巨量資料分析與應用 (4)

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

應用案例與研討 – 大數據資料倉儲

資料倉儲 Data warehouse

- A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions

企業為何需要 Data warehouse

- Increasing customer focus, which includes the analysis of customer buying patterns.
- Repositioning products and managing product portfolios by comparing the performance of sales by time or regions, to fine-tune production strategies
- Analyzing operations and looking for sources of profit
- Managing the customer relationship
- Managing the cost of corporate assets

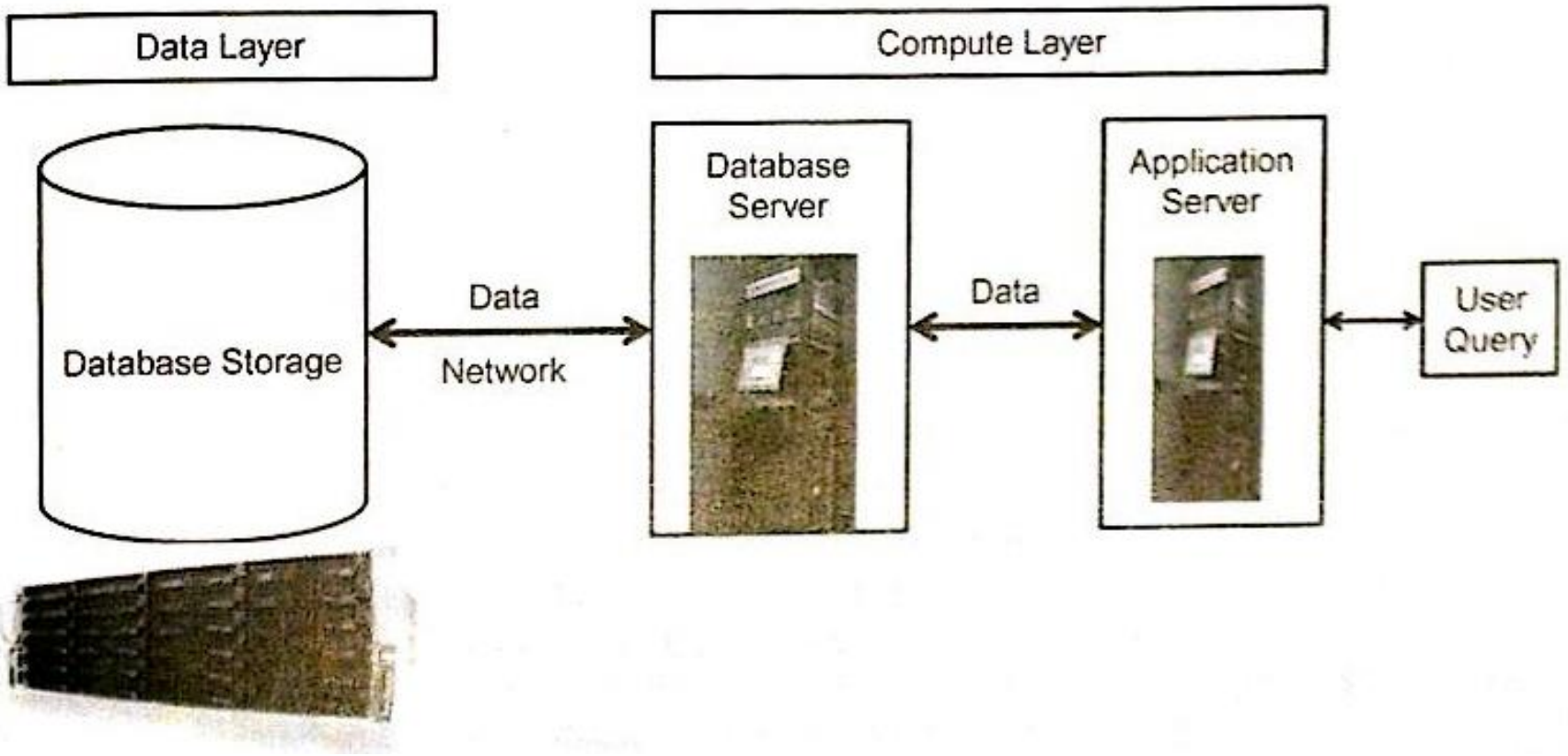
Data warehouse 特性

- for analytical tasks, using data from multiple applications 分析用途為主
- relatively small number of users with relatively long interactions 使用者人數較少，但互動程度高
- read-intensive 讀取為主
- periodically updated 週期性更新資料
- contains current and historical data 新舊資料並存
- Each query frequently results in a large result set and full table scan and multi-table joins 運算量大

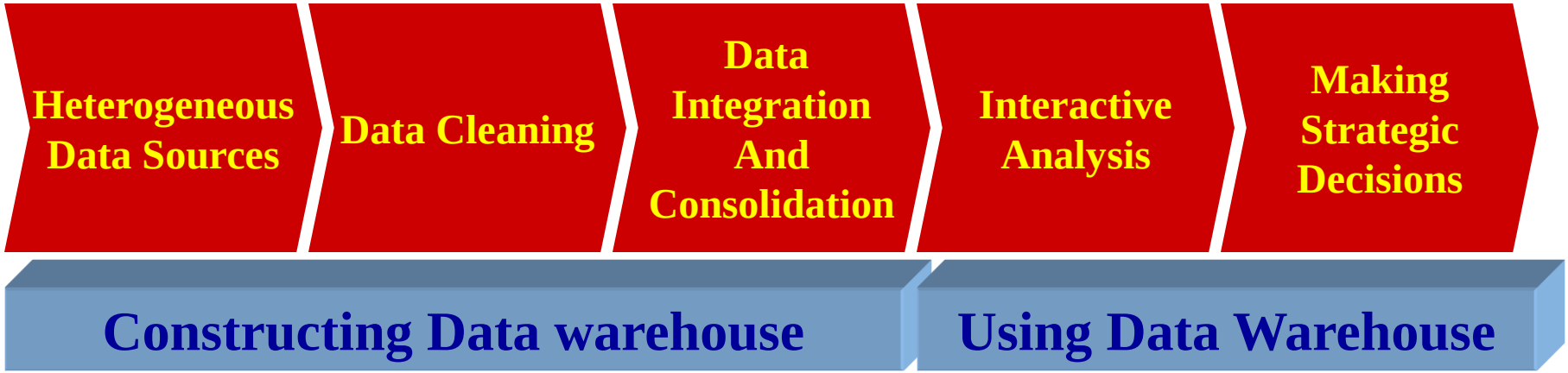
- Why not use Database directly ?
  - update-driven approach is inefficient.
  - Potentially expensive for frequent queries.
- Use Data warehouse instead
  - query-driven approach is enough for making strategic decisions.
  - Separate the operational DBMS for daily and critical operations.

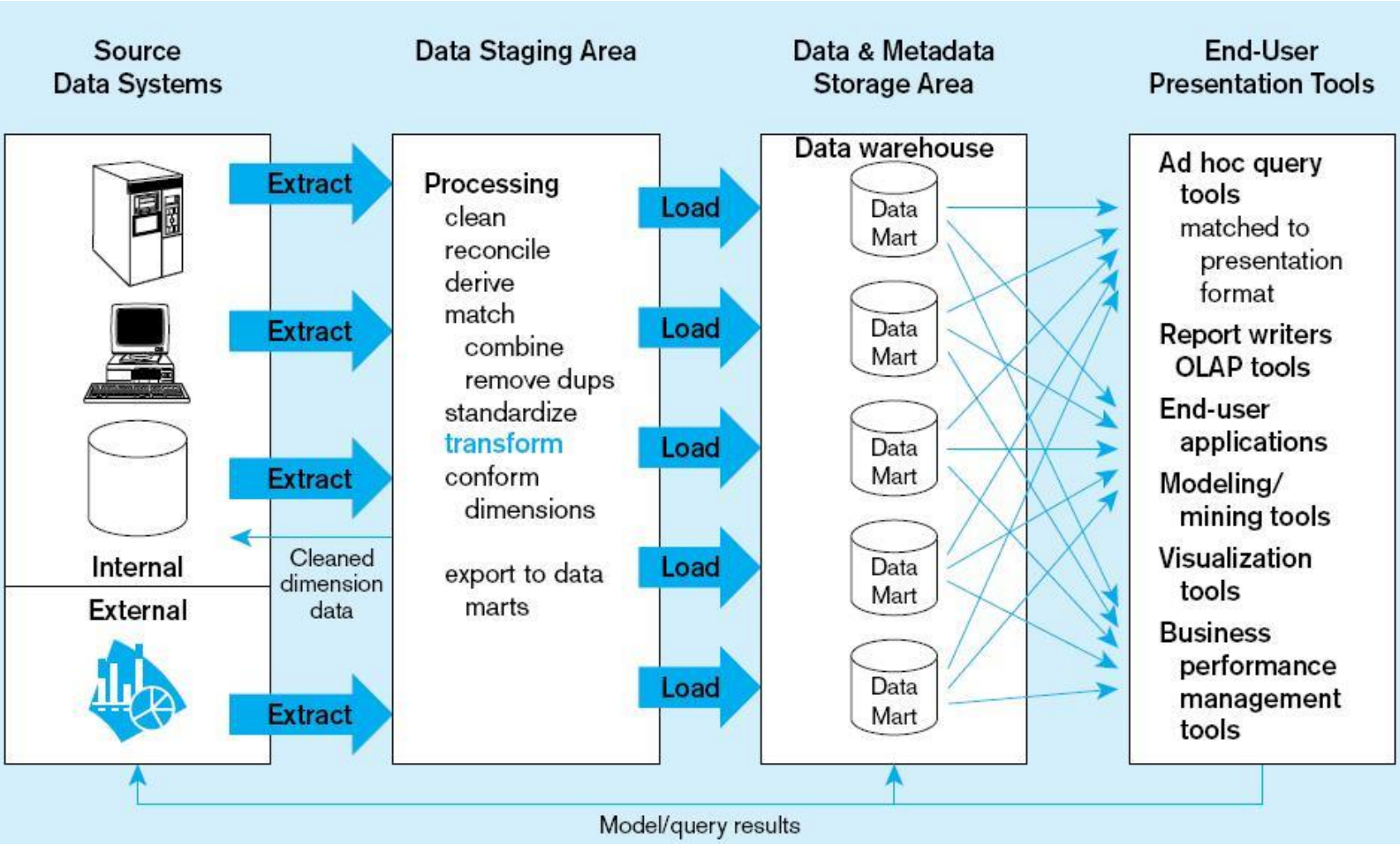
Data warehouse 處理流程

- data warehouse processing (original)



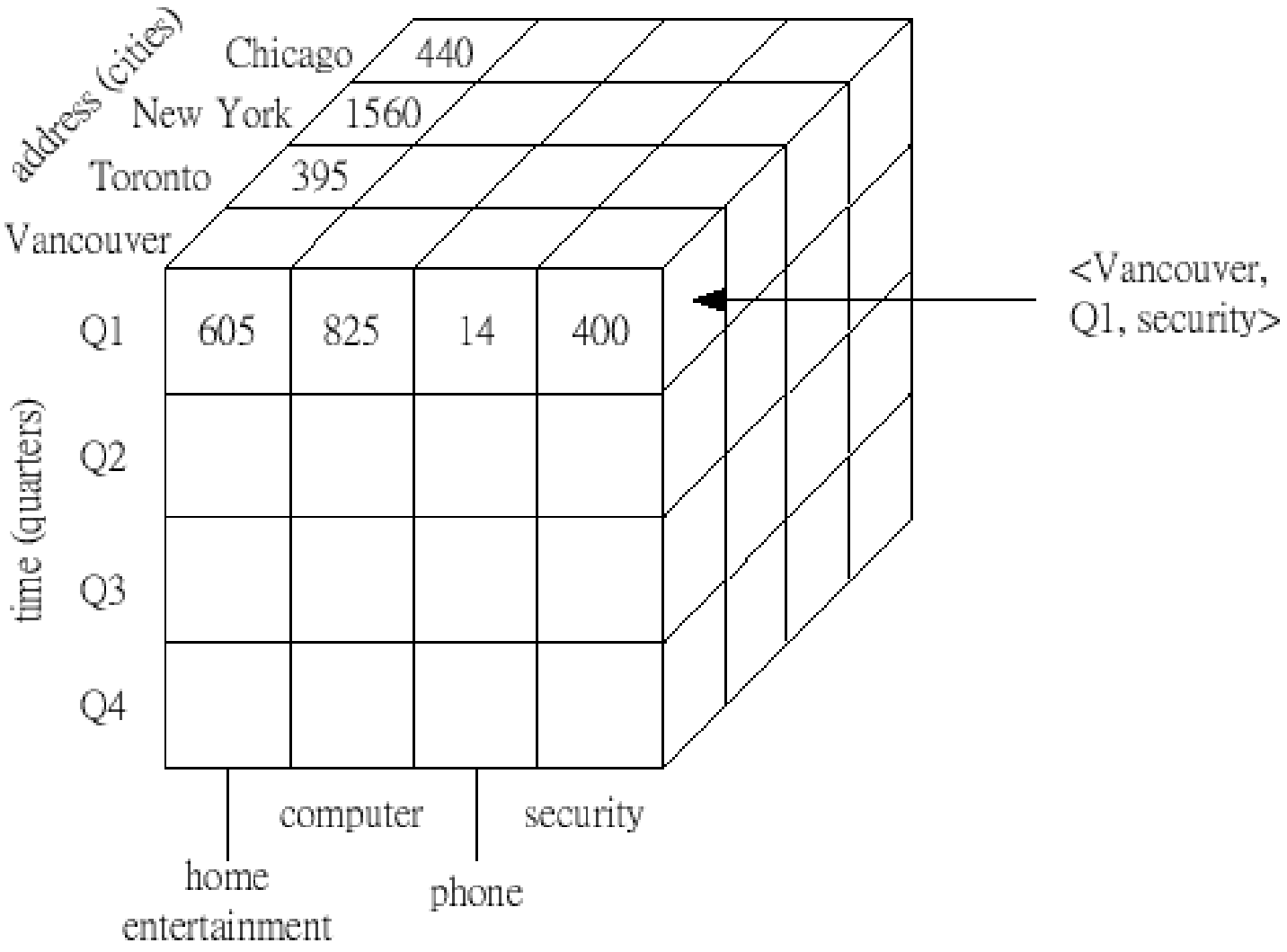
- data warehouse processing (original)



