資料探勘:
概念與技術

— 第三章 —

第三章: 資料倉儲與即時分析處理方法: 簡介

- 何謂資料資料倉儲?
- 多維度資料模型
- 資料倉儲結構
- 資料倉儲製作
- 從資料倉闖到資料探勘
- 總結

什麼是資料倉儲?

- 有許多定義但不明確.
  - 為一決策支援資料庫而它與公司一般性的資料庫是分開的
  - 藉由提供一個強化歷史資料分析平台支援資訊處理.
- "資料倉儲為用於提供管理決策的具主題導向 (subject-oriented)、具整合性 (integrated)、具時間變動 (rime-variant) 與具不變性 (nonvolatile) 的一群資料."—W. H. Inmon
- 資料倉儲化:
  - 建立與使用資料倉儲的過程

資料倉儲—主題導向

- 圍繞在主要主題如客戶、供應商、產品與銷售
- 不著重於記錄日復一日與一般交易，資料倉儲著重於為決策者提供模型與分析資料
- 為特定主題提供精簡的觀點，並排除與決策不相關的資料

資料倉儲—整合性

- 資料倉儲通常整合多個不同性質來源
  - 如關聯式資料庫、檔案與即時交易紀錄等
- 使用資料清除與資料整合方法.
  - 確保命名、編碼結構與屬性測量的一致性
    - 例., 旅館價格: 幣別, 稅, 包含早餐等.
  - 當資料被移至倉儲時進行轉換.

資料倉儲—時間變動

- 資料倉儲的時間軸比一般交易系統要長很多
  - 交易系統: 現在資料值
  - 資料倉儲資料: 提供歷史觀點的資訊 (例., 過去 5-10 年)
- 資料倉儲的每個關鍵結構
  - 每個資料倉儲的關鍵結構很清楚地包含或內含時間元素
  - 但是交易系統的關鍵結構可以包含或不包含時間元素

資料倉儲—不變性

- 資料倉儲與日常交易資料是分開的
- 交易資料更新並不會發生在資料倉儲環境
  - 不需要像交易處理、資料復原與同步控制等步驟
  - 只包含兩個步驟:
    - 開始取出資料 (initial loading of data) 與資料存取 (access of data)

資料倉儲與不同性質資料庫管理系統

- 傳統不同性質資料庫整合: 查詢導向方法
  - 使用包裝器 (wrappers) 與整合器 (integrators或mediators)來進行不同性質資料庫整合
  - 當使用者進行查詢，一個詮釋資料字典 (metadata dictionary) 用於將查詢轉換為各個不同性質的查詢，從各個查詢傳回的結果彙整合成全域答案集合
  - 需要複雜資訊過濾與整合過程，並會競爭區域資源
- 資料倉儲:更新導向, 高效能
  - 不同性質的資料來源事先進行整合並存於資料倉儲，以便進行直接查詢與分析

資料倉儲與操作性資料庫管理系統

- 即時交易處理 (on-line transaction processing)
  - 傳統關聯式資料管理系統主要工作
  - 每天操作如購買、庫存、製造、銀行、工資、註冊與會計.
- 即時分析處理 (on-line analytical processing)
  - 資料倉儲系統主要工作
  - 資料分析與決策制定
- 不同特點 (OLTP vs. OLAP):
  - 使用者與系統導向: 顧客與市場
  - 資料內容: 現在, 詳細與歷史, 強化
  - 資料庫設計:個體相關+ 應用 與星型 + 主題
  - 檢視: 線在, 區域 與 進化, 整合
  - 存取樣式: 更新 與 複雜查詢的唯讀動作

OLTP vs. OLAP

表3.1 OLTP 與 OLAP 比較

特性	即時交易處理	即時分析處理
特徵	交易處理	訊息處理
導向	交易	分析
使用者	職員、資料庫分析師、資料庫專業人員	知識工作者 (例如管理者、執行者、分析師)
功能	每日交易	長時期訊息需求、決策支援
資料庫設計	個體關聯式、應用導向	星狀 / 雪片、主題導向
資料	現在：保證即時	歷史：隨時間正確地維護
總結	原始、高度詳細	總結、合併
檢視	詳細、扁平關聯	總結、多維度
工作單元	簡短、簡單交易	複雜查詢
存取	讀 / 寫	大部分為讀
著重	資料流入	訊息流出
操作	對主要鍵值設定索引 / 雜湊	多次檢視
存取記錄個數	數十個	百萬
使用者個數	數千個	數百
資料庫大小	100 MB 到 GB	100 GB 到 TB
優先順序	高效能、高可用	高度彈性、使用者自治
衡量	交易產出	查詢產出、回應時間

NOTE : 表格部分根據[CD97]。

為什麼需要一個分開的資料倉儲?

- 增進兩者系統的效率
  - 資料庫管理系統— 對OLTP調整: 存取方法, 索引, 同步控制, 回復
  - 倉儲—對OLAP調整: 複雜 OLAP 查詢, 多維度檢視, 整合
- 不同功能與不同資料:
  - 遺失資料: 決策支援需要歷史資料, 而這些一般不會在操作性資料庫中進行維護
  - 資料整合: 決策支援需要整合不同來源資料 (總結與聚合)
  - 資料品質: 不同資料來源通常使用不一致資料表示, 編碼與格式, 而這些需要調和
- 注意: 有越來越多系統直接在關聯式資料庫執行 OLAP

