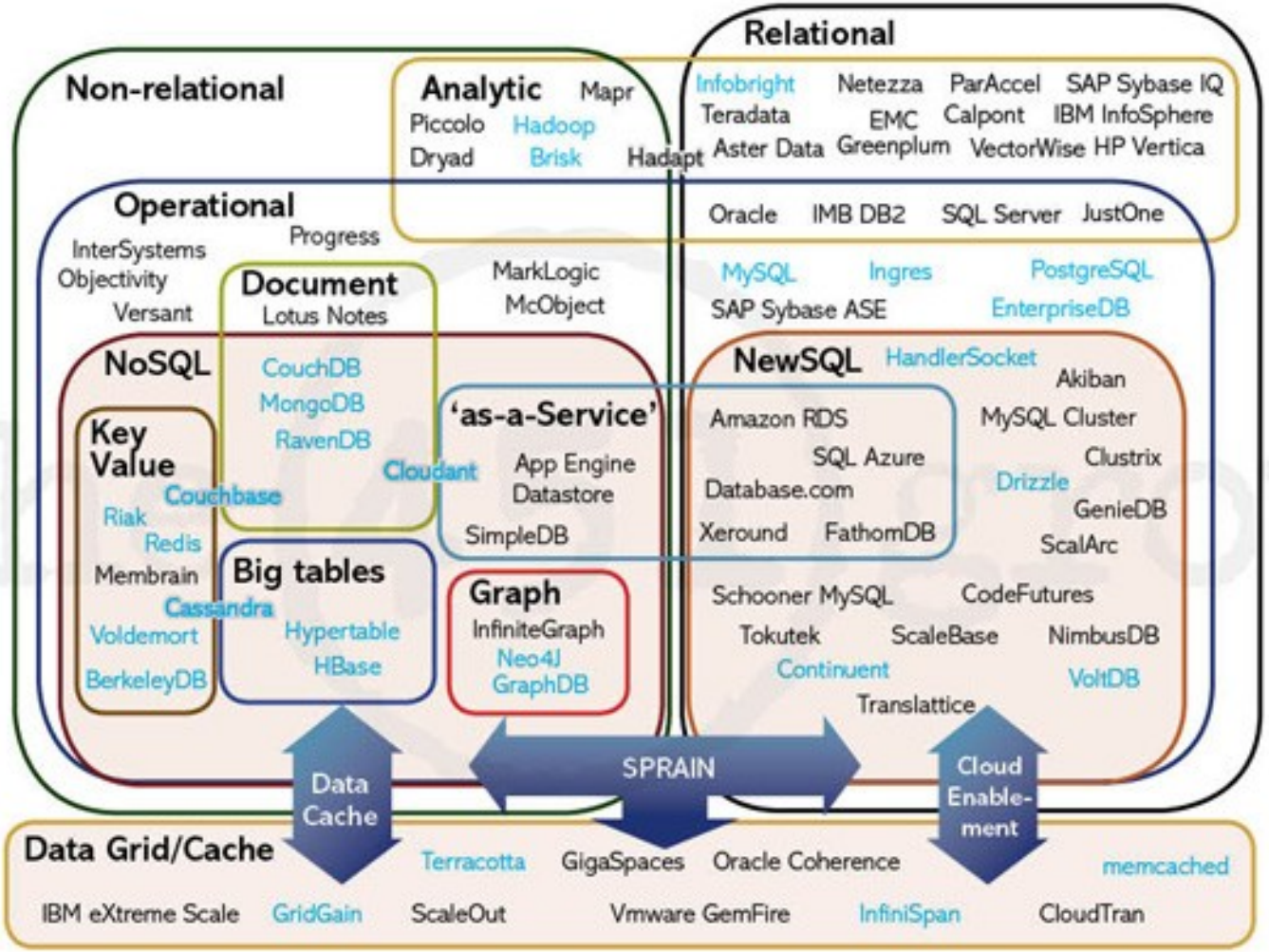
Google 的技術演進 vs Apache 專案

NARLabs

Big Query (JSON, SQL-like)	Dremel (2010)	Apache Drill (2012)
Incremental Index Update (Caffeine)	Percolator (2010)	
Graph Database	Pregel (2009)	Apache Giraph (2011)
Query	BigTable (2006)	Apache HBase (2007)
Map Reduce	MapReduce (2004)	Hadoop MapReduce (2006)
Storage	Google File System (2003)	HDFS (2006)