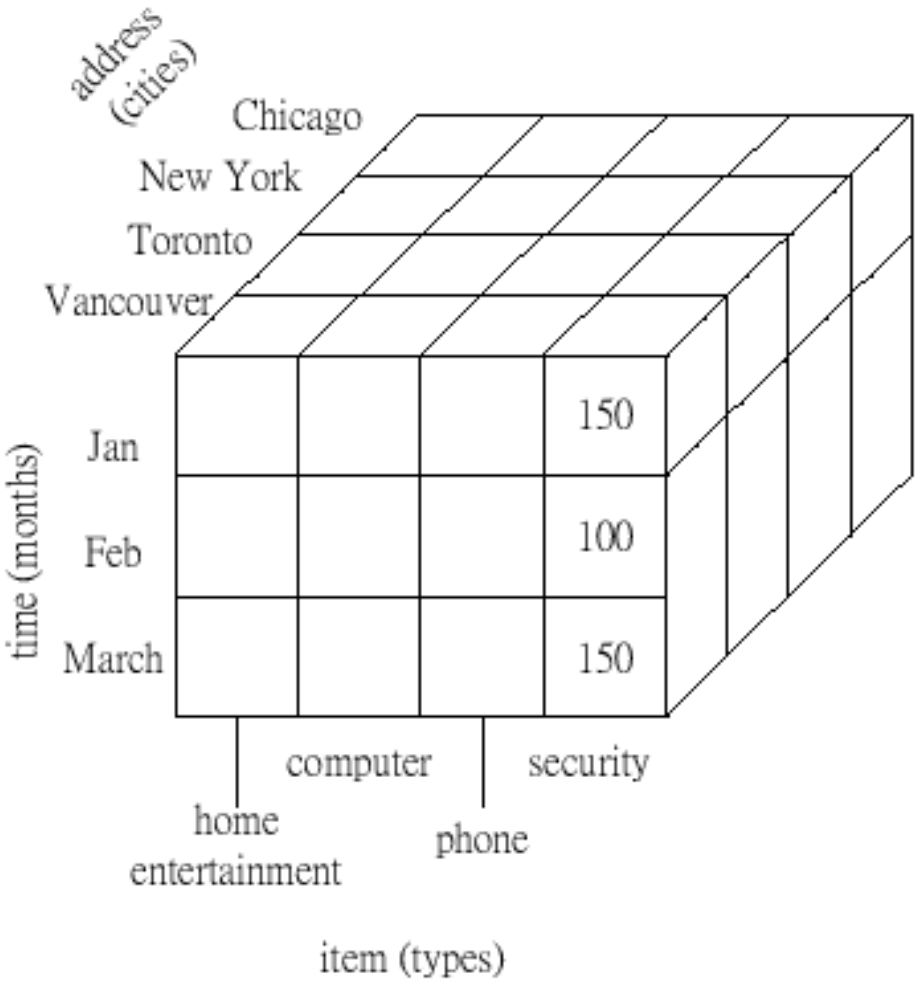


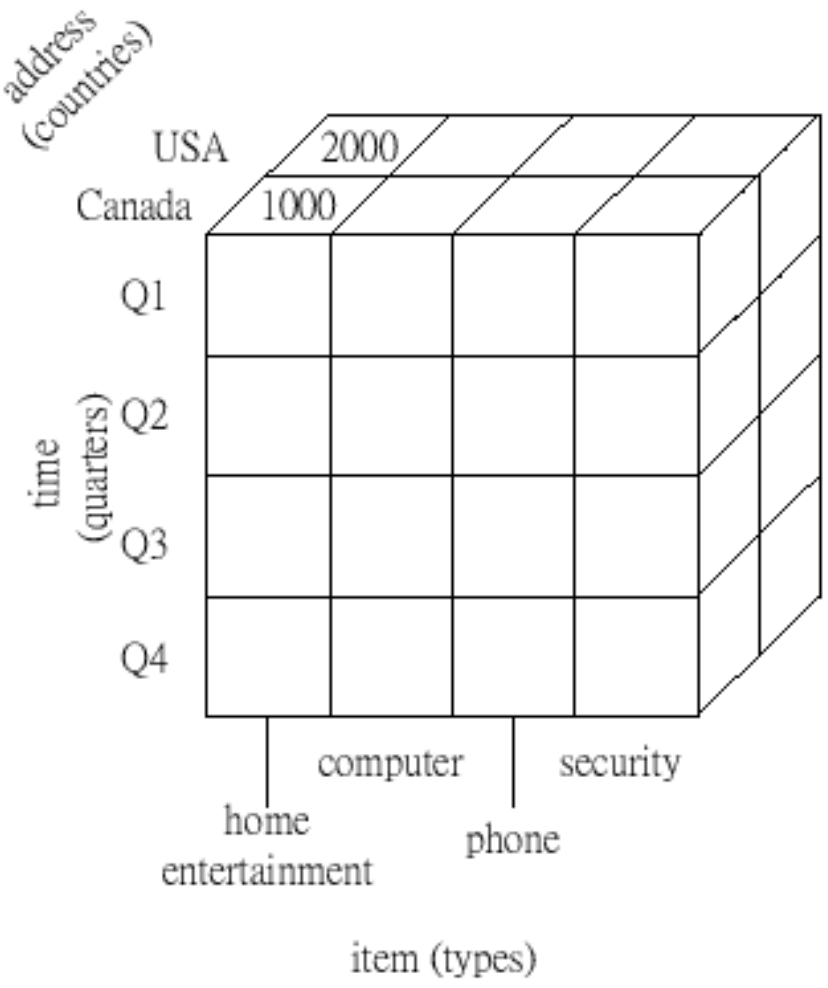
Data warehouse 原理 – Data cube

- A multidimensional, logical view of the data
- Concept hierarchy
  - Multiple data granularity 多重的資料顆粒度
  - Data summarization 資料加總
  - Data generalization 資料一般化

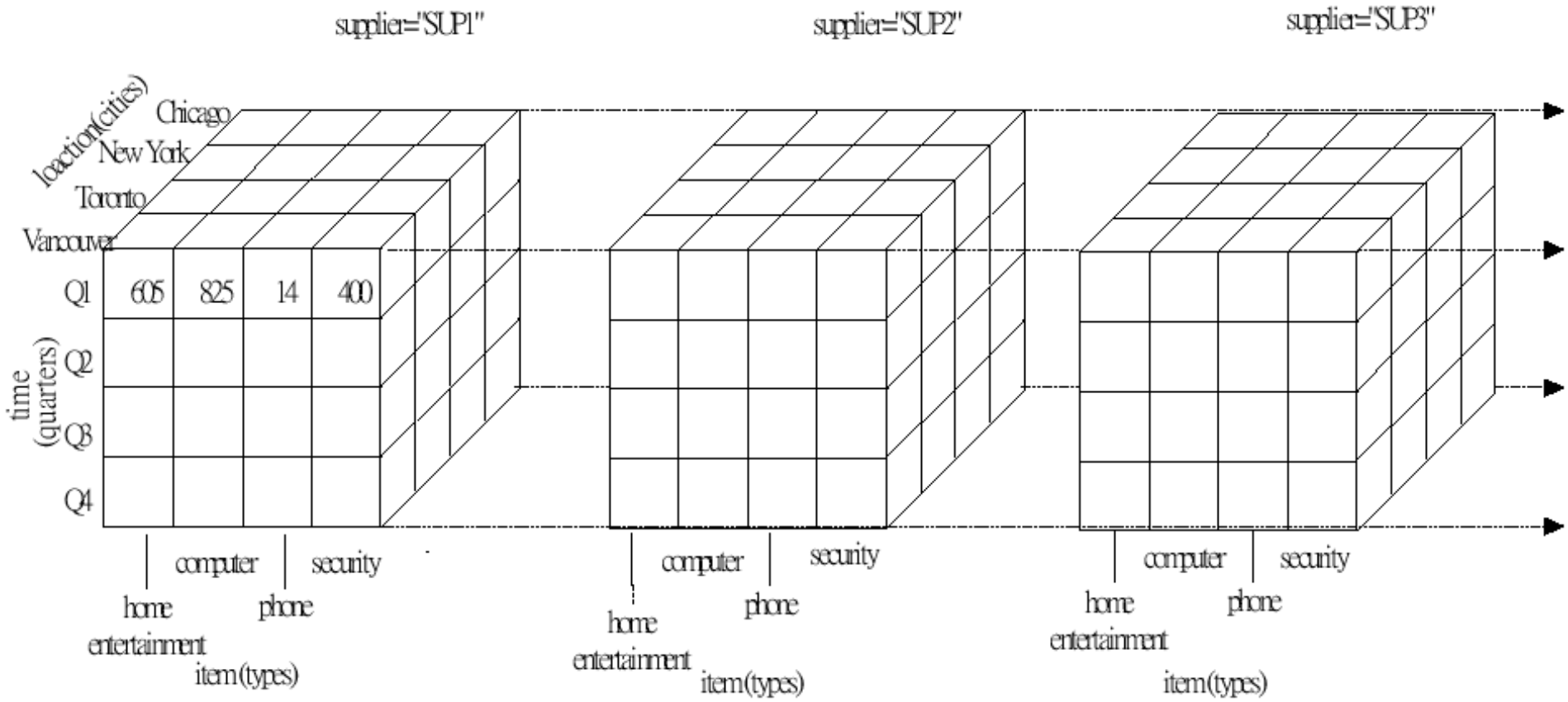


(a)