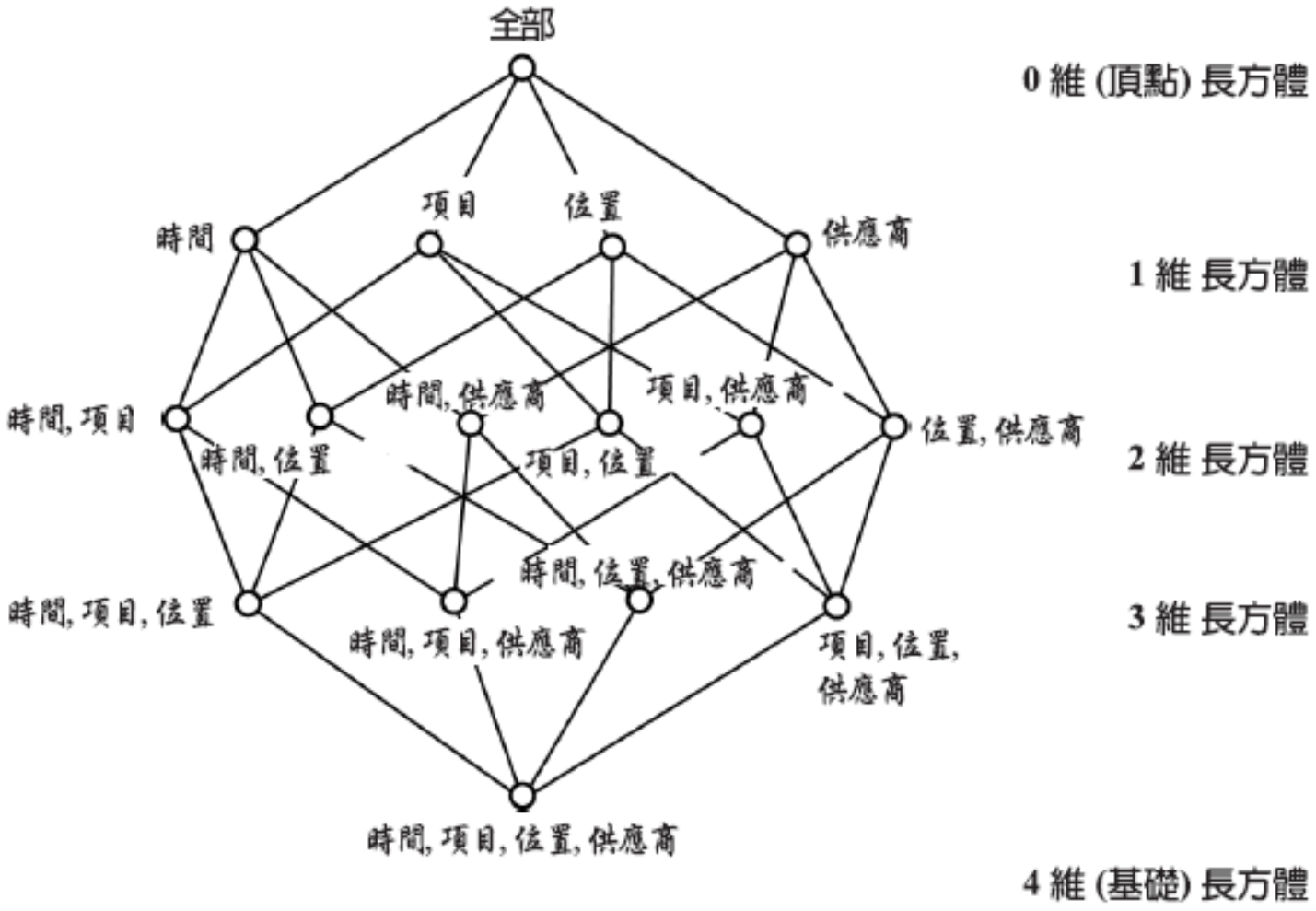
第三章: 資料倉儲與即時分析處理方法: 簡介

- 何謂資料資料倉儲?
- 多維度資料模型
- 資料倉儲結構
- 資料倉儲製作
- 從資料倉闖到資料探勘
- 總結

從表格與試算表到資料方塊

- 資料倉儲是根據多維度資料模型, 它以資料方塊型式來檢視資料
- 一個資料方塊, 例如 銷售, 允許資料以多維度方式進行檢視與模型化
  - 維度表, 例 項目(項目名稱, 品牌, 類型), 或 時間(日, 星期, 月, 季, 年)
  - 事實表包含數值度量與對應至維度表的鍵值
- 在資料倉儲文獻中, 一個 n-維 底部長方體稱為 基礎長方體(base cuboid). 最頂端 0-維 長方體包含粽高層級的匯總稱為 頂點長方體(apex cuboid). 這些長方體的晶格(lattice of cuboids)構成資料方塊

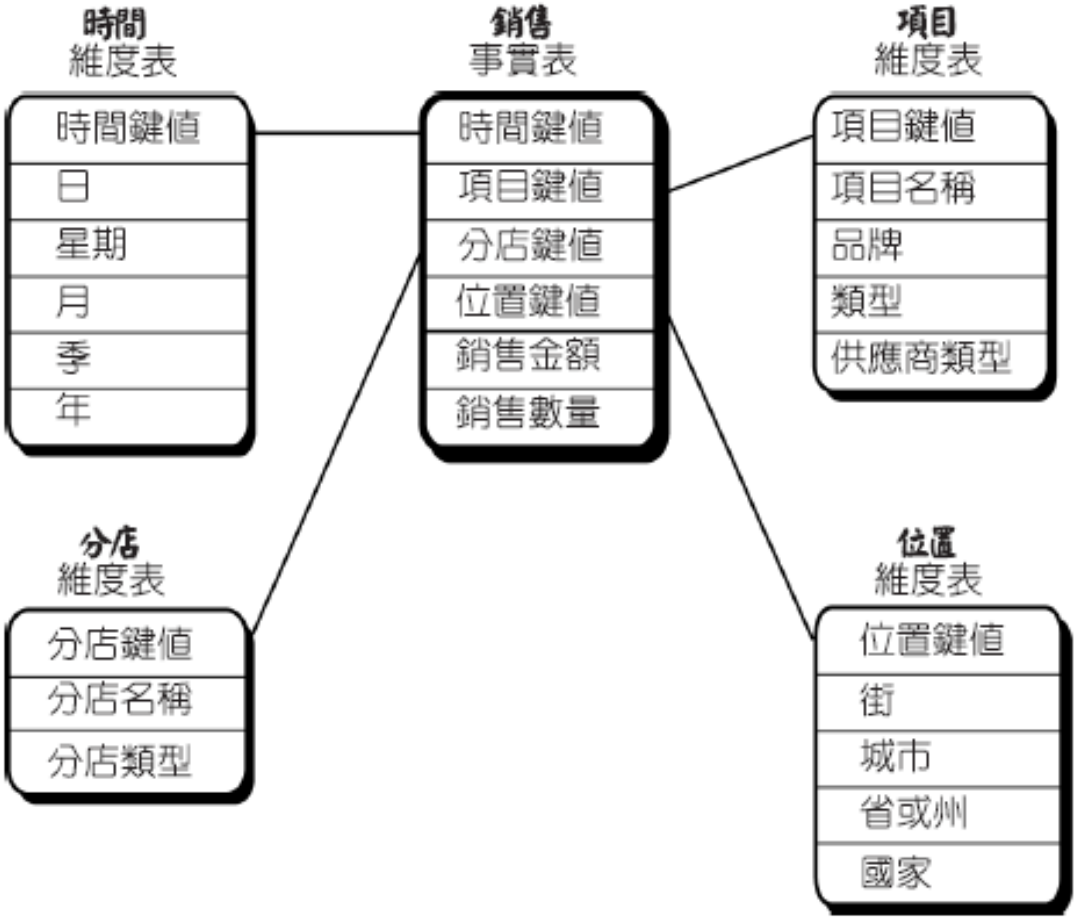
方塊: 晶格長方體


全部
時間
項目
位置
供應商
時間, 項目
時間, 位置
時間, 供應商
項目, 位置
項目, 供應商
位置, 供應商
時間, 項目, 位置
時間, 項目, 供應商
時間, 位置, 供應商
項目, 位置, 供應商
時間, 項目, 位置, 供應商
0 維 (頂點) 長方體
1 維 長方體
2 維 長方體
3 維 長方體
4 維 (基礎) 長方體

