



OLAP和資料倉儲

黃三益
國立中山大學資管系



OLAP

- Overview
- Data Warehouse的建置過程
- Data analysis
- Problems with GROUP BY
- CUBE and ROLLUP
- 支援Data Warehouse的DBMS所具備的特色



Overview

- OLAP (On-Line Analytical Processing) 主要被使用在資料分析的應用上。
- 資料分析包括四個步驟：
 1. 從一個大資料庫抓出想要的資料。
 2. 彙總存成一個檔案或表單。
 3. 將結果以圖形化方式表示出來。
 4. 分析結果後再從資料庫抓出其他想要的資料。
- 試算表MS Excel就是一個這樣的資料分析工具。



Data Warehouse

- 資料倉儲被是用來支援決策技術，以能夠讓知識工作者做更好、更快的決策為目的。
- 資料倉儲通常是一龐大的資料庫，大於任何運作中的資料庫，因為它所包含的資料庫是含歷史資料和部門資料。這全然的資料量可能會以'兆'位元來衡量。
- 資料倉儲的建置程序。

Data Warehouse (Cont'd)

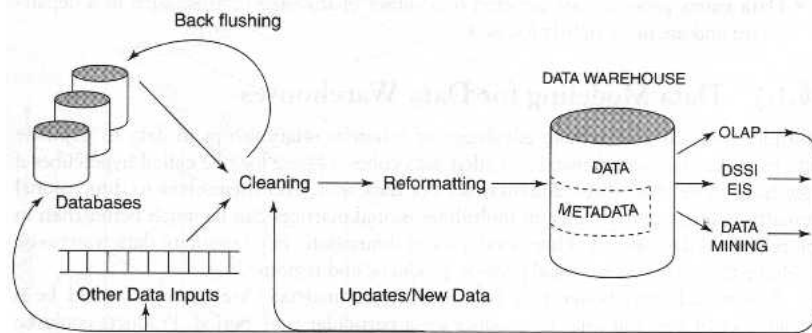


Figure 26.1 The overall process of data warehousing.

[回主頁](#)

[主目錄](#) 5

2006資料庫核心理論與實務

Data analysis

- 資料分析工具視資料集為一個N維的空間。
[Figure1](#)。
- 以關聯模式的觀點來看，就好像一個關聯裡有N+K個屬性，其中N個屬性試用來存'維度值'(dimensions)，其他的K個屬性則用來存'測量值'(measures)。
[Figure2](#)。
- 以這種觀點來看OLAP，便被成為ROLAP，相對於空間維度的觀點 (MOLAP)

[主目錄](#) 6

2006資料庫核心理論與實務



Snowflake/Star Schema

- ROLAP有兩種形態的table。一個fact table 和 數個 dimension tables。
- 一般採用Star Schema，也有採用Snowflake Schema
 - Star Schema裡的dimension table一般並不作正規化，以提高效率。
 - OLAP的資料一般很少改
 - Dimension table裡的主鍵一般是由系統產生，以減少Fact table裡外部鍵的大小。
 - Dimension Table裡的屬性可以形成一個hierarchy或 lattice (e.g., date -> (week, month) -> year)。

2006資料庫核心理論與實務

[主目錄](#) 7



Snowflake/Star Schema(Cont'd)

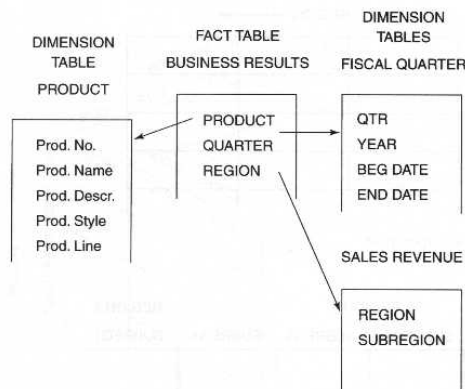


Figure 26.7 A star schema with fact and dimensional tables.