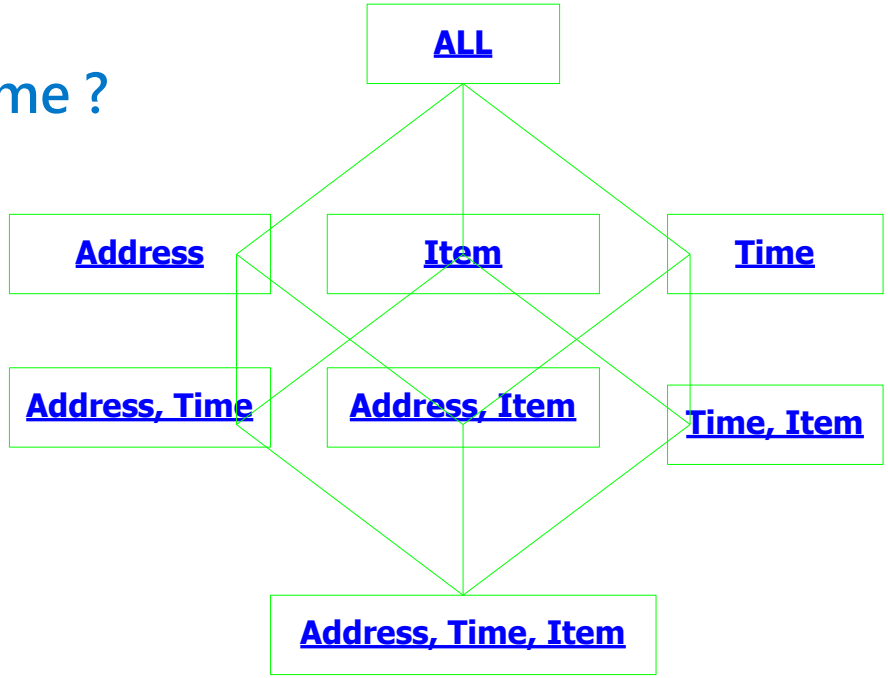




- Roll-up on address



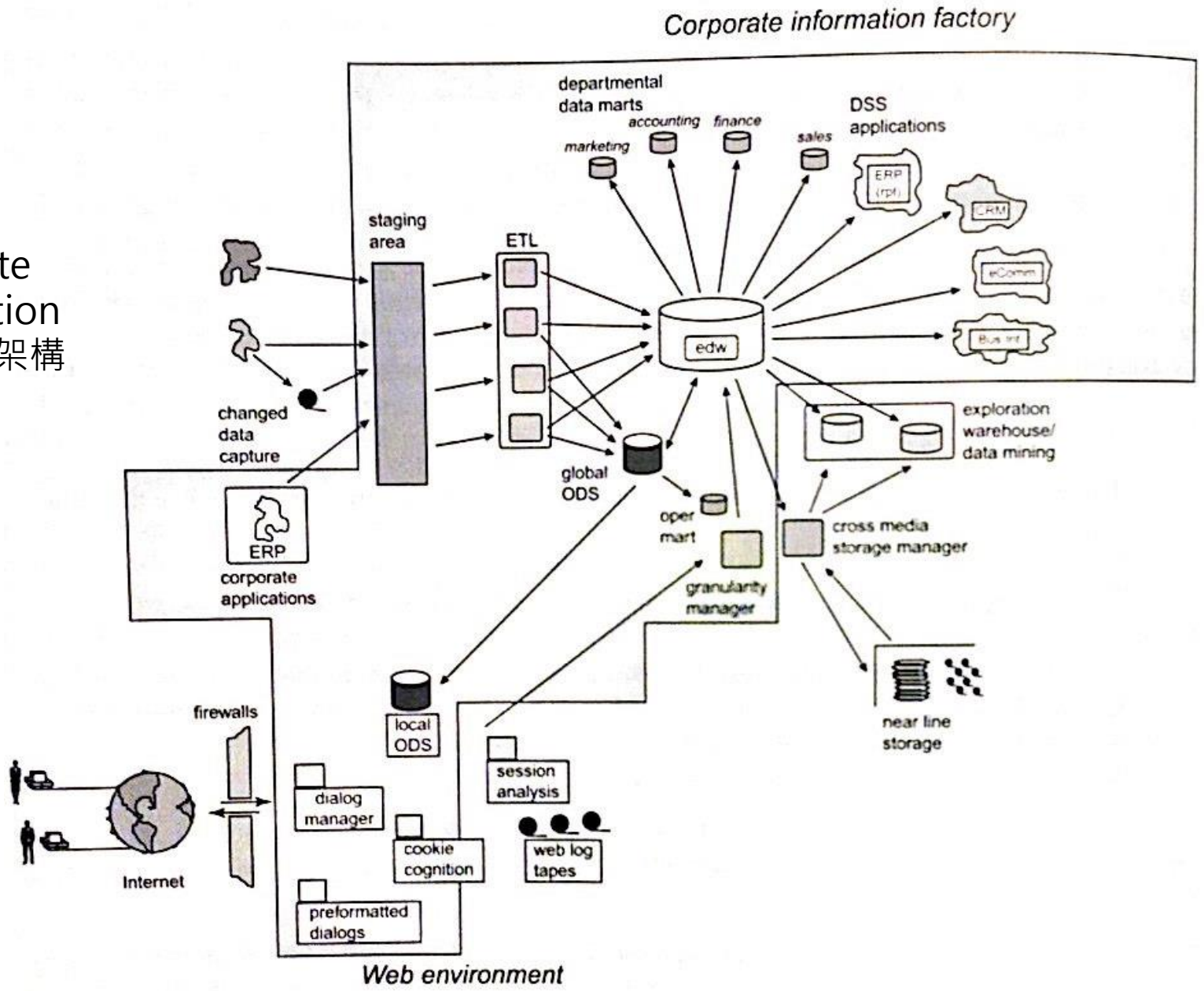
- The challenges
  - There are 2^N combination for data cube computation.
  - Concept hierarchy and Aggregations makes it more complicated
- Efficient data cube computation
  - Materialize every, none, or some ?
  - Algorithms for selection
    - Based on size
    - Based on sharing,
    - Based on access frequency.



On-Line Analytical Processing (OLAP)

- Fast on-line processing of data cubes or multi-dimensional databases
  - Drill down / Roll up、Pivoting 樞紐分析、Slicing / Dicing, etc.
- Provide interactive analysis and quick response to OLAP queries.
  - Summarization and aggregations at every dimension
  - Retrieval and display of data in 2-D or 3-D cross-tabs, charts, and graphs, with easy pivoting of the axes.
  - Analytical modeling : deriving ratios, variance, etc.
  - Use for forecasting, trend, statistical analysis.

Corporate Information Factory 架構

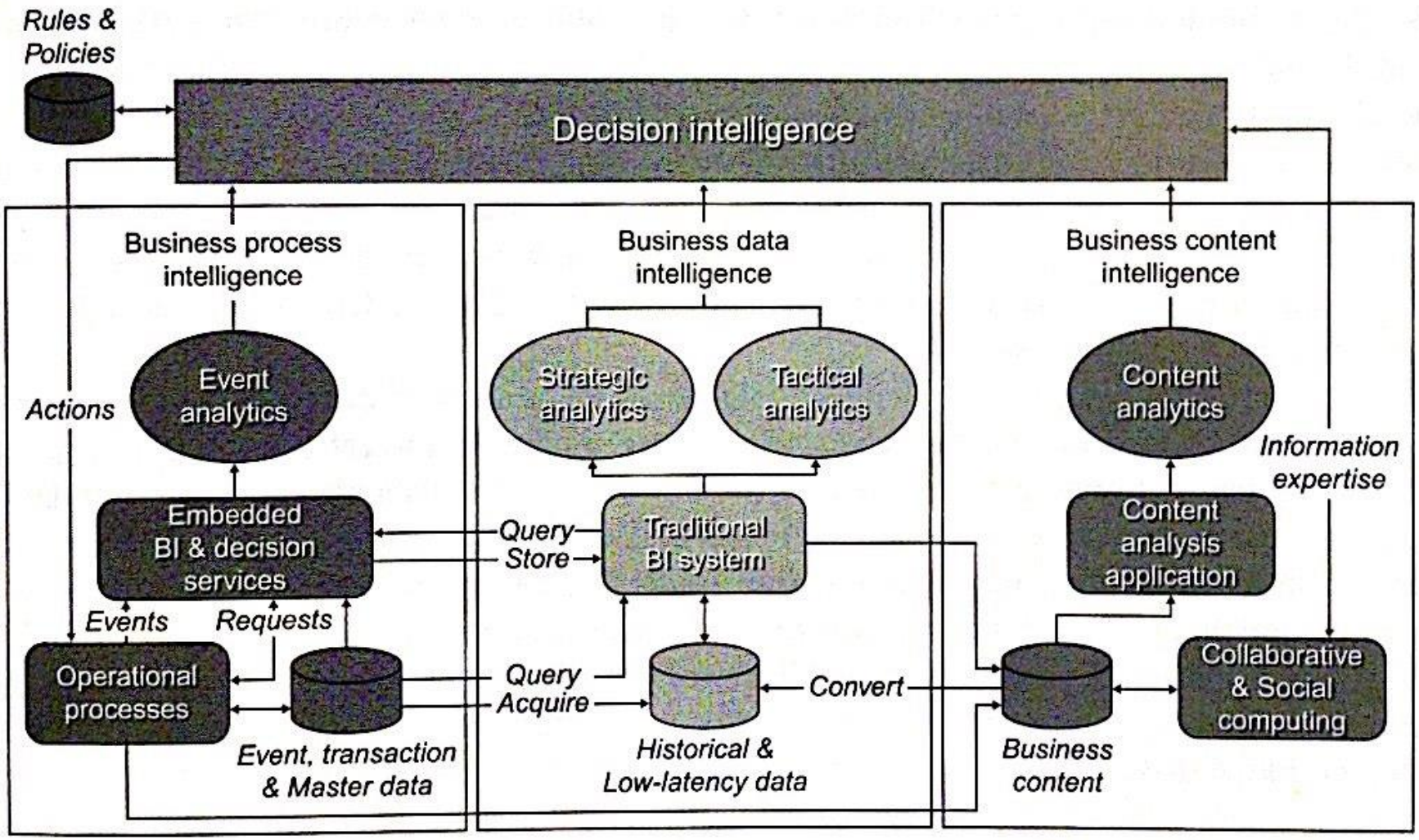


Data warehouse 2.0

- The BI matrix

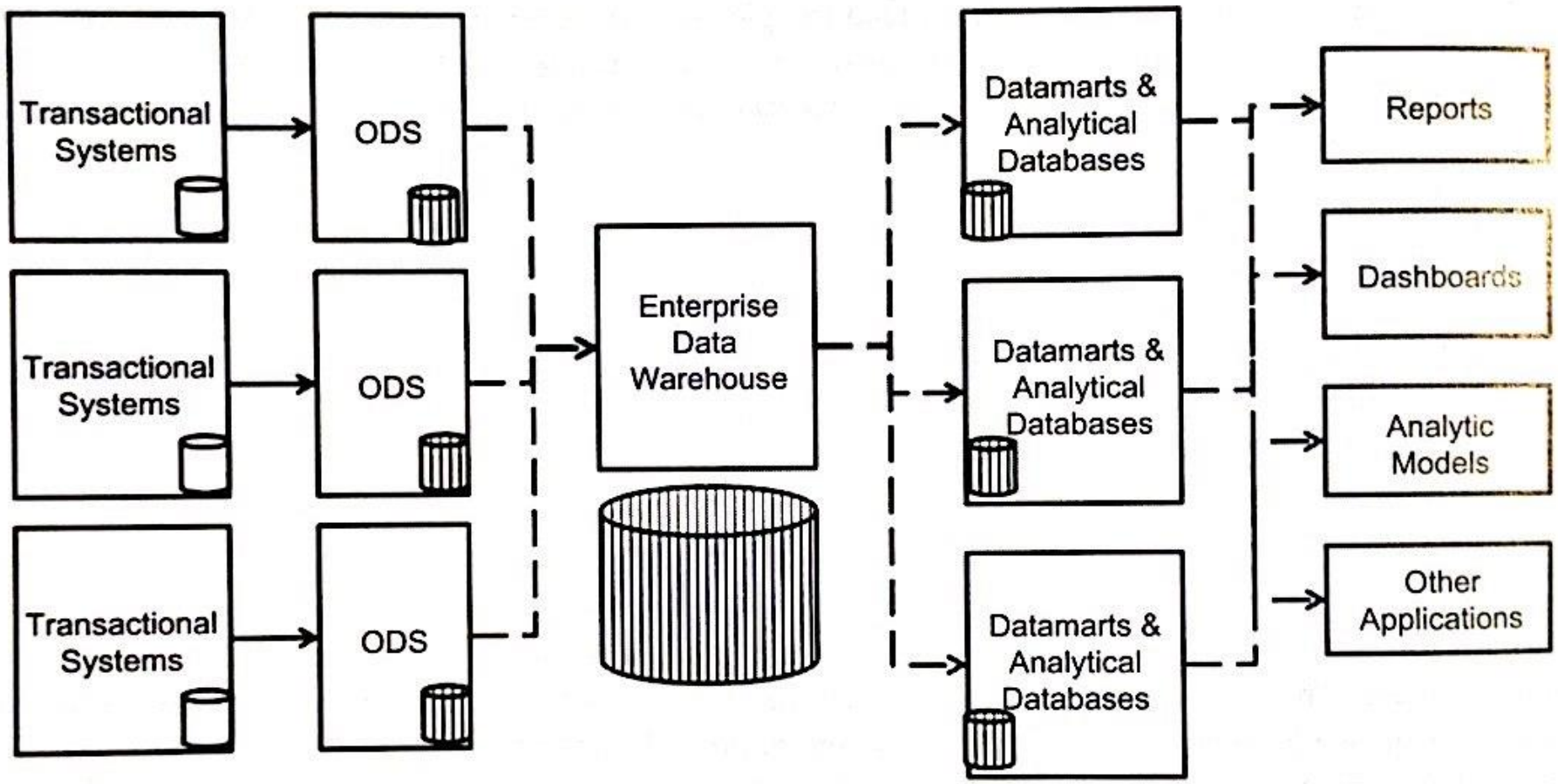
	Strategic BI	Tactical BI	Operational BI
Business focus	Achieve long-term business goals	Manage tactical initiatives to achieve strategic goals	Monitor & optimize operational business processes
Primary users	Executives & business analysts	Business analysts, & LOB managers	LOB managers, operational users & operational processes
Time-frame	Months to years	Days to weeks to months	Intra-day to daily
Data	Historical data	Historical data	Real-time, low-latency & historical data
Mode of operation	User driven Data centric	User driven Data centric	Event driven Process centric

New architecture



目前企業常見的資料倉儲架構

data distribution in a data warehouse today



整合大數據與資料倉儲

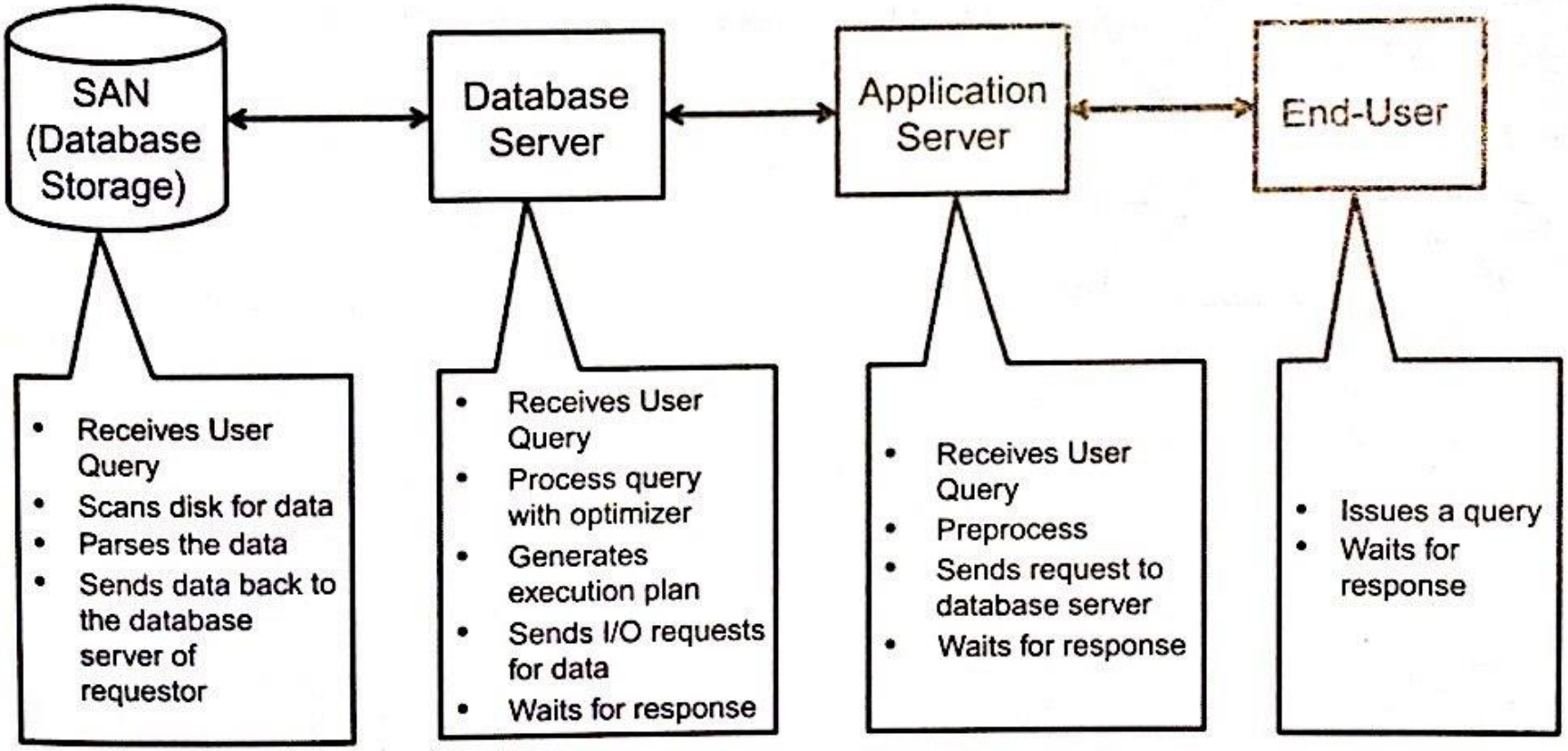
設計前題

- Identify the business objective
- Understand the workload
  - 每次查詢會需要處理的資料量, 以及回傳結果的資料量

