巨量資料處理技術：過去、現在、未來
Big Data Technologies: PAST, NOW and FUTURE

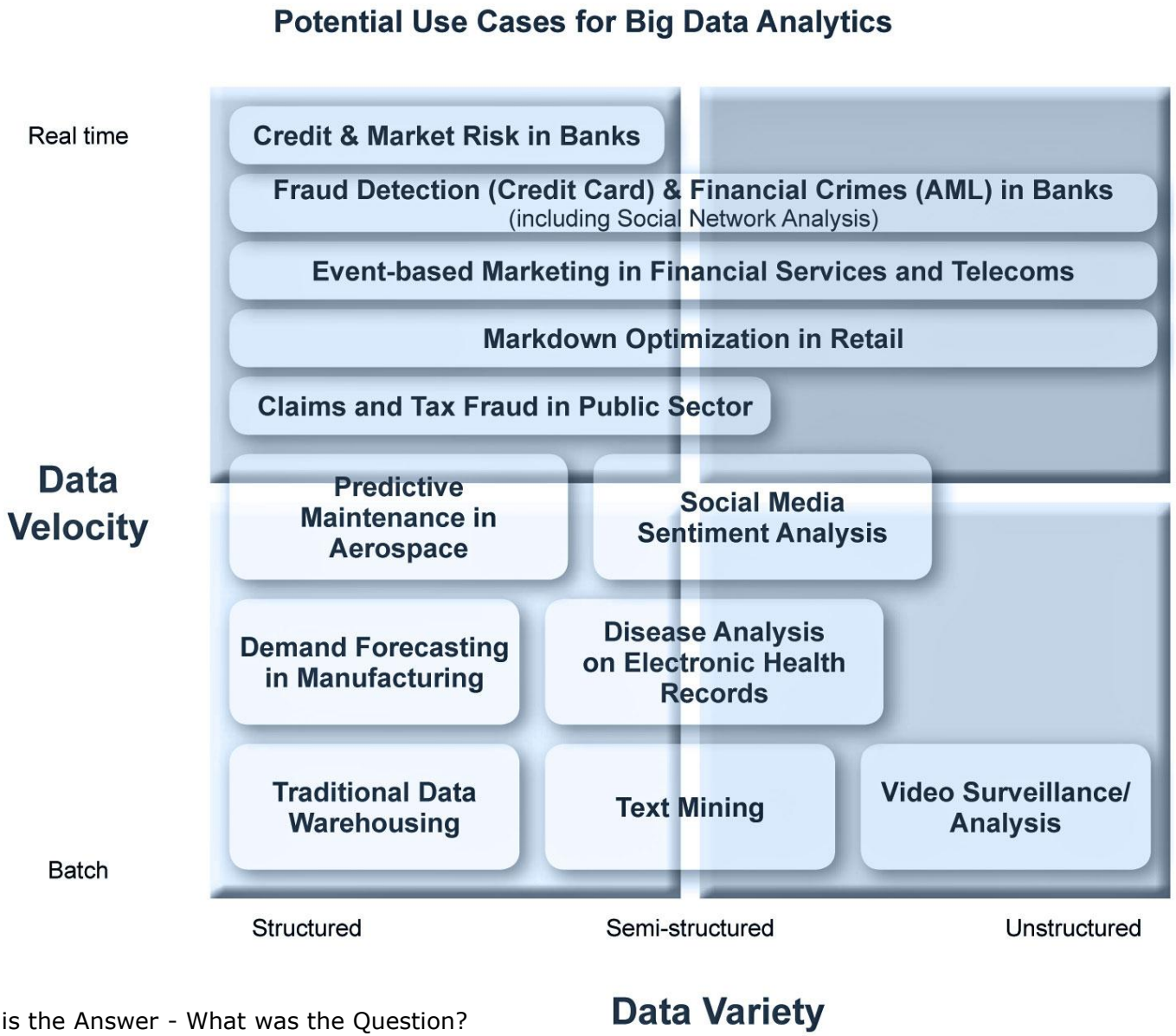
巨量資料處理的資訊架構
The SMAQ stack for big data

LAMP is for Web Services.

未來處理海量資料的人必需知道
You will need a big data stack called 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

不同的應用特性，落在不同的象限



參考來源：Big Data is the Answer - What was the Question?
http://www.saama.com/blog/bid/76211/Big-Data-is-the-Answer-What-was-the-Question

