巨量資料分析與應用 (1)

楊立偉教授

台灣科技大學資管系

2013 Fall

楊立偉教授

- **專長領域為資料庫及語意分析技術、知識管理、數位行銷**

現任	台科大資管系兼任助理教授 2008~
	台大工管系暨商研所兼任助理教授 2006~
	資訊及通信國家標準技術委員
	意藍資訊　董事總經理（創辦人）1999~ 國內規模最大的網路情報與社群口碑自動分析平台
	龍捲風科技　董事總經理 國內企業搜尋引擎市佔率最高；國際檢索競賽第一名
經歷	智威湯遜數位行銷首席顧問、尚藍互動行銷共同創辦人
	2009年獲選100 MVP最有價值經理人，擁有超過20項語意分析專利
	2012年榮獲國家雲端創新獎、數位時代「創業之星」首獎

課程內容

- 目標
  - 了解巨量資料（big data）與相關趨勢
  - 了解巨量資料分析及管理的理論與架構
  - 了解相關技術，以及如何應用於企業營運
  - 了解最新的相關發展議題巨量資料導論
- 對象
  - IT經理或系統網路部門主管
  - 專案經理、系統架構師或系統網路管理人員
  - 對於雲端運算之大量資料處理、分析、應用有興趣者

課程大綱

- 第一部份
  - 巨量資料導論
  - 巨量資料分析與管理架構
  - 巨量資料分析技術
- 第二部份
  - 應用案例與研討 – 企業個案 (1)
  - 應用案例與研討 – 企業個案 (2)
  - 應用案例與研討 – Open Data

參考資料

- 技術相關
  - Data warehousing in the age of big data (2013) by Krish Krishnan.
- 管理相關
  - Big Data : A Revolution that will transform how we live, work, and think (2013) by Viktor Mayer-Schonberger, Kenneth Cukier.

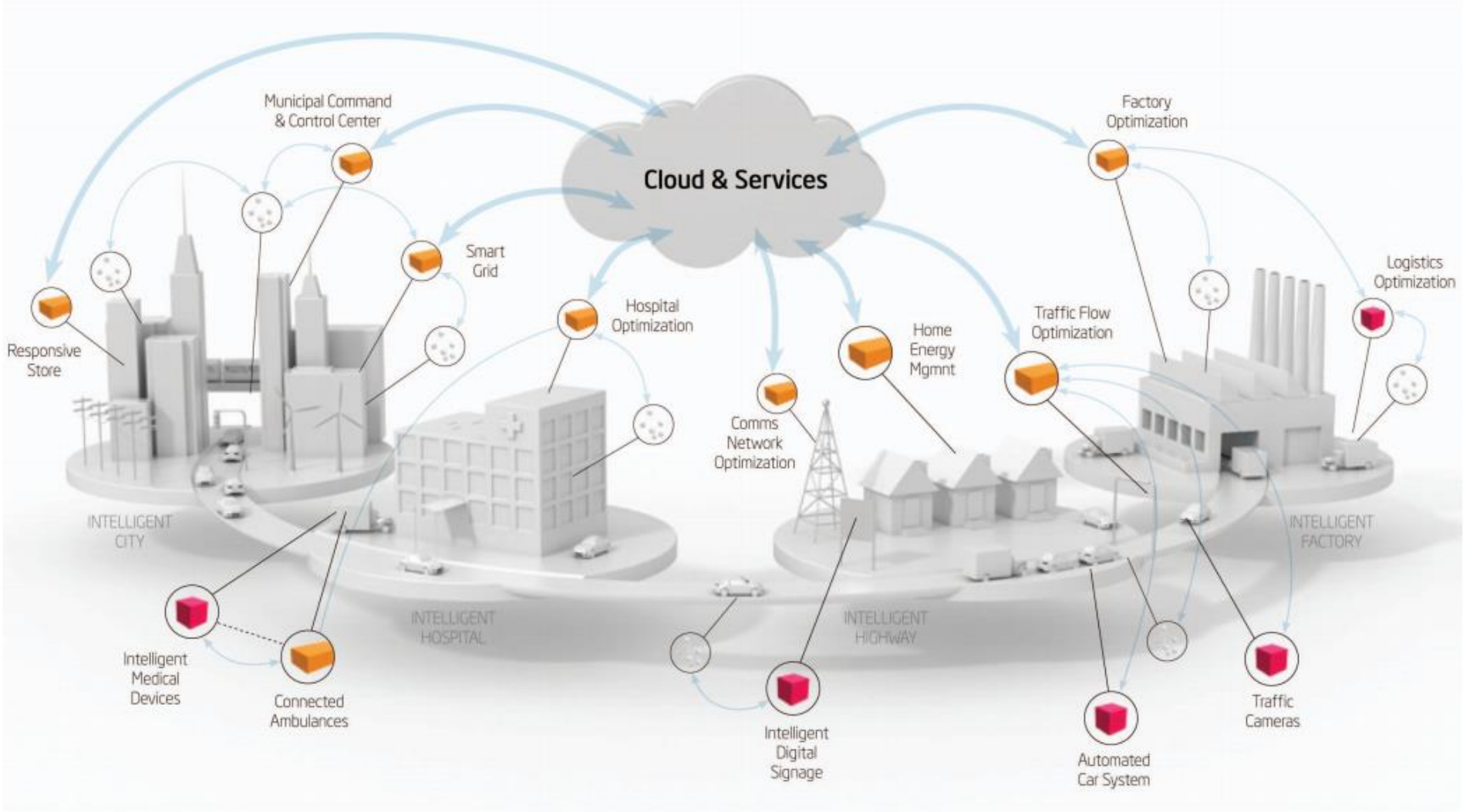
    中譯本：《大數據》，天下文化，2013年5月
  - 《雲端時代的殺手級應用：海量資料分析》，胡世忠 著，天下文化，2013年3月

巨量資料分析導論 (1)

資訊科技的發展趨勢 Trend of IT (1)

- Network 網路
  - Internet and Web 網際網路
    - Internet
    - Internet of Things 物聯網
    - Cloud computing 雲端運算
  - Mobile network 行動(無線)網路
    - 3G, 4G, LTE, WiFi, RFID, NFC, etc.
  - e-Commerce 電子商務
  - Social network 社交網路

雲端運算 與 物聯網 應用情境

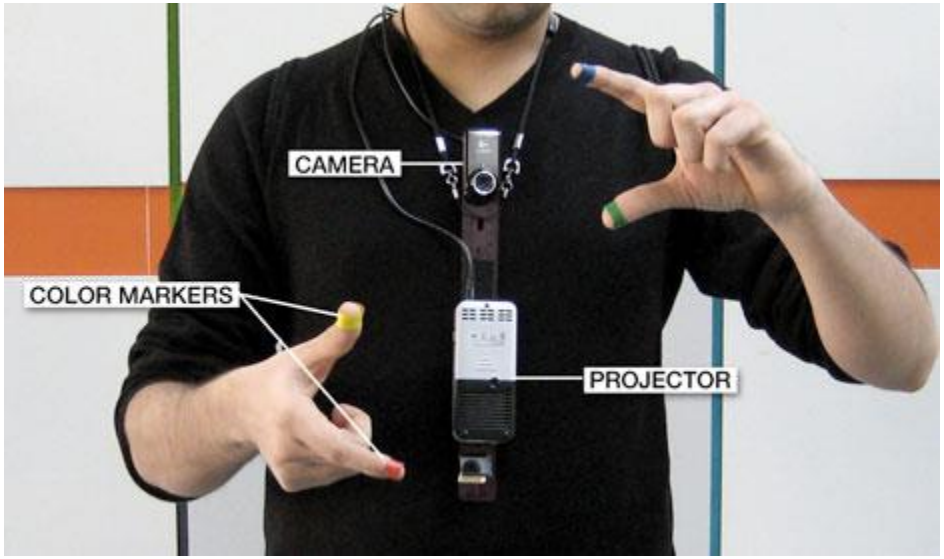


Source : Internet of Things Also a Security Threat by Anthony Myers

資訊科技的發展趨勢 Trend of IT (2)

- **Interface 人機界面**
  - **Multimedia 多媒體**
    - Image, Audio, Video, Virtual Reality, Augmented Reality, etc.
  - **User interface 使用者界面**
    - Touch, Voice, Camera, Motion capturing, etc.
  - **Wearable technology 穿戴式科技**
    - Biometric 生物辨識
    - Biosensor 生物感測

MIT's wearable device turns any surface into an interface



and more wearable devices coming.