Wide Wide	Narrow Wide
Wide Narrow	Narrow Narrow

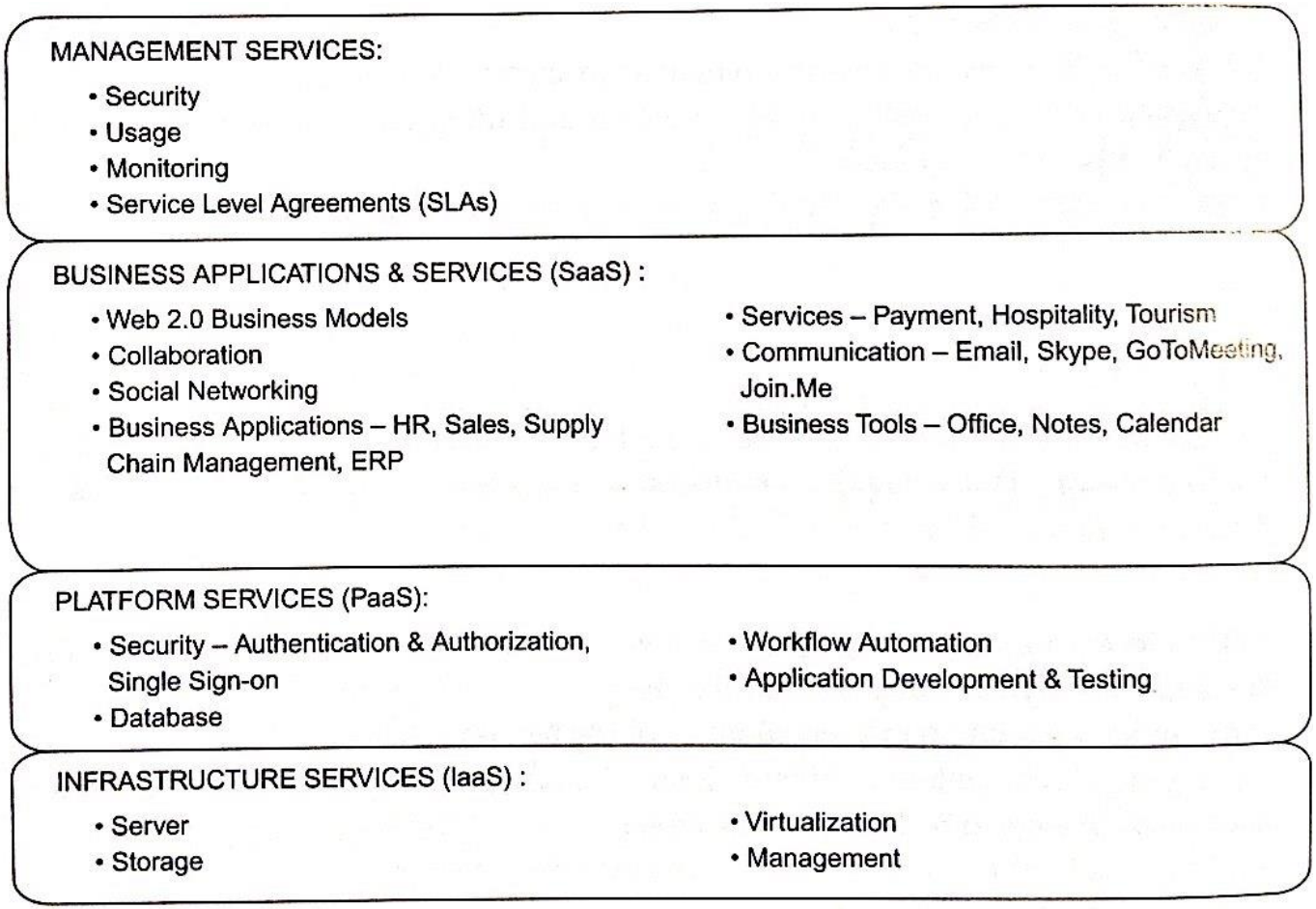
- Understand the data
  - Unstructured
  - Semi-structured

Understand the workload : 傳統的系統架構



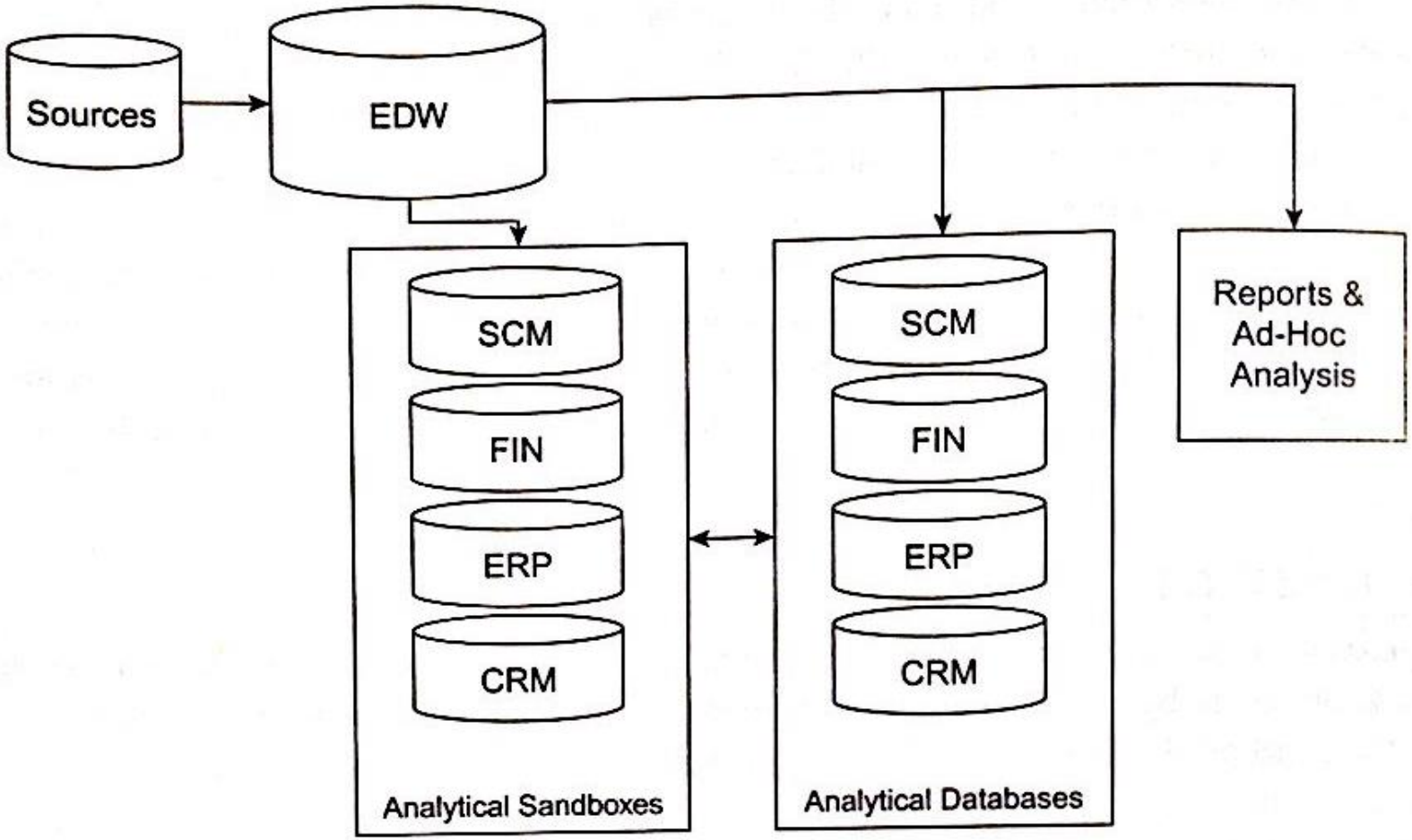
可使用的新技術 (1) 雲端技術

- Cloud computing components



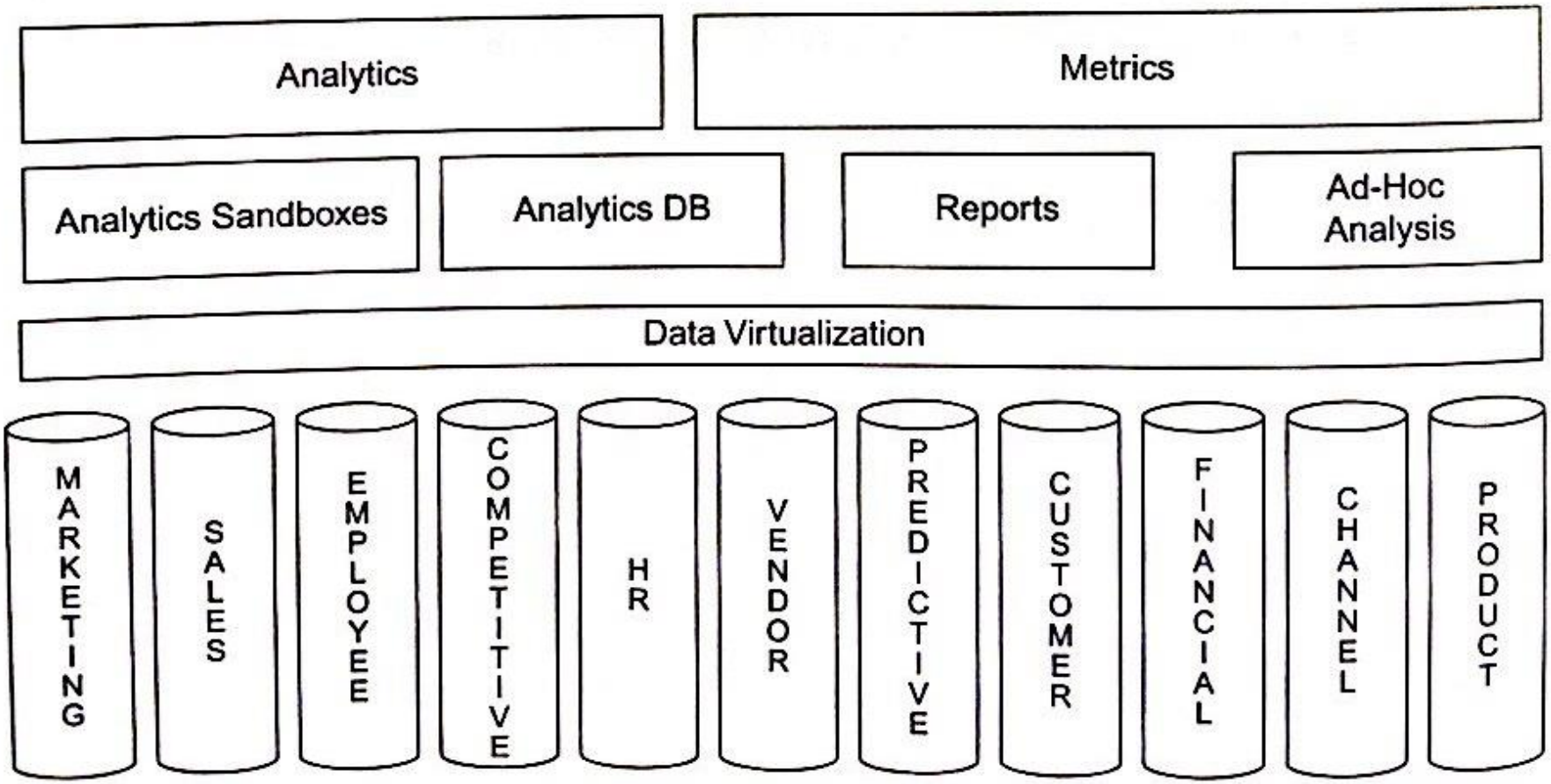
可使用的新技術 (2)