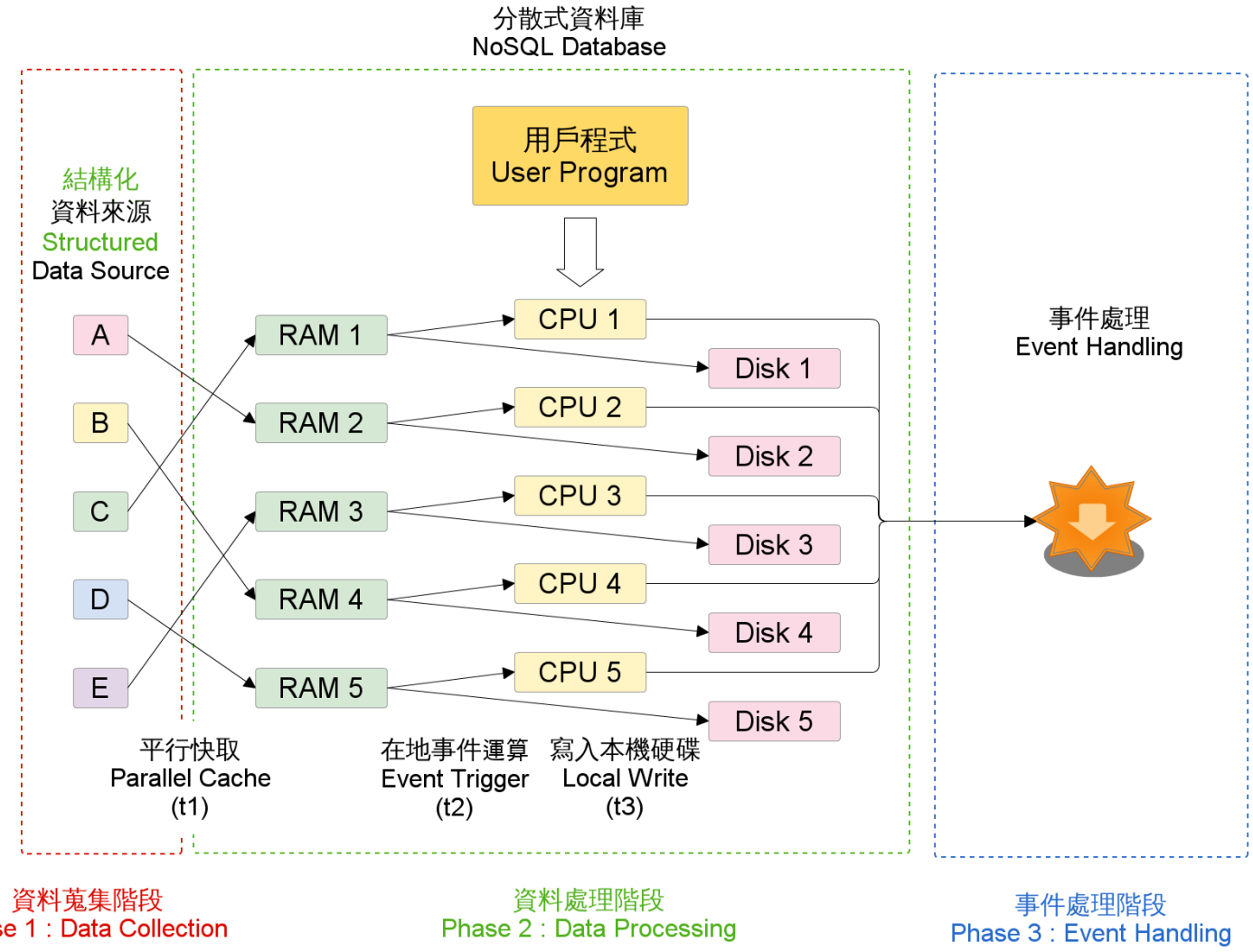
令人眼花撩亂的多樣化資料庫選擇
NoSQL vs NewSQL

NARLabs

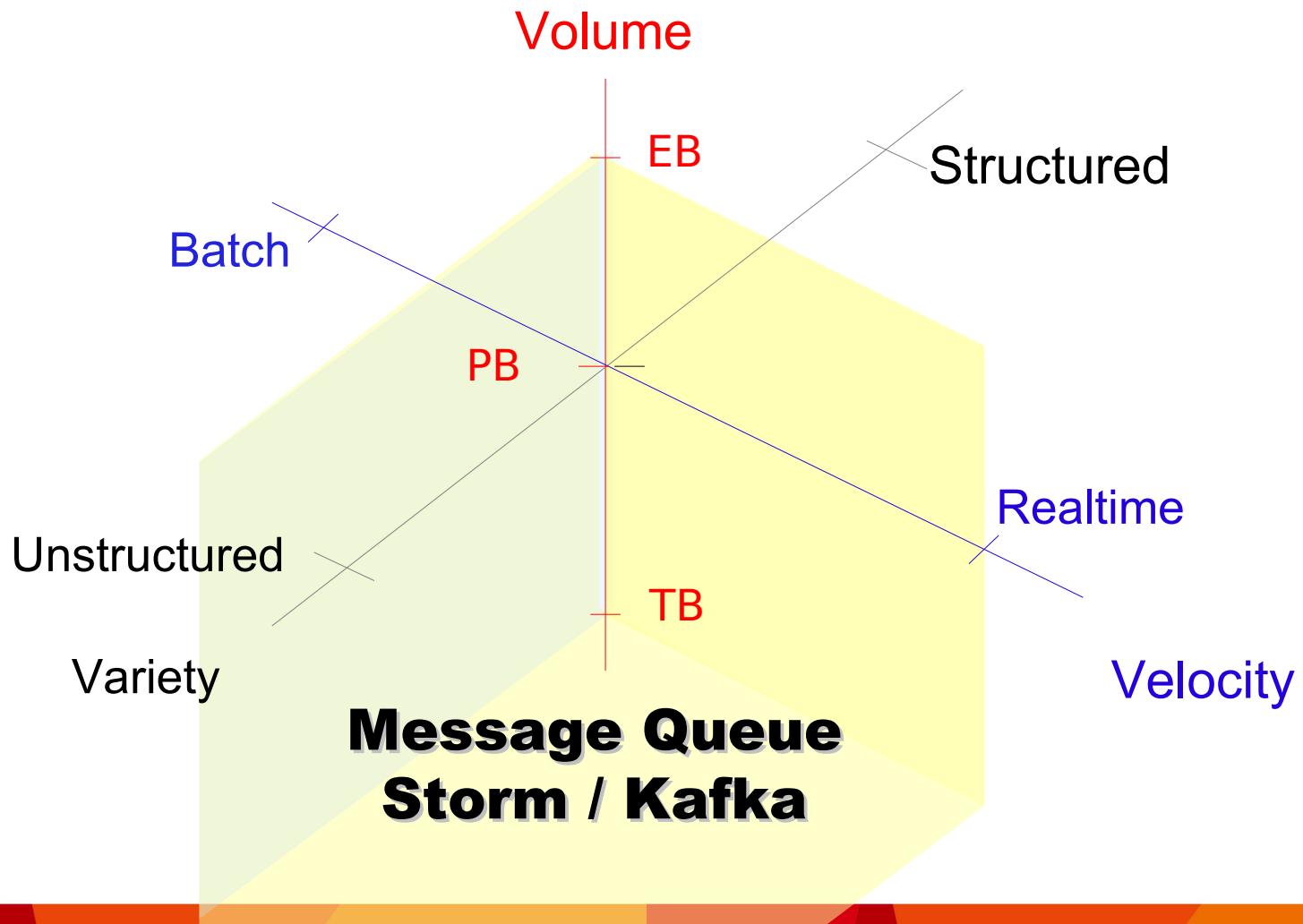