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

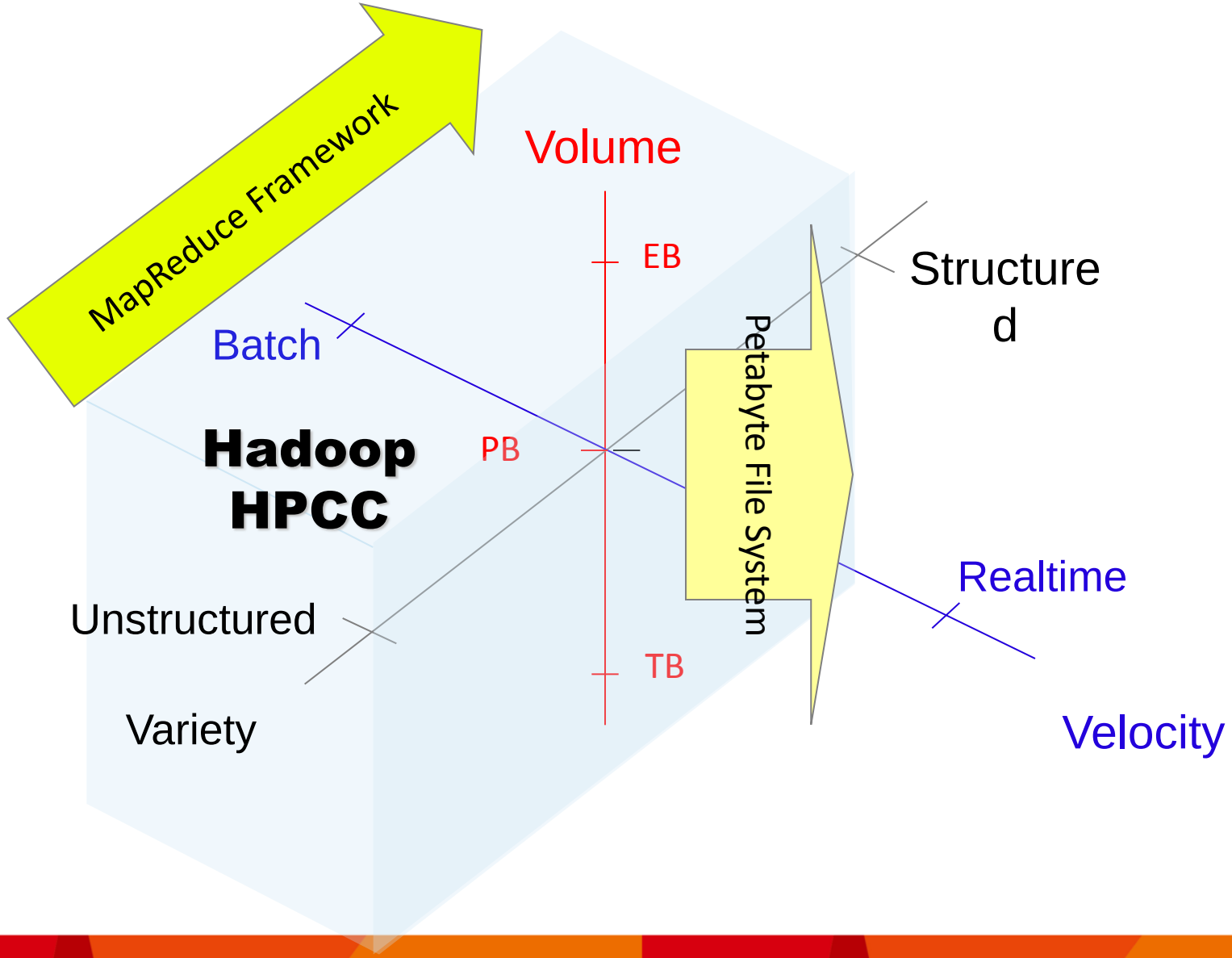
http://www.openpdc.com/

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



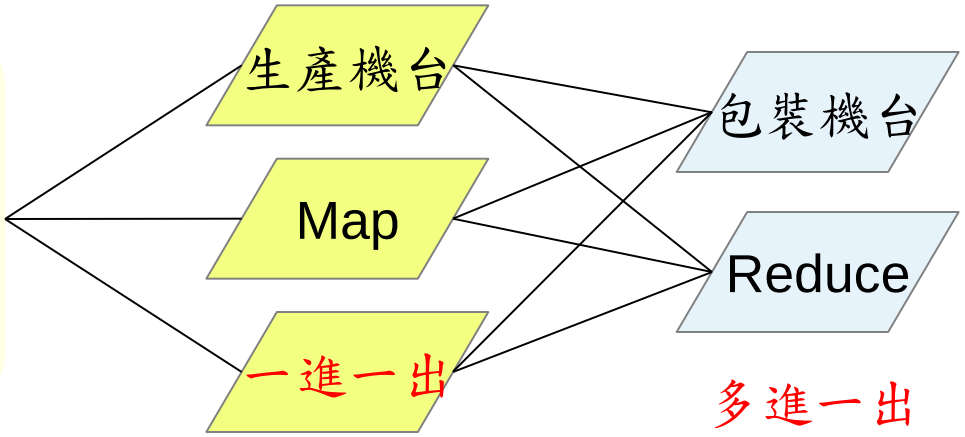
高資料通量處理平台 Hadoop
Key Concept : Data Locality

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。

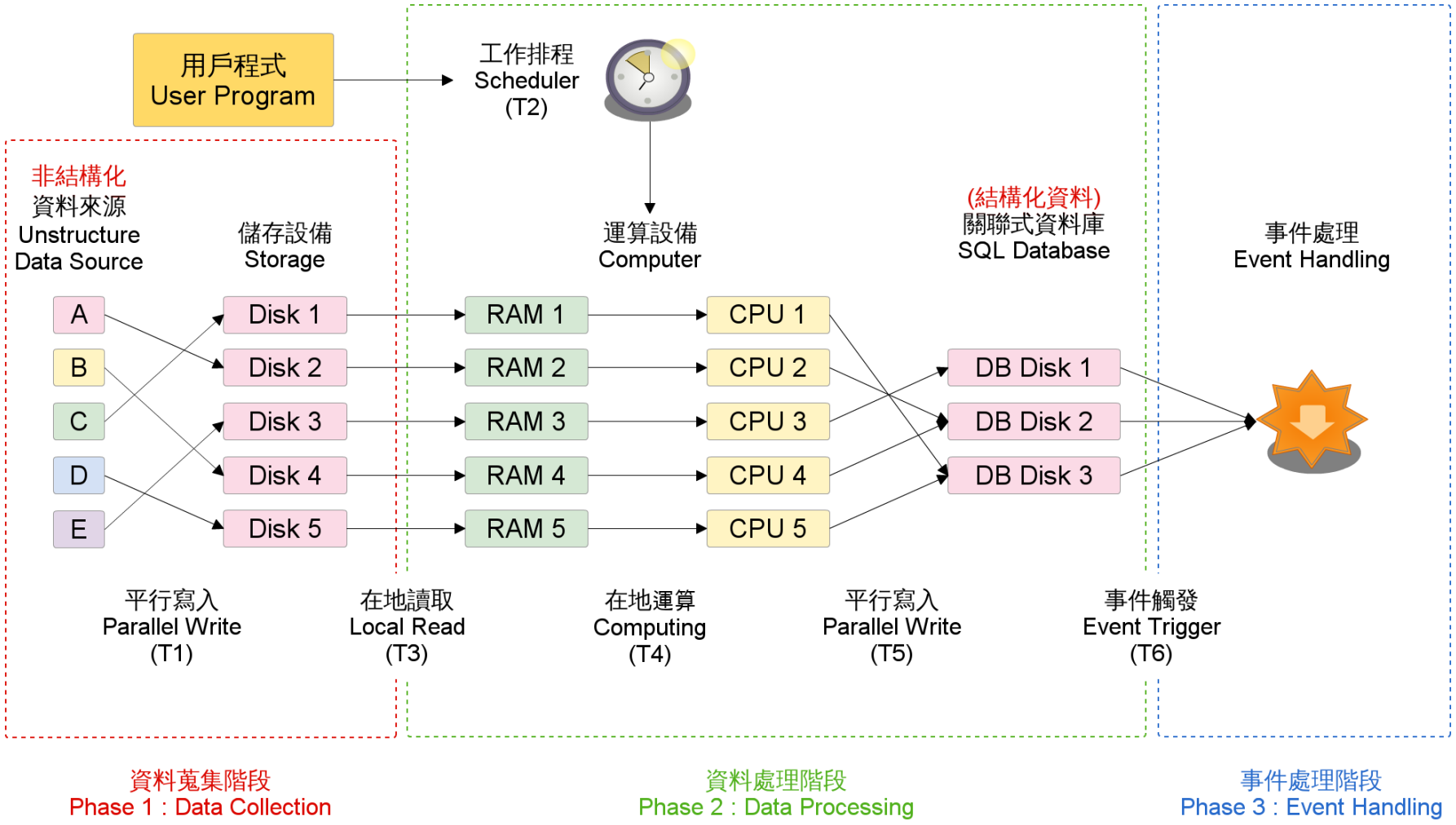
亦可以想像成一個處理海量資料的生產線，只須學會定義 **map** 跟 **reduce** 工作站該做哪些事情。