2006資料庫核心理論與實務

[回主頁](#) [主目錄](#) 8



Snowflake/Star Schema(Cont'd)

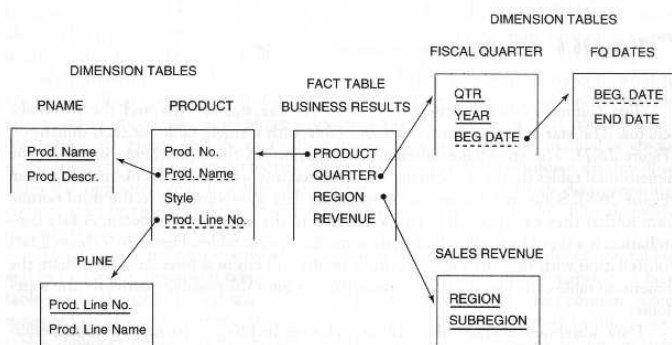
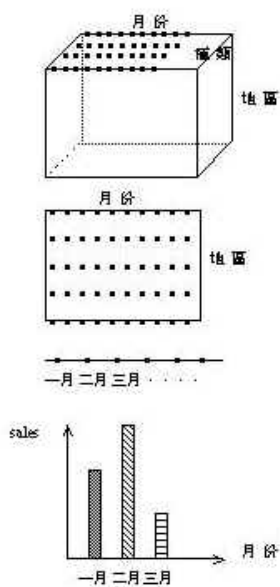


Figure 26.8 A snowflake schema.

[回主頁](#) [主目錄](#)



[回主頁](#)

[主目錄](#)

Data analysis (Con'd)

Dimension Attributes			Measure Attributes
月份	地區	種類	Sales
一月	台北	COFFEE	500
.	.	.	.
.	.	.	.
.	.	.	.

[回主頁](#)

[主目錄](#)₁₁

2006資料庫核心理論與實務

Data analysis (Con'd)

- 資料分析工具為了分析師的方便，廣泛地使用維度縮減(彙總和分群)。關聯資料庫依賴彙總函數和Group By運算子來進行維度縮減。
- SQL的分群彙總愈來愈被廣泛使用，比如TPC-D的設定查詢裡有一6維的group by和三個3維的group bys。See [table2](#)。

[主目錄](#)₁₂

2006資料庫核心理論與實務

Data analysis (Con'd)

Table 2. SQL aggregates in standard benchmarks.

Benchmark	Queries	Aggregates	GROUP BYs
TPC-A, B	1	0	0
TPC-C	18	4	0
TPC-D	16	27	15
Wisconsin	18	3	2
AS ³ AP	23	20	2
SetQuery	7	5	1

2006資料庫核心理論與實務

主目錄¹³

Data analysis (Con'd)

- 除了COUNT, SUM, AVG, MIN, MAX，很多的系統更進一步的提供許多彙總函數，例如：median, 標準差...等等。
- 有一些系統允許使用者去增加彙總函數，例如以下函數((Informix Illustr)
 - Init(&handle)
 - Iter(&handle, value)
 - Value = Final(&handle)
- Red Brick系統，一OLAP的廠商，增加以下的函數以方便應用系統利用。[Figure3](#)。

2006資料庫核心理論與實務

主目錄¹⁴

Data analysis (Con'd)

A1	A2	...	Rank(A1)	N-tile(A1,3)	Ratio-to-Total (A1)	Cumulative (A1)	Running-Sum (A1,3)
2			2	1	2/100	3	2
5			4	2	5/100	12	11
20			8	3	20/100	83	50
13			6	2	.	.	.
1			1	1	.	.	.
8			5	2	.	.	.
4			3	1	.	.	.
30			9	3	.	.	.
17			7	3	.	.	.

2006資料庫核心理論與實務

主目錄¹⁵

Data analysis (Con'd)

- Rank(attribute)：每一筆記錄的attribute屬性值，依其在所有attribute屬性值集合的位置傳回rank(等級)。
- N-tile(attribute, N)：將所有attribute屬性值由大到小分成N個等級。此函數傳回一筆記錄之該屬性值的等級(1..N)。
- Ratio_To_Total(attribute)：一筆記錄之attribute屬性值除以所有attribute屬性值的總和。
- Cumulative(attribute)：小於等於一筆記錄之attribute屬性值的累加值。
- Running_Sum(attribute, n)：小於等於一筆記錄之attribute屬性值最近n個值的累加值。
- Running_Average(attribute, n)：小於等於一筆記錄之attribute屬性值最近n個值的平均值。

2006資料庫核心理論與實務

主目錄¹⁶



Problems with GROUP BY

- SQL的Group By敘述，有三個主要的問題。
 - Histograms
 - roll-up totals and sub-totals for drill-downs
 - cross tabulation
- Histograms
 - 一個histogram是用來顯示一些事物歷史改變的圖。
 - 例如：也許想看每天每個國家的最高溫度，可能會用以下查詢句：
SELECT day, nation, MAX(temp)
FROM Weather
GROUP BY Day(Time) as day,
Nation(Latitude, Longitude) AS nation;