- Different state of analytics and reporting



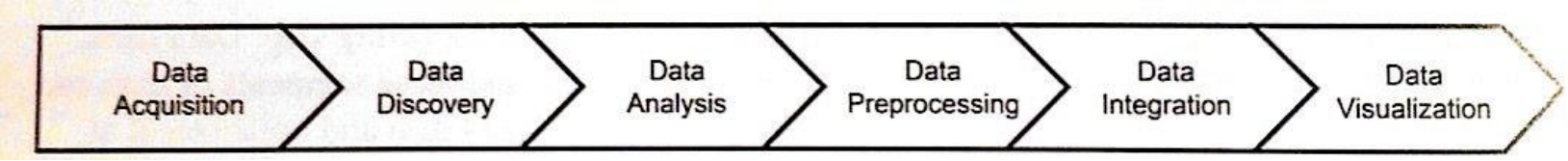
可使用的新技術 (3) 資料虛擬化

- Data virtualization-based architecture



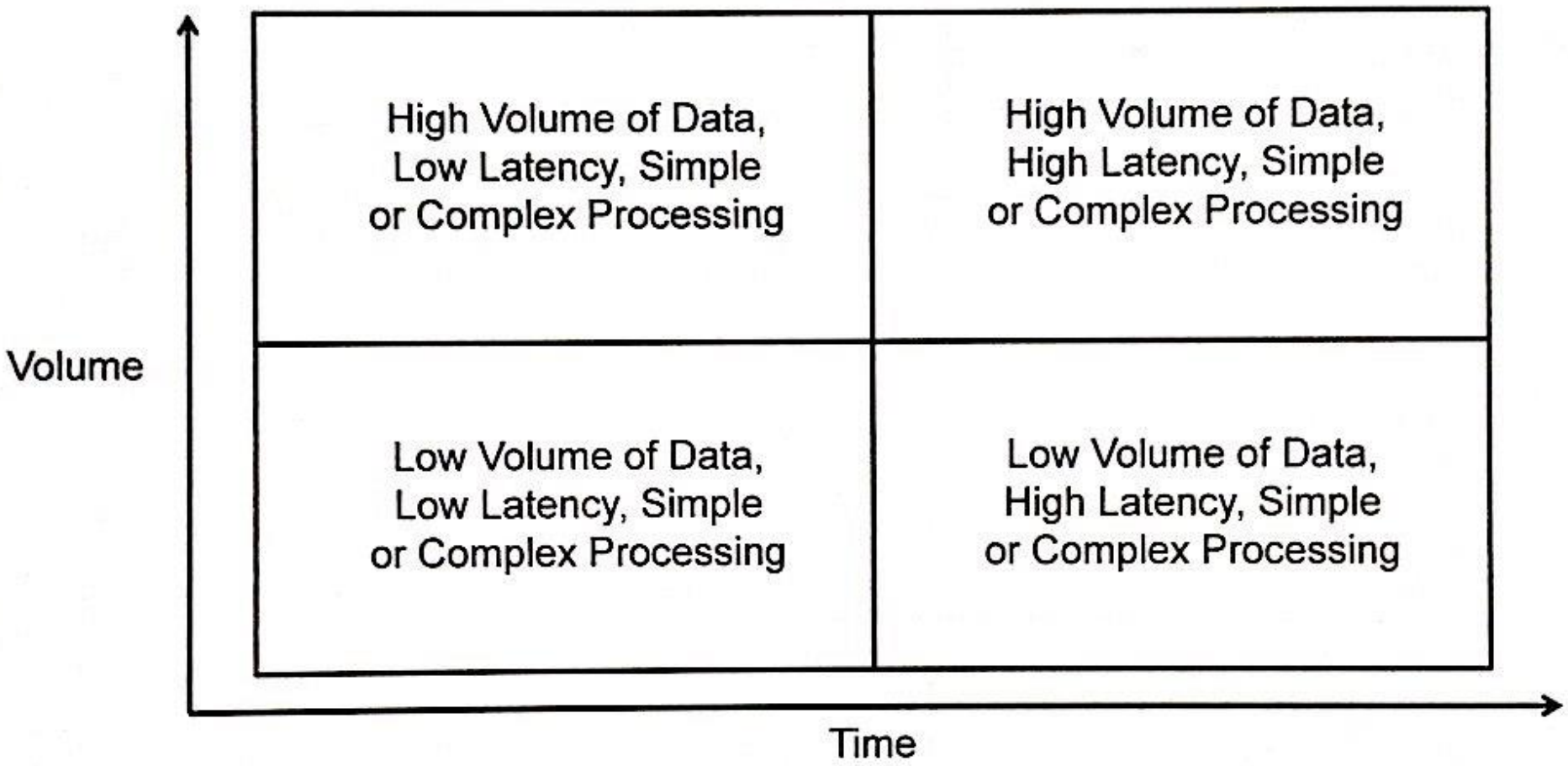
用Big data方式整合Data warehouse

- 依下列流程進行資料處理

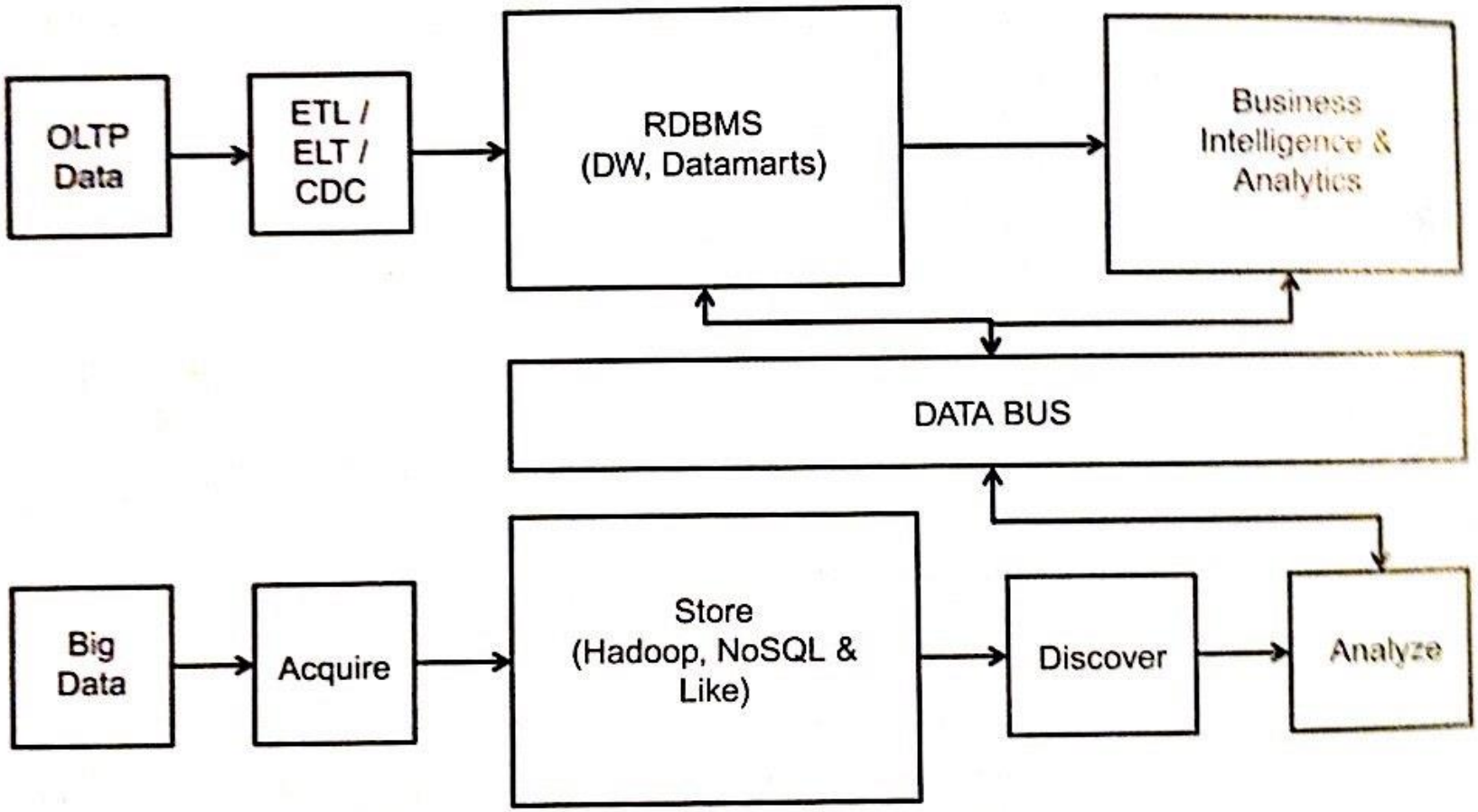


- 整合結構與非結構性資料
- 整合大數據架構與BI / Data warehouse架構

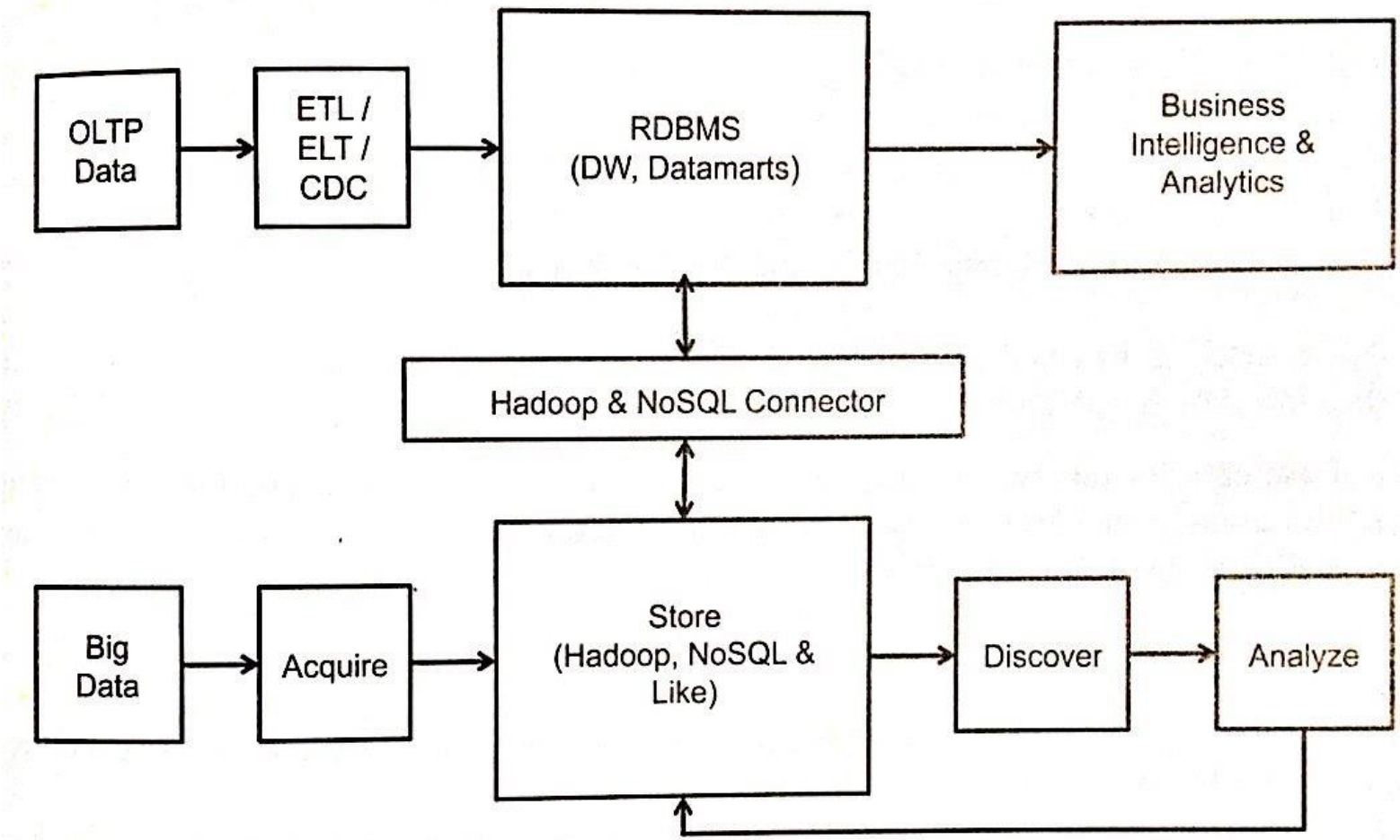
Workload category



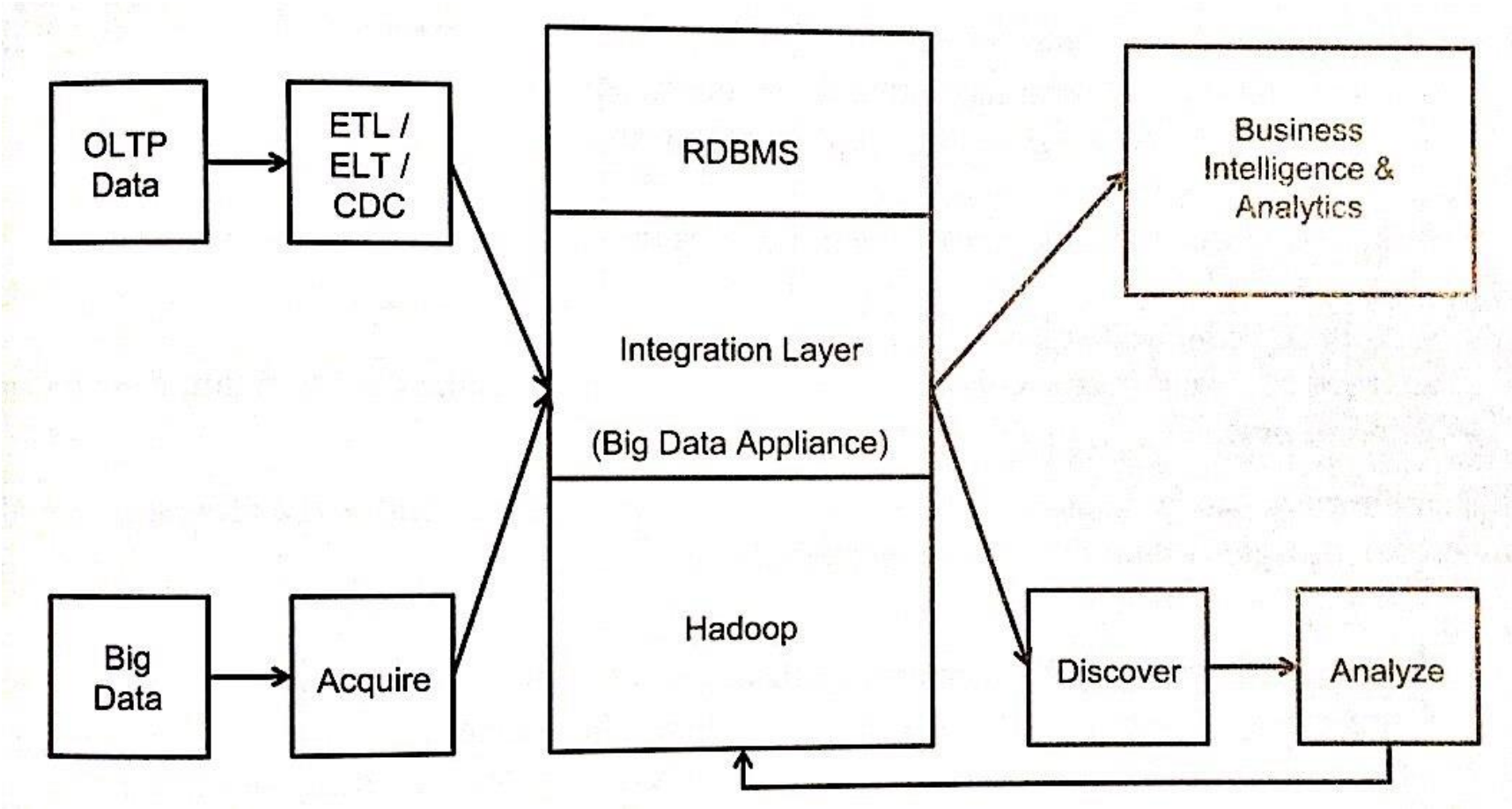
External data integration