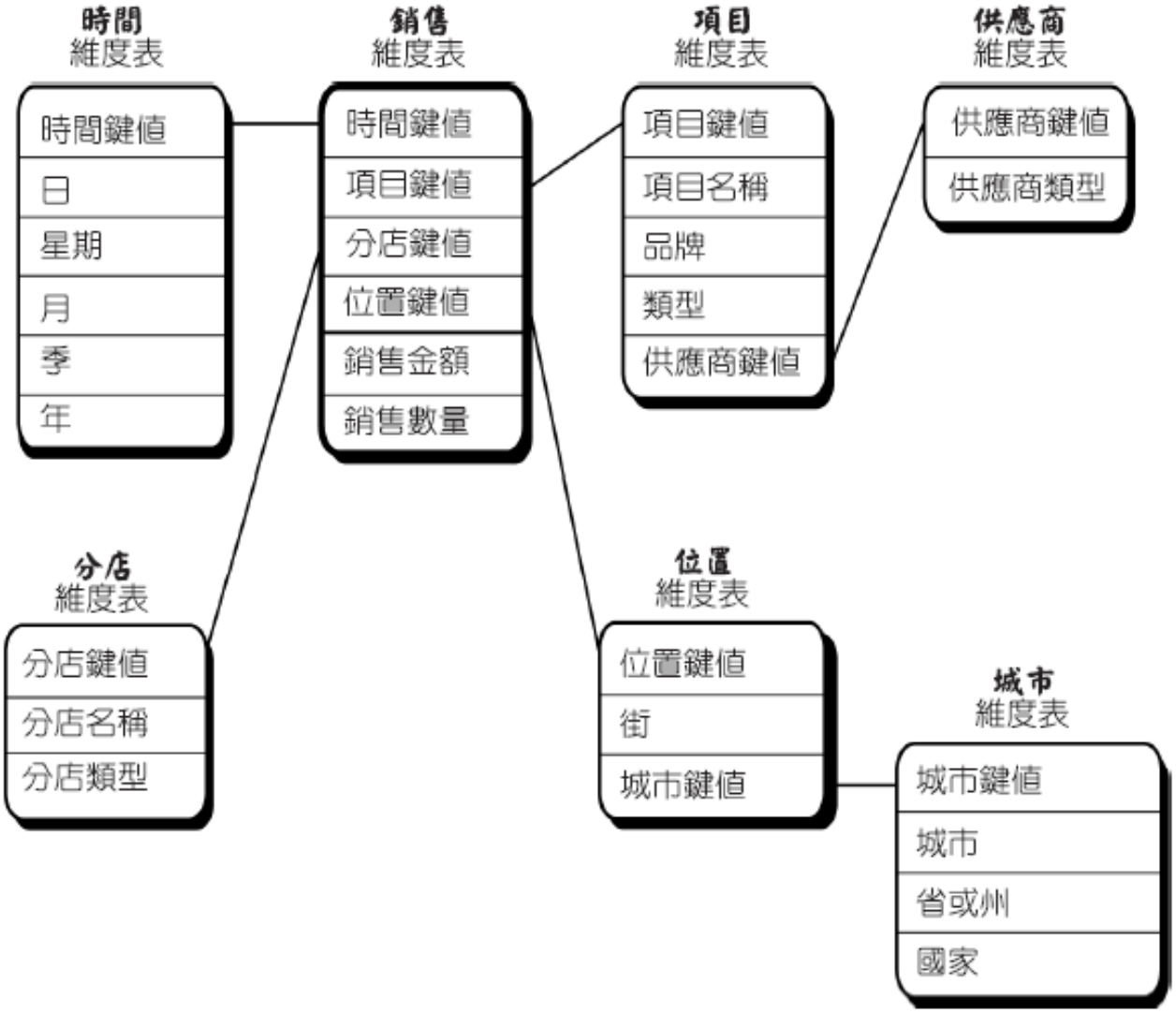
資料倉儲概念模型

- 資料倉儲模型: 維度與度量
  - 星形綱目:中央事實資料表旁邊圍繞維度資料表
  - 雪片綱目:雪片綱目為星形綱目的變異，它透過正規化 (normalization) 將某些維度資料表進行分割，最後綱目的圖形與雪片相似
  - 事實星座綱目:有多個事實表來分享維度表，這種綱目可視為星形綱目的組合，所以稱為銀河綱目 (galaxy schema) 或事實星座綱目 (fact constellation schema)

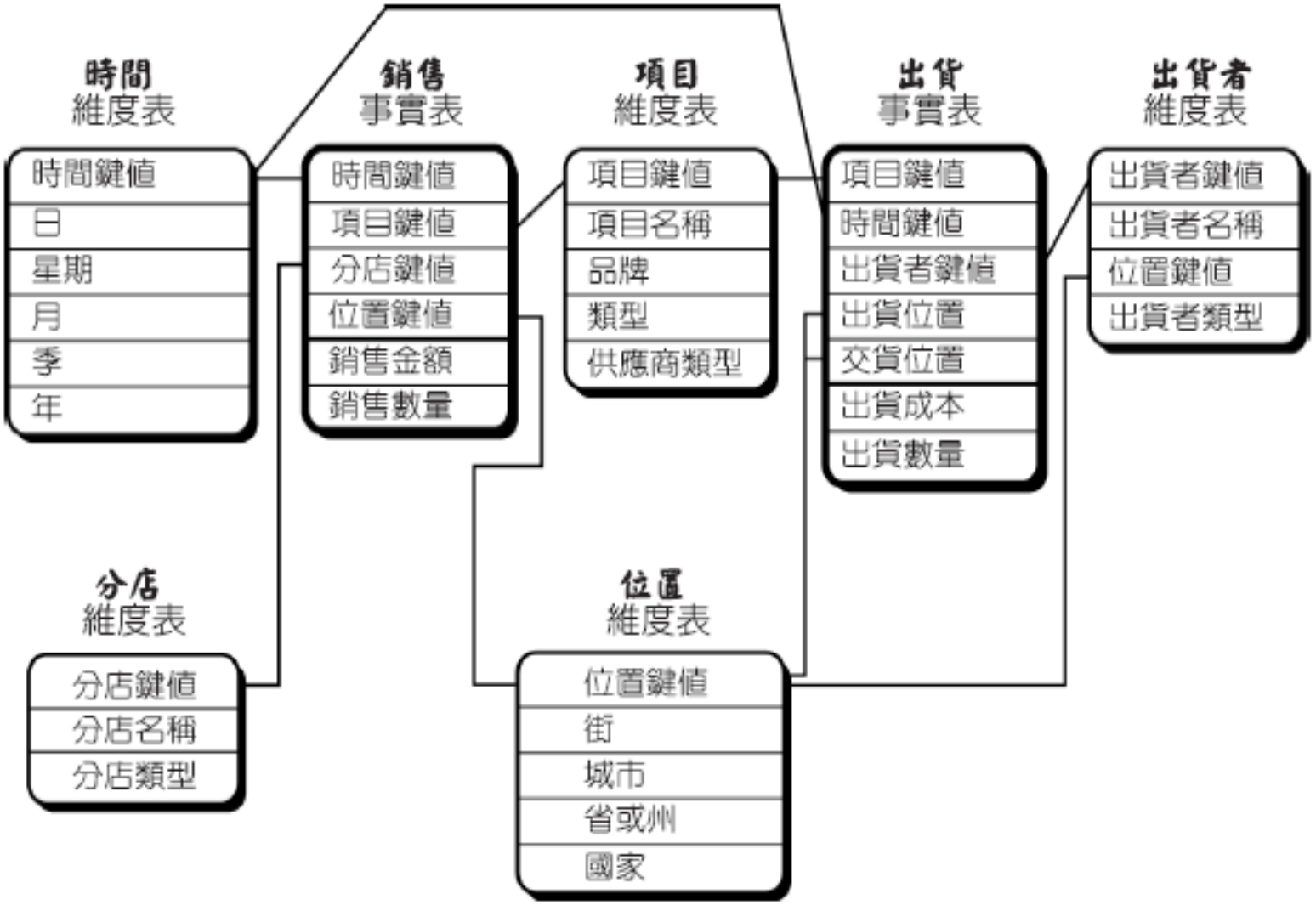
星形綱目範例


時間
維度表
時間鍵值
日
星期
月
季
年
銷售
事實表
時間鍵值
項目鍵值
分店鍵值
位置鍵值
銷售金額
銷售數量
項目
維度表
項目鍵值
項目名稱
品牌
類型
供應商類型
分店
維度表
分店鍵值
分店名稱
分店類型
位置
維度表
位置鍵值
街
城市
省或州
國家

雪片綱目範例


時間
維度表
時間鍵值
日
星期
月
季
年
銷售
維度表
時間鍵值
項目鍵值
分店鍵值
位置鍵值
銷售金額
銷售數量
項目
維度表
項目鍵值
項目名稱
品牌
類型
供應商鍵值
供應商
維度表
供應商鍵值
供應商類型
分店
維度表
分店鍵值
分店名稱
分店類型
位置
維度表
位置鍵值
街
城市鍵值
城市
維度表
城市鍵值
城市
省或州
國家