不合法的SQL！

2006資料庫核心理論與實務

主目錄₁₇



Histograms

- 在OLAP上常需針對日期作進一步的分析，也就是必須分出年，季，月，日等。以下所示，但此查詢句的處理會很沒效率。
SELECT day, nation, MAX(temp)
FROM (**SELECT** Day(Time) **AS** day,
Nation(Latitude, Longitude) **AS** nation,
temp
FROM Weather
) **AS** foo
GROUP BY day, nation

2006資料庫核心理論與實務

主目錄₁₈

Roll-up Totals/subtotals for drill-down

■ Roll-up Totals/subtotals for drill-down

- 以一較粗糙的等級列出共同的彙總資料，然後再連續地以較細的等級列出共同的彙總資料。
- 資料表達：See [Table 3a](#), [3b](#) and [Table4](#)。
- 以table4而言，很清楚的表示出很多個欄位。例如：當一個樞在二欄有M和N個值，此結果將導致有NM個屬性。

Roll-up Totals/subtotals for drill-down(Con'd)

Table 3a. Sales Roll Up by Model by Year by Color.

Model	Year	Color	Sales by Model by Year by Color	Sales by Model by Year	Sales by Model
Chevy	1994	Black	50	90	290
		White	40		
	1995	Black	85		
		White	115		
			200		

[回主頁](#)

Roll-up Totals/subtotals for drill-down(Con'd)

Table 3b. Sales Roll-Up by Model by Year by Color as recommended by Chris Date (Date, 1996).

Model	Year	Color	Sales	Sales by Model by Year	Sales by Model
Chevy	1994	Black	50	90	290
Chevy	1994	White	40	90	290
Chevy	1995	Black	85	200	290
Chevy	1995	White	115	200	290

[回主頁](#)

[主目錄](#) 21

2006資料庫核心理論與實務

Roll-up Totals/subtotals for drill-down(Con'd)

Table 4. An Excel pivot table representation of Table 3 with Ford sales data included.

Sum sales Model	Year/Color						Grand total
	1994		1994 total	1995		1995 total	
	Black	White		Black	White		
Chevy	50	40	90	85	115	200	290
Ford	50	10	60	85	75	160	220
Grand total	100	50	150	170	190	360	510

[回主頁](#)


Roll-up Totals/subtotals for drill-down(Con'd)

- 另一個簡單的roll up範例 [Table5a](#)。
- Table5a能夠被以一個複雜的SQL構成。如 [page36](#)所示。

Roll-up Totals/subtotals for drill-down(Con'd)

Table 5a. Sales summary.

	Model	Year	Color	Units
White	Chevy	1994	Black	50
Total	Chevy	1994	White	40
	Chevy	1994	ALL	90
Black	Chevy	1995	Black	85
White	Chevy	1995	White	115
Total	Chevy	1995	ALL	200
Black	Chevy	ALL	ALL	290



Roll-up Totals/subtotals for drill-down(Con'd)

```
SELECT 'ALL', 'ALL', 'ALL', SUM(Sales)
FROM Sales
WHERE Model = 'Chevy'
UNION
SELECT Model, 'ALL', 'ALL', SUM(Sales)
FROM Sales
WHERE Model = 'Chevy'
GROUP BY Model
UNION
SELECT Model, Year, 'ALL', SUM(Sales)
FROM Sales
WHERE Model = 'Chevy'
GROUP BY Model, Year
UNION
SELECT Model, Year, Color, SUM(Sales)
FROM Sales
WHERE Model = 'Chevy'
GROUP BY Model, Year, Color;
```

This is a simple 3-dimensional roll-up. Aggregating over N dimensions requires N such unions.

2006資料庫核心理論與實務

[回主頁](#) [主目錄](#) 25



Cross-Tabulation (cross tab)

- Cross-Tabulation (cross tab)
 - Rollup是不對稱的。對稱的彙總結果被稱之為cross-tabulation。
 - See [Table 5b](#), [6a](#) and [6b](#).
 - 因為rollup和cross-tab是如此的重要，最好用新的SQL運算子，因為這樣能夠使查詢最佳化更有效率。

2006資料庫核心理論與實務

[主目錄](#) 26



Rollup

Table 5a. Sales summary.