資訊科技的發展趨勢 Trend of IT (3)

- Data
  - Database
    - Data warehouse 資料倉儲
    - Data mining 資料探勘
  - Business intelligence 商業智慧
  - Big data

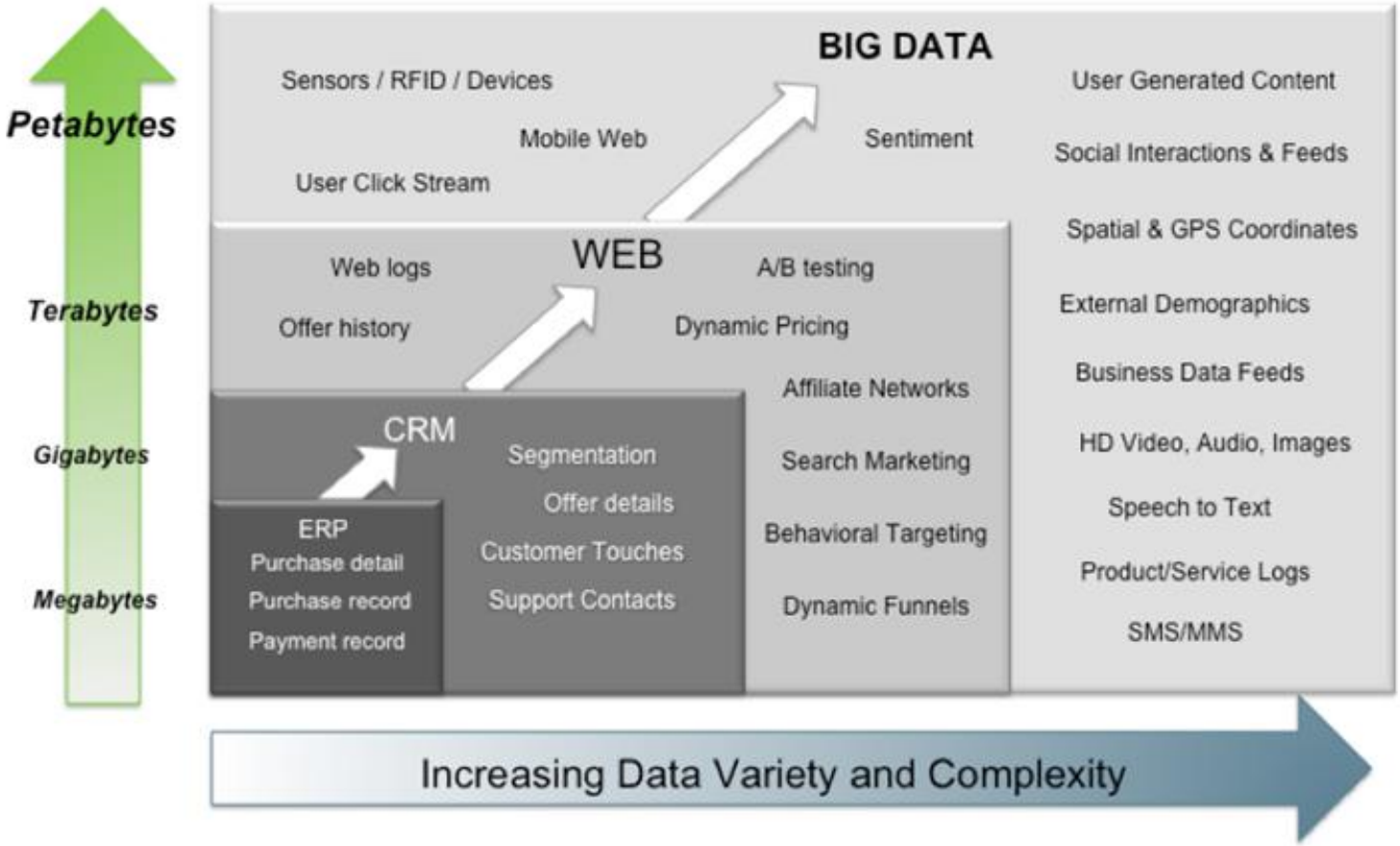
Online in 60 Seconds

A Smarter Planet

By smarter, we mean the world is becoming:

instrumented + interconnected + intelligent

Big Data = Transactions + Interactions + Observations



Source : IBM, Teradata.

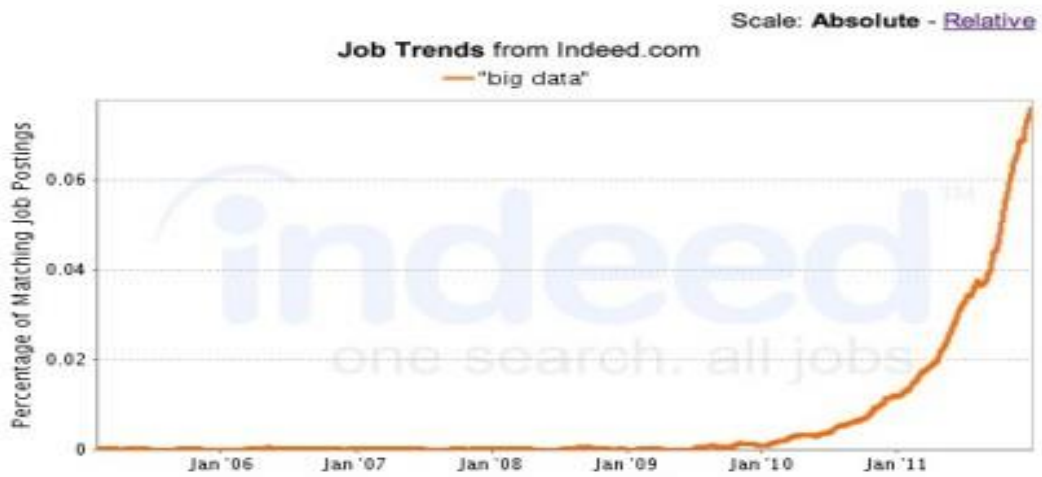
巨量資料分析導論 (2)

Trend of Big Data (1)

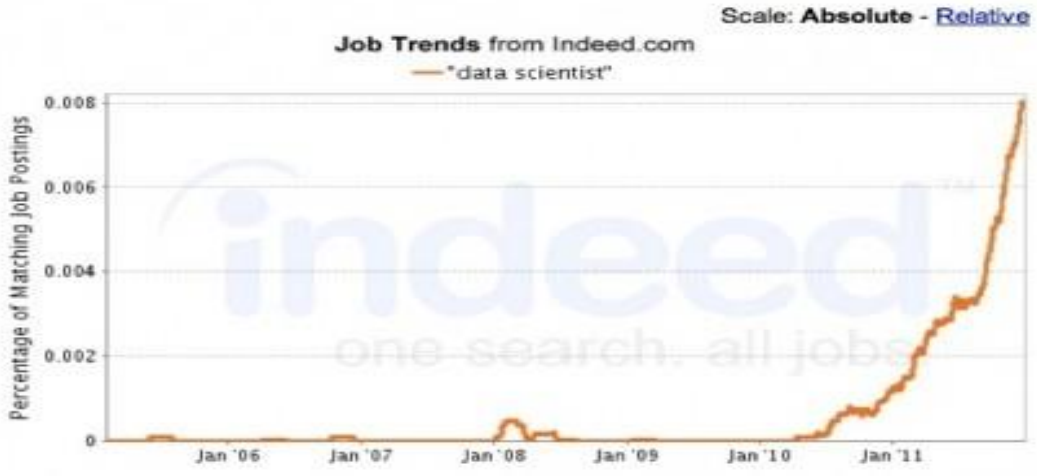
- Big Data 係指資料大量成長
- 根據IBM的研究，全世界90%的資料是在過去2年產生
- Google、Facebook 等，就是站在Big Data上的範例
- 巨大的數據源，將改變整個學術界，商界和政府
- 依賴資料庫工具處理
  - 包括 capture， storage， search， analytics 等

Trend of Big Data (2)