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



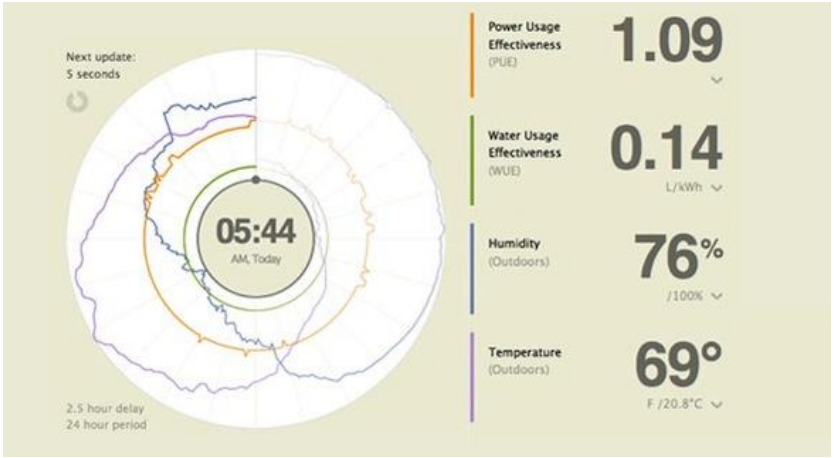
批次作業的運算時間
Processing Time of Batch Jobs



NOW: Big Data in Motion

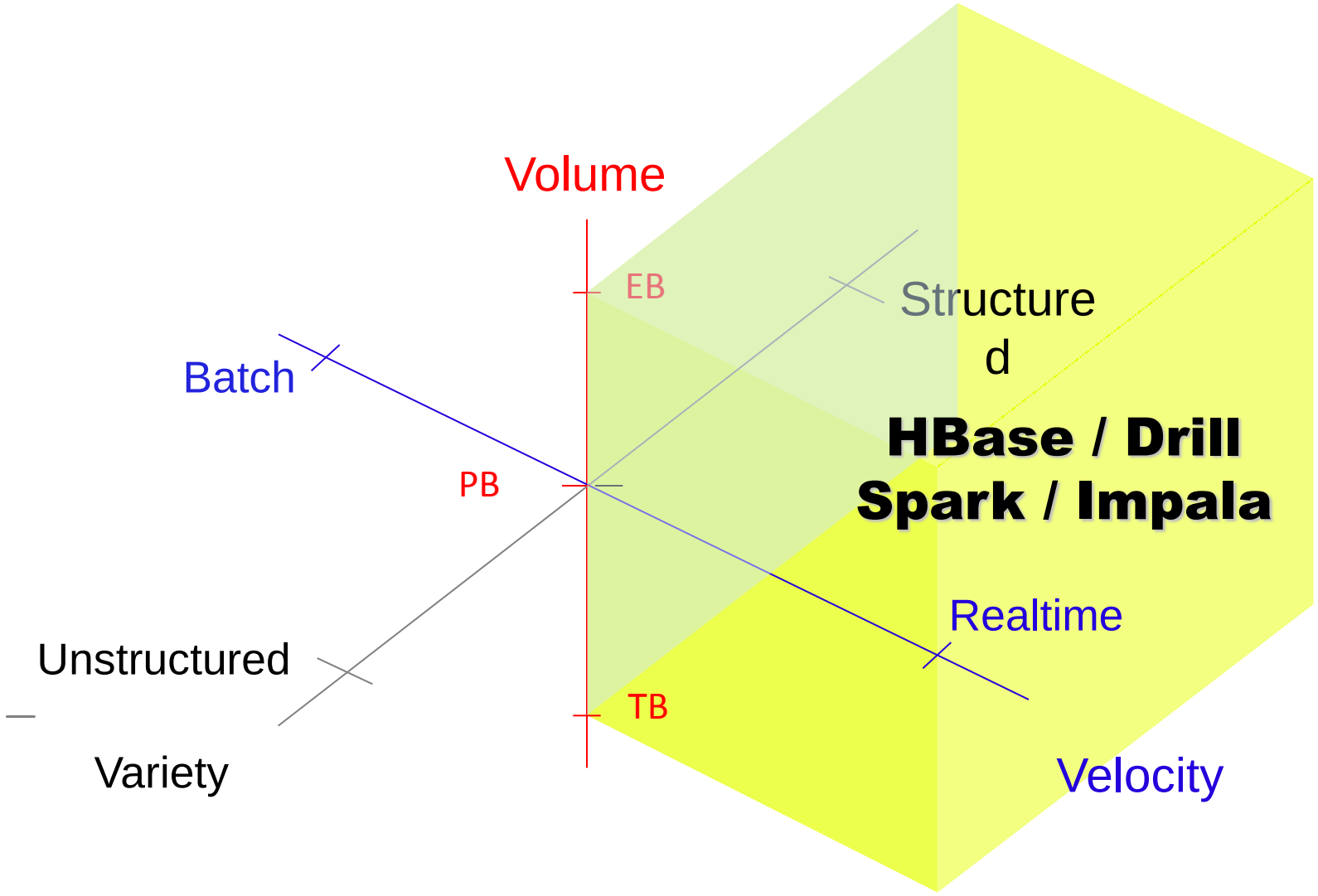
[金融] **Trading Robot**

[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/

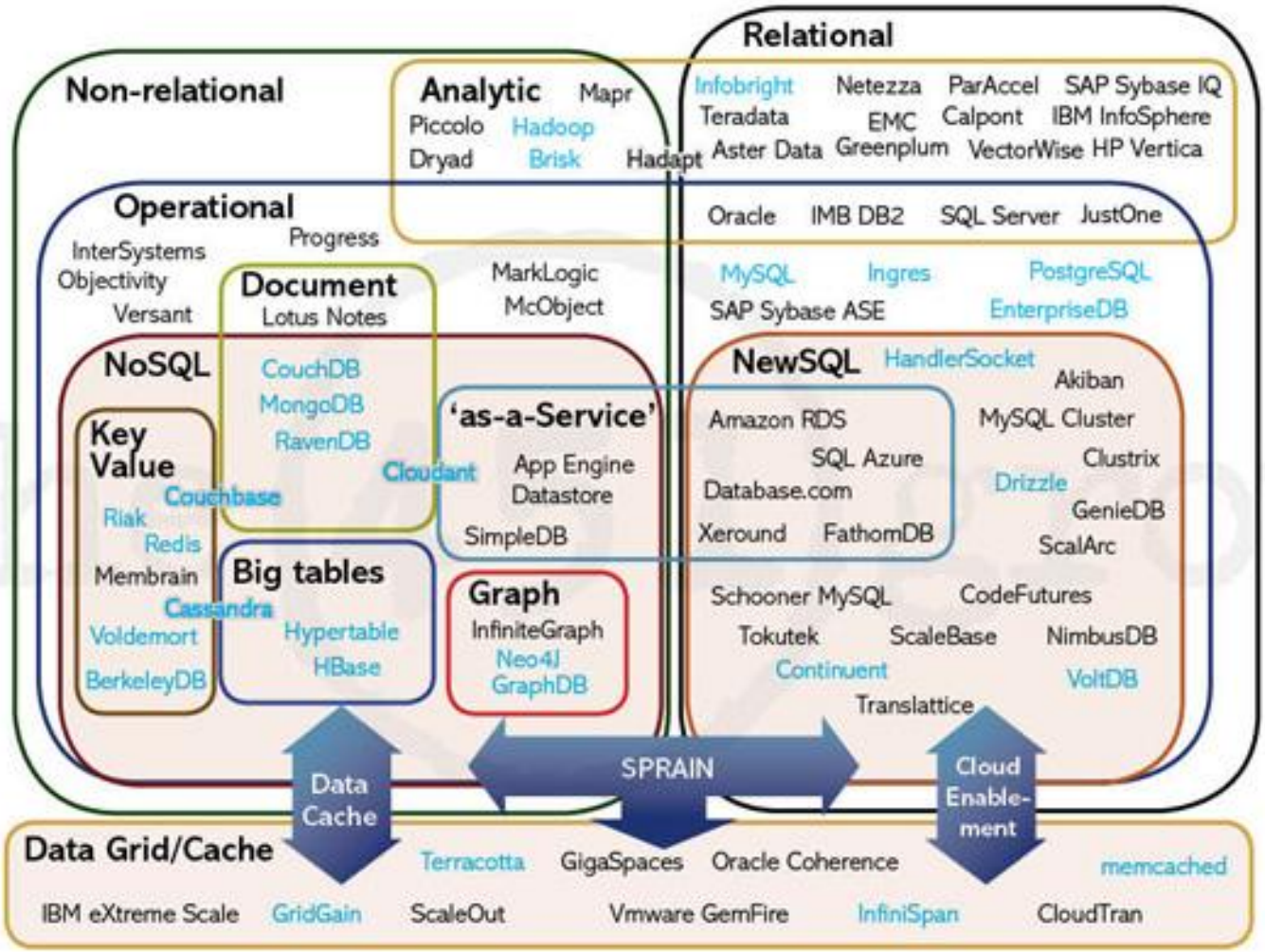
處理巨量資料的三類技術(2)
Data in Motion – In-Memory Processing