Model	Year	Color	Units
Chevy	1994	Black	50
Chevy	1994	White	40
Chevy	1994	ALL	90
Chevy	1995	Black	85
Chevy	1995	White	115
Chevy	1995	ALL	200
Chevy	ALL	ALL	290

2006資料庫核心理論與實務

[回主頁](#) [主目錄](#)

27



Cross Tabulation

Table 5b. Sales summary rows missing from Table 5a to convert the roll-up into a cube.

Model	Year	Color	Units
Chevy	ALL	Black	135
Chevy	ALL	White	155

2006資料庫核心理論與實務

28

Cross-Tabulation (cross tab)(Cont'd)

Table 6a. Chevy sales cross tab.

Chevy	1994	1995	Total (ALL)
Black	50	85	135
White	40	115	155
Total (ALL)	90	200	290

Table 6b. Ford sales cross tab.

Ford	1994	1995	Total (ALL)
Black	50	85	135
White	10	75	85
Total (ALL)	60	160	220

2006資料庫核心理論與實務

[回主頁](#) [主目錄](#)

29

CUBE and ROLLUP

■ CUBE and ROLLUP (SQL99)

- 對於cube的運算子，以下的查詢能夠被構成。
SELECT day, nation, MAX(Temp)
FROM Weather
GROUP BY CUBE
Day(Time) AS day,
Country(Latitude, Longitude) AS nation;
- 假如N個屬性的屬性值有C1,C2,...,C3個，則cube的結果會有(C1+1)(C2+1)...(Cn+1)筆記錄。
- 為支援rollup和drill-down，SQL99提供另一個運算子ROLLUP。ROLLUP在屬性上適合用線性函數關係的應用程式；CUBE則適用在獨立屬性的應用程式上。例如：也許想要在 month, and year做rollup的動作。

2006資料庫核心理論與實務

[主目錄](#)

30

Combining CUBE and ROLLUP

Combining CUBE and ROLLUP

- 語法：

```
GROUP BY [<aggregation list>]
         [ROLLUP <aggregation list>]
         [CUBE <aggregation list>]
```

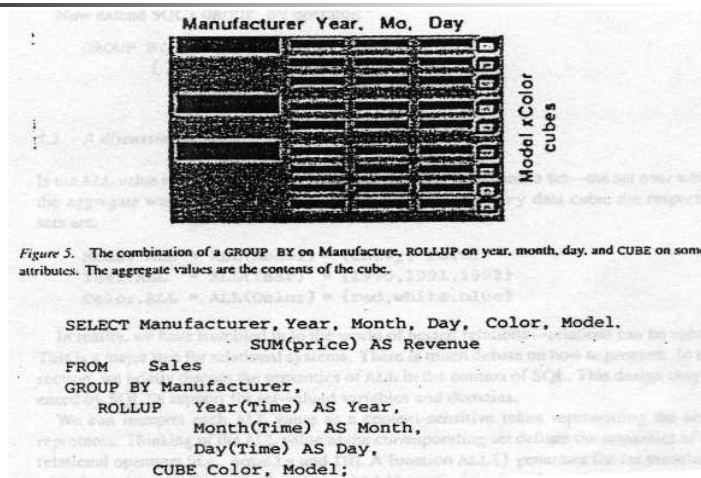
- 例如：下圖可看出以下查詢的結果。

```
SELECT Manufacturer, Year, Month, Day, Color, Model,
       SUM(price) AS Revenue
FROM Sales
GROUP BY Manufacturer
       ROLLUP Year(Time) AS Year,
              Month(Time) AS Month,
              Day(Time) AS Day,
       CUBE Color, Model;
```

2006資料庫核心理論與實務

主目錄 31

Combining CUBE and ROLLUP(Con'd)



2006資料庫核心理論與實務

主目錄 32



WINDOW Frame

- 可以替分群後的每一群設定較大的範圍（成為WINDOW），以方便獲得所要的結果

```
SELECT day, nation, AVERAGE(Temp) OVER W
FROM Weather
WINDOW W AS (PARTITION BY nation
ORDER BY day,
ROWS BETWEEN 3 PRECEDING AND
3 FOLLOWING ;
```



OLAP的DBMS的特色

- 支援特殊的OLAP運算子（如CUBE, ROLLUP, WINDOW)
- 在資料儲存結構和查詢處理上加強其效率
 - 暫存某些彙總資料表（稱為Materialized View），使得複雜的彙總查詢處理速度能加快
 - 支援特殊的索引結構，以加快多維條件的處理。
 - R-Tree