http://www.infoq.com/news/2011/04/newsql

In-Memory Processing 的運算時間
以 HBase 為例

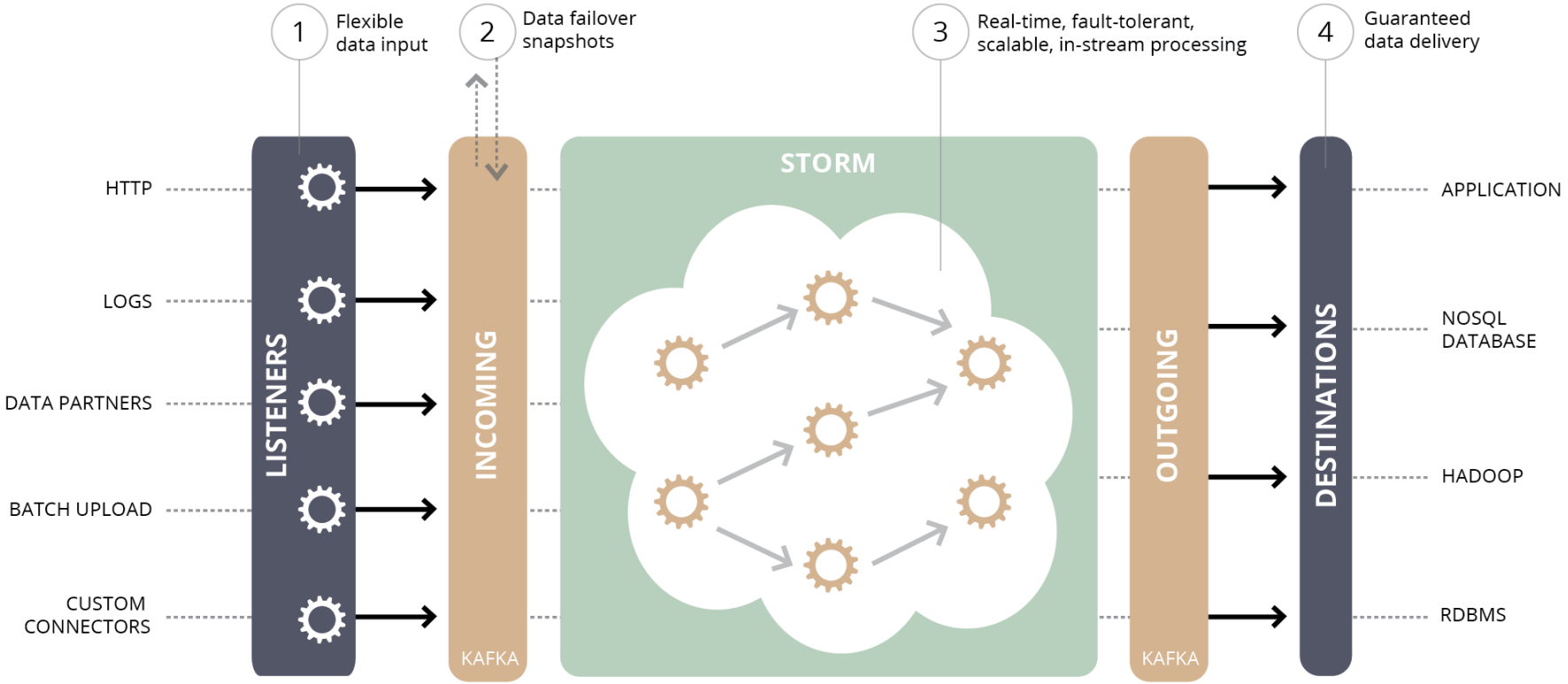


