



# Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals

Jim Gray    Surajit Chaudhuri    Adam Bosworth    Andrew  
Layman    Don Reichart    Murali Venkatrao    Frank Pellow  
Hamid Pirahesh

October 31+1, 2006

Presented by Michael Lawrence





# L<sup>A</sup>T<sub>E</sub>X Beamer

L<sup>A</sup>T<sub>E</sub>X Beamer

<http://latex-beamer.sourceforge.net/>



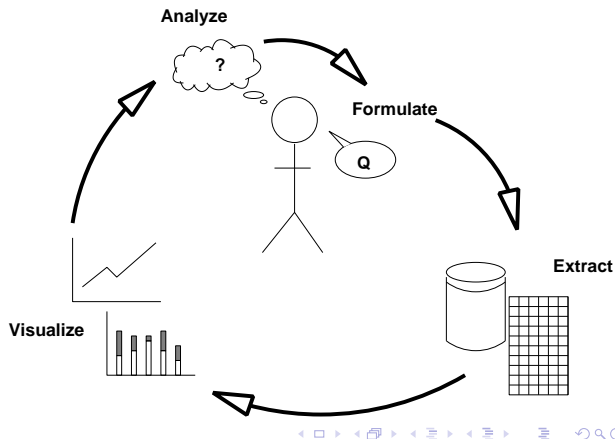
# Outline

- 1 Data Analysis
- 2 Analysis Operations
  - Histograms
  - Roll-Up/Drill-Down
  - Cross-Tab
- 3 The Data Cube
  - The ALL Dummy Value
  - The CUBE Operator

# Data Analysis

## Key features

- Dimensionality reduction
- Aggregation





## Example Data

attrs (dims)			measurements
Model	Year	Colour	Num_Sold
Golf	2001	Black	40
Golf	2001	Silver	65
Golf	2002	Black	70
Golf	2002	Silver	100
Jetta	2001	Black	70
Jetta	2001	Silver	50
Jetta	2002	Black	60
Jetta	2002	Silver	80



## Discussion 1: Multi-dimensional aggregation

- How useful is multi-dimensional aggregation?
- Besides the data warehousing applications mentioned in the paper, can you think of any other applications?



# Problems with SQL GROUP BY

Common analysis operations awkward

- Histograms
- Roll-Up/Drill-Down
- Cross-Tabulations



# Histograms

Categorize models by type

Model	Type
Golf	
Rabbit	Hatchback
⋮	
Jetta	
Passat	Sedan
⋮	





# Histograms

Type	Year	Colour	Num_Sold
Hatchback	2001	Black	40
Hatchback	2001	Silver	65
⋮			
Sedan	2001	Black	70
Sedan	2001	Silver	50
⋮			

Problem: not directly supported by SQL



# Histograms

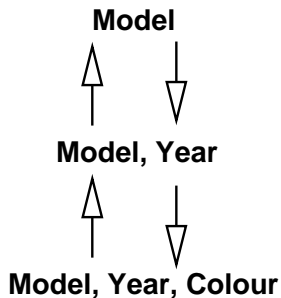
Type	Year	Colour	Num_Sold
Hatchback	2001	Black	40
Hatchback	2001	Silver	65
⋮			
Sedan	2001	Black	70
Sedan	2001	Silver	50
⋮			

Problem: not directly supported by SQL



# Roll-Up/Drill-Down

View data at decreasing/increasing levels of granularity





# Roll-Up/Drill-Down (1)

Model	Year	Colour	Num_MCY	Num_MY	Num_M
Golf	2001	Black	40		
		Silver	65		
				105	
	2002	Black	70		
		Silver	100		
				170	
				275	

Problem: not relational



## Roll-Up/Drill-Down (1)

Model	Year	Colour	Num_MCY	Num_MY	Num_M
Golf	2001	Black	40		
		Silver	65		
				105	
	2002	Black	70		
Silver		100			
			170		
					275

Problem: **not relational**



## Roll-Up/Drill-Down (2)

Chris Date (1996)

Model	Year	Colour	Num_Sold	Num_MY	Num_M
Golf	2001	Black	40	105	275
Golf	2001	Silver	65	105	275
Golf	2002	Black	70	170	275
Golf	2002	Silver	100	170	275