Google的技術演進 vs Apache 專案

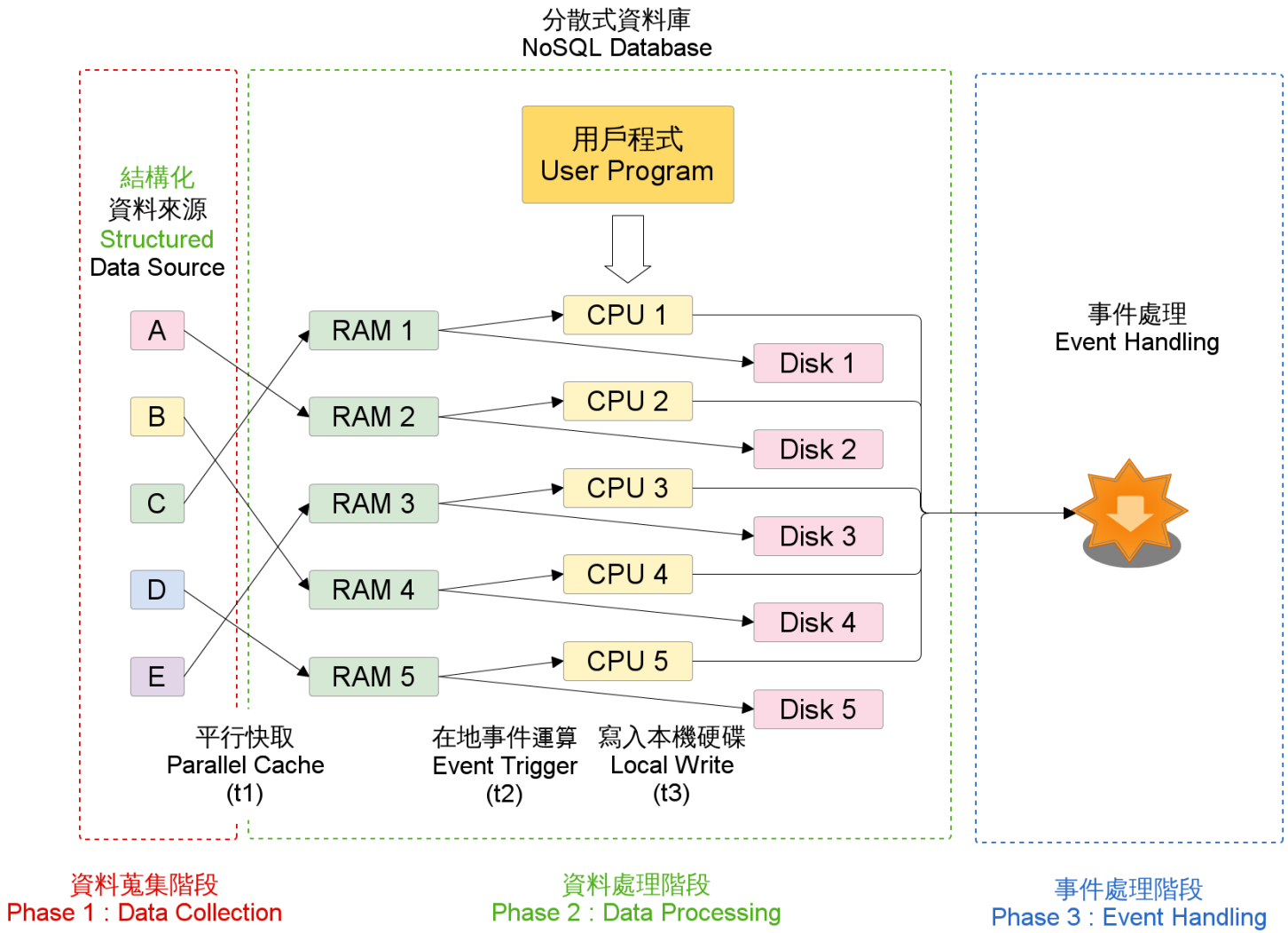
Big Query (JSON, SQL-like)	Dremel (2010)	Apache Drill (2012)
Incremental Index Update (Caffeine)	Percolator (2010)	
Graph Database	Pregel (2009)	Apache Giraph (2011)
Query	BigTable (2006)	Apache HBase (2007)
Map Reduce	MapReduce (2004)	Hadoop MapReduce (2006)
Storage	Google File System (2003)	HDFS (2006)