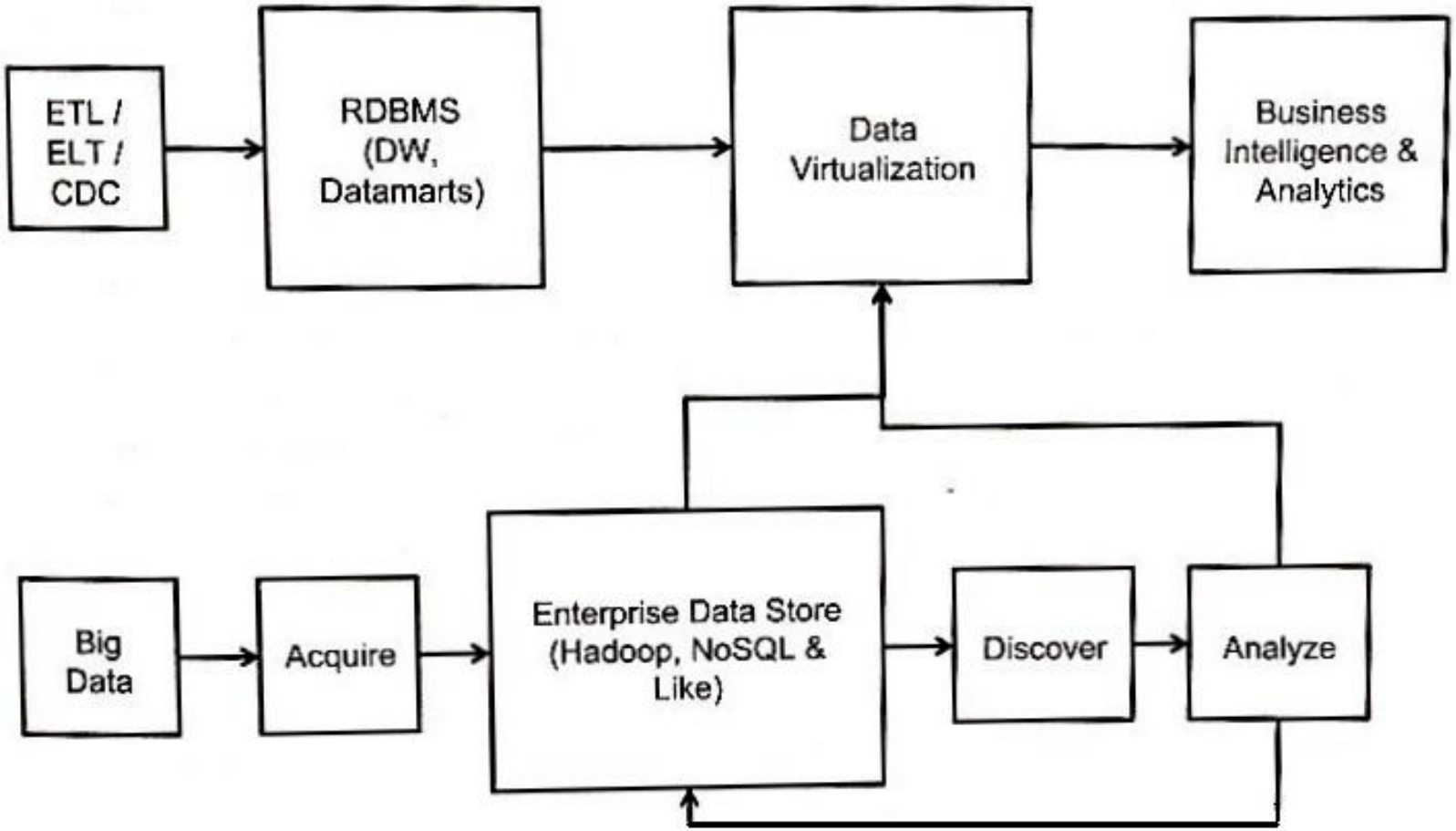
Integration-driven approach



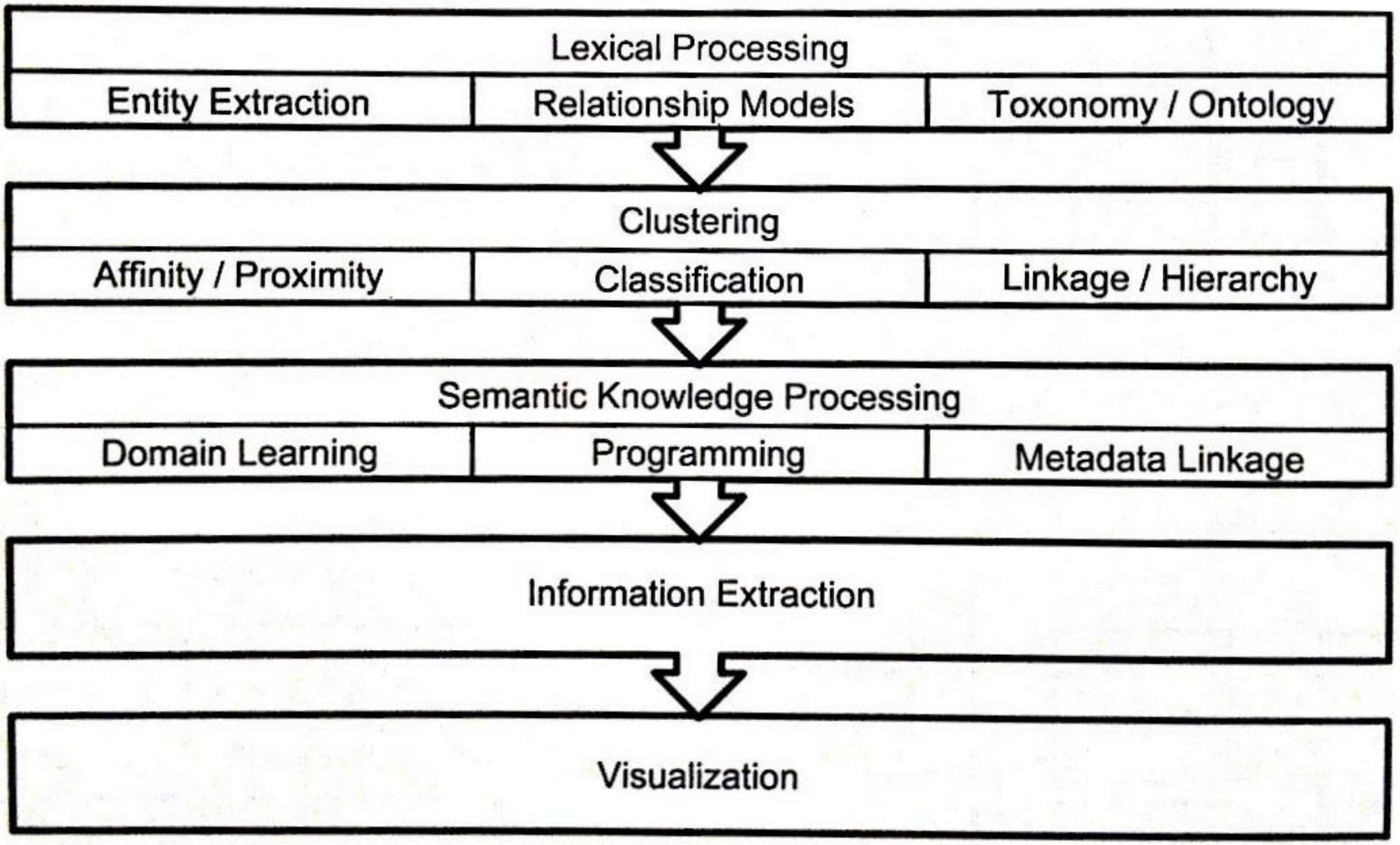
Conceptual big data appliance



Data virtualization-based big data integration

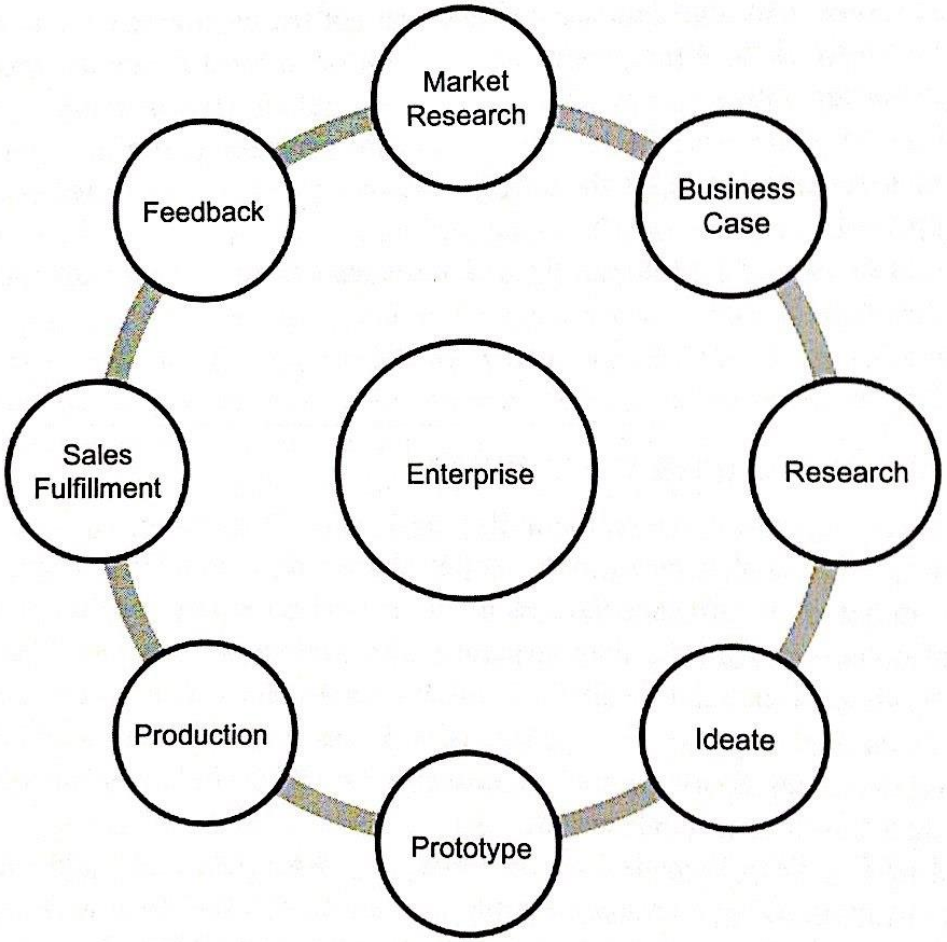


Data virtualization-based : semantic framework

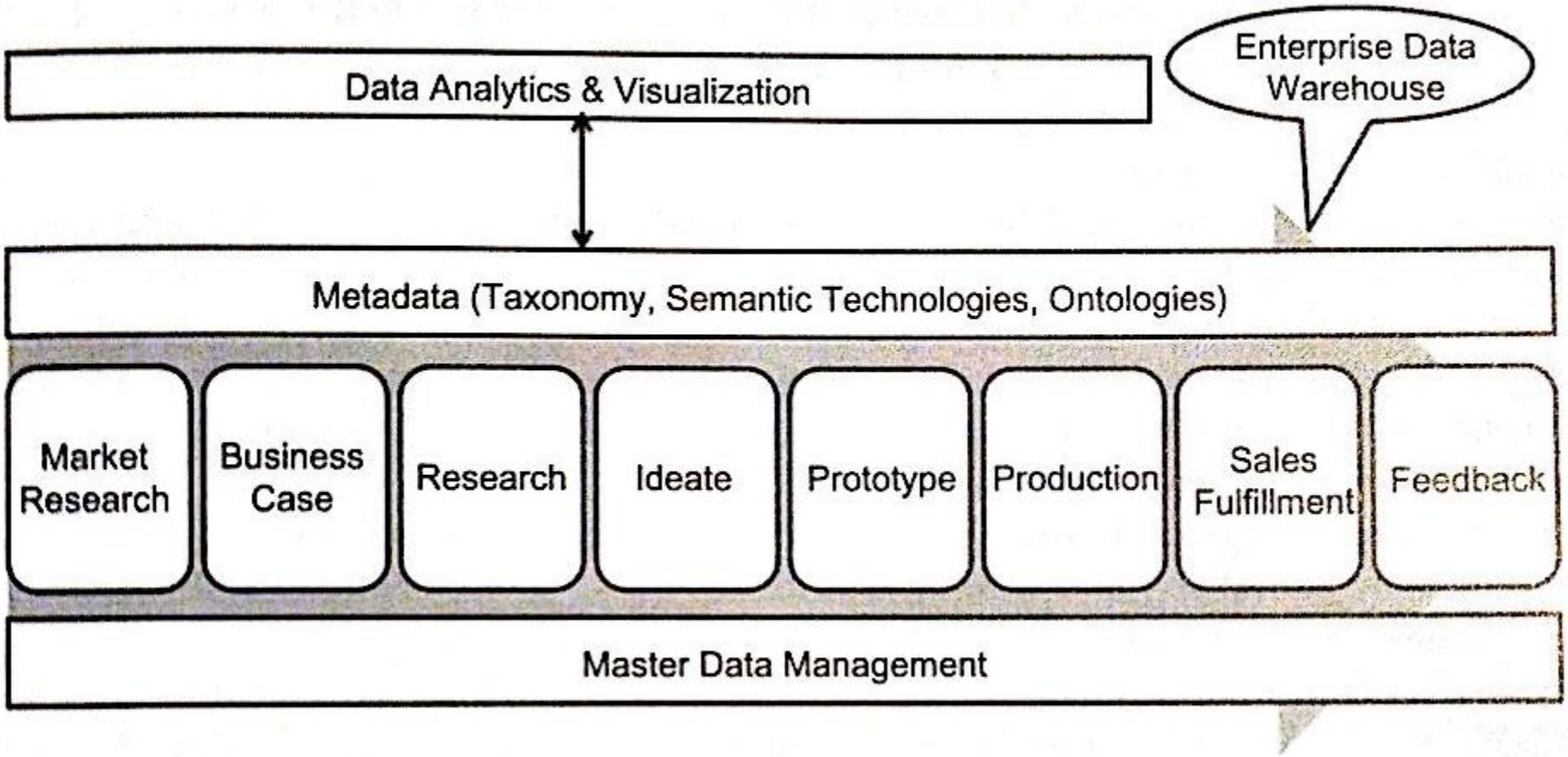


Data-driven architecture for big data

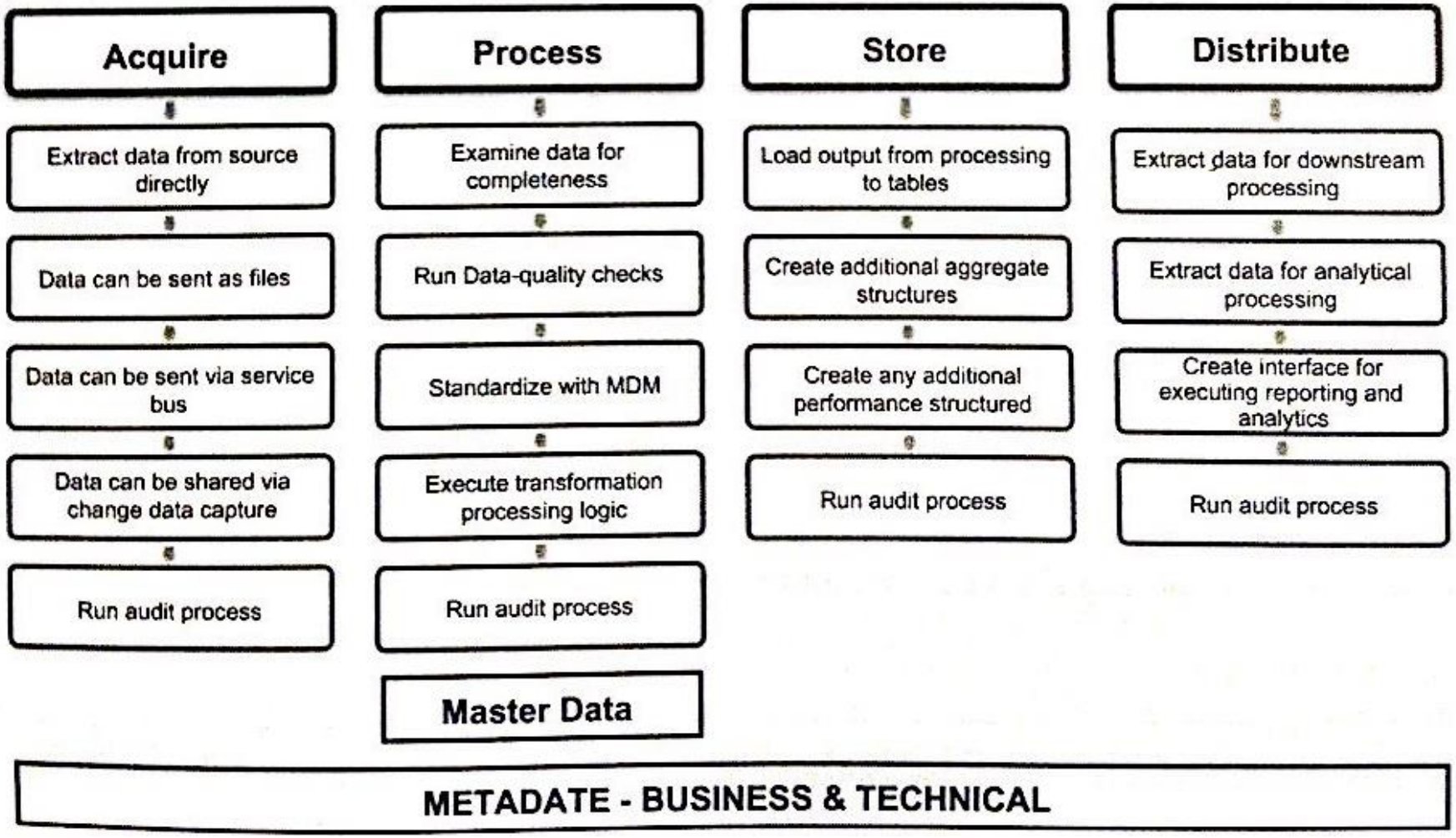
- Enterprise use of data before big data : 分散的子系統