"big data" Job Trends



"data scientist" Job Trends

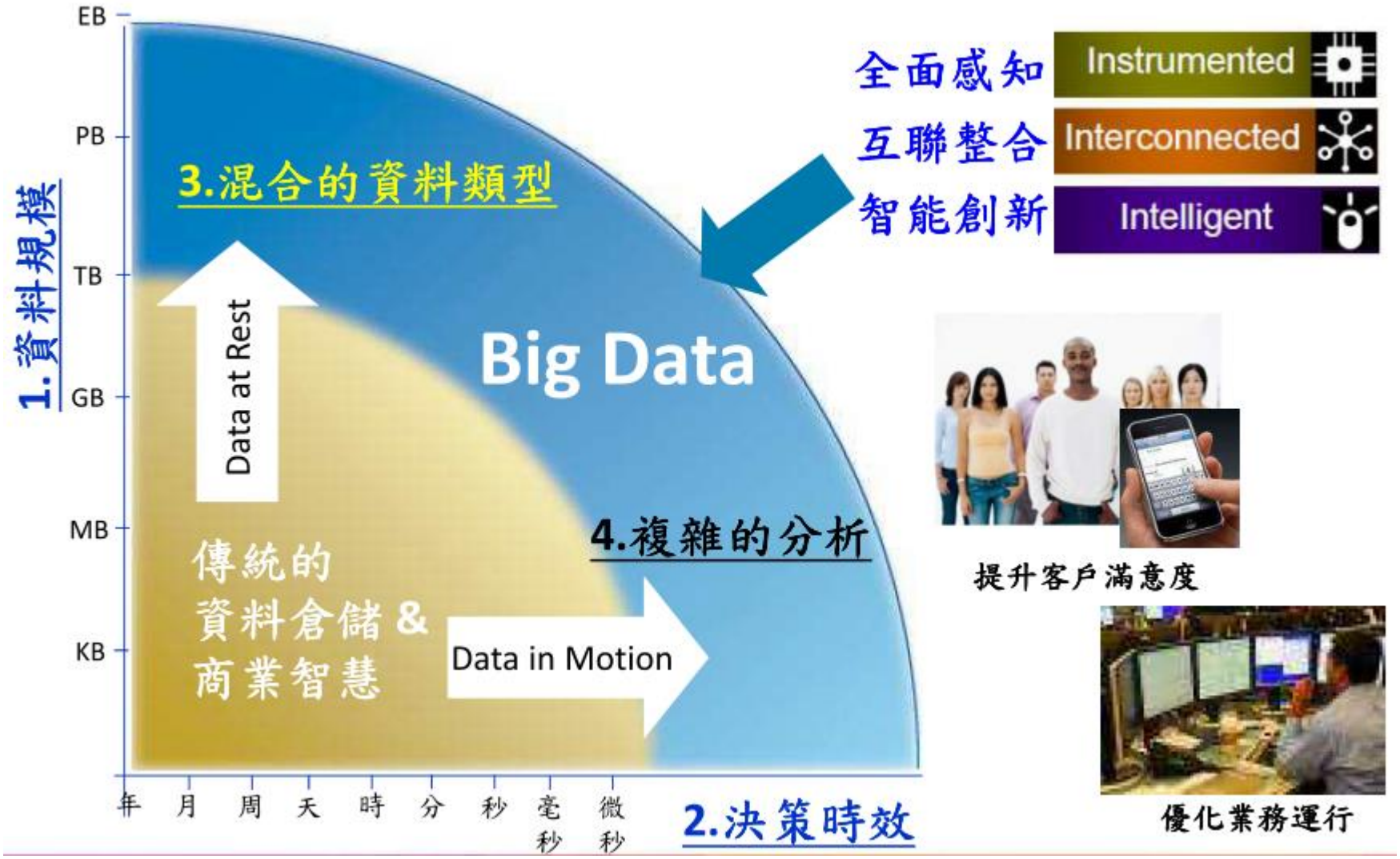


資料分析人才
需求大幅增加

Trend of Big Data (3) 每天產生的資料

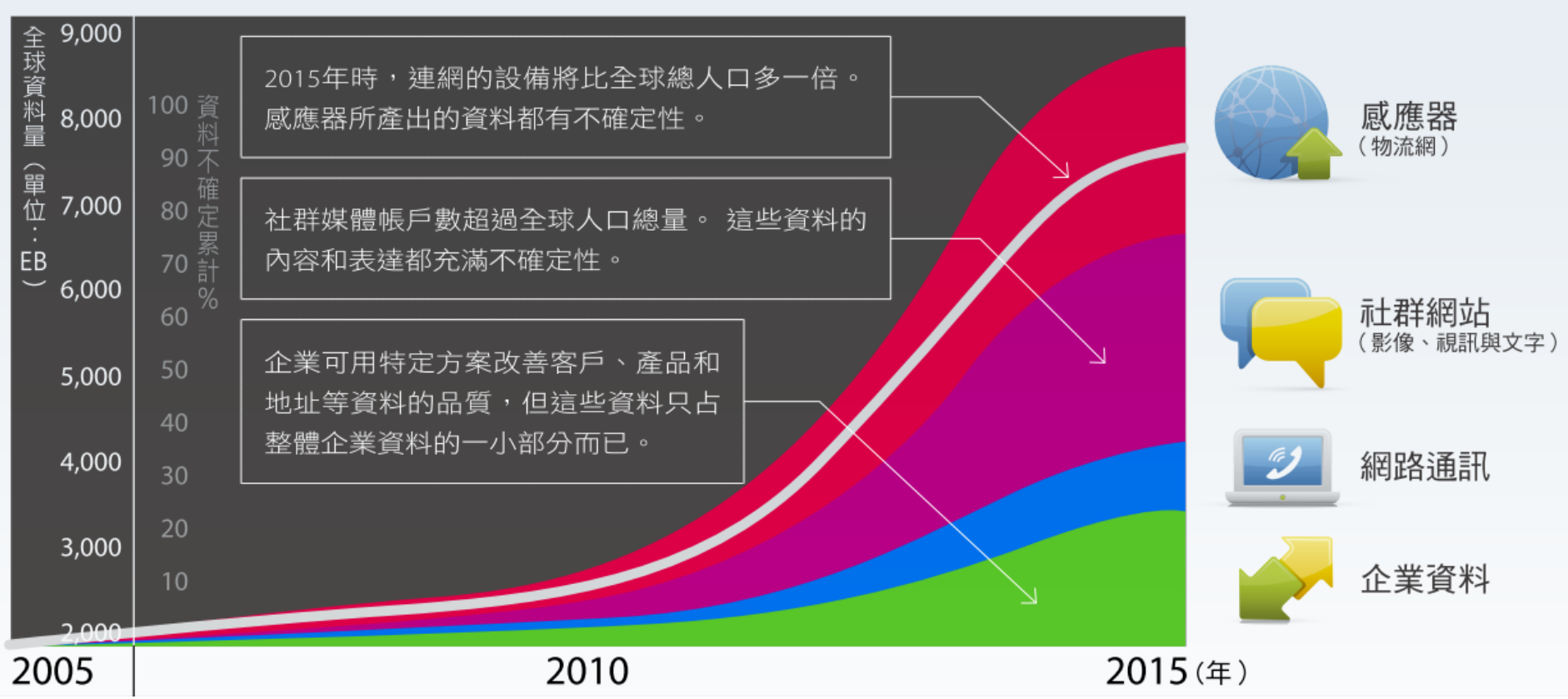
- 搭公車、搭捷運、刷悠遊卡
- 到便利商店買飲料
- 用手機上臉書、按個讚、打個卡
- 送個LINE、拍張照片上傳
- 上網瀏覽、刷卡購物、給個評價
- 過個馬路被數個鏡頭拍下...

Big Data 巨量資料分析的應用緣起



Source : IBM 2012 智慧科技論壇

Big Data 的主要來源



Source : IBM 2012全球CEO調查報告
https://www-07.ibm.com/tw/blueview/2012oct/8.html

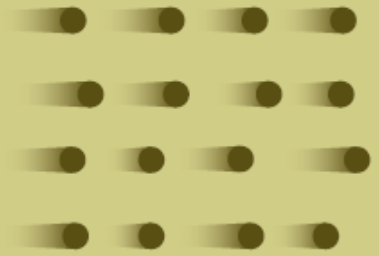
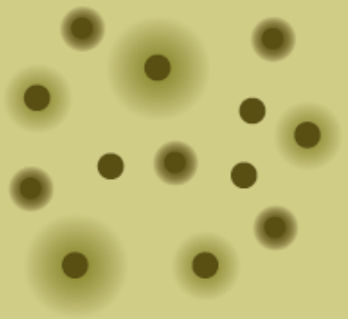
Big Data 的應用方式

- 運用資料與演算，達成智慧決策

Source : IBM 2012全球CEO調查報告
https://www-07.ibm.com/tw/blueview/2012oct/8.html

Big Data 的特性

- 數量大、產生速度快、多樣性、可能存有誤差資料

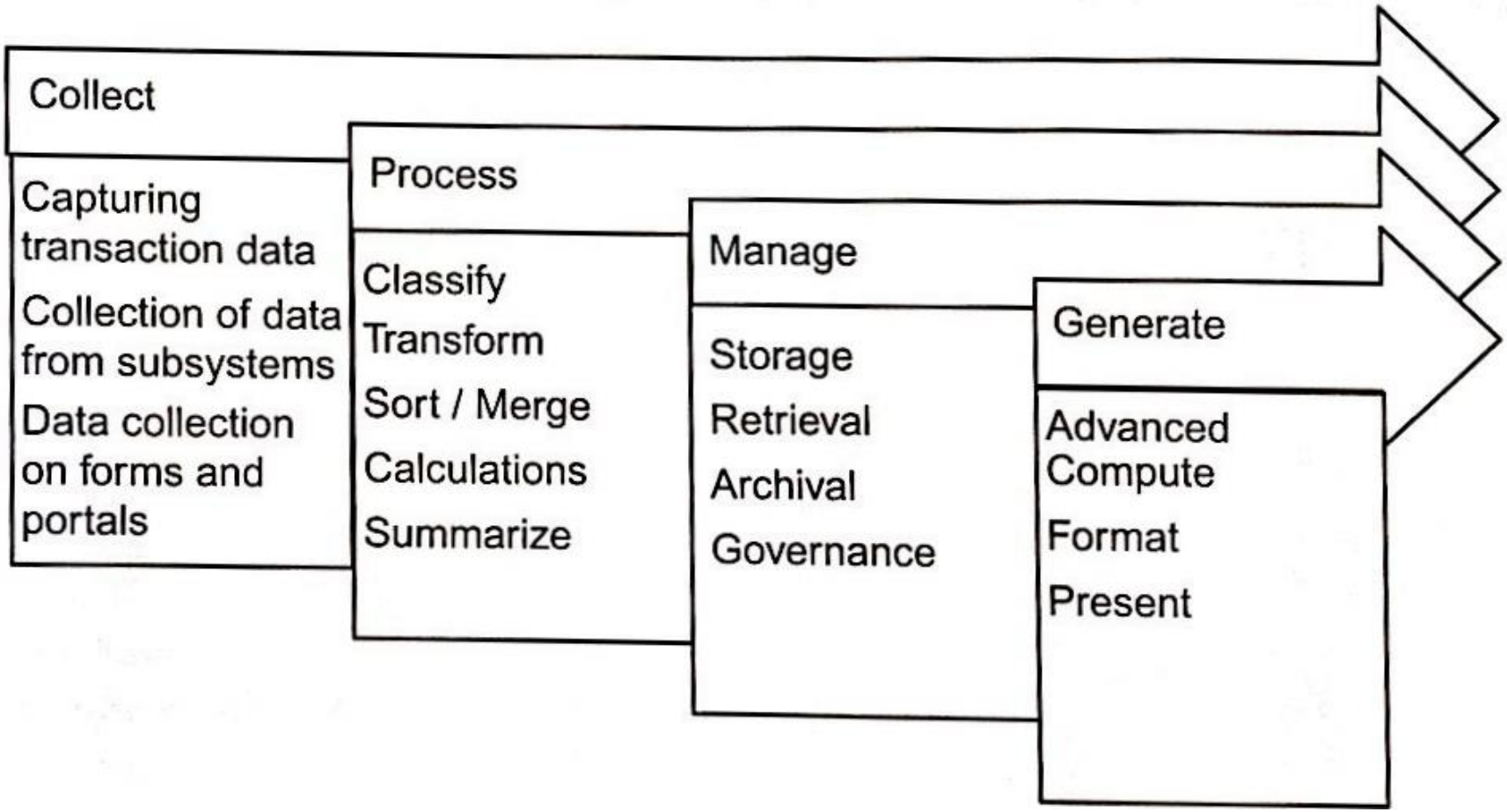
Volume	Velocity	Variety	Veracity
			
Data at rest	**Data in motion**	**Data in many forms**	**Data in doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text and multimedia	Uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception and model approximations