Problem: *2D* columns



## Roll-Up/Drill-Down (2)

Chris Date (1996)

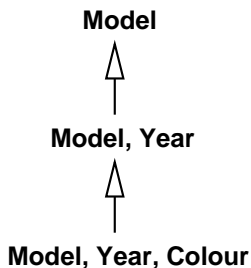
Model	Year	Colour	Num_Sold	Num_MY	Num_M
Golf	2001	Black	40	105	275
Golf	2001	Silver	65	105	275
Golf	2002	Black	70	170	275
Golf	2002	Silver	100	170	275

Problem: *2D* columns

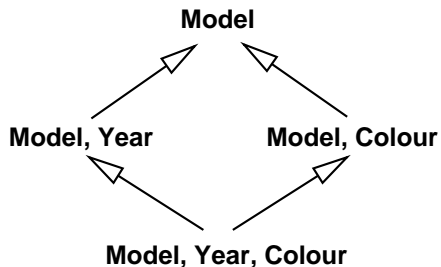


# Cross-Tab

## Roll-Up



## Cross Tab







# Pivot Table (Excel)

Model	2001		2001 Total	2002		2002 Total	Grand Total
	Black	Silver		Black	Silver		
Golf	40	65	105	70	100	170	275
Jetta	70	50	120	60	80	140	260
Total	110	115	225	130	180	310	535

Problem:  $N \times M$  values



# Pivot Table (Excel)

## Model, Colour, Year

Model	2001		2001 Total	2002		2002 Total	Grand Total
	Black	Silver		Black	Silver		
Golf	40	65	105	70	100	170	275
Jetta	70	50	120	60	80	140	260
Total	110	115	225	130	180	310	535

Problem:  $N \times M$  values



# Pivot Table (Excel)

Model, Year

Model	2001		2001 Total	2002		2002 Total	Grand Total
	Black	Silver		Black	Silver		
Golf	40	65	105	70	100	170	275
Jetta	70	50	120	60	80	140	260
Total	110	115	225	130	180	310	535

Problem:  $N \times M$  values



# Pivot Table (Excel)

## Colour, Year

Model	2001		2001 Total	2002		2002 Total	Grand Total
	Black	Silver		Black	Silver		
Golf	40	65	105	70	100	170	275
Jetta	70	50	120	60	80	140	260
Total	110	115	225	130	180	310	535

Problem:  $N \times M$  values



# Pivot Table (Excel)

## Model

Model	2001		2002			Grand Total	
	Black	Silver	2001 Total	Black	Silver		2002 Total
Golf	40	65	105	70	100	170	275
Jetta	70	50	120	60	80	140	260
Total	110	115	225	130	180	310	535

Problem:  $N \times M$  values



# Pivot Table (Excel)

Year

Model	2001		2002			Grand Total	
	Black	Silver	2001 Total	Black	Silver		2002 Total
Golf	40	65	105	70	100	170	275
Jetta	70	50	120	60	80	140	260
Total	110	115	225	130	180	310	535

Problem:  $N \times M$  values



# Pivot Table (Excel)

Total

Model	2001		2002			Grand Total	
	Black	Silver	2001 Total	Black	Silver		2002 Total
Golf	40	65	105	70	100	170	275
Jetta	70	50	120	60	80	140	260
Total	110	115	225	130	180	310	535

Problem:  $N \times M$  values



# Pivot Table (Excel)

Model	2001		2001 Total	2002		2002 Total	Grand Total
	Black	Silver		Black	Silver		
Golf	40	65	105	70	100	170	275
Jetta	70	50	120	60	80	140	260
Total	110	115	225	130	180	310	535

Problem:  $N \times M$  values





## Solution: Overload Column Values

Add special ALL value

Model	Year	Colour	Num_Sold
Golf	2001	Black	40
Golf	2001	Silver	65
Golf	2001	ALL	105
Golf	2002	Black	70
Golf	2002	Silver	100
Golf	2002	ALL	175
Golf	ALL	ALL	275