事實星座綱目範例


時間
維度表
時間鍵值
日
星期
月
季
年
銷售
事實表
時間鍵值
項目鍵值
分店鍵值
位置鍵值
銷售金額
銷售數量
項目
維度表
項目鍵值
項目名稱
品牌
類型
供應商類型
出貨
事實表
項目鍵值
時間鍵值
出貨者鍵值
出貨位置
交貨位置
出貨成本
出貨數量
出貨者
維度表
出貨者鍵值
出貨者名稱
位置鍵值
出貨者類型
分店
維度表
分店鍵值
分店名稱
分店類型
位置
維度表
位置鍵值
街
城市
省或州
國家

在 DMQL方塊定義與法

- 方塊定義 (事實表)

  define cube <cube_name> [<dimension_list>]:
  <measure_list>

- 維度定義(維度表)

  define dimension <dimension_name> as
  (<attribute_or_subdimension_list>)

- 特殊例子 (共享維度表)
  - 首先如同 "方塊定義"
  - define dimension <dimension_name> as
    <dimension_name_first_time> in cube
    <cube_name_first_time>

用DMQL定義星形綱目

define cube sales_star[時間, 項目, 分店, 位置];

銷售金額 = sum (銷貨金額) , 銷售數量 = count (*)

define dimension 時間 as (時間鍵值, 日, 星期, 月, 季, 年)

define dimension 項目 as (項目鍵值, 項目名稱, 品牌, 類型, 供應商類型)

define dimension 分店 as (分店鍵值, 分店名稱, 分店類型)

define dimension 位置 as (位置鍵值, 街, 城市, 省或州, 國家)

用DMQL定義雪片綱目

define cube sales_snowflake[時間, 項目, 分店, 位置]:
銷售金額= sum (銷貨金額) , 銷售數量 = count (*)
define dimension 時間 as (時間鍵值, 日, 星期, 月, 季, 年)
define dimension 項目 as (項目鍵值, 項目名稱, 品牌, 類型, 供應商 (供
應
商鍵值, 供應商類型))
define dimension 分店 as (分店鍵值, 分店名稱, 分店類型)
define dimension 位置 as (位置鍵值, 街, 城市 (城市鍵值, 城市, 省或州,
國家))

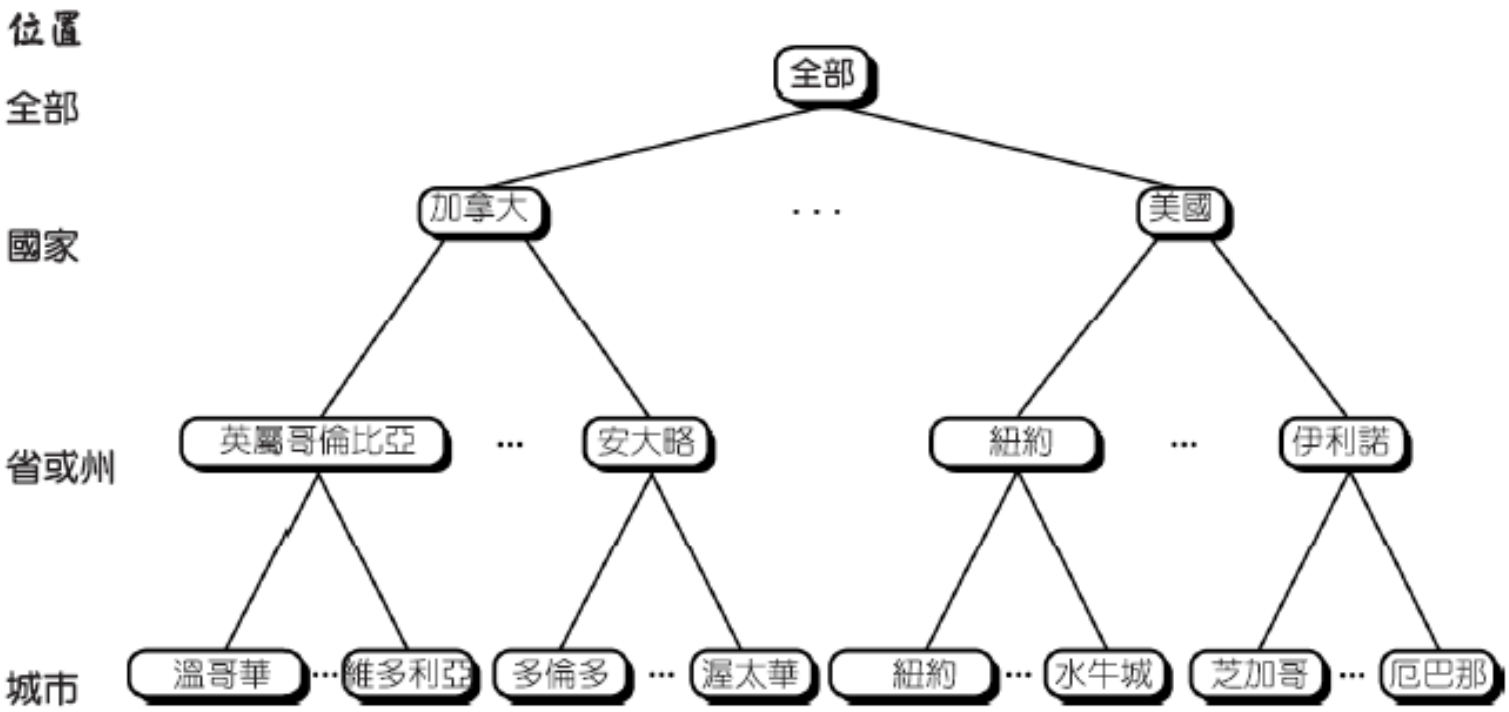
用DMQL定義事實星座綱目

define cube sales[時間, 項目, 分店, 位置]:
銷售金額 = sum (銷貨金額) , 銷售數量 = count (*)
define dimension 時間 as (時間鍵值, 日, 星期, 月, 季, 年)
define dimension 項目 as (項目鍵值, 項目名稱, 品牌, 類型, 供應商類型)
define dimension 分店 as (分店鍵值, 分店名稱, 分店類型)
define dimension 位置 as (位置鍵值, 街, 城市, 省或州, 國家)
define cube shipping[時間, 項目, 出貨者, 出貨位置, 交貨位置]:
出貨成本 = sum (出貨金額) , 出貨數量 = count (*)
define dimension 時間 as 時間 in cube sales
define dimension 項目 as 項目 in cube sales
define dimension 出貨者 as (出貨者鍵值, 出貨者名稱, 位置 as
位置 in cube sales, 出貨者類型)
define dimension 出貨位置 as 位置 in cube sales
define dimension 交貨位置 as 位置 in cube sales

資料方塊度量: 三類

- 分散式:如果將這個值代入函數的結果與利用整體資料的答案相同時，則這個函數可以用分散式來處理
  - 例., count(), sum(), min(), max()
- 代數式:當一個聚合函數透過個參數代數函數進行計算，其為代數式,是具上限的正數，每個參數是透過分散式的聚合函數得之
  - 例., avg(), min_N(), standard_deviation()
- 整體式:當一個聚合函數描述一個子聚合 (subaggregate) 時，儲存大小沒有一個固定限制
  - 例., median(), mode(), rank()