Source : IBM Big Data Hub
http://www.ibmbigdatahub.com/

Big Data 對企業的整體資訊供應鍊

Source : IBM 2012全球CEO調查報告
https://www-07.ibm.com/tw/blueview/2012oct/8.html

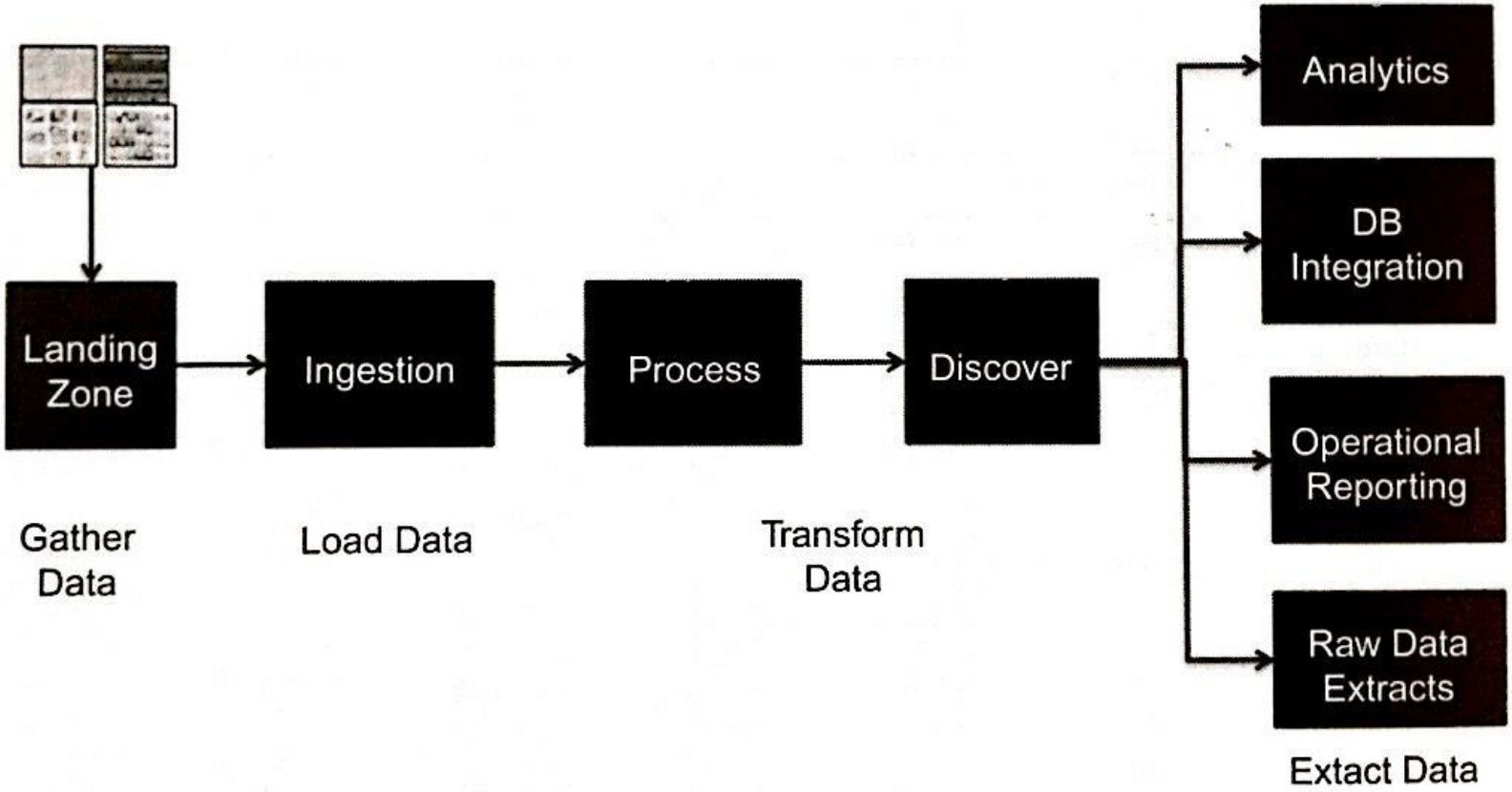
Data Processing Cycle



Source : Data warehousing in the age of big data by Krish Krishnan. Morgan Kaufmann, 2013

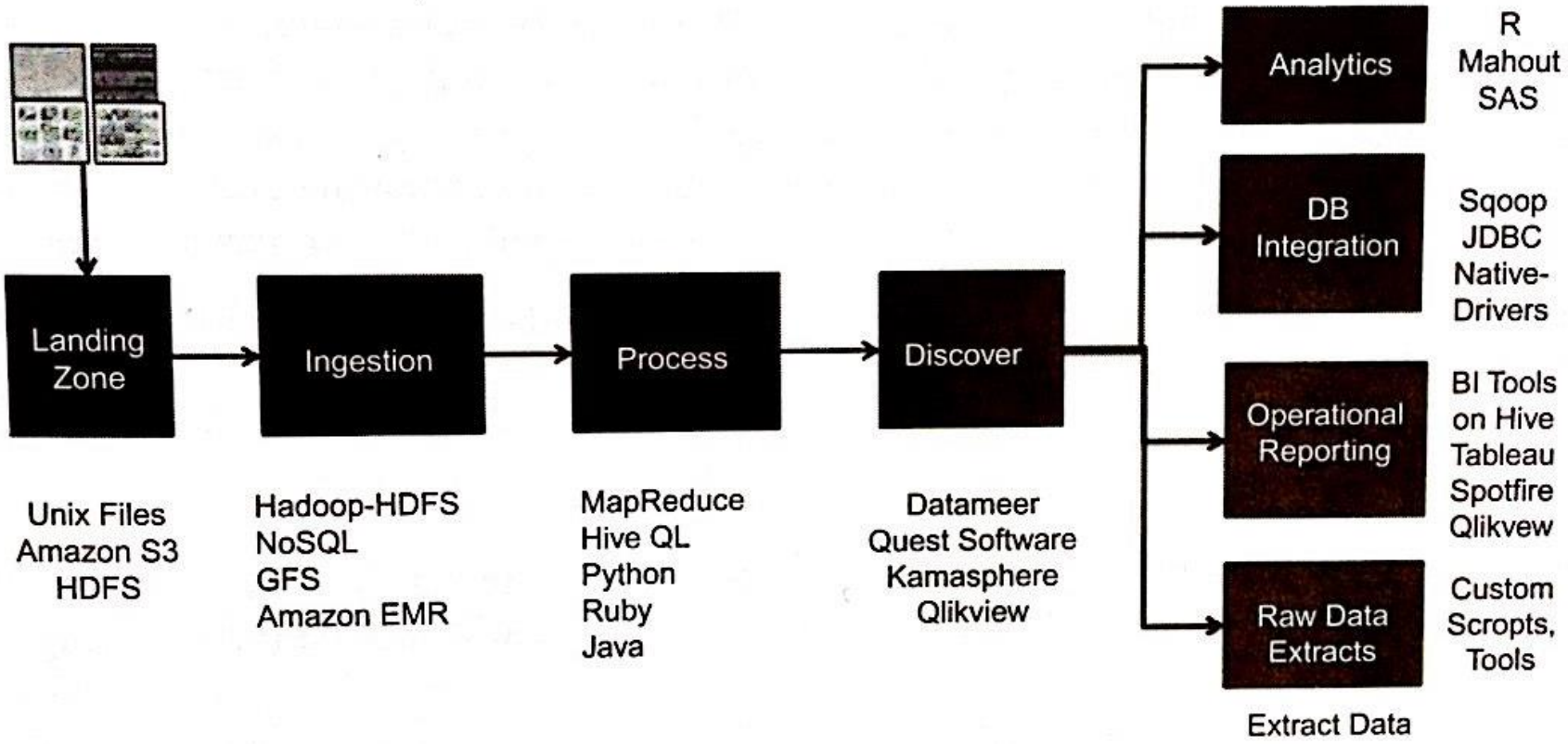
巨量資料分析技術 (1)

Big Data Processing Flow



Source : Data warehousing in the age of big data by Krish Krishnan. Morgan Kaufmann, 2013

