淺談巨量資料專案的導入規劃與相關經驗分享

Jazz Yao-Tsung Wang

BigThing Camp @ 2016/02/27

About Me

- 王耀聰 Jazz Yao-Tsung Wang

- Hadoop.TW 共同創辦人
- HadoopCon 社群年會總召
- Hadoop The Definitive Guide 譯者
- Hadoop Operations 譯者
- 自由軟體愛好者 / 推廣者 / 開發者

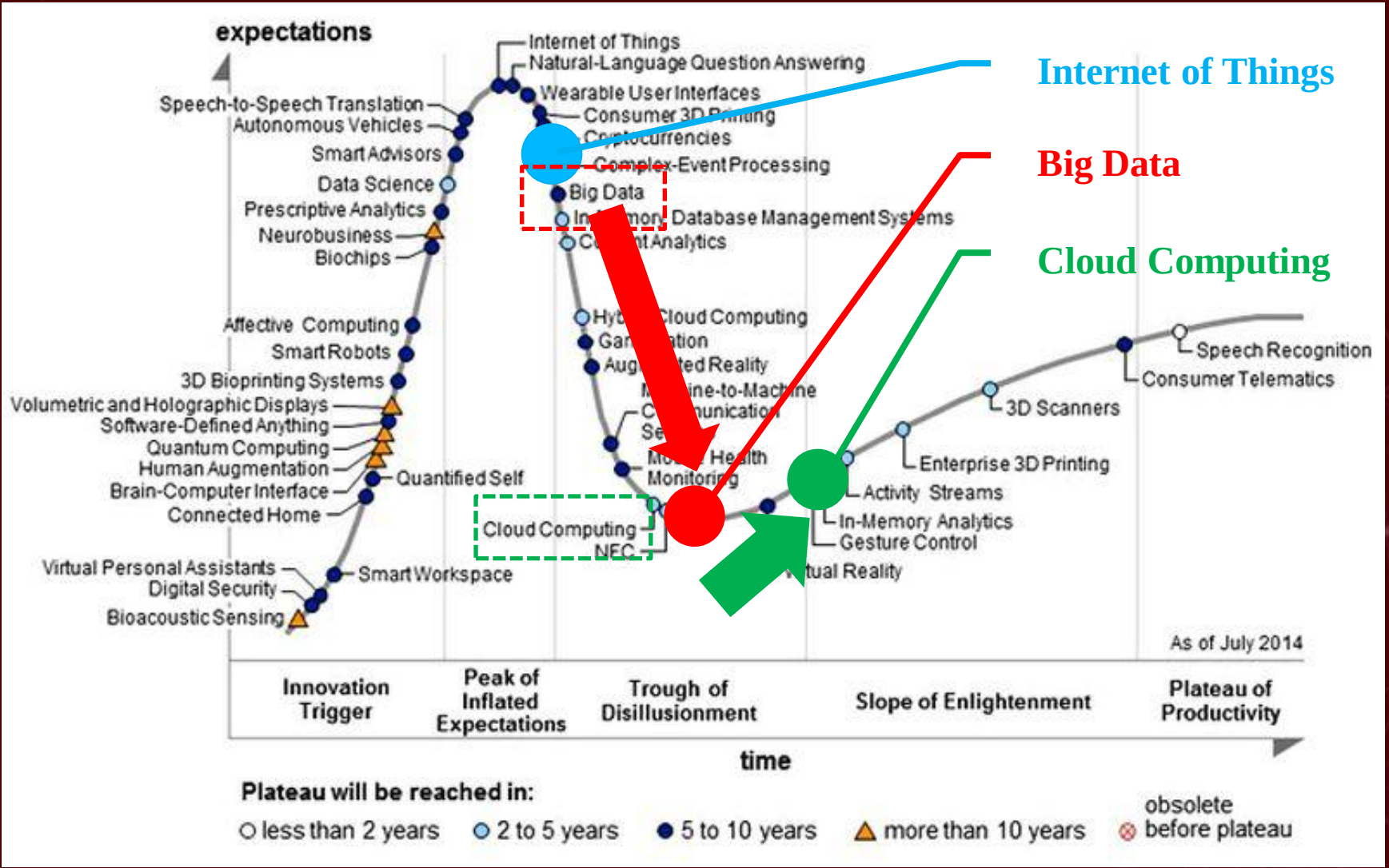
- http://about.me/jazzwang - slideshare, github, etc.

Agenda

- 論 Big Data 退燒 與 規劃六思考帽
- 企業導入 Big Data 的 執行心法
- Big Data Stack 的 過去、現在與未來
- 壓寶新興技術的 線圖指標

預言：Gartner Hype Cycle 2015

萌芽期　夢幻期　幻滅期　平原期　高原期



政府採購平台 首納雲端產品

2015年07月10日 04:10 記者呂雪彗 / 台北報導

A A A

點閱 708 2/10 我要評比

工業局推動軟體（雲端）創新採購概況

年度	預算金額	品項	立約商	實際採購
103年度軟體產品	9億7,017萬元	188項	215家	3億3,604萬元
104年4月軟體產品	34億5,000萬元(全年)	1,614項	540家	2億667萬元
104年7月雲端產品	8億3,150萬元	67項	11家	--

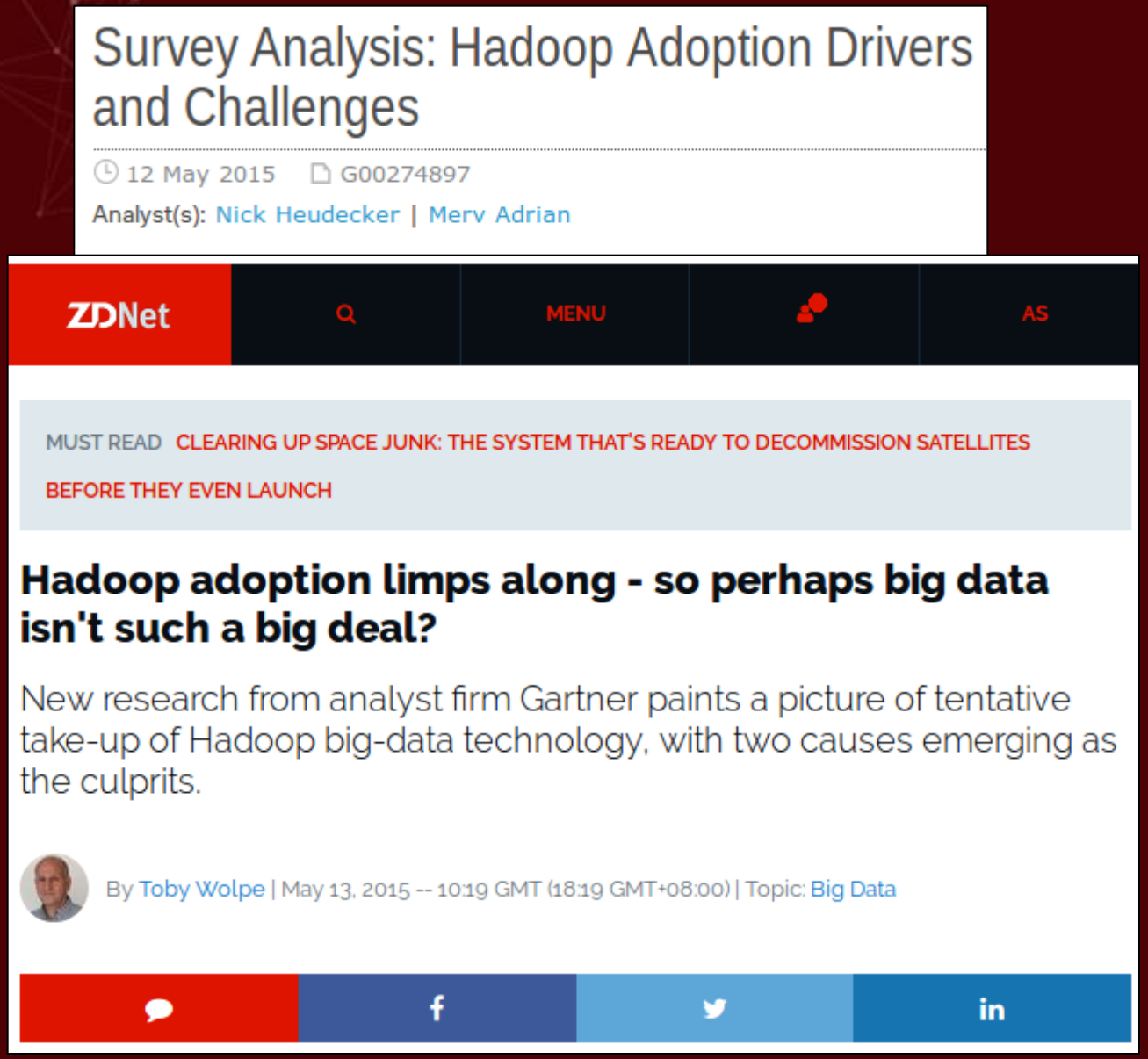
工業局推動軟體（雲端）創新採購概況

http://www.chinatimes.com/newspapers/20150710000158-260205

在地趨勢觀察：

中信標開始有「雲端產品」，使經費核銷與會計科目能解套，象徵著「雲端」正式進入各位每天的工作環境，**代表「雲端」回來了~很快將進入平原期。**

https://www.gartner.com/doc/3051617/survey-analysis-hadoop-adoption-drivers



Survey Analysis: Hadoop Adoption Drivers and Challenges

12 May 2015 G00274897

Analyst(s): Nick Heudecker | Merv Adrian

ZDNet MENU AS

MUST READ CLEARING UP SPACE JUNK: THE SYSTEM THAT'S READY TO DECOMMISSION SATELLITES BEFORE THEY EVEN LAUNCH

Hadoop adoption limps along - so perhaps big data isn't such a big deal?

New research from analyst firm Gartner paints a picture of tentative take-up of Hadoop big-data technology, with two causes emerging as the culprits.