- Enterprise data-driven architecture : 統一的使用方式

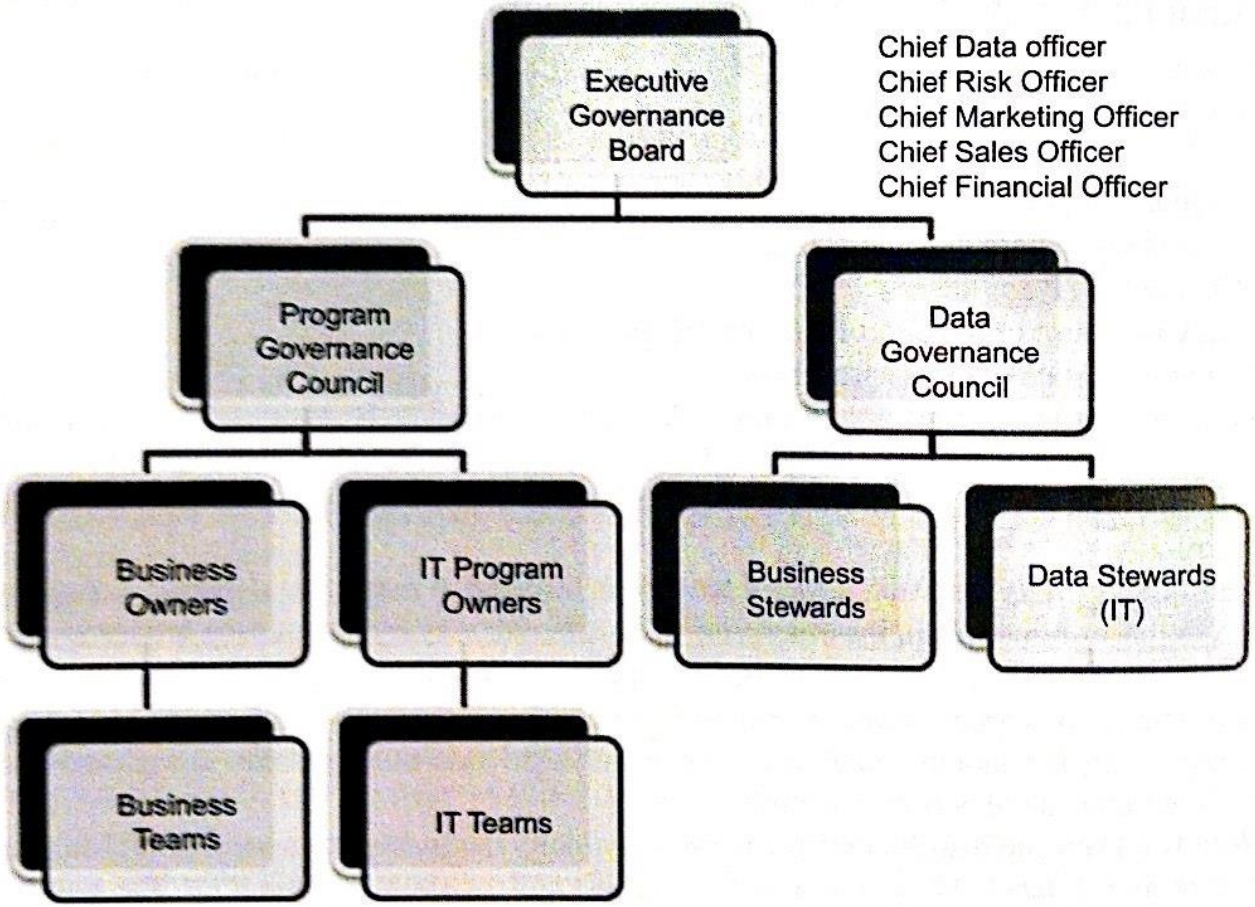


Data processing cycles with integration of MDM and metadata



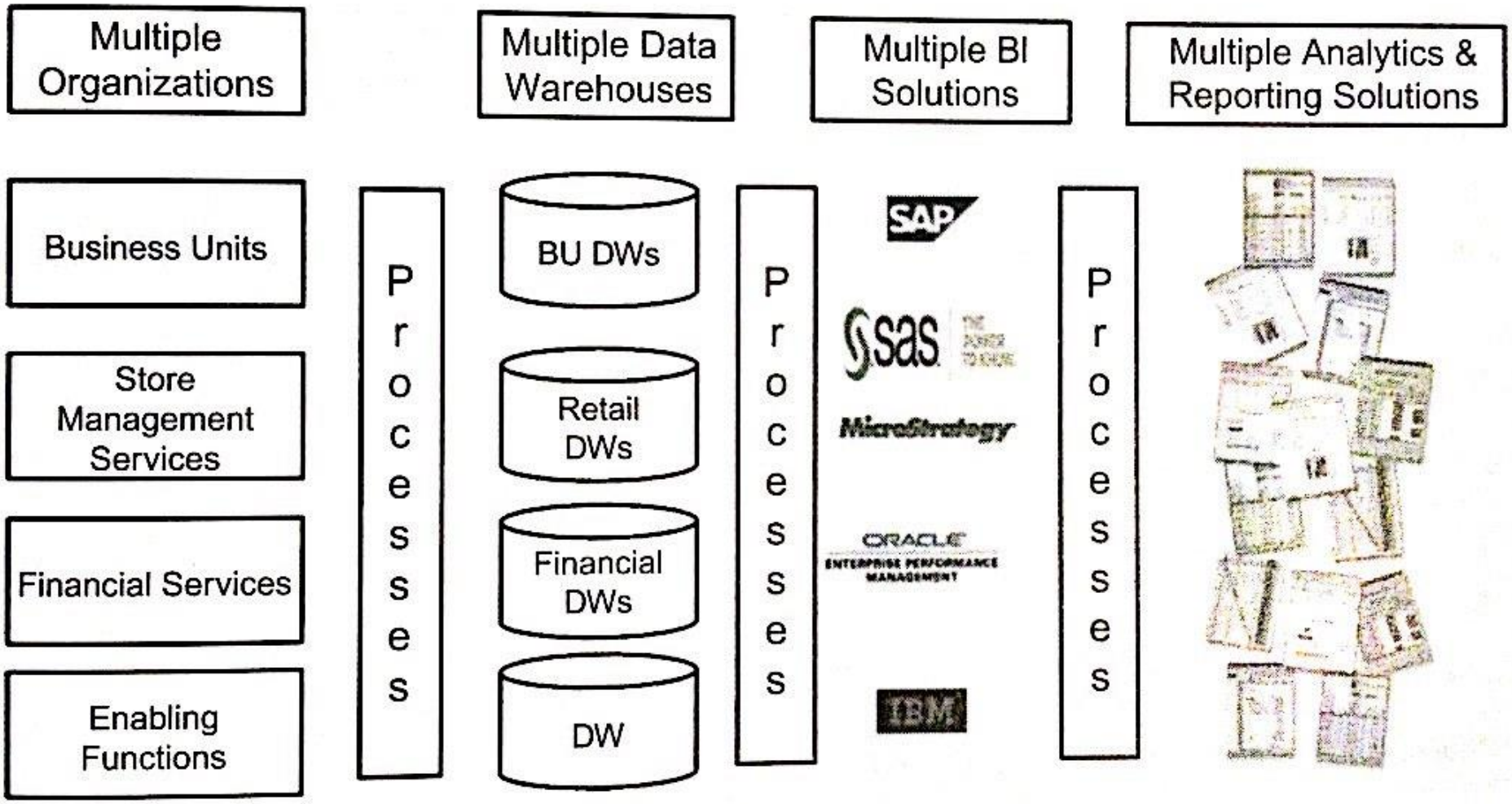
資訊管理與Life cycle of big data

- Data governance teams

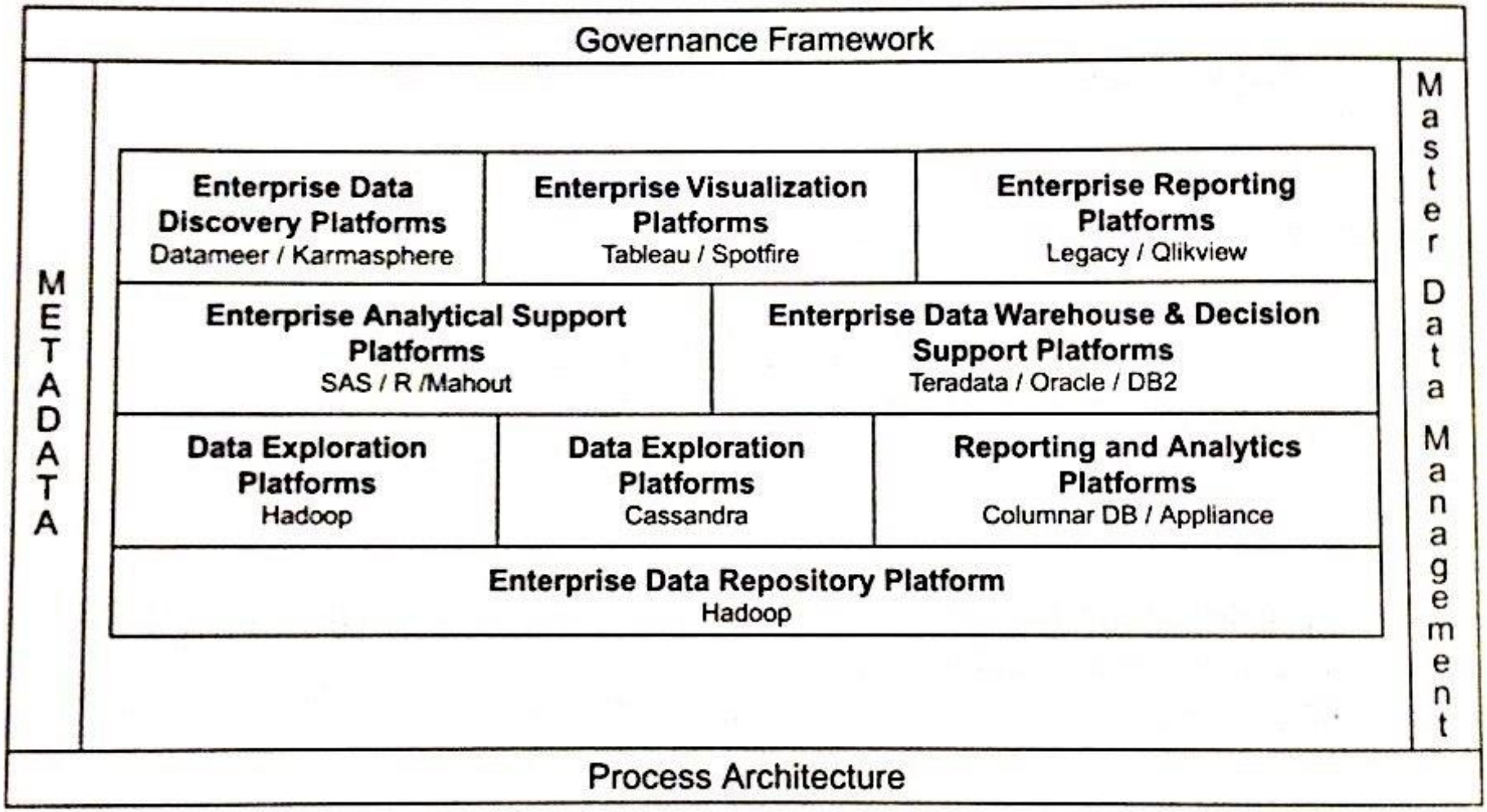


Implement the big data/DW in real life

- Current state architecture

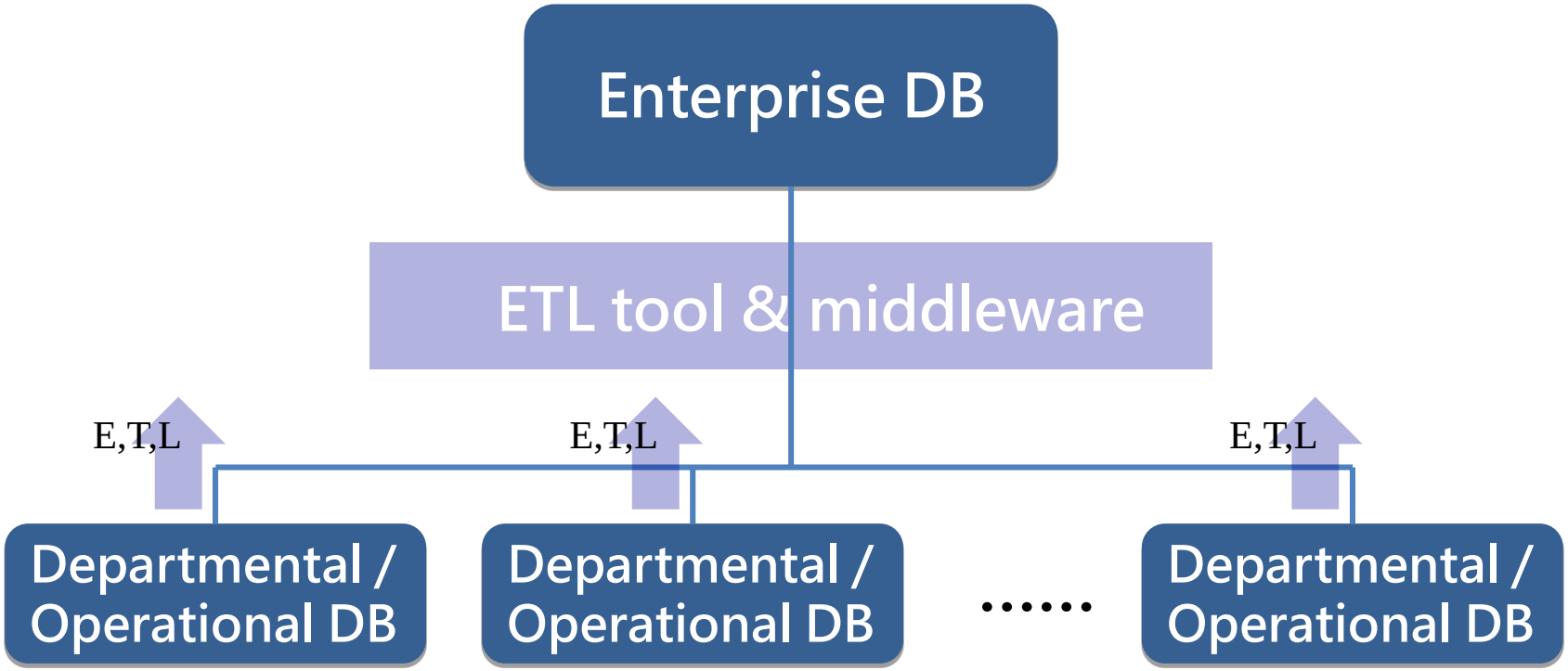


Big data modular architecture conceptual model



整合大數據與資料倉儲 – 實務案例

知識整合 (1) Database approach

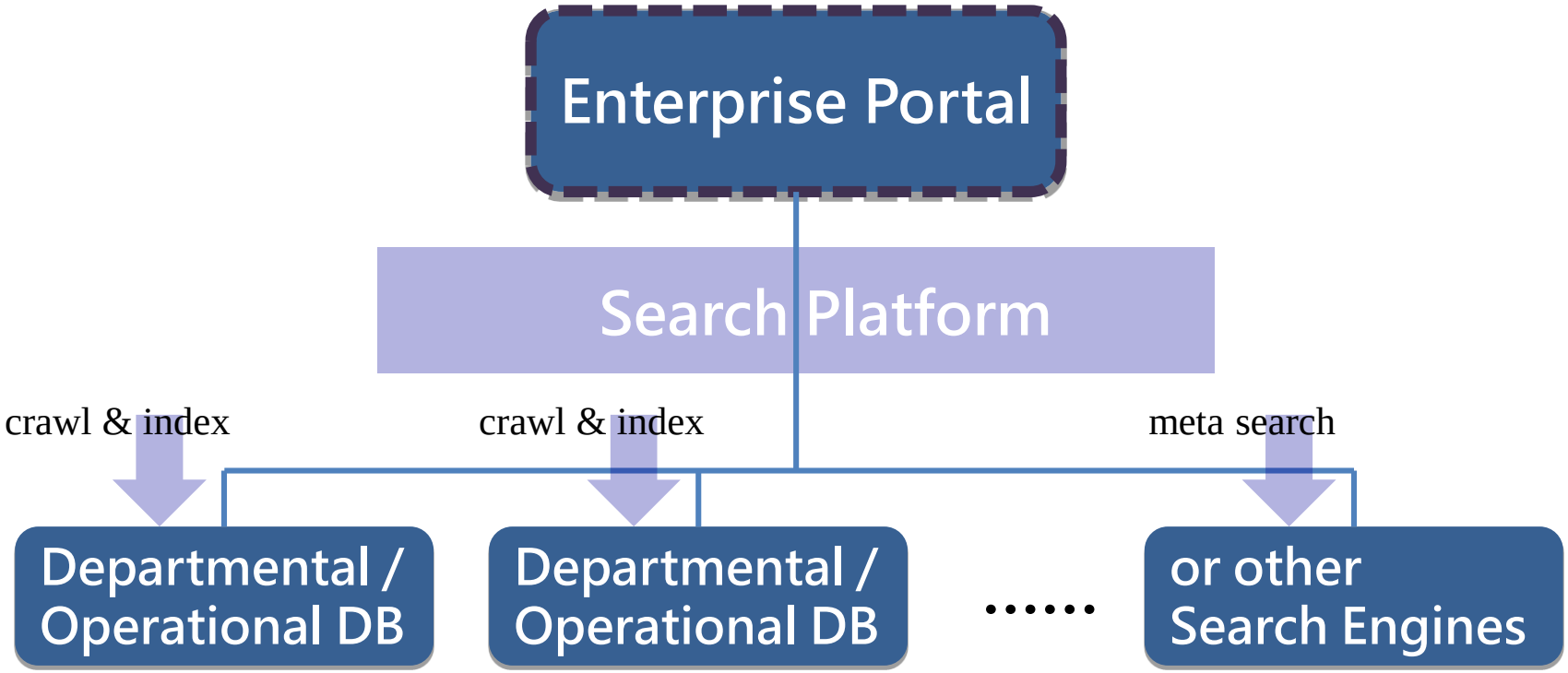


For structured data :

heterogeneous sources → ETL

data integration → middleware

知識整合 (2) Big data approach



For (un)structured data :

heterogeneous sources → crawl & index

data integration → query & result

資訊整合與重構 – 以旅遊產業為例

- Travel Vertical Portal founded in 2007
- 國內最大的旅遊搜尋入口

國內最大的旅遊商品及旅遊情報資料庫

旅遊資訊入口

旅遊產品比價

搜尋技術匯集資訊

FUNTIME Beta
全國最大旅遊搜尋網站

加入最愛 | 首頁 | 登入/加入會員 | 忘記密碼

國內訂房比價 | 國際訂房比價 | 國際機票比價 | 今日最新折扣 | 團體/自由行 | 找行程安排 | facebook | Plurk | FunTime 部落格

旅行，一定要先比價！

共89家航空公司、443, 439筆資料，票價、稅金一目瞭然！

⊙ 依目的地搜尋 BY CITY ▸ 航班與機位查詢

*出發地 不限

*目的地 東南亞 泰國 曼谷

*出發日期 2011-03-21 ~ 回程日期 2011-03-24