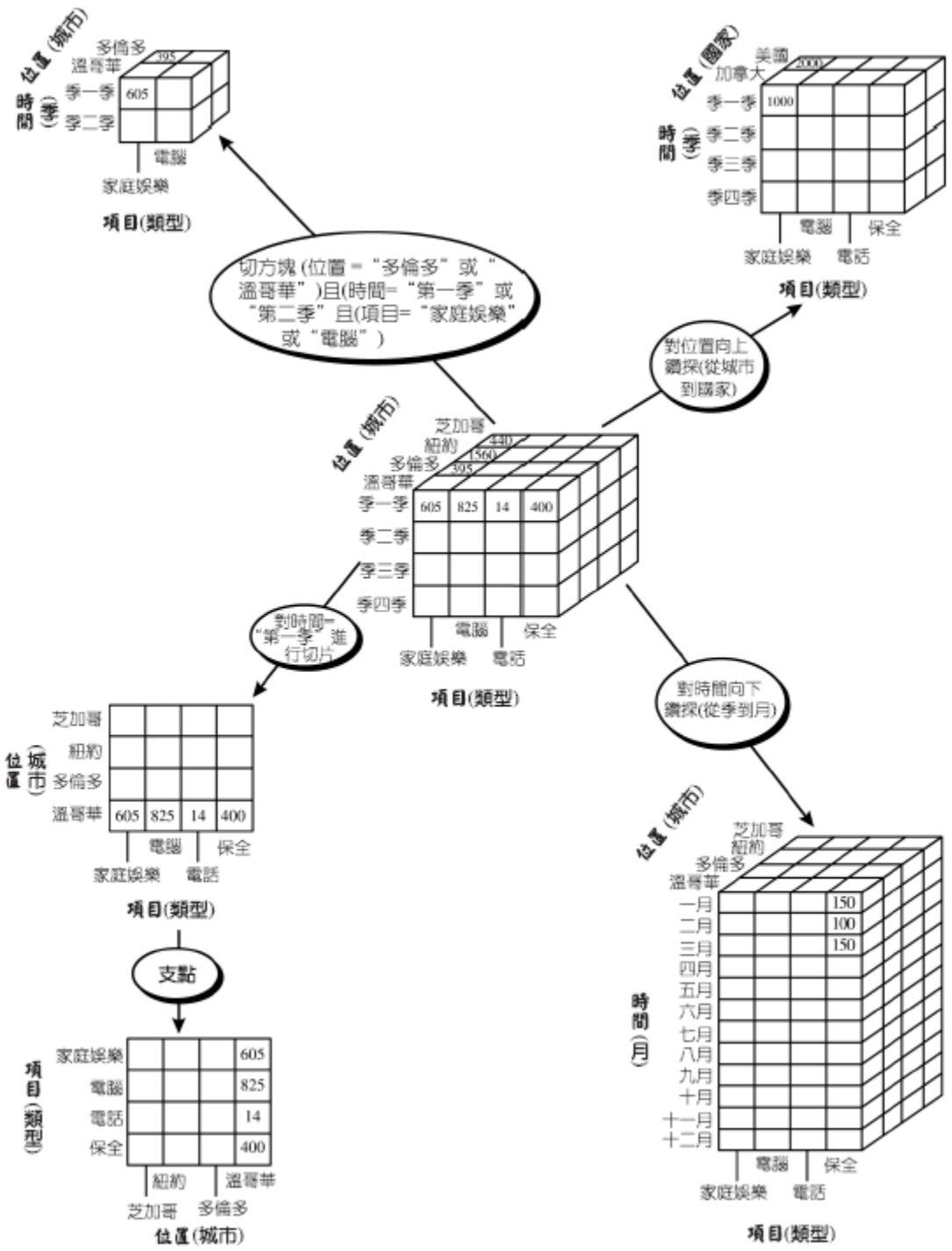
概念階層: 維度 (位置)


位置
全部
國家
省或州
城市
全部
加拿大
...
美國
英屬哥倫比亞
...
安大略
紐約
...
伊利諾
溫哥華
維多利亞
多倫多
渥太華
紐約
水牛城
芝加哥
厄巴那

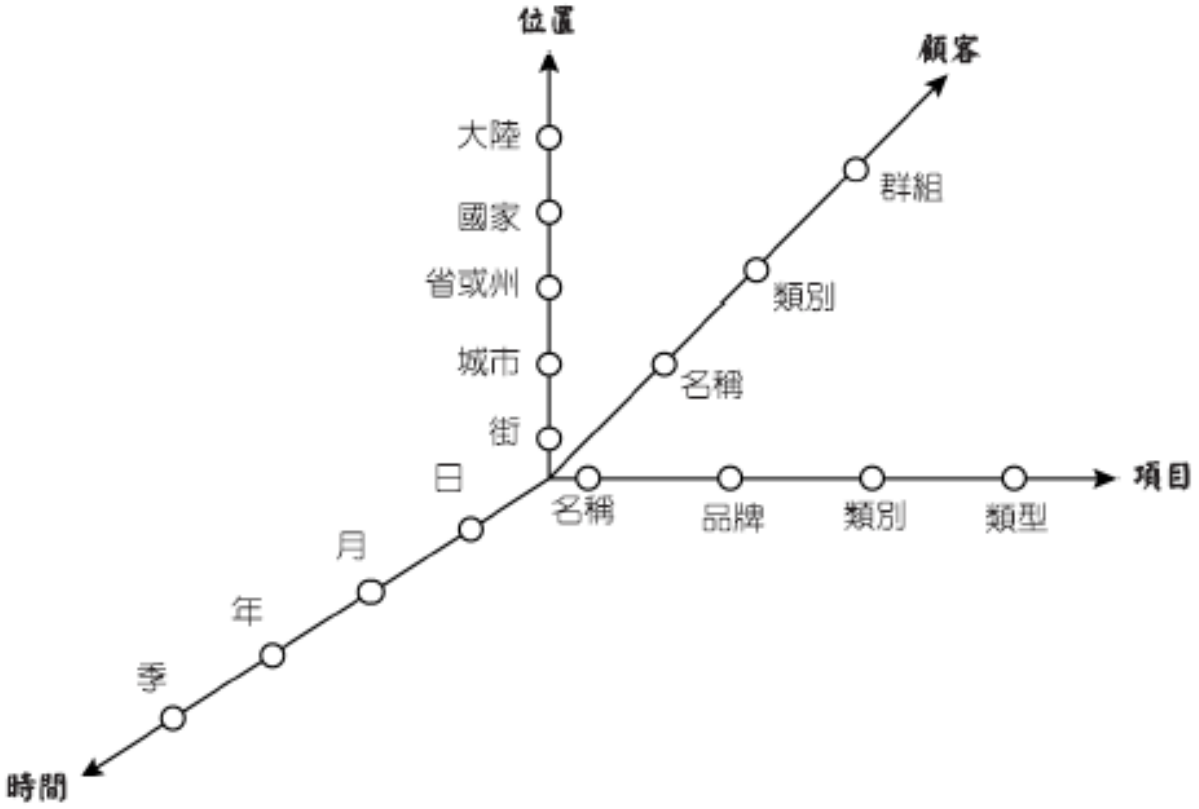
傳統OLAP的運算

- 往上鑽探 (drill-up): 資料彙總
  - 讓特定維度的概念階層往上移動或進行維度刪減
- 往下鑽探 (roll down):為往上鑽探的相反動作
  - 透過由特定維度的概念階層往下移動或進行維度增加來完成
- 切片與切方塊: 對應與選擇
- 支點 (rotate):
  - 藉由旋轉資料軸來提供不同的資料檢視
- 其他運算
  - 橫向鑽探:執行包含一個以上事實表的查詢
  - 穿透鑽探:使用關聯式SQL，它會穿越資料方塊的底層到它後端的關聯資料表