Big Data Processing Platform



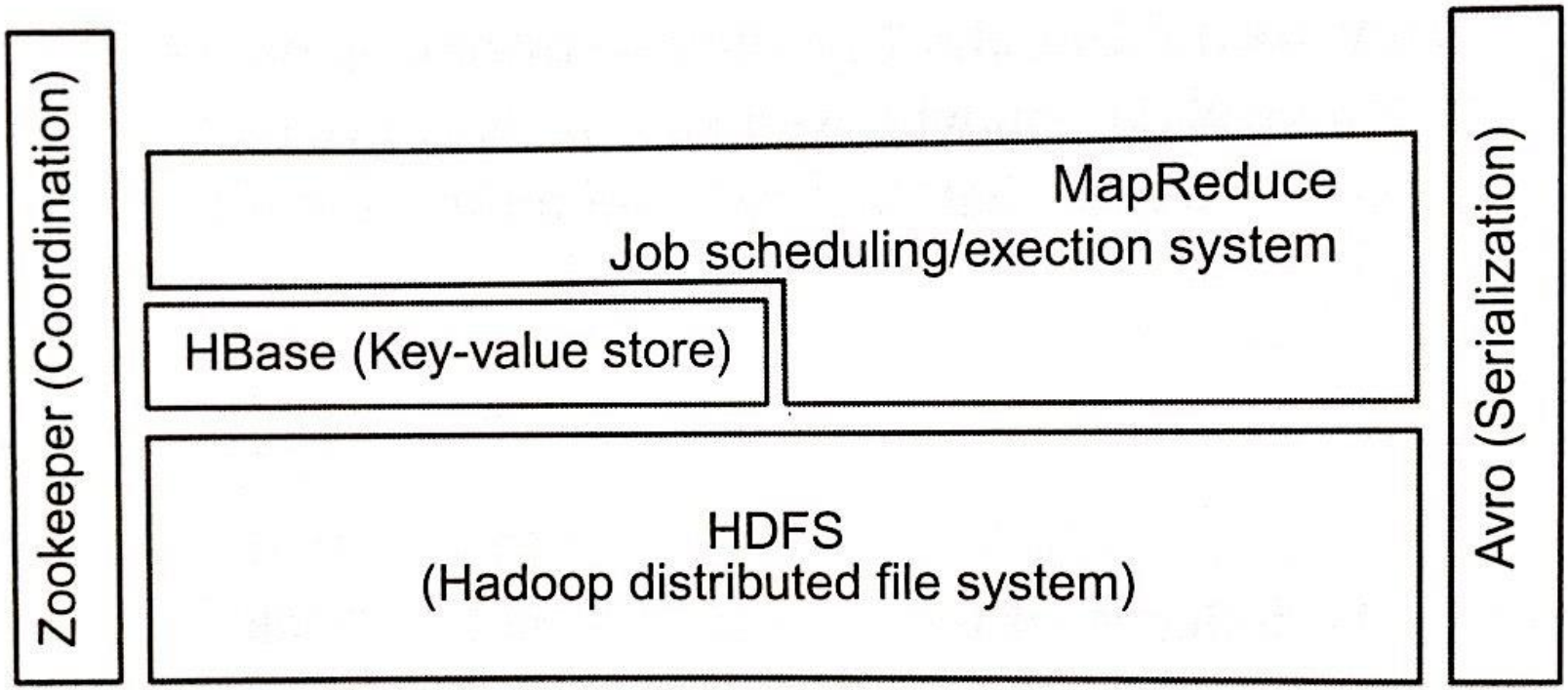
Source : Data warehousing in the age of big data by Krish Krishnan. Morgan Kaufmann, 2013

Big Data 新架構的基本需求

- Extreme parallel processing 儘量平行化處理
- Minimal database usage 降低資料庫用量
- Distributed file-based storage 分散式檔案儲存
- Linearly scalable infrastructure 可線性擴充的架構
- Programmable APIs 提供開發界面
- High-speed replication 可高速複製
- High availability 高可用性
- Localized processing of data and storage of results
- Fault tolerance 容錯性

相關技術架構 – 以Hadoop為例

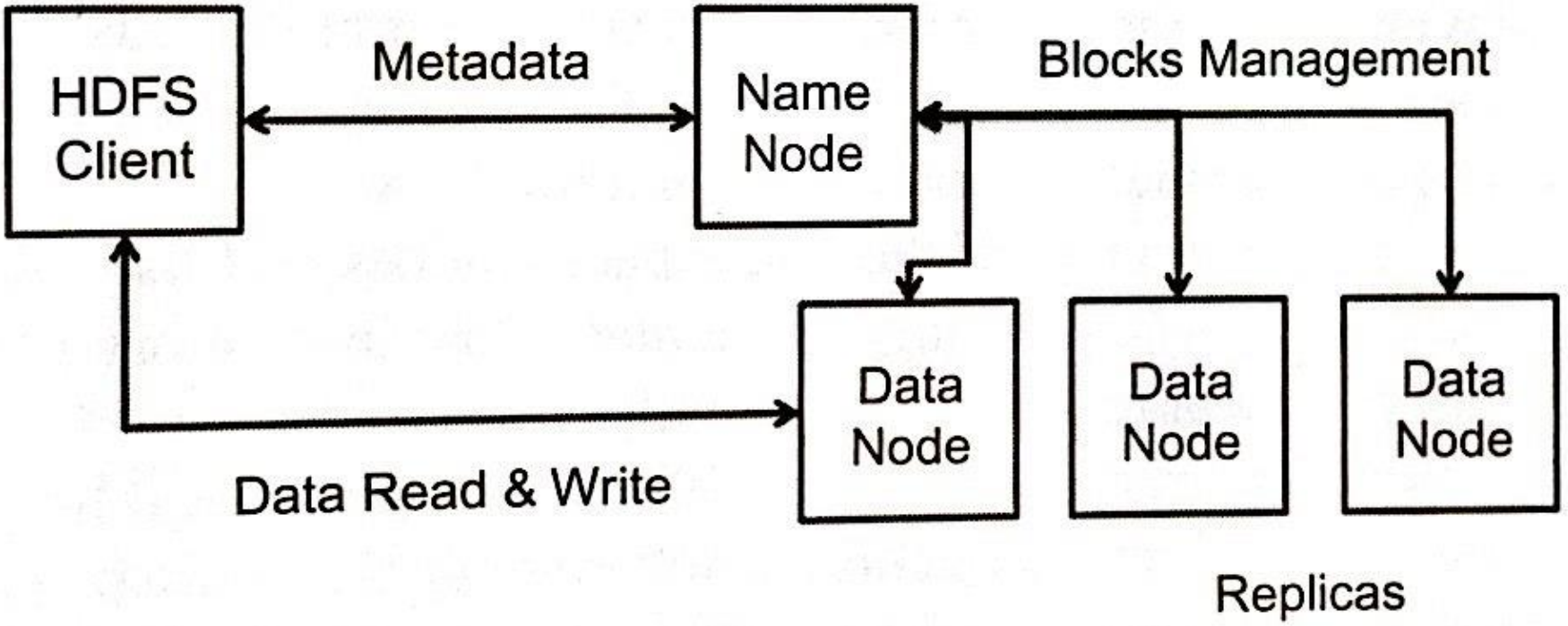
- Hadoop 核心元件



Source : Data warehousing in the age of big data by Krish Krishnan. Morgan Kaufmann, 2013