艙等 ○ 不限 ◉ 經濟艙 ○ 商務艙 ○ 頭等艙

航班 ○ 不限 ◉ 來回 ○ 單程

比價來源網站: 吉常旅游

約 313,736 筆資料 比 價

我最低價

揪團飛香港
湊團票含稅$6, 379

01 02 03

本週限時精選 ★★★★★

到

01 ▼比價省689▲ 首爾最便宜機票 含稅只要$8,036
02 【最低價只剩4位】直飛上海虹橋 $9,192含稅
03 【飛香港不到8千】全市場最低價 (還有位子哦!)

NO	航空公司	訂購網站	航段	轉機	艙等	票種	出發效期	未稅票價	稅金	機票總價▲	查機位/訂購	說明	轉寄
1	菲律賓航空	玉山	來回	轉機	T經濟艙	外勞票 0天-1年	02-06 03-31	6,667	3,300	9,967			
2	國泰航空	易飛	來回	轉機	N經濟艙	3天-30天外勞票	02-05 06-30	6,634	3,489	10,123			
3	國泰航空	玉山	來回	轉機	N經濟艙	外勞票 3天-30天	02-05 06-30	6,526	3,599	10,125			
4	菲律賓航空	吉帝	來回	轉機	T經濟艙	1年/外勞票	02-06 03-31	6,465	3,670	10,135			
5	菲律賓航空	雄獅	來回	轉機	T經濟艙	1年外勞票年票	02-06 03-31	6,567	3,647	10,214			
6	菲律賓航空	燦星	來回	轉機	T經濟艙	0天-12月外勞票	02-06 03-31	6,567	3,647	10,214			
7	長榮航空	易遊	來回	直飛	U經濟艙	1個月外勞票	02-11 06-24	7,944	2,321	10,265			
8	長榮航空	雄獅	來回	直飛	U經濟艙	1月外勞票旅遊票	02-11 06-24	7,944	2,321	10,265			
9	長榮航空	玉山	來回	直飛	U經濟艙	外勞票 3天-1月	02-11 06-24	7,979	2,321	10,300			
10	長榮航空	易飛	來回	直飛	U經濟艙	3天-1個月外勞票	02-11 06-24	8,025	2,321	10,346			
11	國泰航空	雄獅	來回	轉機	N經濟艙	30天外勞票旅遊票	02-05 06-30	6,567	3,795	10,362			

航班與機位查詢

- Data integration from various sources, both structured and unstructured data.

Innovation - reorganize all travel information

依主題或景點，將相關資訊進行重構

搜尋結果就像一本有圖文「活的旅遊雜誌」

問題討論