令人眼花撩亂的多樣化資料庫選擇
NoSQL vs NewSQL

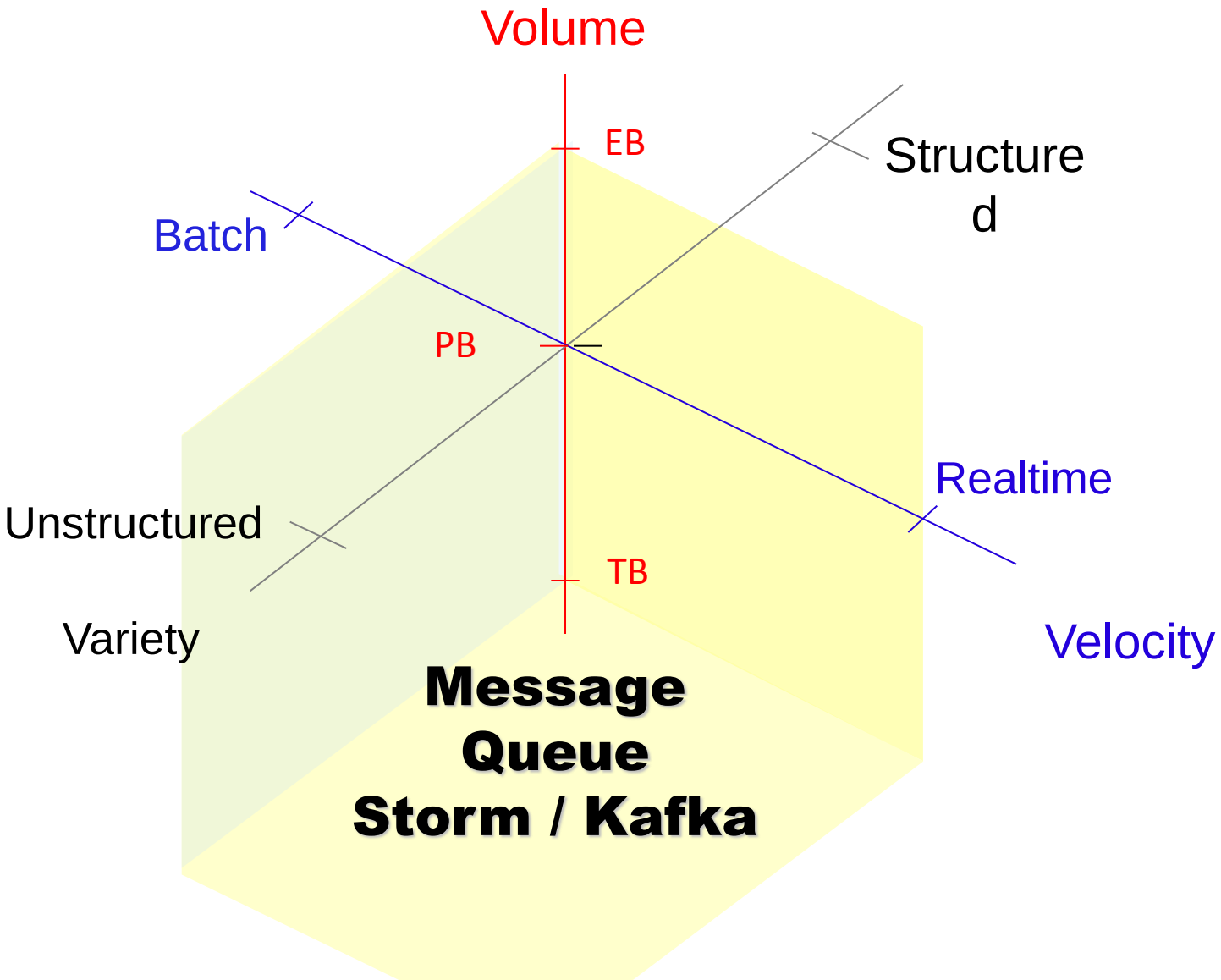


http://www.infoq.com/news/2011/04/newsql

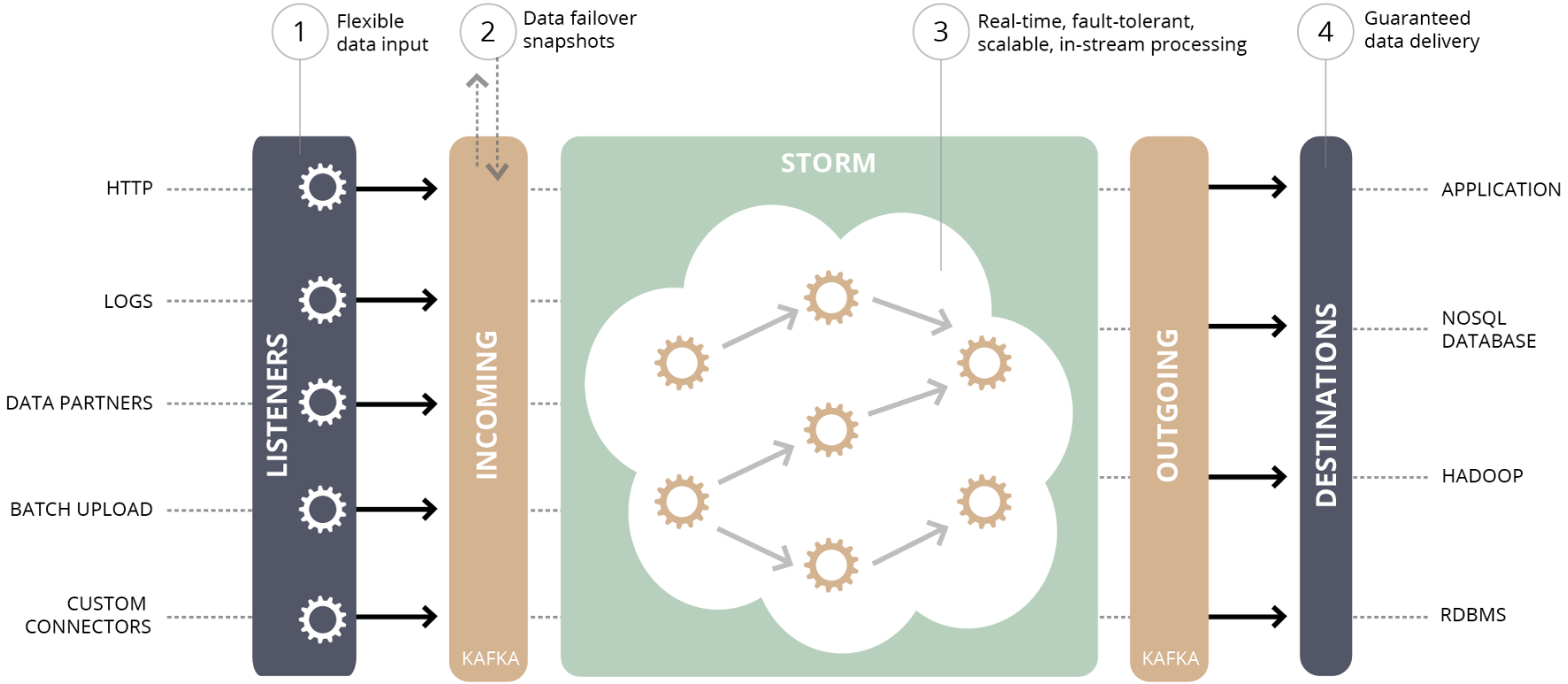
In-Memory Processing的運算時間
以HBase為例



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



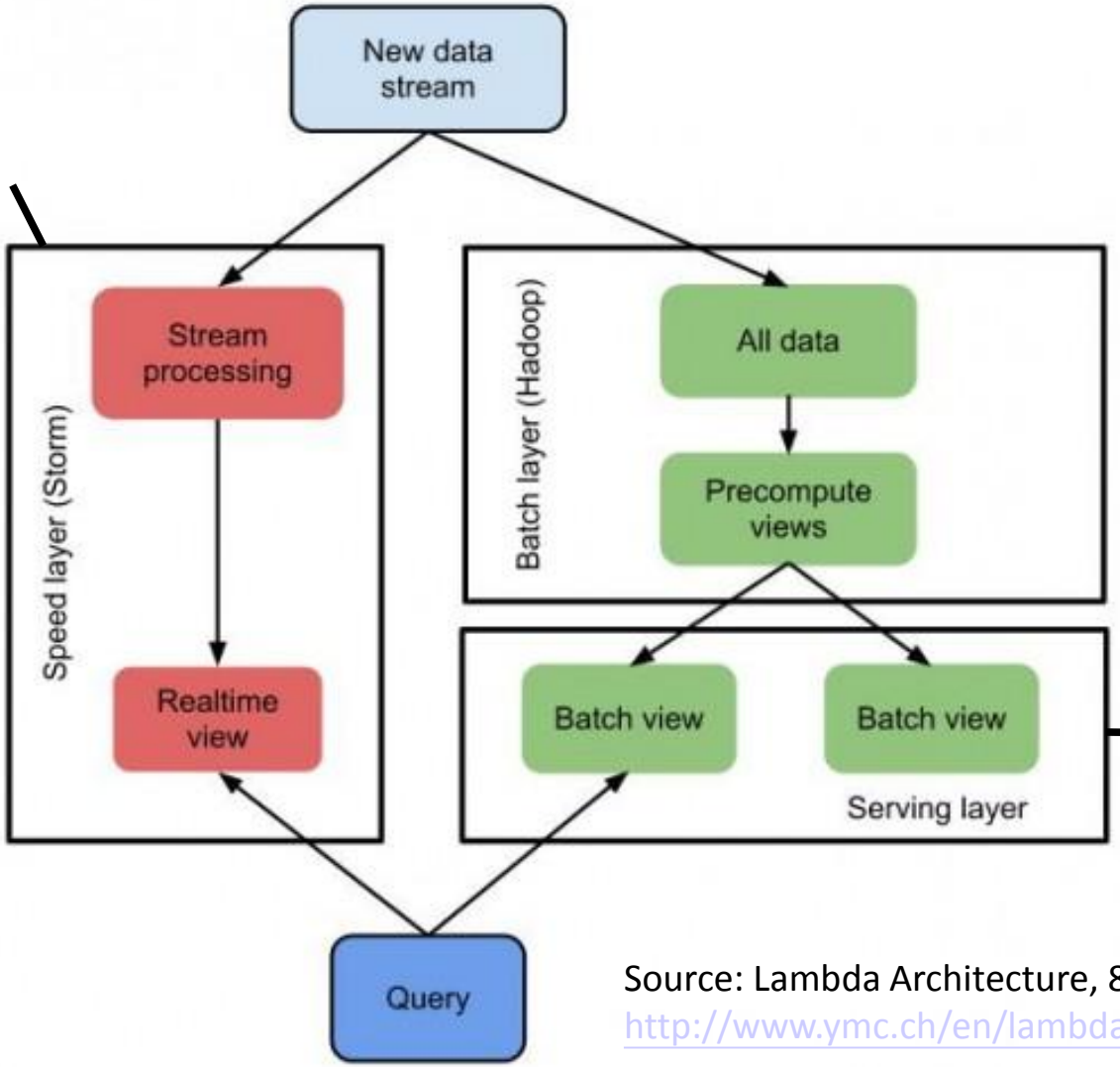
Twitter Storm
+ Apache Kafka