範例. 3.10 傳統OLAP 運算



星網查詢模型



第三章: 資料倉儲與即時分析處理方法: 簡介

- 何謂資料資料倉儲?
- 多維度資料模型
- 資料倉儲結構
- 資料倉儲製作
- 從資料倉闖到資料探勘
- 總結

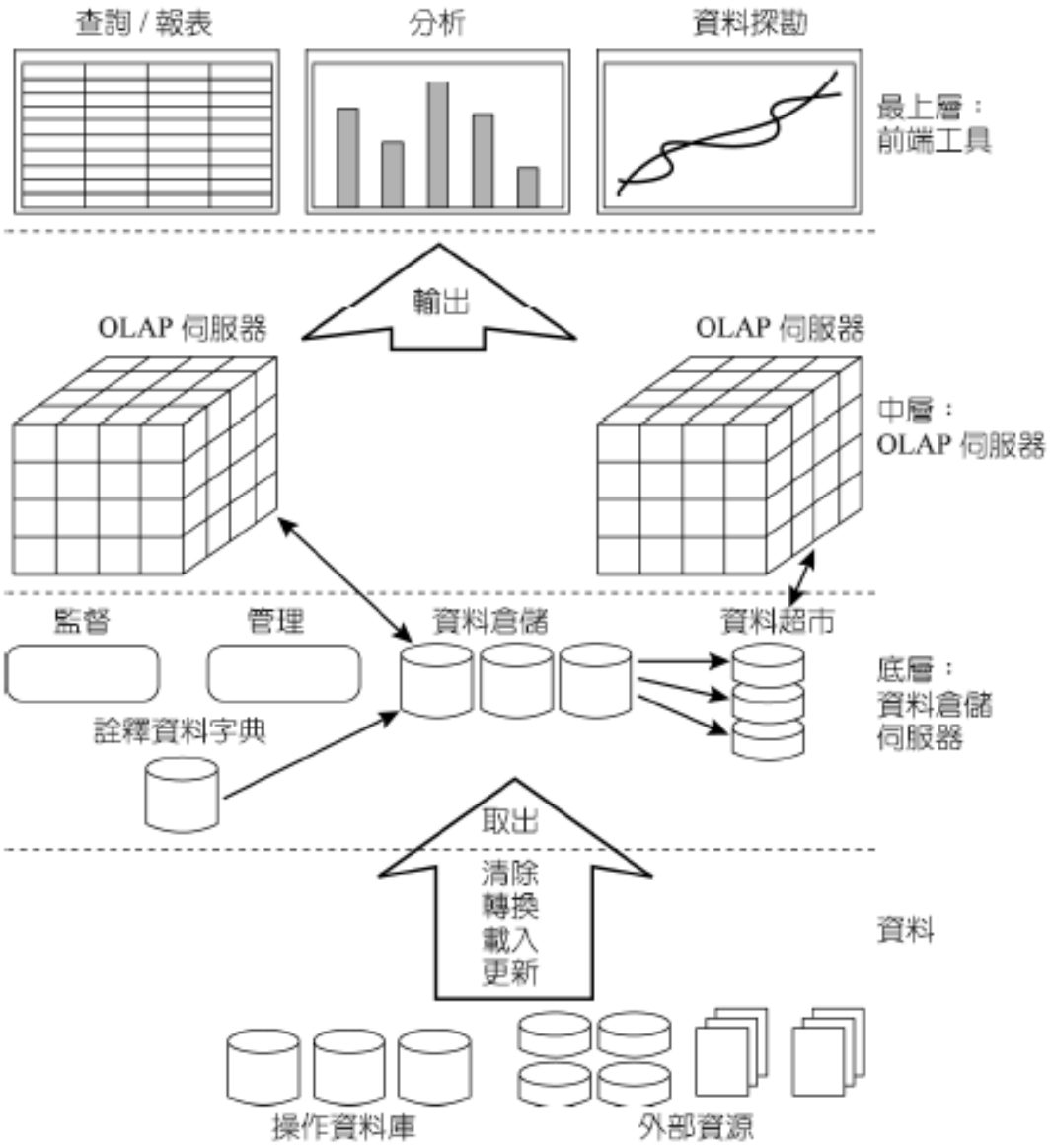
設計資料倉儲:商業分析架構

- 設計資料倉儲要考量四種不同觀點
  - 由上而下觀點
    - 用於選擇與資料倉儲相關的訊息
  - 資料來源觀點
    - 顯示操作系統所獲得、儲存與管理訊息
  - 資料倉儲觀點
    - 包含事實表與維度表
  - 商業查詢觀點
    - 依據使用者觀點的資料倉儲的資料觀點

資料倉儲設計過程

- 由上而下、由下而上或是兩種整合的方式來建立
  - 由上而下:由全部設計與計畫開始 (成熟)
  - 由下而上:由實驗與雛型開始 (快速)
- 從軟體工程的觀點
  - 瀑布法:進行下一個步驟前，對本步驟進行結構與對稱性的分析
  - 螺旋法:在兩個短時間的成功版本內快速產生函數系統
- 一般資料倉儲設計的步驟
  - 選擇要進行模型的商業過程，例如訂單、發票等
  - 選擇商業過程粒子 (grain)
  - 選擇每個事實表紀錄的維度
  - 選擇每個事實表紀錄的度量

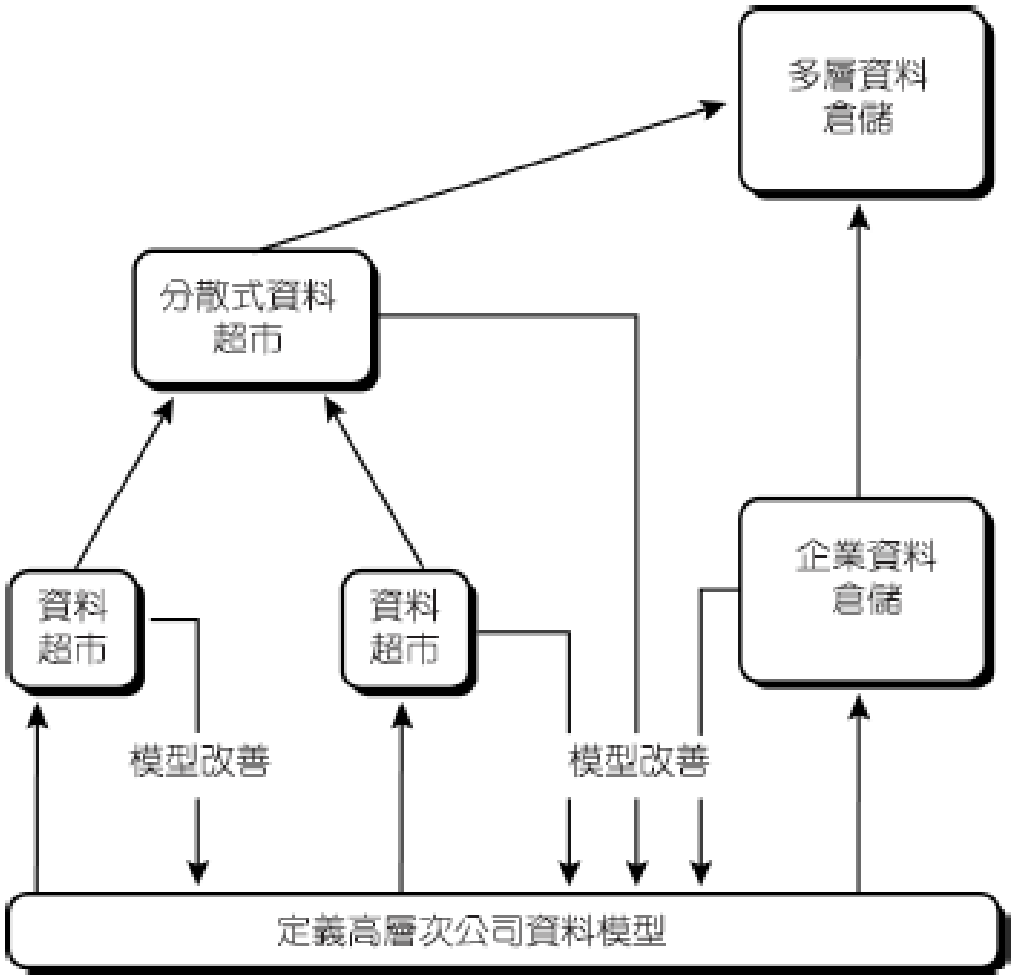
資料倉儲:多層次資料倉儲架構



三種資料倉儲模型

- 企業倉儲
  - 收集涵蓋整個組織的資訊
- 資料超市
  - 資料超市包含公司特定使用族群資料，它的範圍侷限於特定主題，例如侷限於客戶、項目與銷售
    - 資料超市可分為相依或不相依
- 虛擬倉儲
  - 虛擬倉儲為一組在操作資料庫的觀點
  - 僅有部分的觀點能實施