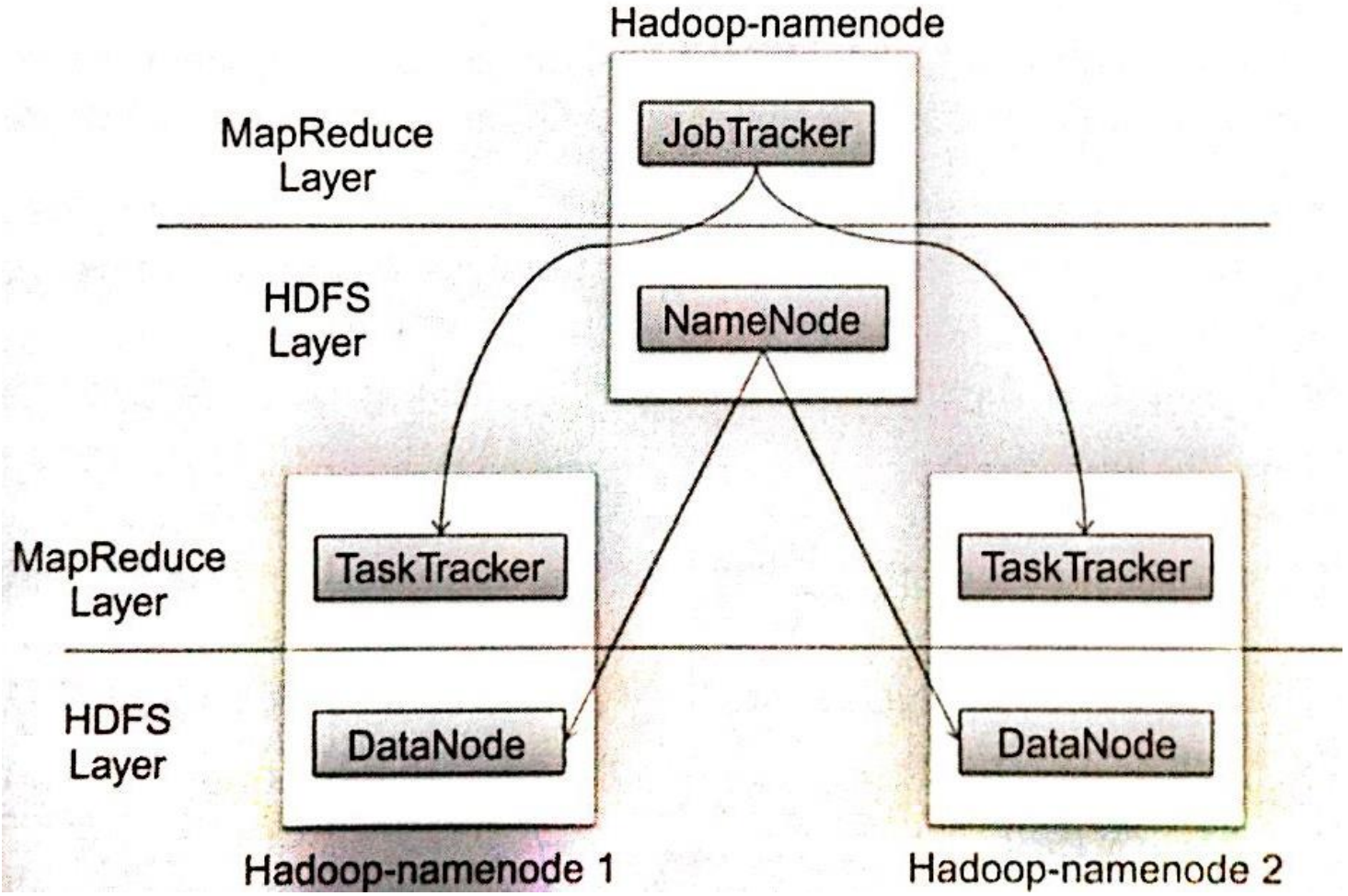
(1) HDFS 架構

- Hadoop Distributed File System
  - Provide high throughput access to application data.
  - Redundancy, Scalability, Fault tolerance, Cross-platform
    - 一般硬體即可, 可適用大檔(GB to PB), 適用快速連續讀取