By Toby Wolpe | May 13, 2015 -- 10:19 GMT (18:19 GMT+08:00) | Topic: Big Data

http://www.zdnet.com/article/hadoop-adoption-limps-along-so-perhaps-big-data-isnt-such-a-big-deal/

全球趨勢觀察：

根據 Gartner 2015 年五月份的調查報告，只有 26% 受訪者表示未來 12~24 個月內將投資導入 Hadoop ，**代表「巨量資料」將進入幻滅期谷底。**

[關鍵原因]

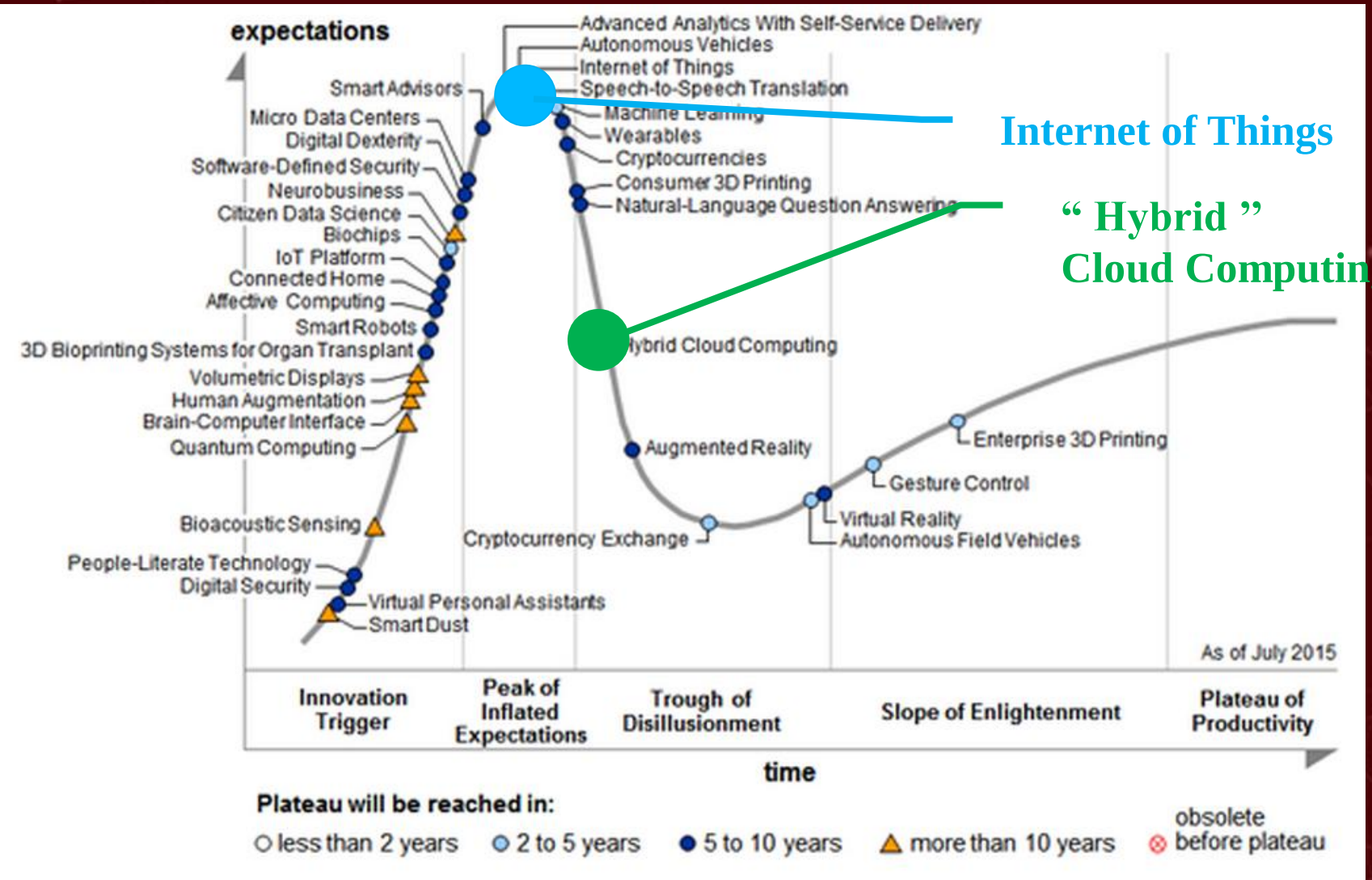
企業內部缺乏對應的技術人才

[絃外之音]

1~2 年後進入平原期

開獎：Gartner Hype Cycle 2015

萌芽期　夢幻期　幻滅期　平原期　高原期



Big Data ~~退燒~~畢業了!!

- 隱身進入以下領域：
  - Internet of Things (物聯網)
  - Business Intelligence and Analytics (商業智慧)
  - Enterprise Architecture
  - Web-Scale IT
  - Digital Banking Transformation (數位金融)
  - Utility Industry IT
  - CRM Customer Service and Customer Engagement
  - CRM Marketing Applications
  - Digital Commerce (電子商務)

回歸基本面：想要?需要?

國際金價	**提供給客戶的價值**	產品通路	1
開採成本	總擁有成本	軟硬體投資	6
提煉廠	分析平台與工具軟體	SMAQ	5
含金度	資料鑑價？	**商業模式**	4
開採權	分析資料的合法性	個資法	3
金礦	資料集	Open Data	2

Big Data 專案的規劃六思考帽

- 問題一：組織想要解決什麼商業問題 ?? (Value)
  可以用資料解決嗎 ?? (降低成本 or 增加收益)
- 問題二：這些資料哪些是內部資料 ?? 哪些是外部資料??
  該如何獲得 ?? 有哪些型態 ?? (Variety)
- 問題三：分析這些資料是否合乎法規需求 ??
  有無需要事先聲明的保護條款 ?? (Legality)
- 問題四：驗證答案真的在這堆資料裡 ?? 資料是否可靠 ??
  需要多少資料才能找到答案 ?? (Volume , Veracity)
- 問題五：挑選合理的資料處理/分析平台 – 人、流程、技術
  定義多快找到答案才能解決商業問題 (Velocity)
- 問題六：定義效益評量指標 (怎麼算 ROI ?? 或 KPI 是什麼 ??)
  持續改善的時程藍圖 (Validation , Roadmap)

Agenda

- 論 Big Data 退燒 與 規劃六思考帽
- 企業導入 Big Data 的 執行心法
- Big Data Stack 的 過去、現在與未來
- 壓寶新興技術的 線圖指標

導入藍圖 Roadmap 4

企業內部的人力資源盤點 People 1

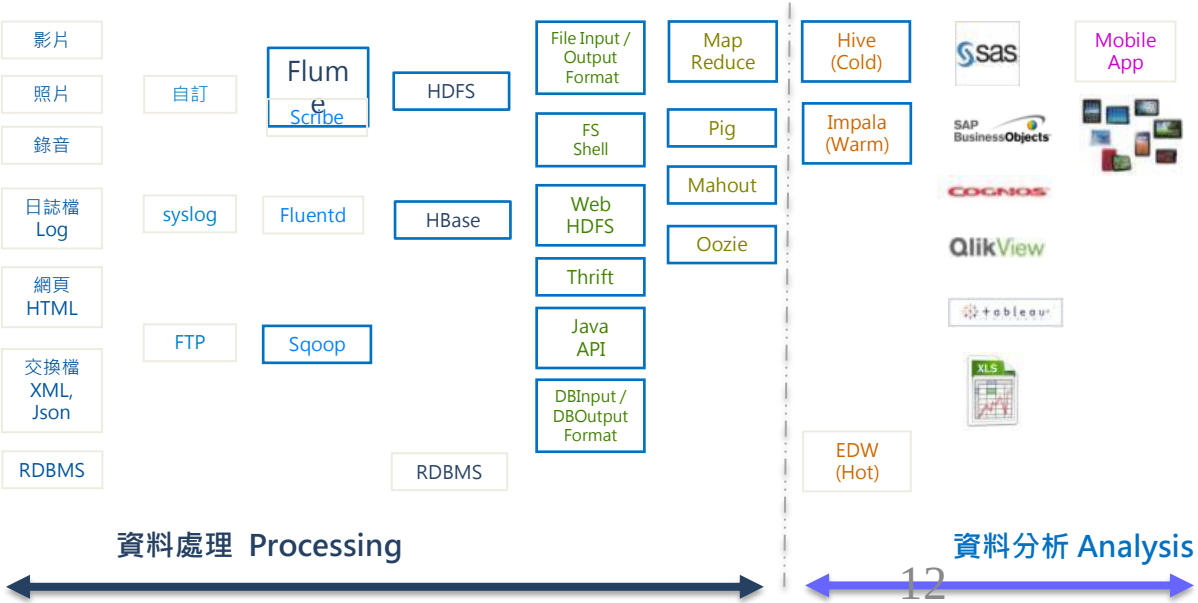
Engineer (電機) | Network (網通) | System Admin | Programmer (資工) | DBA (資管) | Analyst (統計) | Decision Maker

處理巨量資料的常見流程 Process 2

生 流 蒐 存 取 算 析 用 看 變

資料源 網路協定 前處理 儲存方式 存取方式 資料處理 資料分析 視覺化 解讀 行動

處理巨量資料的技術盤點 Technology 3



進入專案執行的第一步：Task Force

重新定義 R&R
重新規範權責分工
溝通・溝通・溝通

- 切記！您需要的是一個團隊，不是一個英雄！
- 導入 Big Data 也有 DevOps 的文化衝擊問題！
  - Hadoop 由 Ops 角度出發，Spark 由 Devs 角度出發
- 請確認每個成員知道這個專案的意義與價值
- 角色衝突：接受它、處理它、放下它
- 協作文化：善用協作平台，維持資訊通透

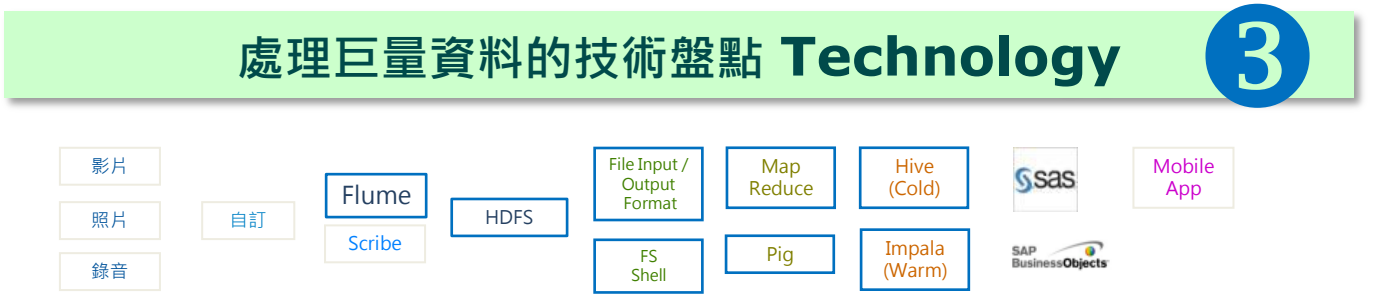
專案執行的第二步：設計資料流

這裡的十字箴言
只是一個非常概略
的參考範本

- 將規劃六思考帽的結果拿來這裡用
- 驗證每個項目的負責人與利害關係人
- 由架構師協助確認每個流程之間的技術可行性
- 提醒：
  - 網路架構常是一個盲點/限制 (外網、DMZ、內網)
  - 資料是否落地會影響整個處理/分析流程的速度