SQL: UNION of GROUP-BYs on ALL



## Solution: Overload Column Values

Add special ALL value

Model	Year	Colour	Num_Sold
Golf	2001	Black	40
Golf	2001	Silver	65
Golf	2001	ALL	105
Golf	2002	Black	70
Golf	2002	Silver	100
Golf	2002	ALL	175
Golf	ALL	ALL	275

**SQL:** UNION of GROUP-BYs on ALL



## Solution: Overload Column Values

Add special ALL value

Model	Year	Colour	Num_Sold
Golf	2001	Black	40
Golf	2001	Silver	65
Golf	2001	ALL	105
Golf	2002	Black	70
Golf	2002	Silver	100
Golf	2002	ALL	175
Golf	ALL	ALL	275

SQL: UNION of GROUP-BYs on ALL



## Solution: Overload Column Values

Add special ALL value

Model	Year	Colour	Num_Sold
Golf	2001	Black	40
Golf	2001	Silver	65
Golf	2001	ALL	105
Golf	2002	Black	70
Golf	2002	Silver	100
Golf	2002	ALL	175
Golf	ALL	ALL	275

SQL: UNION of GROUP-BYs on ALL



## Solution: Overload Column Values

Add special ALL value

Model	Year	Colour	Num_Sold
Golf	2001	Black	40
Golf	2001	Silver	65
Golf	2001	ALL	105
Golf	2002	Black	70
Golf	2002	Silver	100
Golf	2002	ALL	175
<b>Golf</b>	<b>ALL</b>	<b>ALL</b>	<b>275</b>

SQL: UNION of GROUP-BYs on ALL



# Is ALL Enough?

- $2D$  UNIONS for Roll-Up/Drill-Down
- $2^D$  UNIONS for Cross-Tab
- Awkward
- Inefficient



# Is ALL Enough?

- $2D$  UNIONS for Roll-Up/Drill-Down
- $2^D$  UNIONS for Cross-Tab
- Awkward
- Inefficient



# Is ALL Enough?

- $2^D$  UNIONS for Roll-Up/Drill-Down
- $2^D$  UNIONS for Cross-Tab
- Awkward
- Inefficient





# Is ALL Enough?

- $2D$  UNIONS for Roll-Up/Drill-Down
- $2^D$  UNIONS for Cross-Tab
- Awkward
- Inefficient



# Is ALL Enough?

- $2D$  UNIONS for Roll-Up/Drill-Down
- $2^D$  UNIONS for Cross-Tab
- Awkward
- Inefficient



## Discussion 2: The CUBE Operator

- How hard did you find it to understand the CUBE operator?
  
- As a query writer, would you feel comfortable using it? Or, would you rather use the "solutions" described in the previous slides?



# The CUBE Operator

```
SELECT <attribute-list> <aggregated-measurements>  
FROM <table>  
WHERE <selection-cons>  
GROUP BY CUBE <attribute-list>
```

Generates equivalent result as UNION of GROUP BY on all subsets of <attribute-list> using ALL.



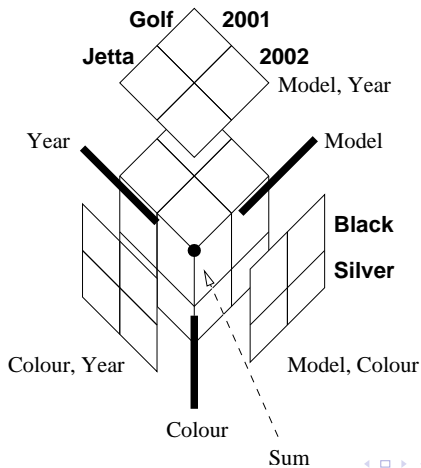
# The CUBE Operator

```
SELECT <attribute-list> <aggregated-measurements>  
FROM <table>  
WHERE <selection-cons>  
GROUP BY CUBE <attribute-list>
```

**Generates equivalent result as UNION of GROUP BY on all subsets of <attribute-list> using ALL.**



# The CUBE Operator





## Discussion 3: Overloading NULL

The authors intended to reduce special-case handling code by “overloading” NULL with the “ALL” value.

- Do you think overloading NULL is a good idea?
- Does this actually reduce special-case handling code, or is special-case handling code still required, but even more difficult to deal with because of the overloading?