處理巨量資料的三類技術 (3)
Streaming Data Collection



Twitter Storm + Apache Kafka

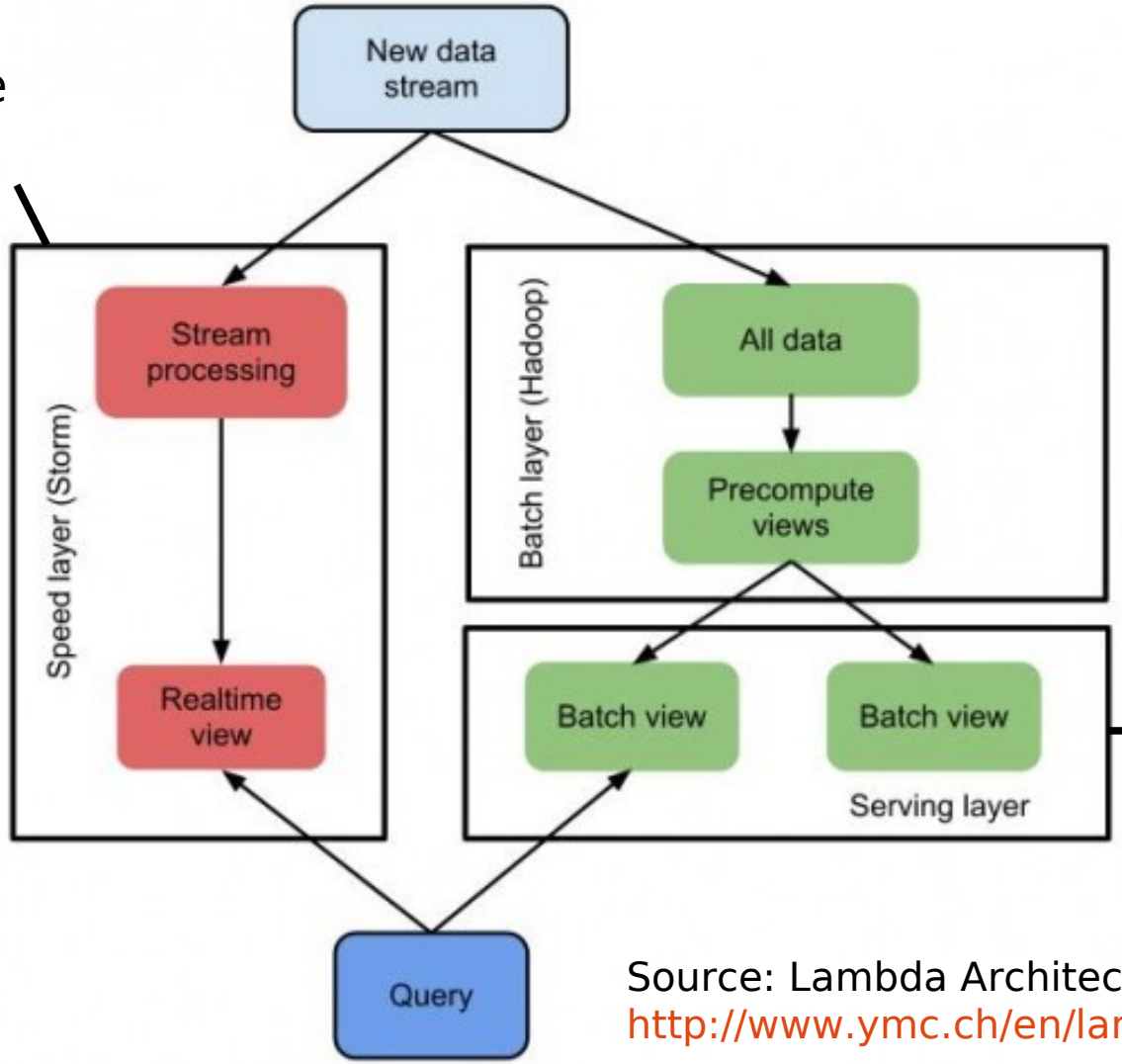
NARLabs



http://blog.infochimps.com/2012/10/30/next-gen-real-time-streaming-storm-kafka-integration/