專案執行的第三步：挑選合適堆疊

自建・採購・外包
各有優劣，考驗著
各位 CXO 的智慧



- 由「人+流程-風險」推定「技術」
- 軟體堆疊(Software Stack)的組成無絕對好壞
  往往跟組織成員的學習歷程與商業目標有關
- 趨勢：
  - 異質資料整合是過去幾年較常聽到的新興需求
  - 資料倉儲(DW)是過去幾年較常聽到的切入點

導入藍圖 Roadmap 4

專案的潛在風險：資訊安全與資料品質

- Hadoop 剛滿 10 歲 (2006年生的童工)
  別跟 38 歲的老員工 RDBMS 比較
  - 請記得為何您需要童工，不用老員工
  - 童工與老員工各有地位，互補非取代
- 別好高騖遠，期待一次到位
  - 先求可行 – 可以解決商業問題
  - 二求安全 – 可以滿足 ISO 規範 (AAA)
  - 再求更好 – 資料的六個標準差 (6S)
- 趨勢：
  - 2015年，少數業界先驅的位置在「安全」
  - 開始有少數新創在討論 Data Governance

Agenda

- 論 Big Data 退燒 與 規劃六思考帽
- 企業導入 Big Data 的 執行心法
- Big Data Stack 的 過去、現在與未來
- 壓寶新興技術的 線圖指標