資料倉儲建立的建議方式



資料倉儲後端工具與公用程式

- 資料取出
  - 從多個外部不同性質的來源集結資料
- 資料清除
  - 找出資料的錯誤並盡力矯正
- 資料轉換
  - 進行不同資料格式轉換
- 取出
  - 進行排序、總計、合併、設定電腦觀點、整合檢查與建立索引與分割
- 更新
  - 將更新資料傳遞給資料倉儲

詮釋資料字典

- 在資料倉儲中，詮釋資料用於定義倉儲個體:
- 資料倉儲結構的描述
  - 包含資料倉儲的綱目、觀點、維度、階層、導出資料定義與資料超市位置及內容
- 操作詮釋資料
  - 包含資料血統 (資料轉移與資料轉換的歷史)、資料現狀 (現行、歷史或刪除) 與監視訊息 (倉儲使用統計、錯誤報告與稽核追蹤)
- 總結使用方法
- 從操作環境到資料倉儲的對應
- 系統效能相關資料
  - 包含增進資料存取效能的索引與設定檔
- 商業詮釋資料
  - 包含商業專有名詞與定義、資料擁有權資訊與收費政策

OLAP伺服器類型

- 關聯式OLAP (ROLAP)
  - 使用關聯式或延伸關聯式資料庫管理系統，來儲存管理資料倉儲的資料並支援OLAP
  - 包含後端關聯式資料庫管理系統最佳化、聚合與瀏覽邏輯製作、其他工具與服務
  - 具可量度性
- 多維度OLAP (MOLAP)
  - 稀疏矩陣式多維度儲存引擎
  - 預先計算總結資料進行快速索引
- 混合OLAP (HOLAP) (e.g., Microsoft SQLServer)
  - 具彈性, 例., 低層次: 關聯式, 高層次: 陣列
- 特殊SQL (例., Redbricks)
  - 在唯讀環境的星型與雪片綱目中提供進階查詢語言

第三章: 資料倉儲與即時分析處理方法: 簡介

- 何謂資料資料倉儲?
- 多維度資料模型
- 資料倉儲結構
- 資料倉儲製作
- 從資料倉闖到資料探勘
- 總結

資料方塊運算

- 用DMQL定義方塊與計算

  define cube sales_cube[城市, 項目, 年]: sum(銷售金額)

  compute cube sales_cube

- 轉換成類似SQL語言 (藉由新運算元 cube by, introduced by Gray et al.'96)

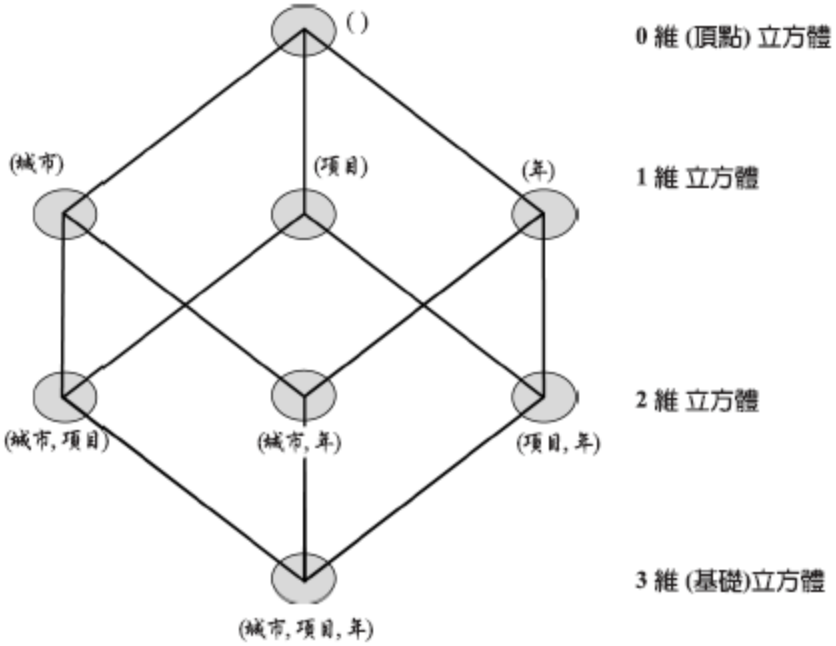
  SELECT 城市, 項目, 年 SUM (金額)

  FROM SALES

  CUBE BY城市, 項目, 年

- 須計算下列群組

  (城市, 項目, 年) , (城市, 項目) , (城市, 城市, 年) , (項目, 年) , (城市) , (項目) , (年) , ()



OLAP資料索引:位元對應索引法

- 對特定欄進行所引
- 位元對應索引對每個屬性值會包含一個不同的位元向量 bit-op is fast
- 如果屬性包含n個不同值，則位元對應索引會包含n位元
- 如果資料表的某一列屬性值為v，則將對應屬性值為v的位元設為1，其餘設為0
- 不適用於高資料範圍基數

基礎資料表

記錄識別值	項目	城市
R1	H	V
R2	C	V
R3	P	V
R4	S	V
R5	H	T
R6	C	T
R7	P	T
R8	S	T

項目位元對應索引表

記錄識別值	H	C	P	S
R1	1	0	0	0
R2	0	1	0	0
R3	0	0	1	0
R4	0	0	0	1
R5	1	0	0	0
R6	0	1	0	0
R7	0	0	1	0
R8	0	0	0	1

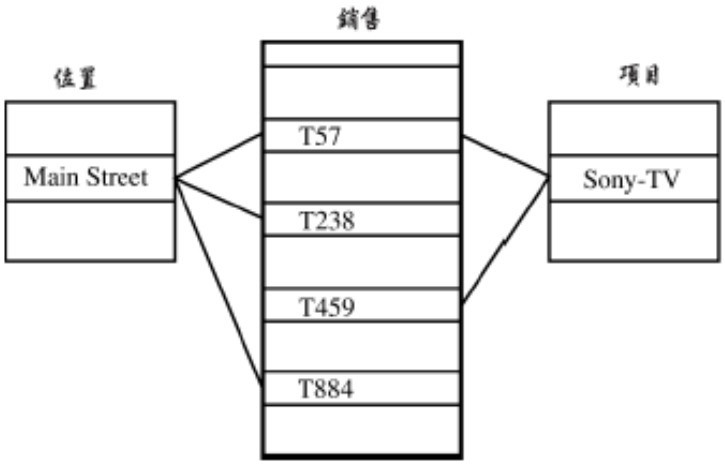
城市位元對應索引表

	V	T
R1	1	0
R2	1	0
R3	1	0
R4	1	0
R5	0	1
R6	0	1
R7	0	1
R8	0	1

請注意：H代表家庭娛樂、C代表電腦、P代表電話、S代表保全、V代表溫哥華、T代表多倫多

OLAP資料索引:連結索引法

- 連結索引法: JI(R-id, S-id) 當 R (R-id, …) ▷◁ S (S-id, …)
- 傳統索引的方法將資料鏈的值對應到包含這個值的列
  - 連結索引法記錄兩個關聯的連結列
- 資料倉儲中利用星狀綱目進行跨資料表的搜尋時，連結索引法是相當有用的，因為事實表與相關維度表的關聯包含事實表的外來鍵值與維度表的主要鍵.
  - 例. 事實表: 銷售與兩個維度位置與項目
    - 位置的連結索引紀錄銷售中不同城市的值組的紀錄識別值
  - 連結索引可延伸至多維度