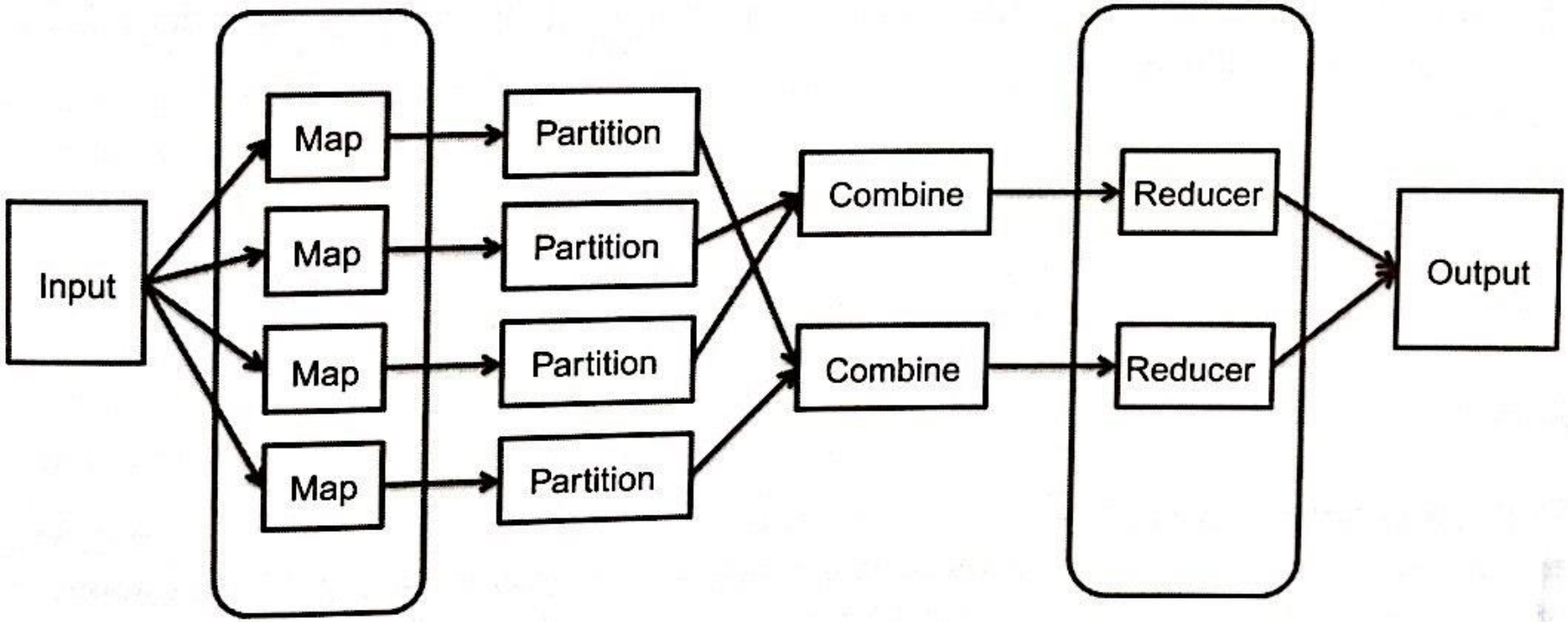
(2) Hadoop 的平行處理方式

- Job executions in Hadoop

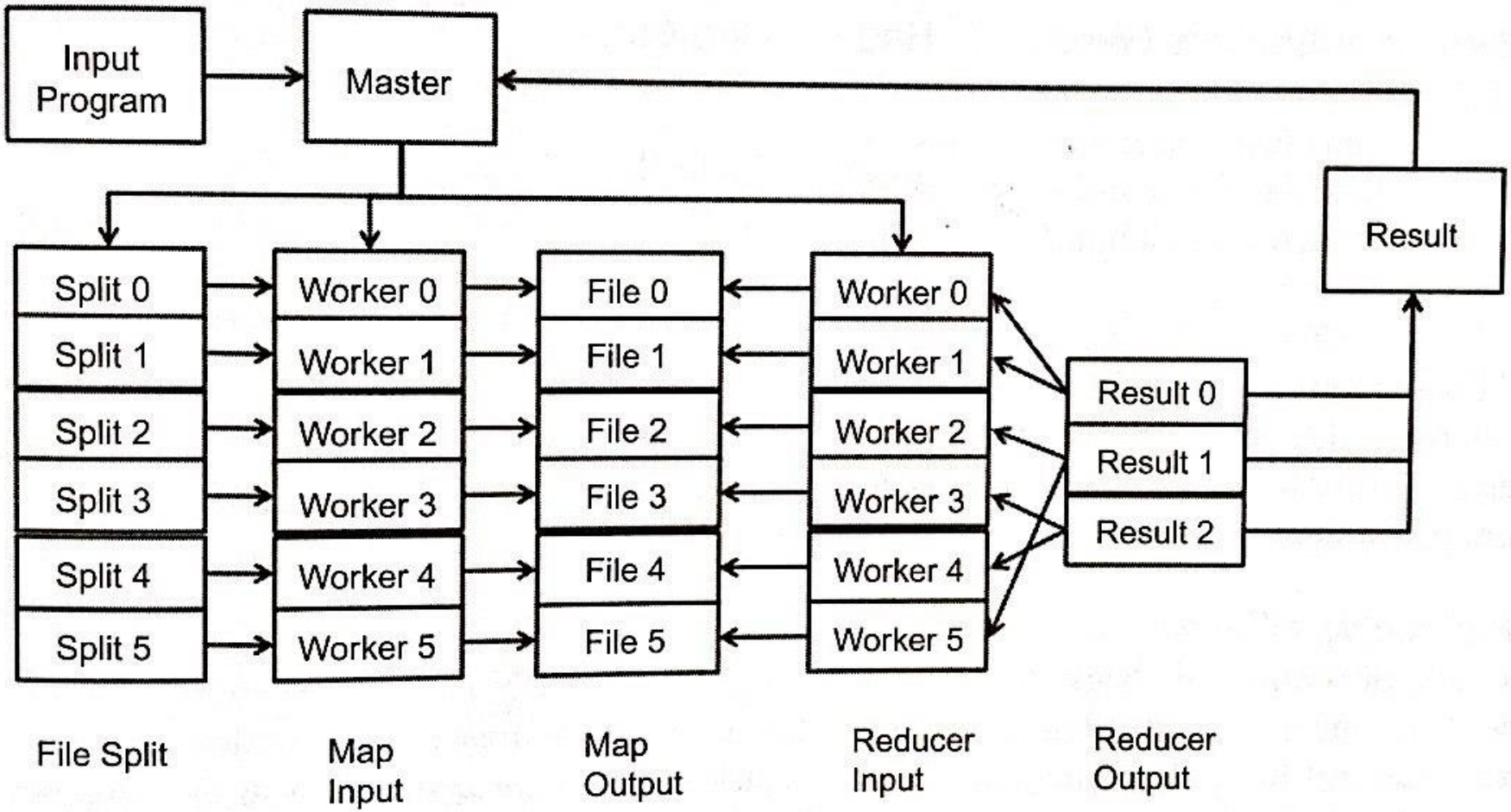


(3) MapReduce 的架構

- a programming model : divide, conquer, merge

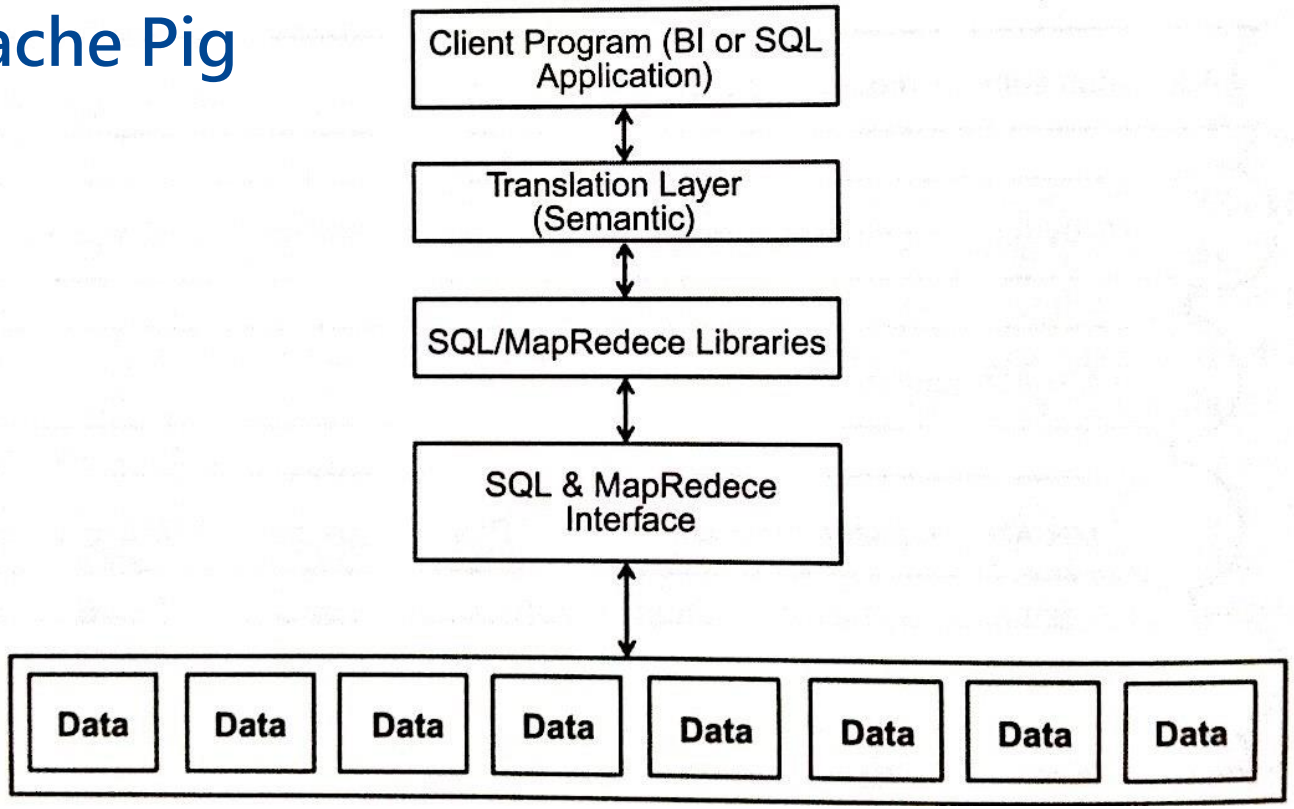


MapReduce implementation



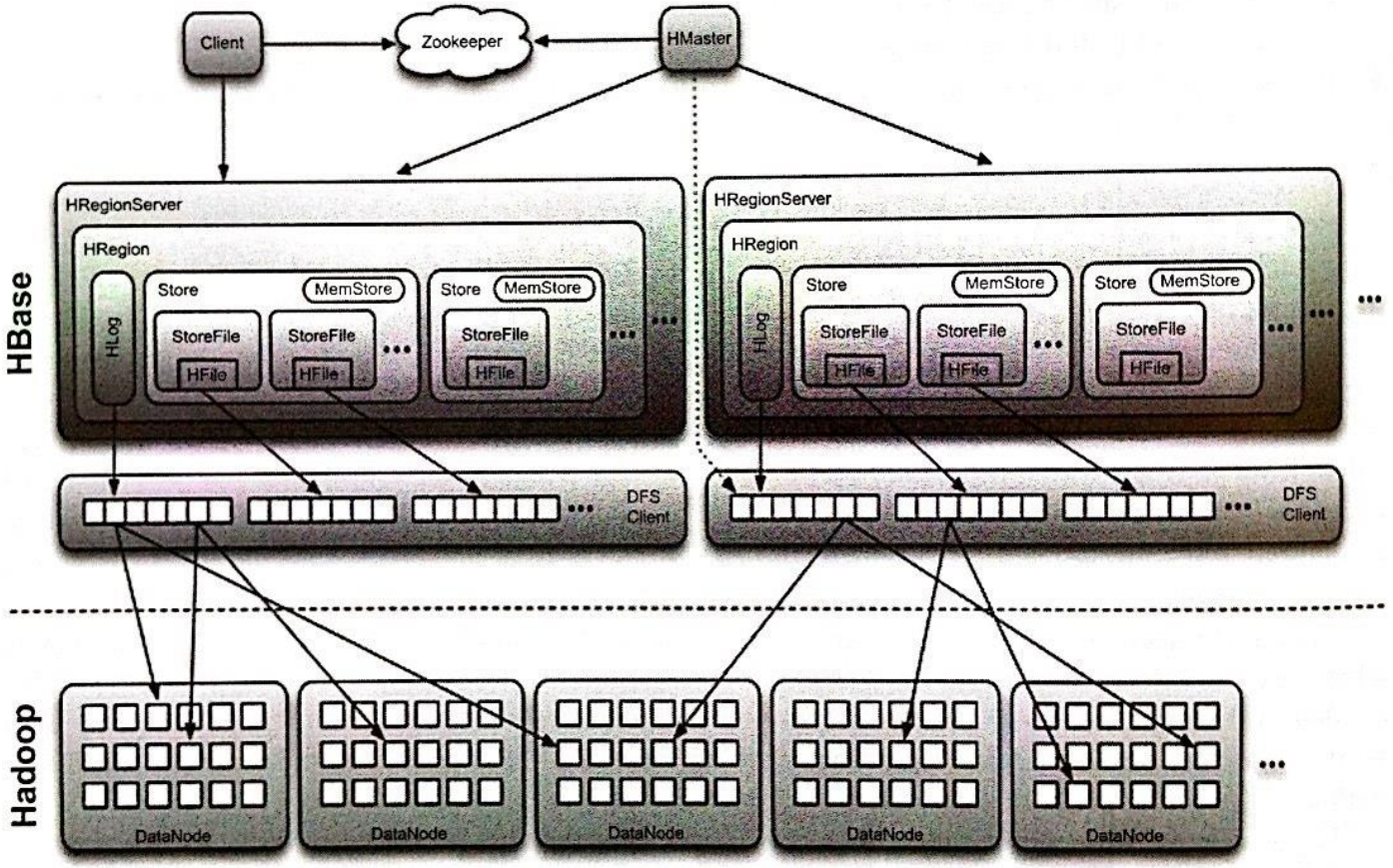
(4) 結合SQL與MapReduce的架構

- 結合MapReduce處理大規模結構與非結構資料，和廣為熟悉的SQL後結果處理能力
- 參考 : Apache Pig



(5) HBase 的架構

- non-relational (key-value), column-oriented, multidimensional, distributed database