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/

How big data can curb the world's energy consumption

http://www.openpdc.com/

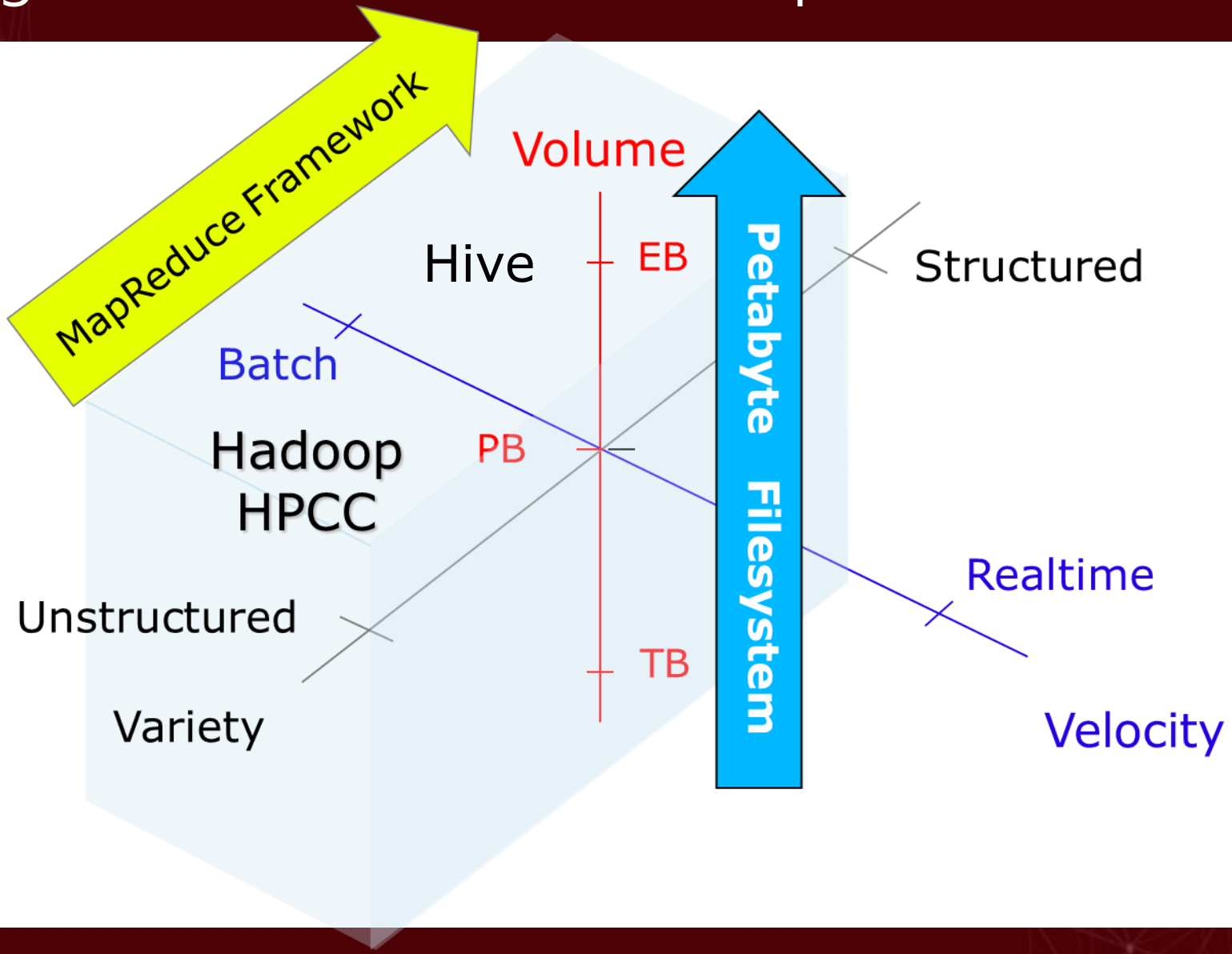
facebook

NETFLIX

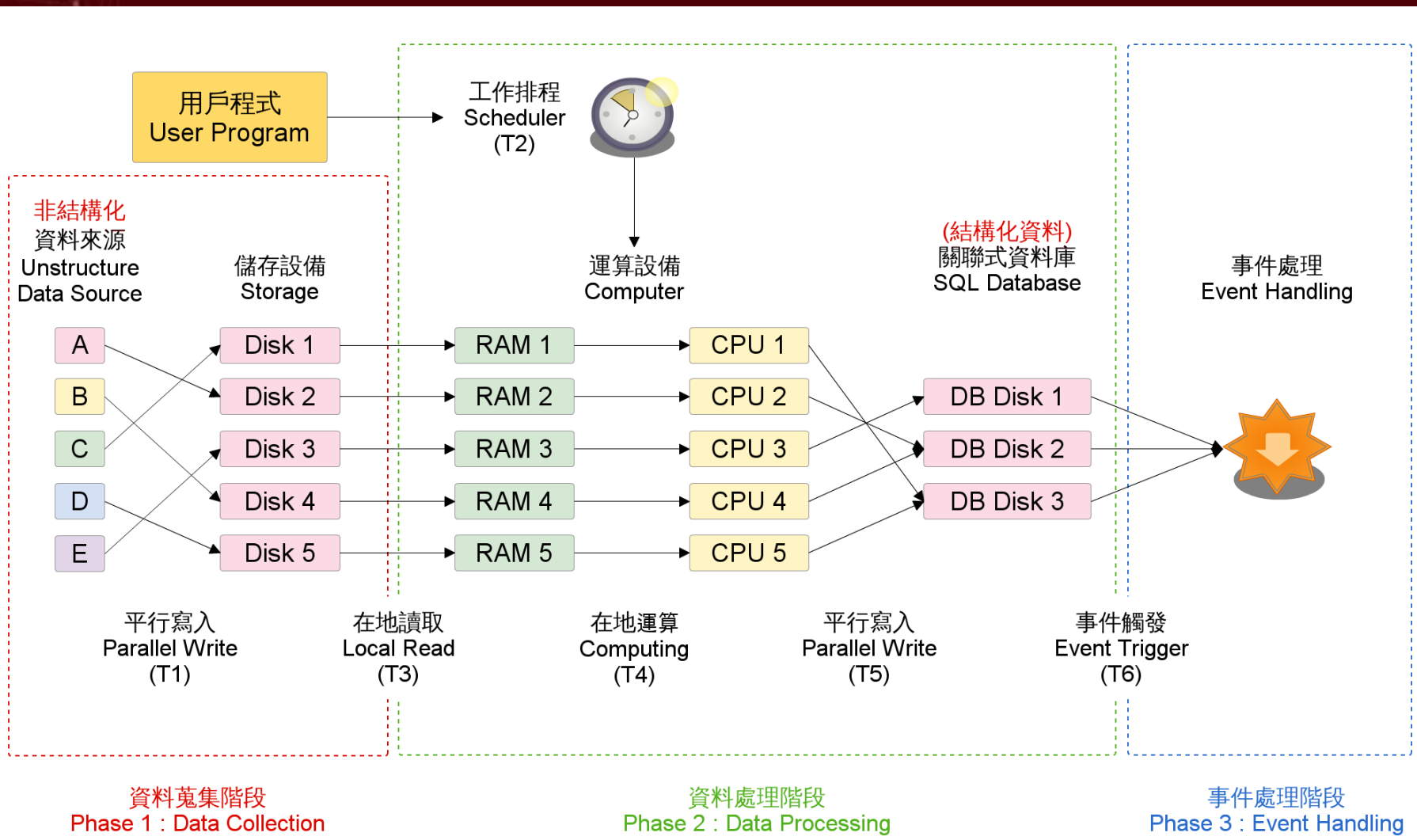
One hospital's embrace of big data

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

Big Data at Rest 代表技術：MapReduce



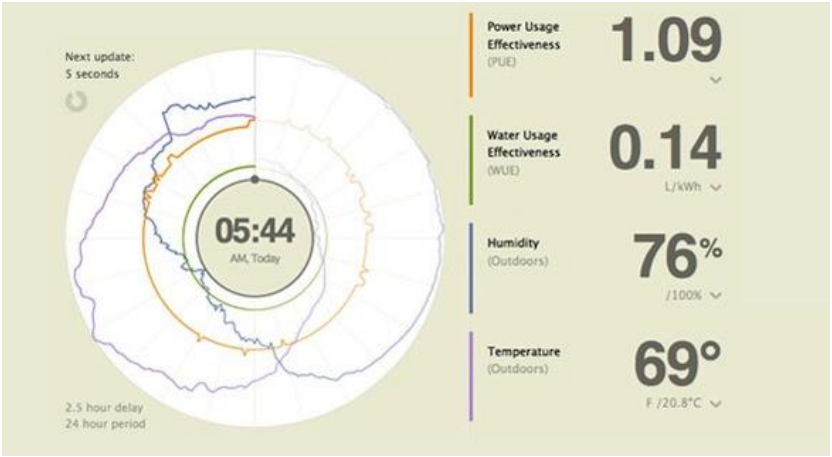
Processing Time of Batch Jobs



NOW: Big Data in Motion

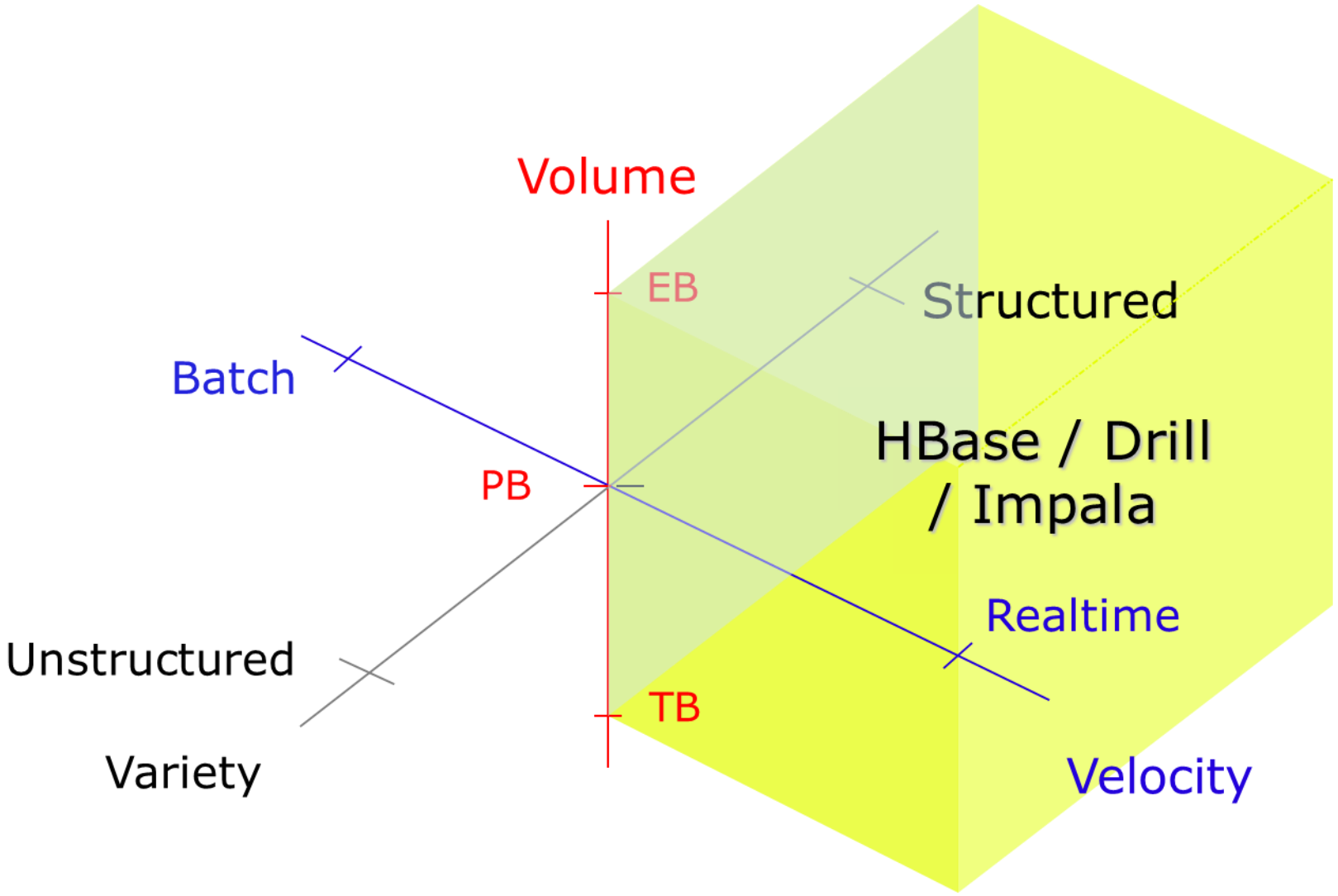
[金融] Trading Robot

[災防] 海嘯、土石流
Disaster Prevention
Tsunami Forecast

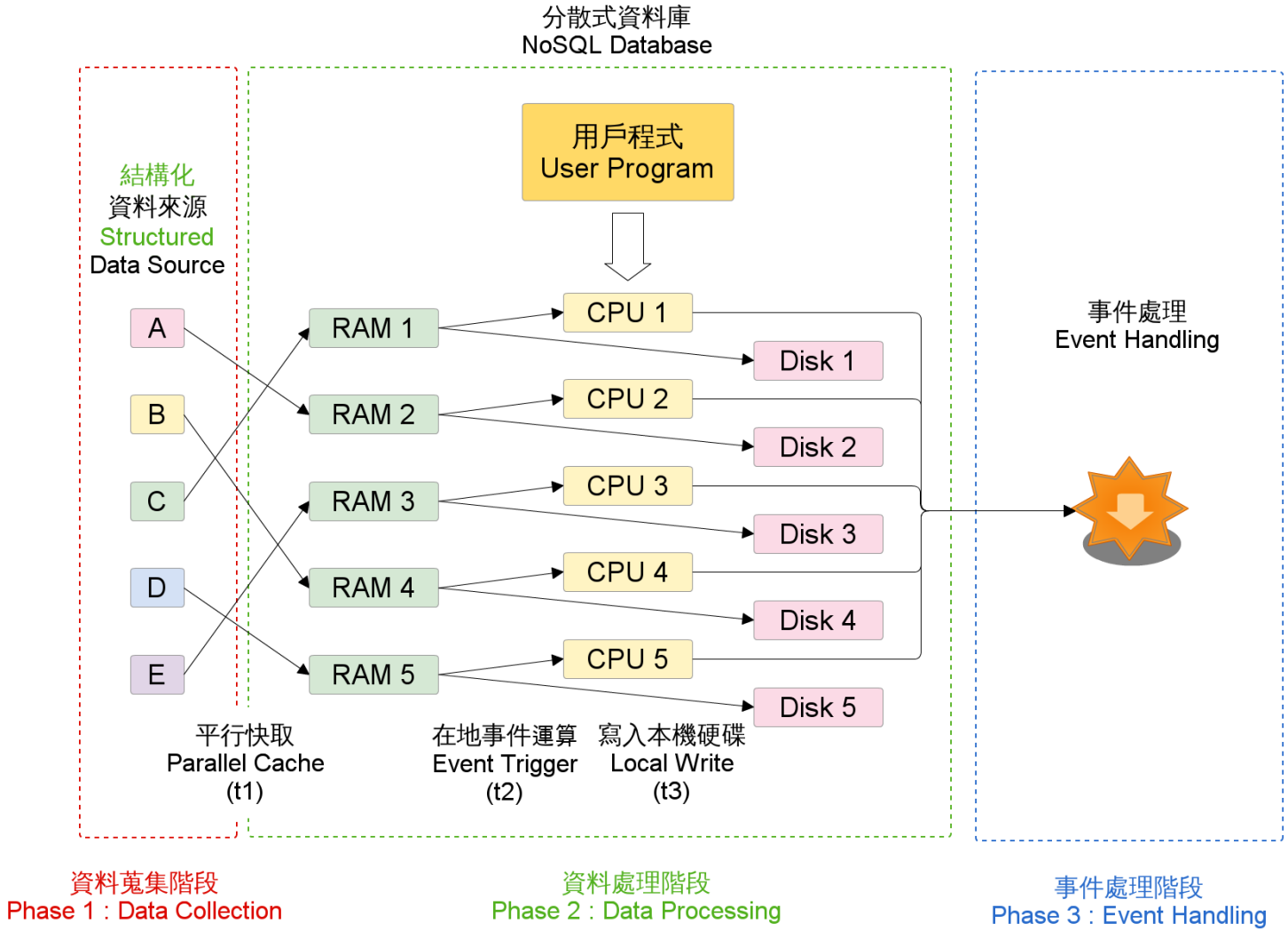


[資訊] 機房即時用電資訊監控、警訊
Realtime Data Center Power Usage and related notifications
http://www.newmobilelife.com/2013/04/21/facebook-pue-real-time-charts/