混合模式的巨量資料處理架構
Lambda Architecture

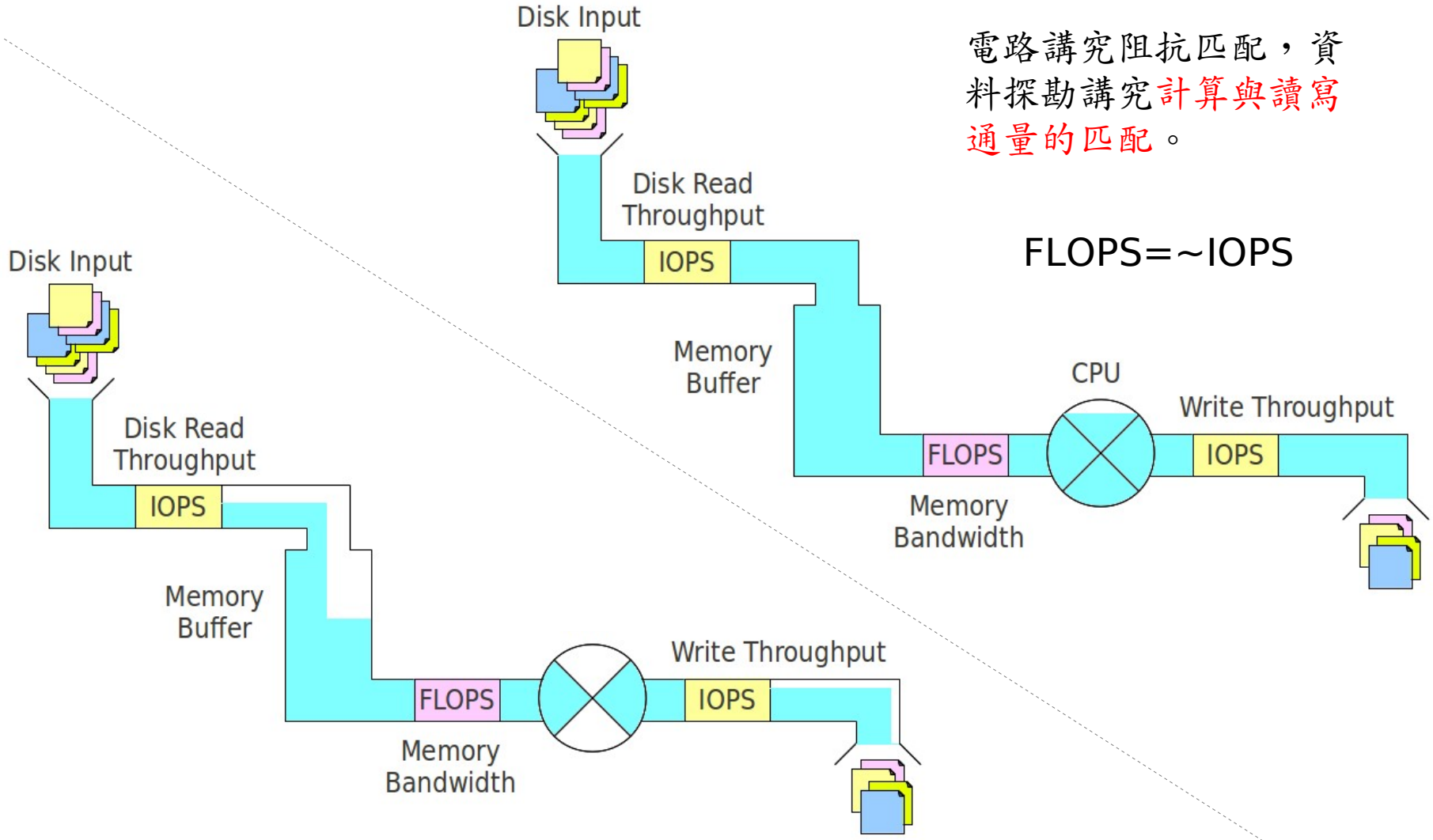
NARLabs



ElephantDB
Or
Voldemort

Source: Lambda Architecture, 8. March 2013
http://www.ymc.ch/en/lambda-architecture-part-1