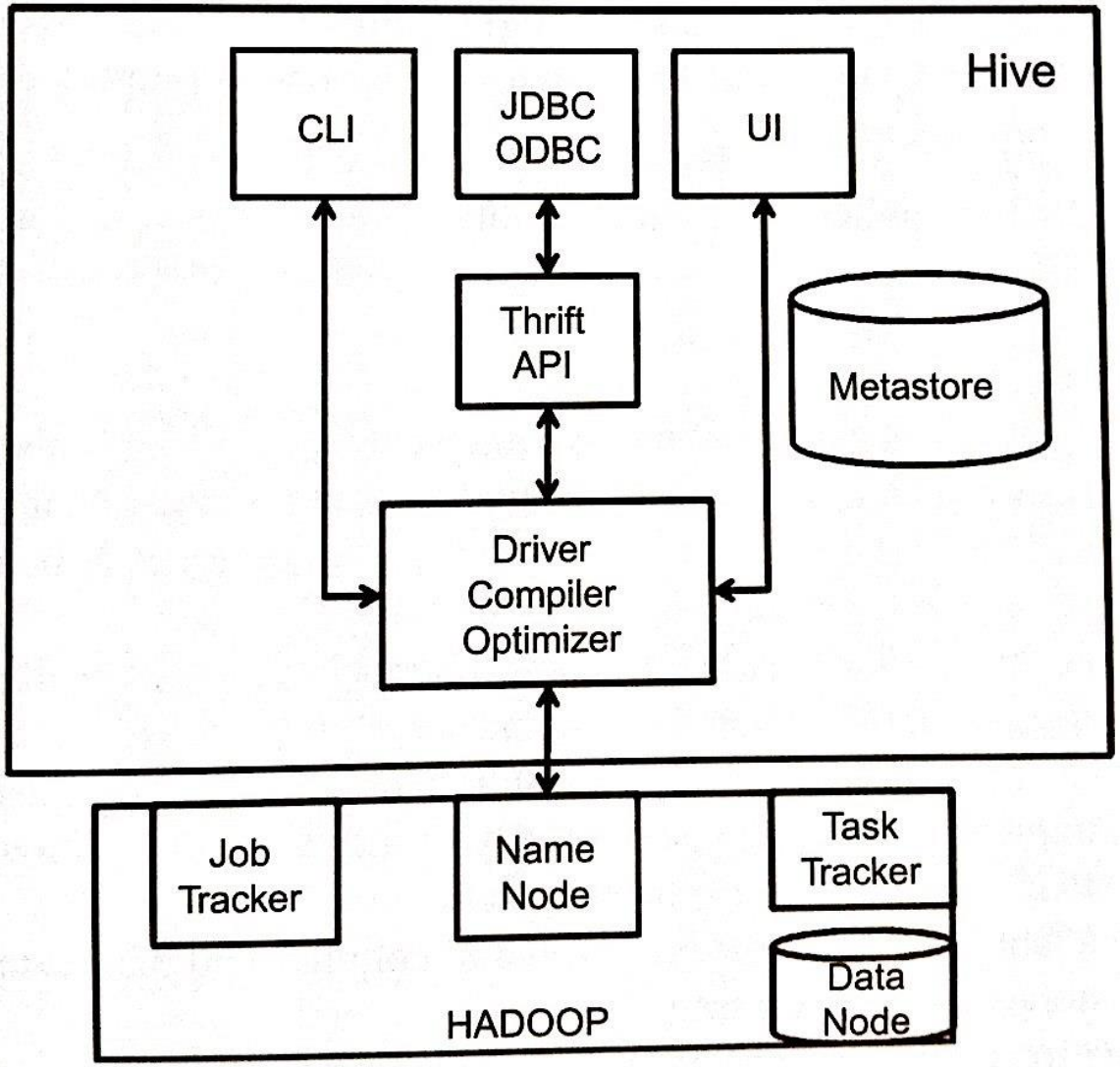
(6) HBase 的儲存方式

- 新增為主, 支援多版本, 可自定schema, 有層狀結構

Row key	TS	Column "recipe:"	
"www.foodie.com"	t10	"recipe: foodie.com"	"FOODIE"
"www.foodtv.com"	t9	"recipe: foodtv.com"	"FOODTV.COM"
	t8	"recipe: foodtv.com/ spicy/curry"	"FOODTV.COM"

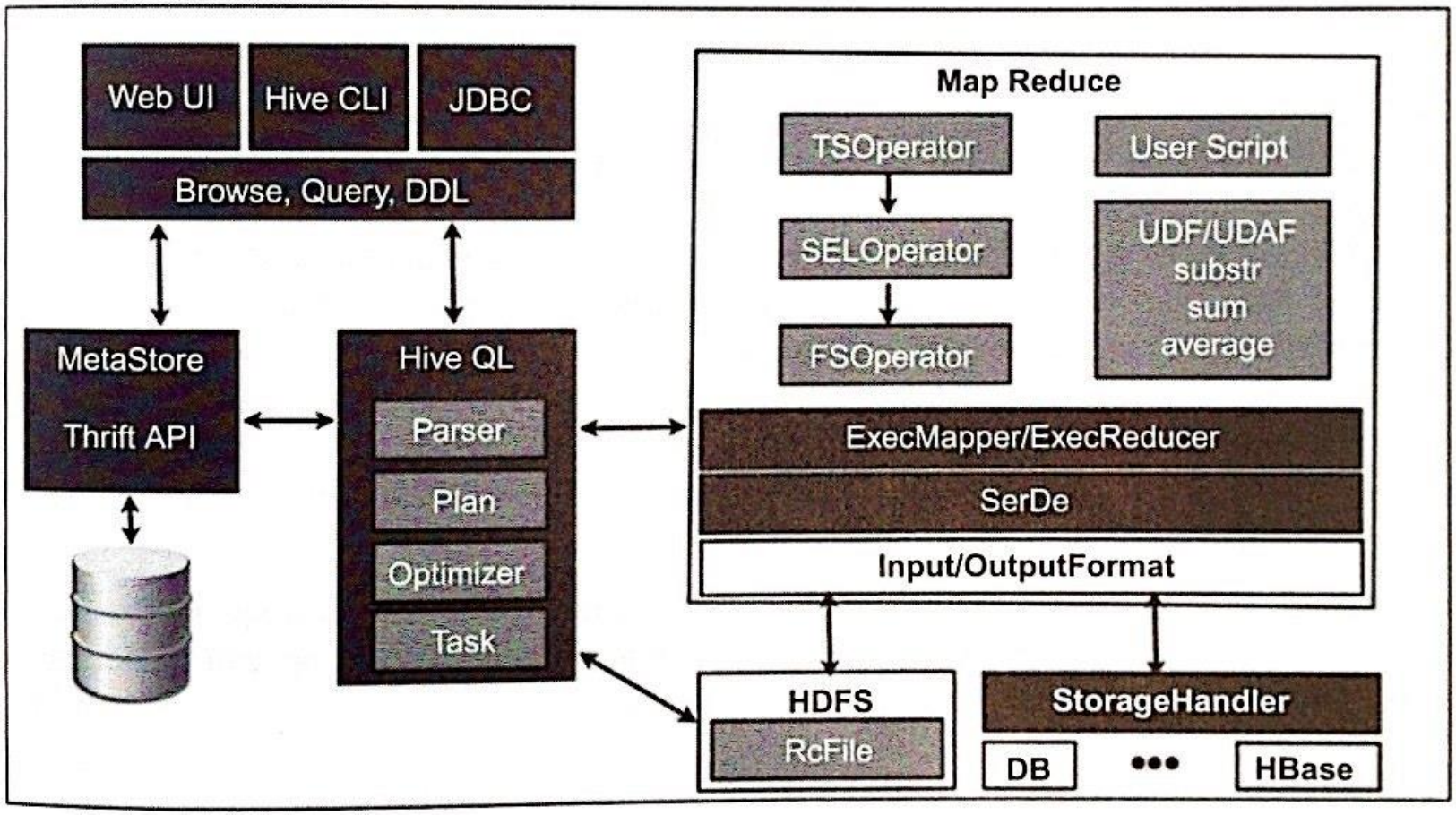
(7) Hive的架構

- 以Hadoop為基礎的資料倉儲方案
  - extend SQL interface



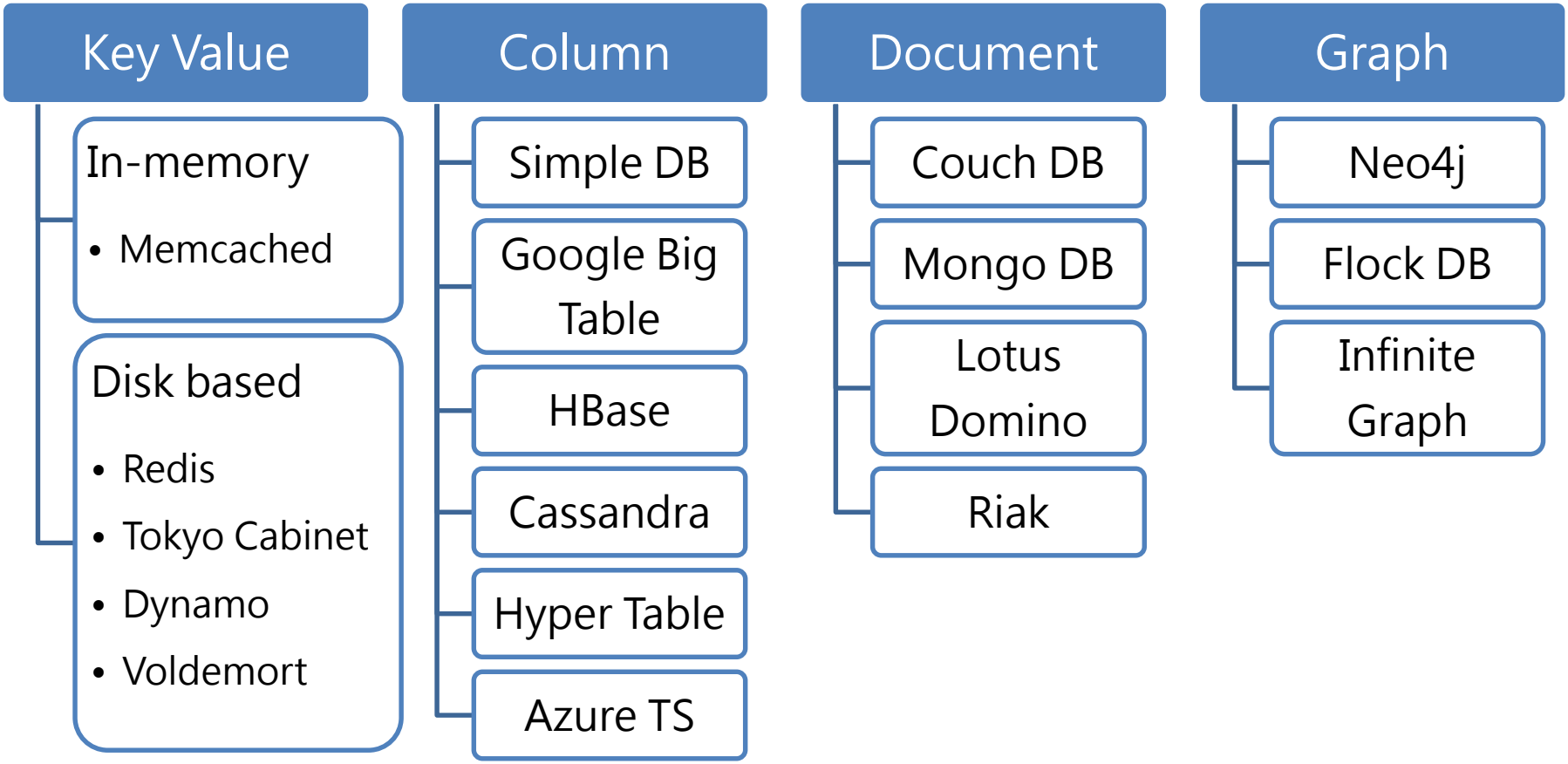
(7) Hive 的處理流程

- 可在整合入企業商業智慧 (business intelligence) 架構



相關技術架構 – 以NoSQL為例

- Not only SQL 關聯式資料庫以外的選擇



問題討論