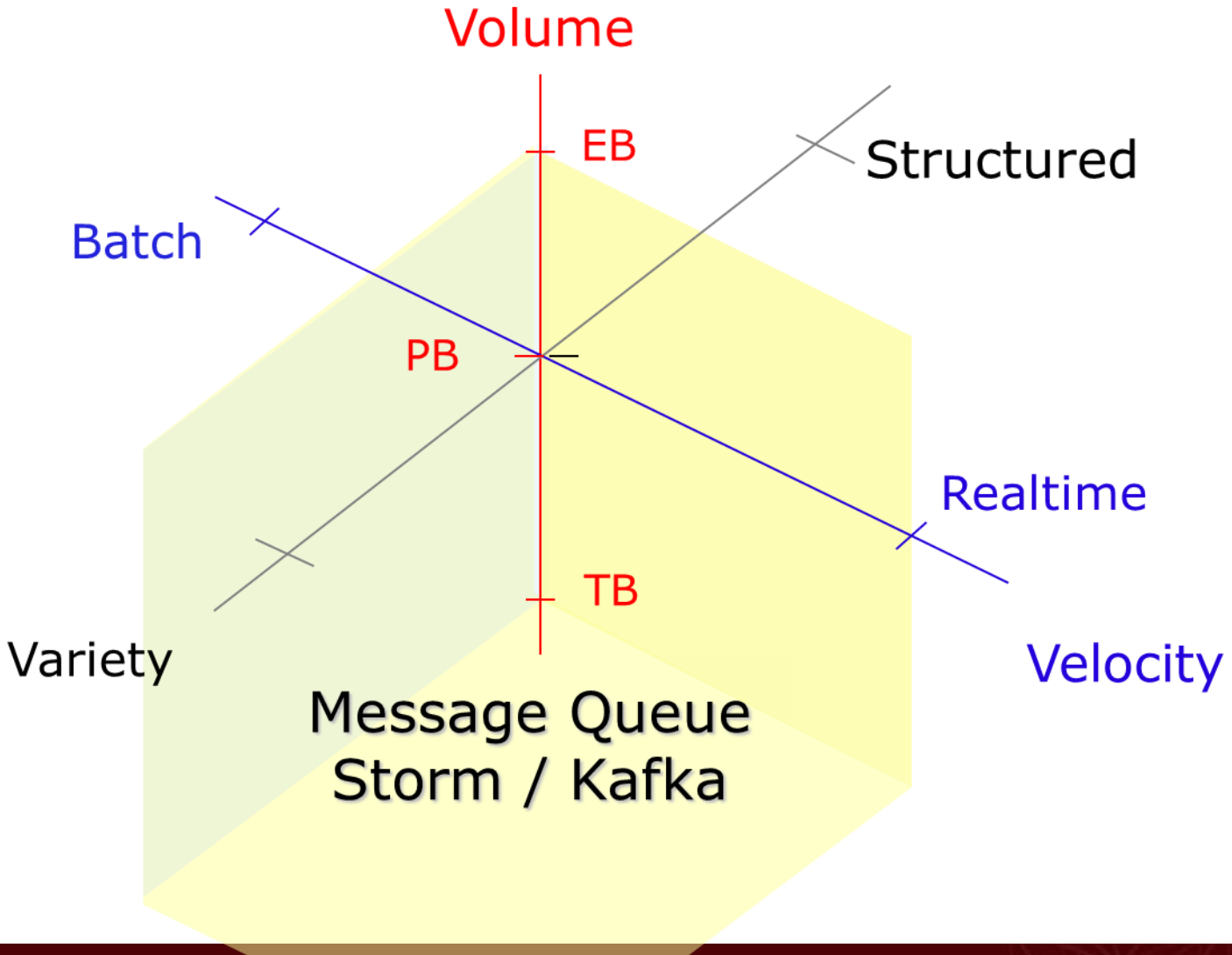
Big Data in Motion 代表技術 (1)
In-Memory Processing