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 for Big Data after 2013

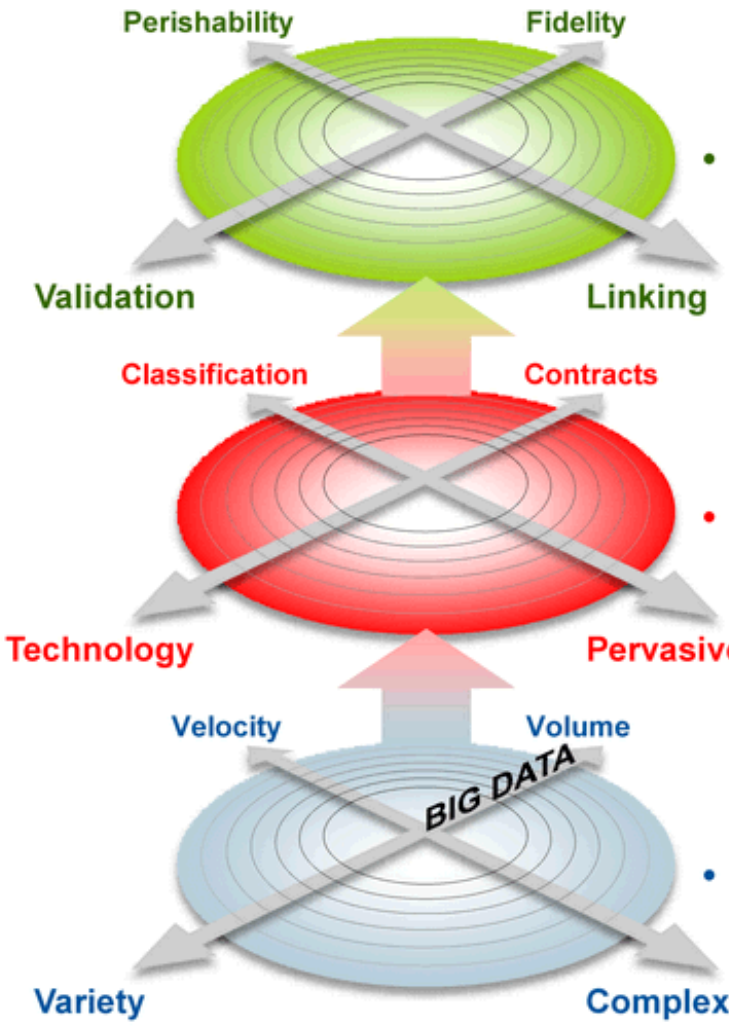
HBase
Storm



ElephantDB
Or
Voldemort

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

Future: 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

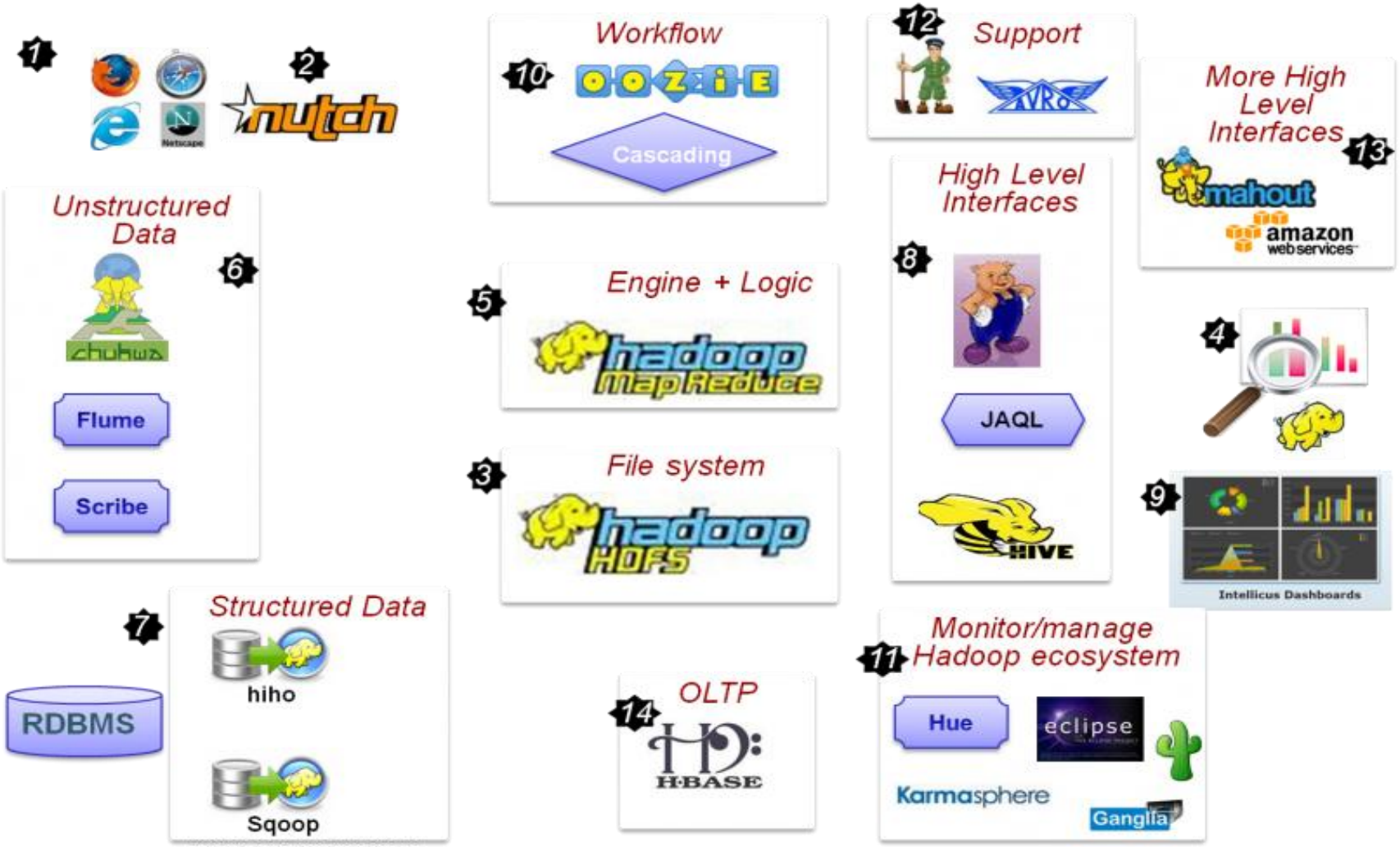
國家安全：美國**PRISM**計劃
(網軍**!**終極警探**4.0)**