有效率處理OLAP查詢

- 決定現有的長方體應執行何種運算
  - 包含選擇、投射、向上鑽探 (群組) 與向下鑽探轉換為對應的SQL或OLAP運算。例如切片與切方塊會對應至成形長方體的選擇與投射動作
- 決定相關的動作要套用至哪些成形長方體
  - 假設要對{品牌，省或州}進行查詢當“年 = 2004”，並且四個可用的成形長方體:

    1) {年, 項目名稱, 稱市}

    2) {年, 品牌, 國家}

    3) {年, 品牌, 省或州}

    4) {項目名稱, 省或州} where年 = 2004

    應選擇上述哪個長方體進行查詢?
- 在MOLAP中探索索引,壓縮與密集陣列結構

第三章: 資料倉儲與即時分析處理方法: 簡介

- 何謂資料資料倉儲?
- 多維度資料模型
- 資料倉儲結構
- 資料倉儲製作
- 從資料倉闖到資料探勘
- 總結

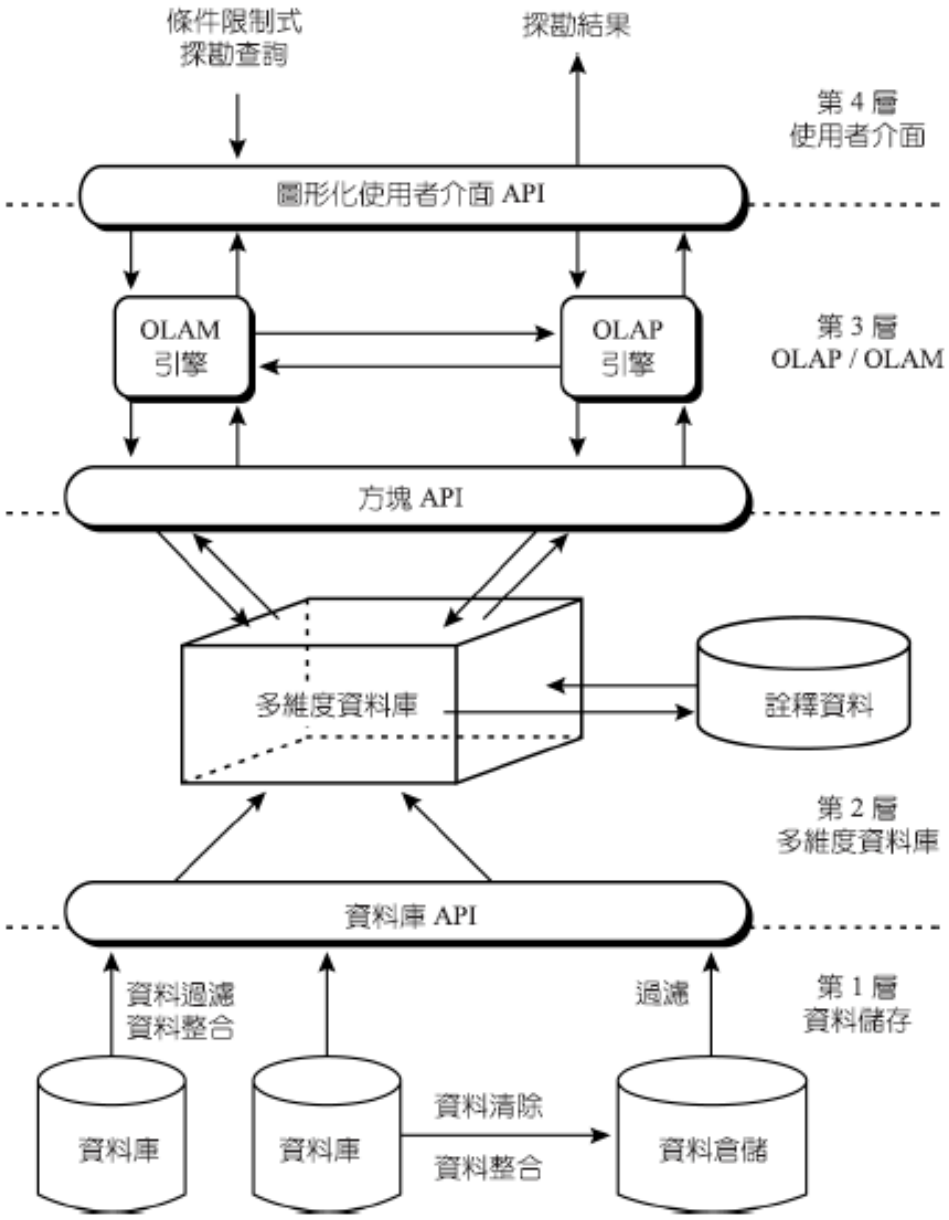
資料倉儲使用

- 資料倉儲的應用有三類
  - 資訊處理
    - 支援查詢、基本統計分析與利用相互表格、圖表與圖形
  - 分析處理
    - 資料倉儲的多維度資料分析
    - 支援基本OLAP運算，包含切片、切方塊、往下鑽探、往上鑽探與支點
  - 資料探勘
    - 尋找隱藏樣式
    - 支援關聯、建立分析模型、進行判別與預測、提供對探勘結果的顯示工具

從即時分析處理到即時分析探勘

- 為何要即時分析探勘?
  - 資料倉儲中高品質的資料
    - 資料倉儲包含整合、一致與乾淨資料
  - 圍繞資料倉儲相關訊息處理的基礎結構
    - ODBC/OLE DB連接、網路存取服務、報表與OLAP服務
  - OLAP式資料分析探索
    - 進行鑽探、支點、過濾、切方塊與切片的探勘
  - 資料探勘函數的即時選擇
    - 藉由OLAP與探勘函數的整合，提供使用者適當的探勘函數

OLAM 架構


條件限制式
探勘查詢
探勘結果
第 4 層
使用者介面
圖形化使用者介面 API
OLAM
引擎
OLAP
引擎
第 3 層
OLAP / OLAM
方塊 API
多維度資料庫
詮釋資料
第 2 層
多維度資料庫
資料庫 API
資料過濾
資料整合
過濾
第 1 層
資料儲存
資料庫
資料庫
資料清除
資料整合
資料倉儲

第三章: 資料倉儲與即時分析處理方法: 簡介

- 何謂資料資料倉儲?
- 多維度資料模型
- 資料倉儲結構
- 資料倉儲製作
- 從資料倉闖到資料探勘
- 總結

總結:資料倉儲與即時分析處理方法

- 為何要資料倉儲?
- 資料倉儲多維度模型
  - 星形、雪片與事實星座綱目
  - 資料方塊包含維度與度量
- OLAP運算包含往上鑽探、往下 (橫向、穿透) 鑽探、切片與切方塊、支點
- 資料倉儲架構
- OLAP伺服器:關聯式OLAP (ROLAP)、多維度OLAP (MOLAP) 與混合OLAP
- 有效計算資料方塊
  - 部份, 全部, 沒有成形
  - OALP 資料索引:位元對應與連結索引法
  - OLAP 查詢處理
- 從即時分析處理到即時分析探勘 (on-line analytical mining)