In-Memory Processing的運算時間
以HBase為例

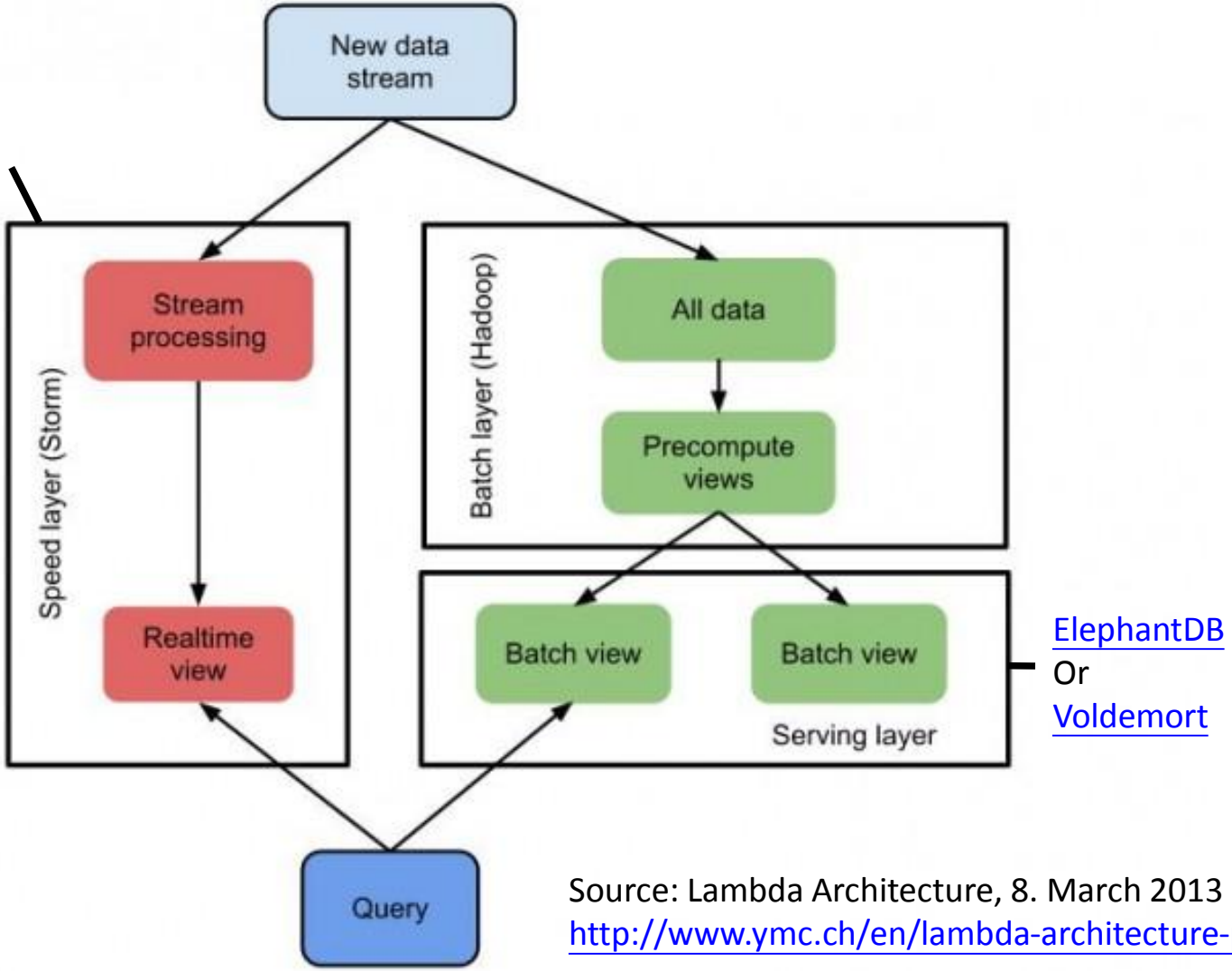


Big Data in Motion 代表技術 (2)
Streaming Data Collection



混合模式的巨量資料處理架構
Lambda Architecture

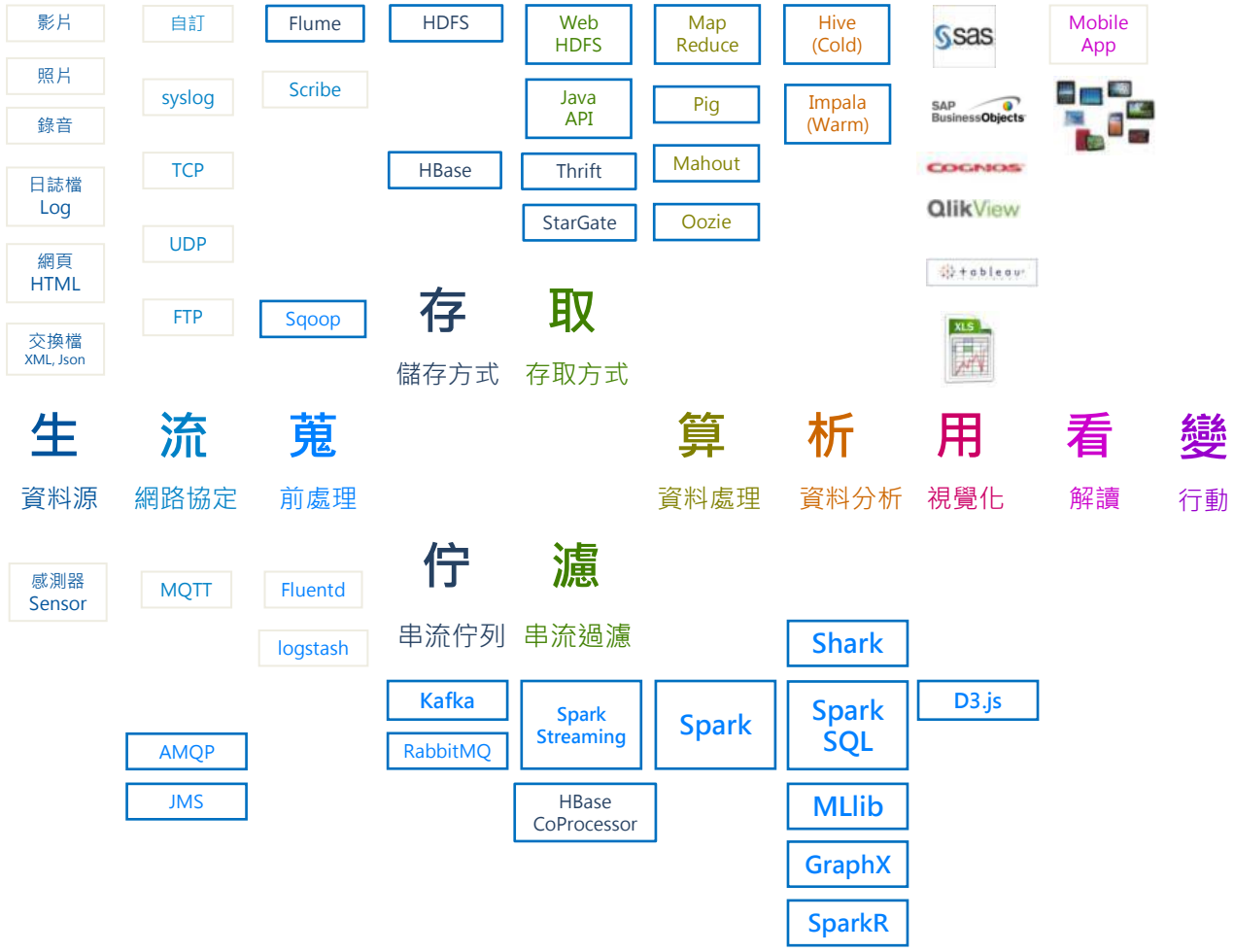
HBase
Storm



ElephantDB
Or
Voldemort

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

符合 Lambda Architecture 的十二字箴言



靜止的巨量資料
Big Data at Rest

更新頻率較低
可接受批次處理

分鐘到小時

流動的巨量資料
Big Data in Motion

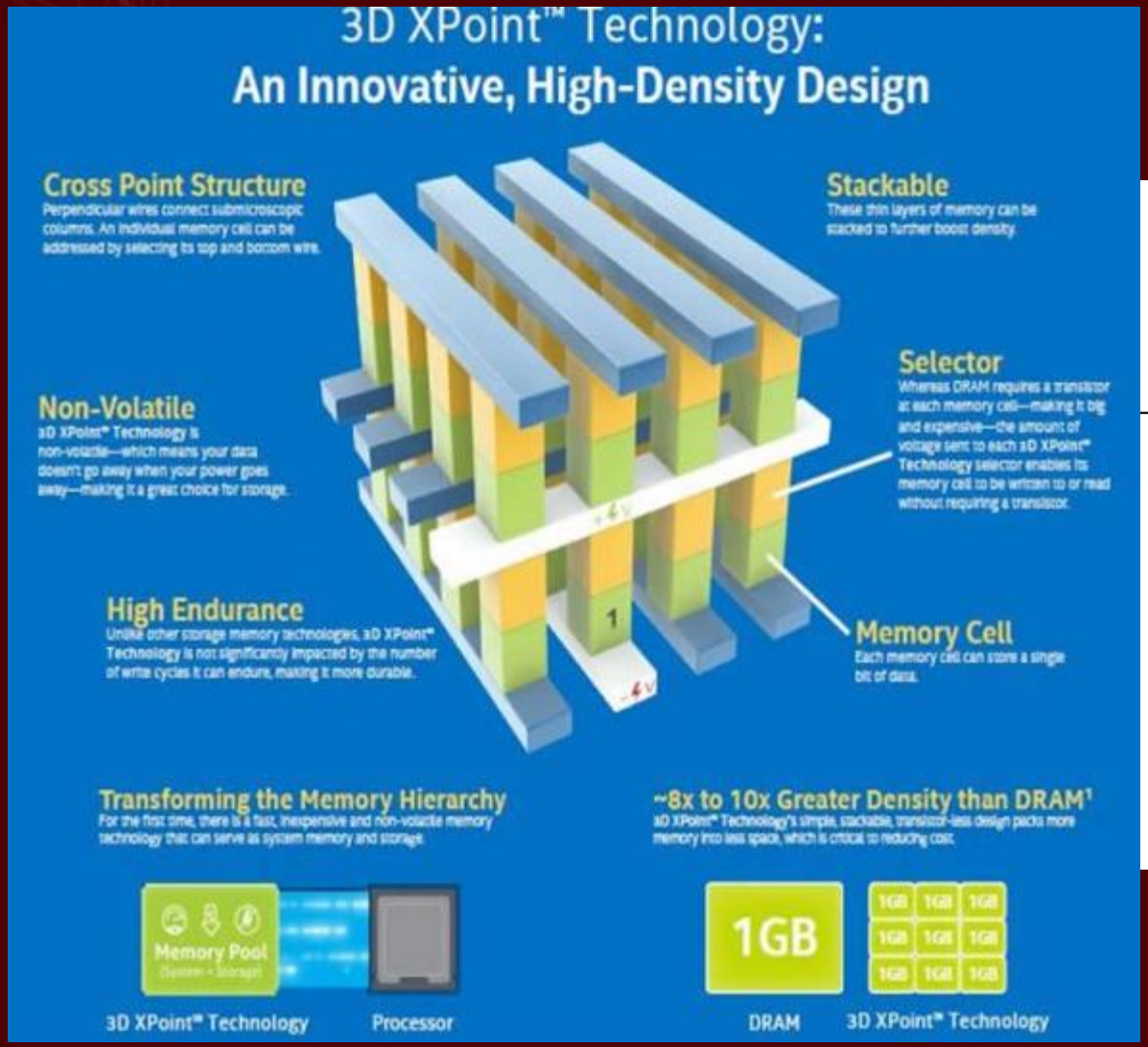
更新頻率很高
Micro Batch
"微批次"

毫秒到秒

Big Data Stack 的未來 (1)：必要之痛 Security

- Hadoop Security 的四大範疇：
  - **Authentication** – 帳號密碼認證
  - **Authorization** – 基於帳號身分，管理讀寫權限 → Sentry
  - **Auditing** – 稽核讀寫的紀錄
  - **Encryption** – 資料的加密、通訊的加密 (運算過程的加密?)
- 那 Spark 呢?? 還在經歷生長痛中....
  - Authentication – 1.3 剛支援 Kerberos
    https://issues.apache.org/jira/browse/SPARK-5493
  - Authorization – 目標做到 Spark SQL column-level 管控
  - Auditing – 是否有工具??還在找
  - Encryption – 進行中
    https://issues.apache.org/jira/browse/SPARK-5682

Big Data Stack 的未來 (2)：NVM 記憶體崛起



http://pmem.io/

pmem.io
Persistent Memory Programming

Home | Glossary | Documents | NVM Library | Blog | About

This site is focused on making *persistent memory programming* easier. The current focus is on the Linux NVM Library, which is a library (set of libraries, actually) designed to provide some useful APIs for server applications wanting to use persistent memory. You can read more about the NVM Library or go directly to the source. Contributions are welcome!

Note: The NVM Library is still under development and is not yet ready for production use.

The Linux NVM Library builds on the **Direct Access** (DAX) changes under development in Linux. Check out the PRD repo for the latest snapshot of this work.

What Is It?

For many years computer applications organize their data between two tiers: memory and storage. We believe the emerging *persistent*

Recent Blog

An introduction
replication
Posted November 23,
Replication is a
raising the relia
your pmemobj
applications. Y
basically think

https://youtu.be/IWsjbqbkqh8

Agenda

- 論 Big Data 退燒 與 規劃六思考帽
- 企業導入 Big Data 的 執行心法
- Big Data Stack 的 過去、現在與未來
- 壓寶新興技術的 線圖指標

面對如此複雜的 Apache Big Data Stack

APACHE PROJECT LIST

BY CATEGORY

Overview
All Projects
Attic
Big Data
Build Management
Cloud
Content
Databases
FTP
Graphics
HTTP
HTTP-module
Incubating
JavaEE
Labs
Libraries
Mail
Mobile
Network-client
Network-server
OSGi
RegExp
Retired
Testing
Virtual-machine
Web-framework
XML
FAQ

BY NAME

HTTP Server

A
Abdera
Accumulo
ACE
ActiveMQ
Airavata
Allura
Ambari
Ant
Any23
APR
Archiva
Aries
Aurora
Avro
Axis

B
Bigtop
Bloodhound
BookKeeper
Buildr
BVal

C
Camel
Cassandra
Cayenne
Celix
Chemistry
Chukwa
Clerezza
CloudStack
Cocoon
Commons
Continuum
Cordova
CouchDB
Creadur
Crunch
cTAKES
Curator
CXF