結論二：購買伺服器的黃金比例
1 core : 2+ GB RAM : 1 SSD Disk



電路講究阻抗匹配，資料探勘講究計算與讀寫通量的匹配。

FLOPS=~IOPS

NARLabs

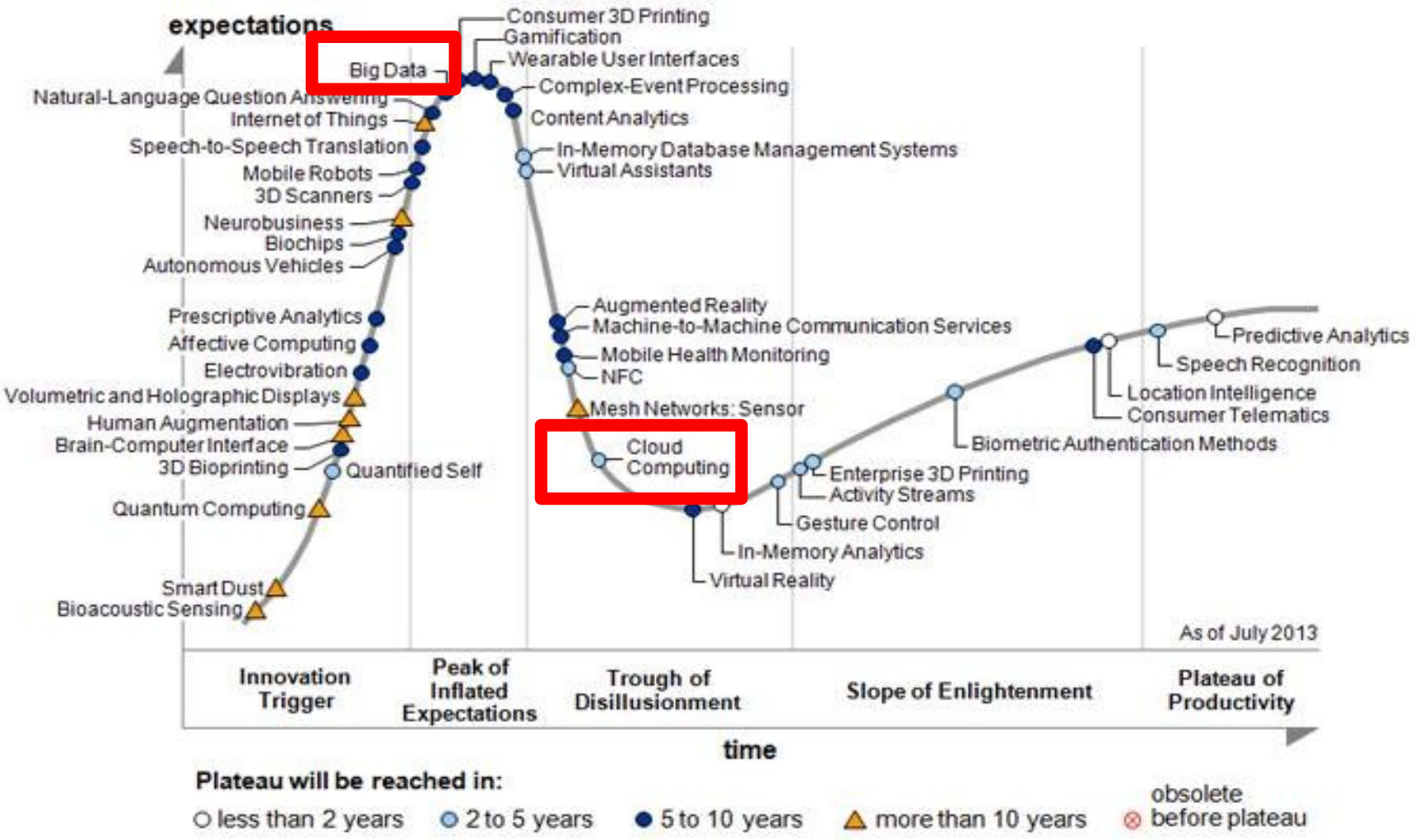
演講大綱 Agenda

- Linux is everywhere 開放無所不在
- What is Big Data ? 何謂巨量資料
- Big Data in Motion ! 即時巨資應用
- The Next Big Thing ? 下半場的重點
- Conclusion 三大結論回顧

NARLabs

市場現況： Gartner Hype Cycle 2013

萌芽期 夢幻期 幻滅期 平原期 高原期



NEXT: Big Data Security

品質管控

- Qualification and Assurance

權限管控

- Access Enablement and Control

數量管控

- Quantification

當我們緊密相連 .....

世界政經：歐盟想分 **Tweeter**
找出經濟、政治的脈動

國家安全：美國 **PRISM** 計劃
(網軍**!** 終極警探 **4.0)**

組織如何因應 **APT ?**
Big Data 平台本身的安全性**?**

有太多安全的問題等待解決！

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

結論三：購買伺服器時不妨留意資料儲存與資料傳輸能否善用硬體加密

NARLabs

Power Hardware	PowerVM	PowerSC	AIX Operating System
	PowerVM™	PowerSC™	AIX®
Systems	Virtualization	Security & Compliance	Systems Software
Power Hardware with security built in & hardware assists for encryption	Secure Virtualization Platform ensuring isolation integrity	Virtualization Centric Security Extensions to protect the Cloud and Virtual Data Center	Secure Operating Systems– defense in depth: role based access, trusted execution, encrypted file system

NARLabs

演講大綱 Agenda

- Linux is everywhere 開放無所不在
- What is Big Data ? 何謂巨量資料
- Big Data in Motion ! 即時巨資應用
- The Next Big Thing ? 下半場的重點
- Conclusion 三大結論回顧

結論：

1. 選擇 Linux 有效降低總擁有成本 (34%) ！

2. 買硬體時，可考慮採用 SSD ，以加速靜態巨量資料的處理速度。為了動態巨量資料的 In-Memory Processing 需求，別忘了多買記憶體 (至少預留多一點插槽) ！

3. 為了保障您寶貴的資料，請選用支援加密的硬體與作業系統！

NARLabs

問題與討論

Questions?