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！

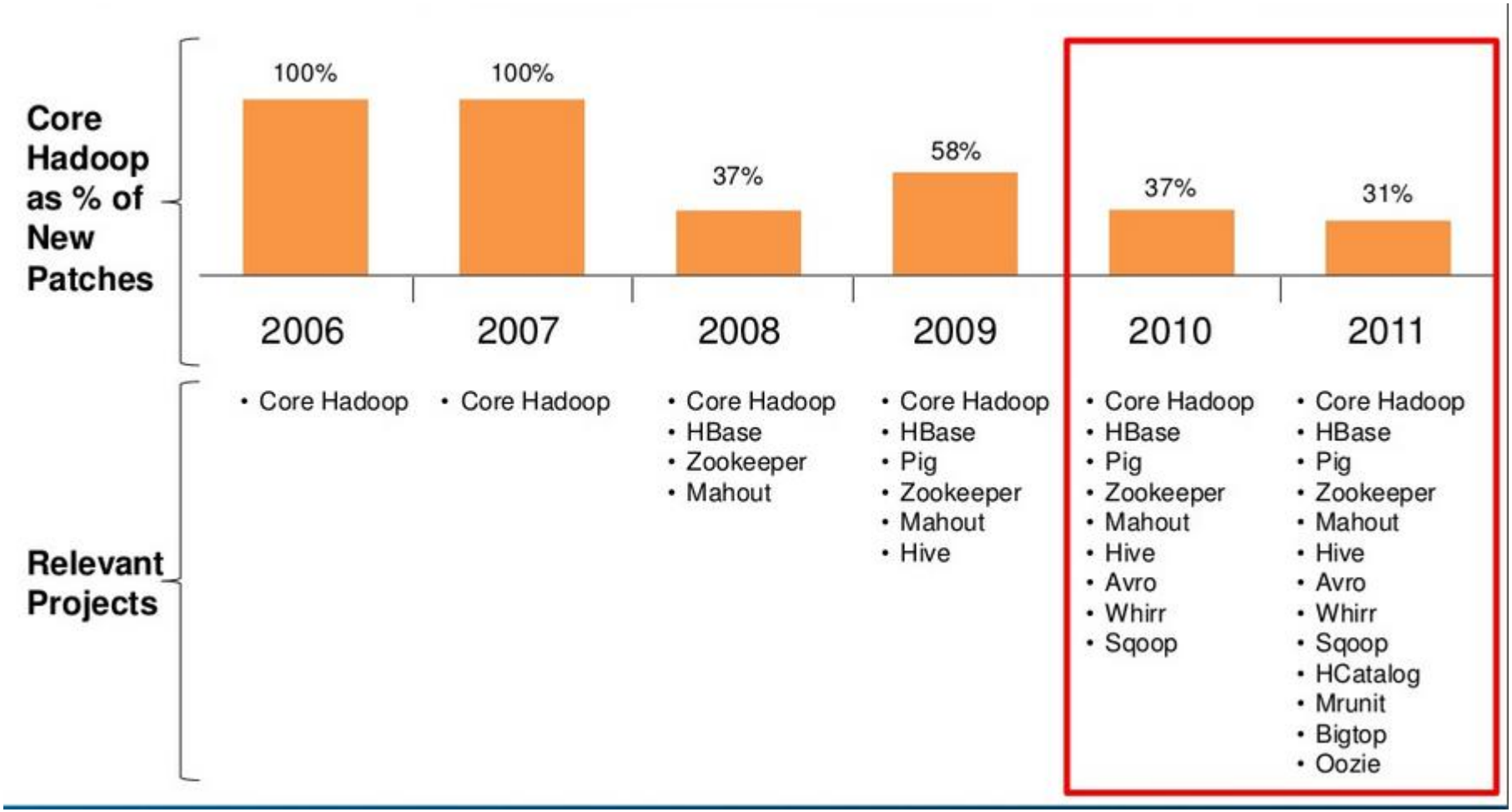
Source: Gartner (March 2011), 'Big Data' Is Only the Beginning of Extreme Information Management, April 2011, http://www.gartner.com/id=1622715

Future: Better Hadoop Ecosystem



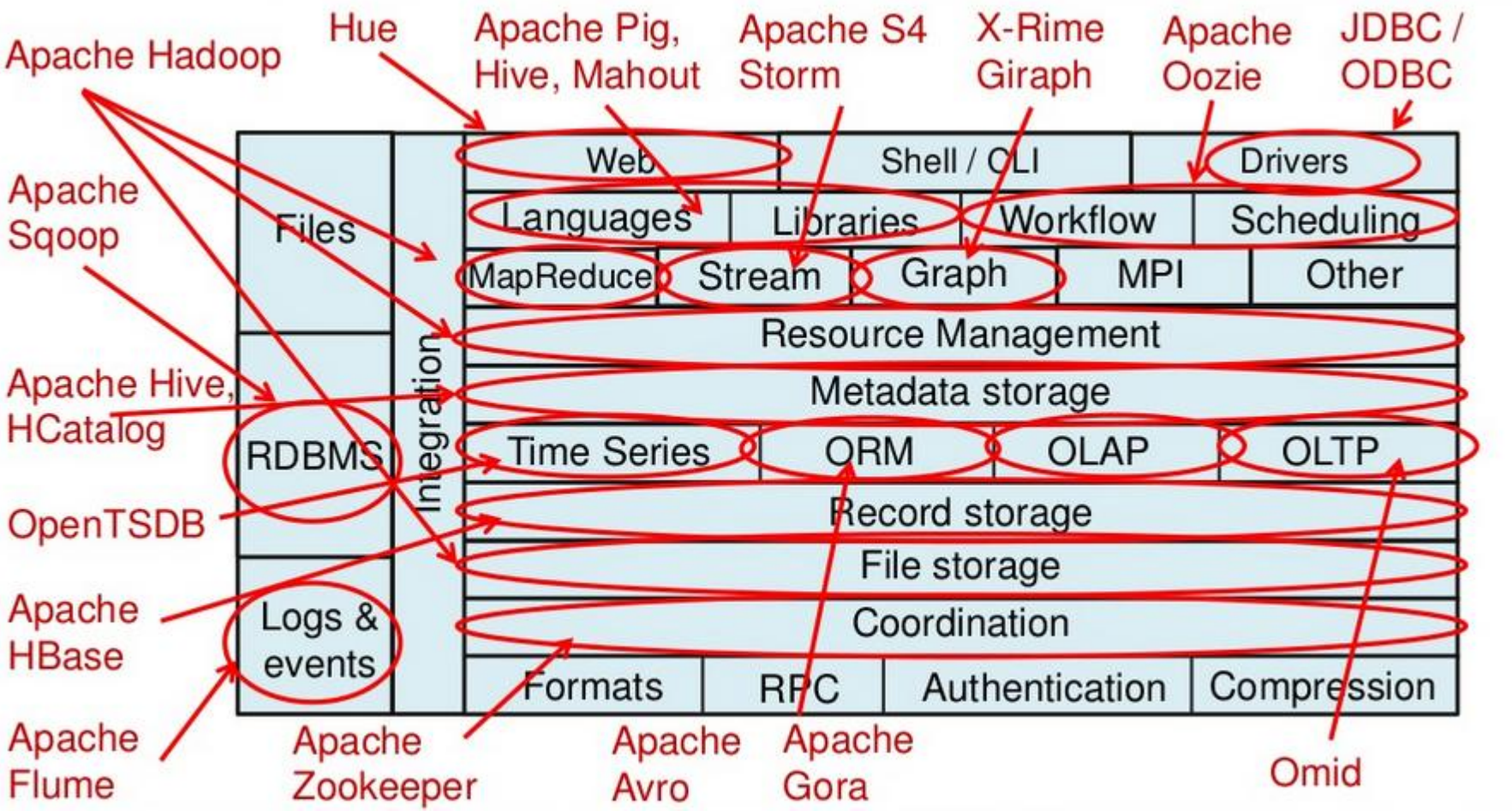
http://rationalintelligence.com/wp_log/?p=104

Evolution of Apache Hadoop Ecosystem



Hadoop World 2011: The Hadoop Stack - Then, Now and in the Future
http://www.slideshare.net/slideshow/embed_code/10110006

Complexity of Apache Big Data Stack



Hadoop World 2011: The Hadoop Stack - Then, Now and in the Future
http://www.slideshare.net/slideshow/embed_code/10110006