D
DB
DeltaSpike
DeviceMap
Directory
Drill

E
Empire-db
Etch

F
Falcon
Felix
Flex
Flink
Flume
Forrest

G
Geronimo
Giraph
Gora
Gump

H
Hadoop
Hama
HBase
Helix
Hive
HttpComponents

I
Isis
Ignite

J
Jackrabbit
James
jclouds
Jena
JMeter
JSPWiki
jUDDI

K
Kafka
Karaf
Knox

L
Lens
Libcloud
Logging
Lucene
Lucene.Net
Lucy

M
Mahout
ManifoldCF
Marmotta
Maven
Mesos
MetaModel
MINA
MRUnit
MyFaces

N
Nutch
Nifi

O
ODE
OFBiz
Olingo

Pivot
POI
Portals

Q
Qpid

R
Rave
River
Roller

S
Samza
Santuario
Serf
ServiceMix
Shindig
Shiro
SIS

Tez
Thrift
Tika
Tiles
Tomcat
TomEE
Traffic Server
Turbine
Tuscany

U
UIMA
Usergrid

V
VCL
Velocity
VXQuery

W

▲ http://apache.org/index.html#projects-list

▼ http://incubator.apache.org/

List of all current Incubator projects

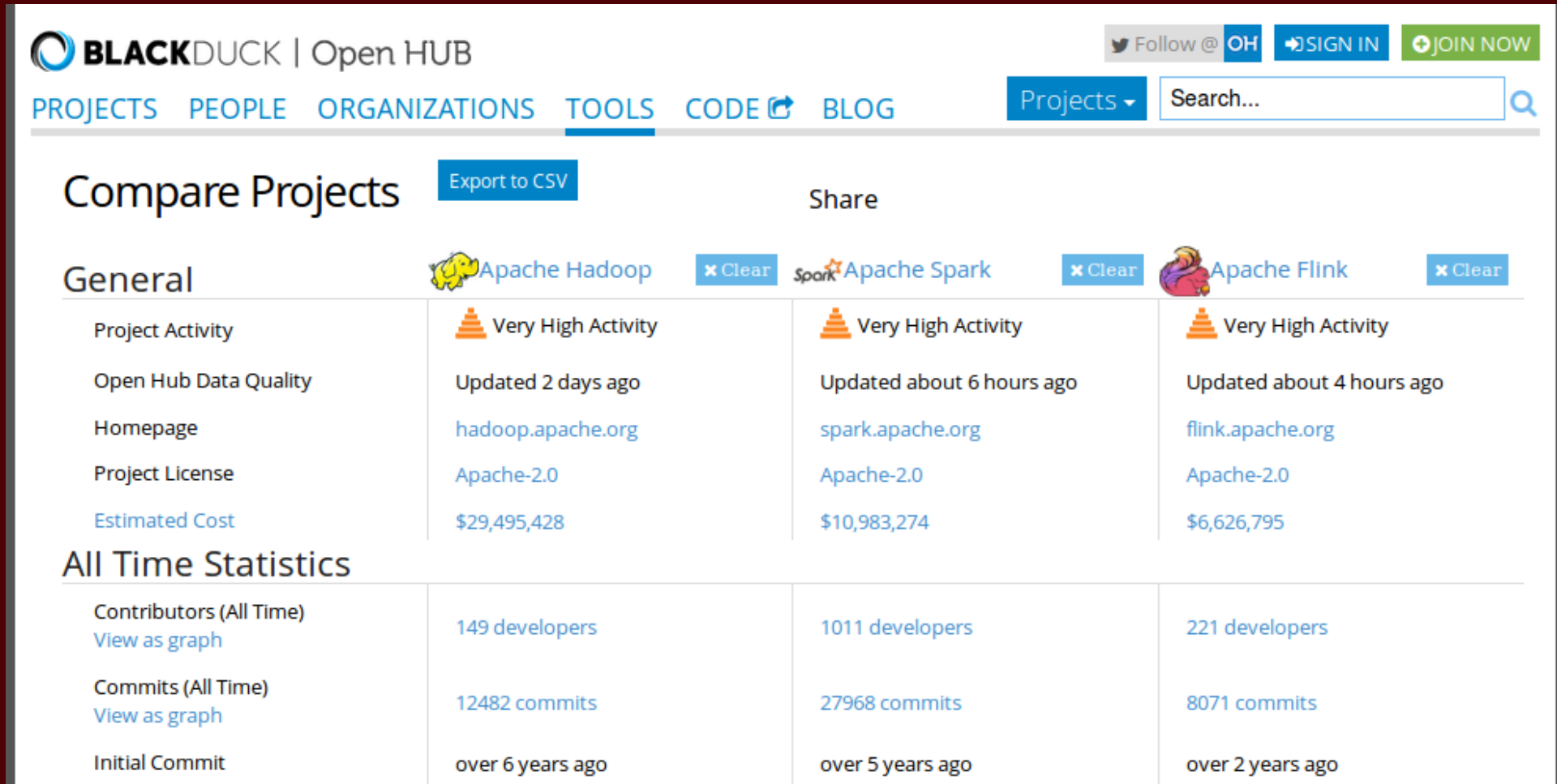
Apex	AsterixDB	Atlas	BatchEE	Blur	Climate Model Diagnostic Analyzer
CommonsRDF	Concerted	Corinthia	Cotton	DataFu	Eagle
FreeMarker	Geode	Groovy	HAWQ	HORN	HTrace
Impala	Johnzon	Kudu	Kylin	log4cxx2	MADlib
Metron	MRQL	Mynewt	Myriad	ODF Toolkit	OpenAz
Ranger	REEF	Ripple	Rya	S2Graph	SAMOA
Sentry	Singa	Sirona	Slider	Streams	SystemML
Tamaya	Taverna	TinkerPop	Trafodion	Twill	Unomi
Wave	Zeppelin				

如何去蕪存菁挑中明日之星??

- 讓數據說話吧!!
- 以下是爵士派密傳自由軟體選股評選心法
  - 基本面 – 供應端(開發者)的公開資訊
  - 籌碼面 – 供應端財務報表的公開資訊
  - 消息面 – 產品曝光度的公開資訊
  - 技術面 – 接收端(使用者)的公開資訊
  - 法人面 – 第三方的產業分析報告

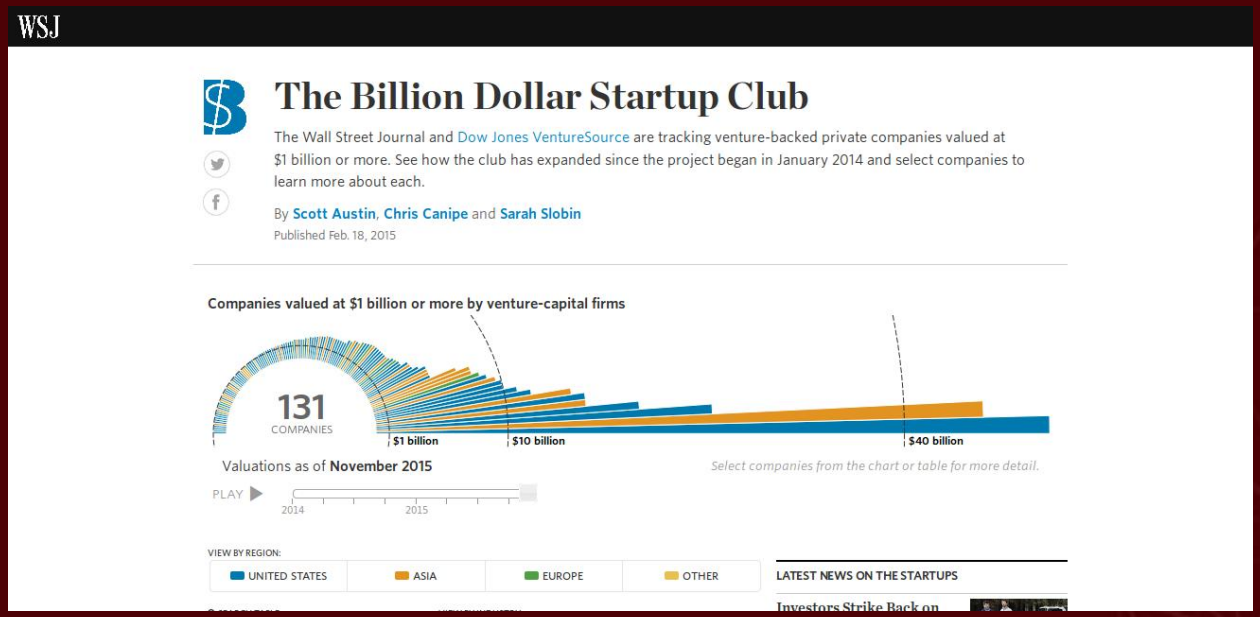
開發者的公開資訊

- 工具一：http://github.com
- 工具二：https://www.openhub.net/p/_compare
- 健康指標：開發者人數、提交數、更新頻率、更新行數



籌碼面 – 誰有比較多銀彈?

- 自由軟體的商業模式是「技術服務」，賣知識與人時
- 想避免服務斷炊的風險，現實問題是公司銀彈夠不夠
  - 該技術背後有公司 – 好的開始
  - 只有一間的話，小心壟斷或斷炊
  - 有兩間以上，更好，但要持續觀察「市佔率」
- 工具：http://graphics.wsj.com/billion-dollar-club/

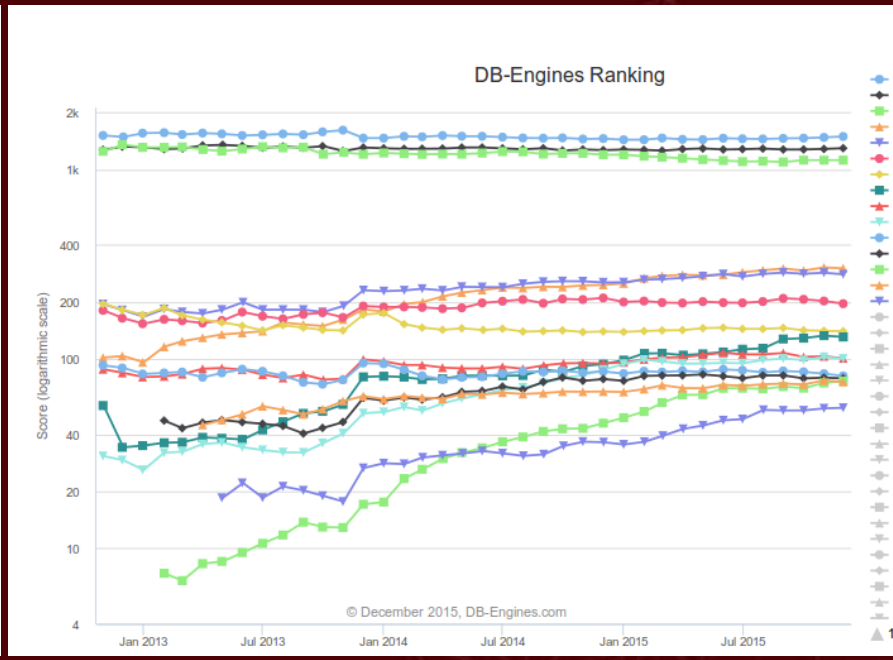


產品曝光度的公開資訊

- 許多軟體都會遇到「叫好不叫座」的狀況 -- 自我行銷量 vs 口碑行銷量
- 工具：
  - https://www.google.com/trends (關鍵字查詢量)
  - https://www.google.com/#q=技術&tbm=nws (新聞發行量)
  - http://alexa.com (網站流量)
  - http://db-engines.com/en/ranking

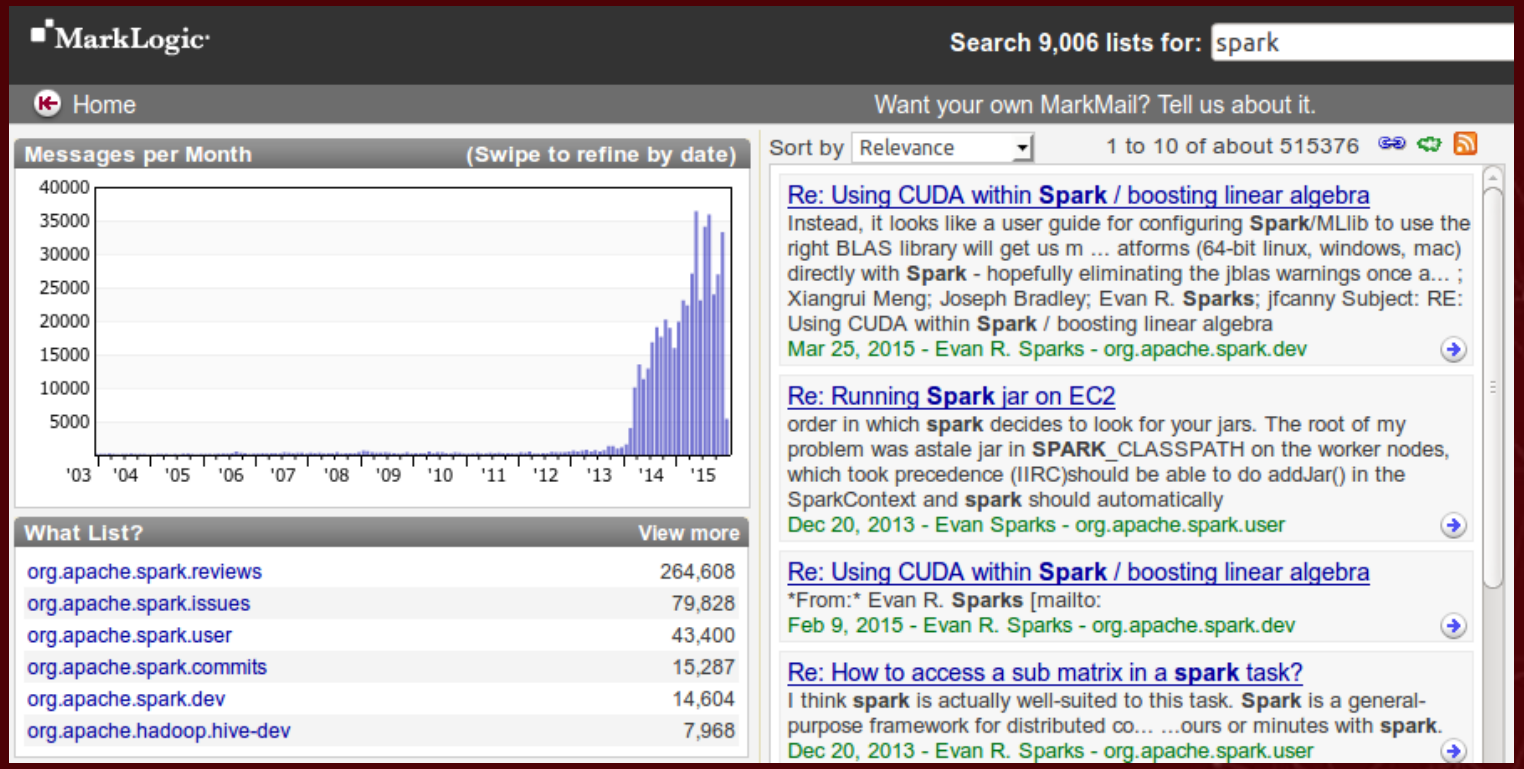
289 systems in ranking, December 2015

Rank Dec 2015	Rank Nov 2015	Rank Dec 2014	DBMS	Database Model	Score Dec 2015	Score Nov 2015	Score Dec 2014
1.	1.	1.	Oracle	Relational DBMS	1497.55	+16.61	+37.76
2.	2.	2.	MySQL	Relational DBMS	1298.54	+11.70	+29.96
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1123.16	+0.83	-76.89
4.	4.	↑ 5.	MongoDB +	Document store	301.39	-3.22	+54.87
5.	5.	↓ 4.	PostgreSQL	Relational DBMS	280.09	-5.60	+26.09
6.	6.	6.	DB2	Relational DBMS	196.13	-6.40	-14.13
7.	7.	7.	Microsoft Access	Relational DBMS	140.21	-0.75	+0.31
8.	8.	↑ 9.	Cassandra +	Wide column store	130.84	-2.08	+36.78
9.	9.	↓ 8.	SQLite	Relational DBMS	100.85	-2.60	+6.15
10.	10.	10.	Redis +	Key-value store	100.54	-1.87	+12.66
11.	11.	11.	SAP Adaptive Server	Relational DBMS	81.47	-2.24	-4.52
12.	12.	12.	Solr	Search engine	79.15	-0.63	+0.73
13.	↑ 14.	↑ 16.	Elasticsearch	Search engine	76.57	+1.79	+30.67
14.	↓ 13.	↓ 13.	Teradata	Relational DBMS	75.72	-1.37	+8.32
15.	↑ 16.	↑ 17.	Hive	Relational DBMS	55.27	+0.36	+18.90
16.	↓ 15.	↓ 15.	HBase	Wide column store	54.25	-2.21	+3.17



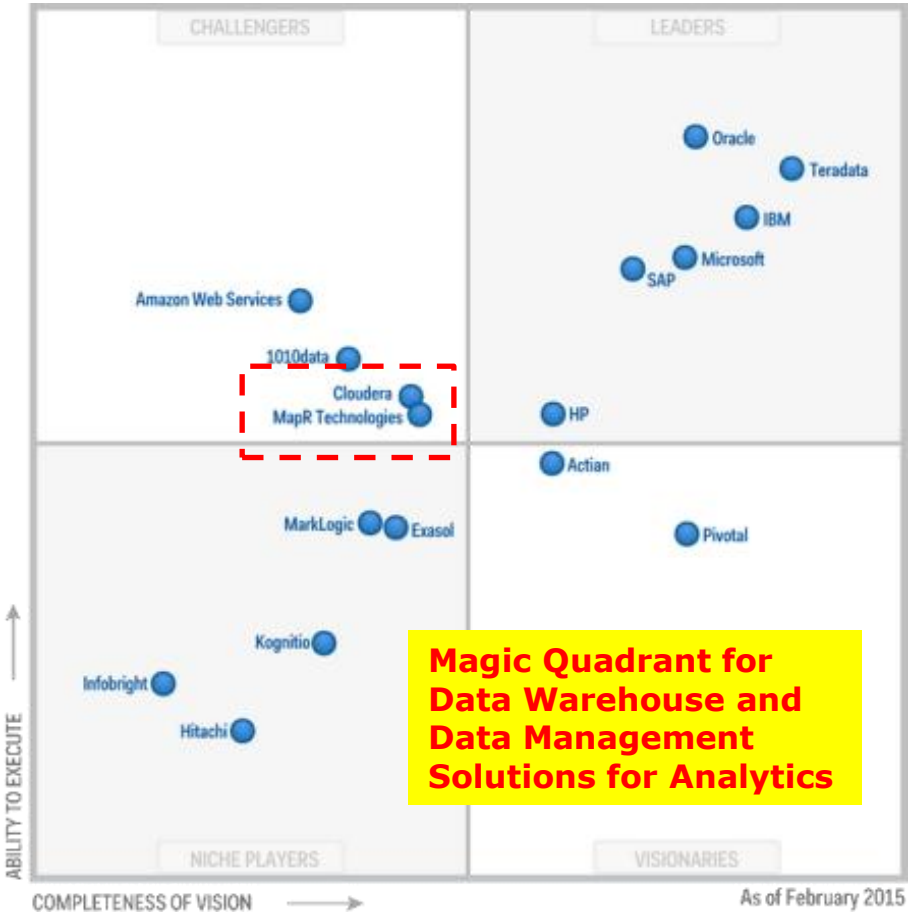
使用者的公開資訊

- 討論區、社群都是反映使用者數量的關鍵指標
- 工具：
  - http://stackoverflow.com/
  - http://markmail.org/

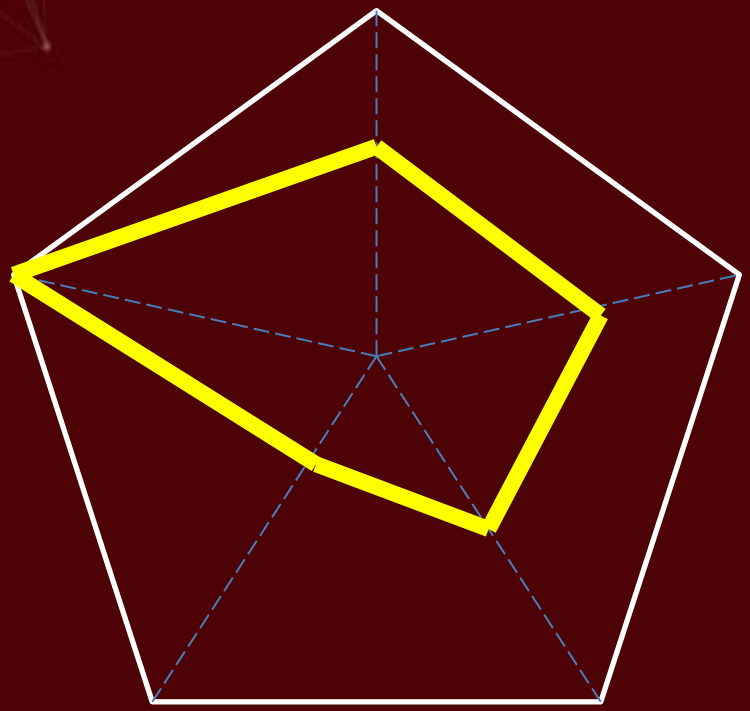


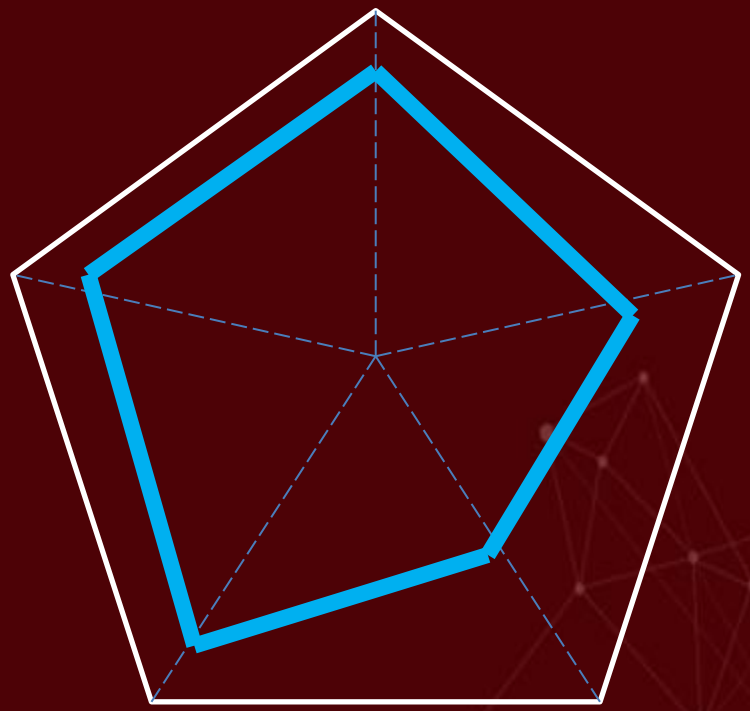
產業分析師的報告

- 工具：Gartner Magic Quadrant



透過不同面向的瞭解，再做結論





最合適貴公司的技術跟「人」、「流程」、「商業問題」有關，
以上分析只供「風險預測」。

結論

- 請使用**規劃六思考帽**，評估企業是否需要導入 Big Data
- 若確定要導入，再參考**執行心法**
  - **人 + 流程 - 風險 → 技術**
- 技術現況：符合 Lambda Architecture 的十二字箴言

- 風險評估：五個線圖指標
- 預見未來：當「記憶體」變身「儲存」，將引爆軟體新革命!!