Lection 9

Learning Objectives

- Definition of OLAP
- Data cubes
- Cube aggregations and the Cube operator
- OLAP operations
- OLAP servers

What is OLAP?

OLAP has two immediate consequences: online part requires the answers of queries to be fast, the analytical part is a hint that the queries itself are complex

i.e., Complex questions with Fast Answers!

Why OLAP?

- Empowers end users to do own analysis
- Frees up IS backlog of report requests
- Ease of use
- No knowledge of tables or SQL required

From Tables and Spreadsheets to Data Cubes

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - Dimension tables, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - Fact table contains measures (such as dollars_sold) and keys to each of the related dimension tables

Ex: 2-d Cube

Fact table view:

sale	prodId	storeId	amt
	p1	c1	12
	p2	c1	11
	p1	c3	50
	p2	c2	8

Multi-dimensional cube:

	c1	c2	c3
p1	12		50
p2	11	8	

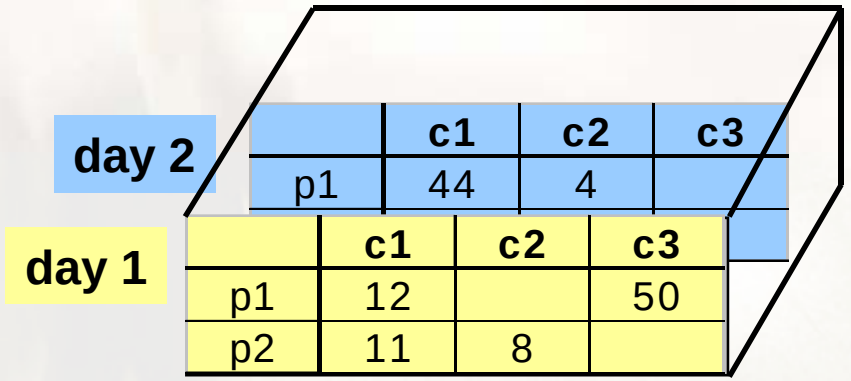
dimensions = 2

Ex: 3-d Cube

Fact table view:

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

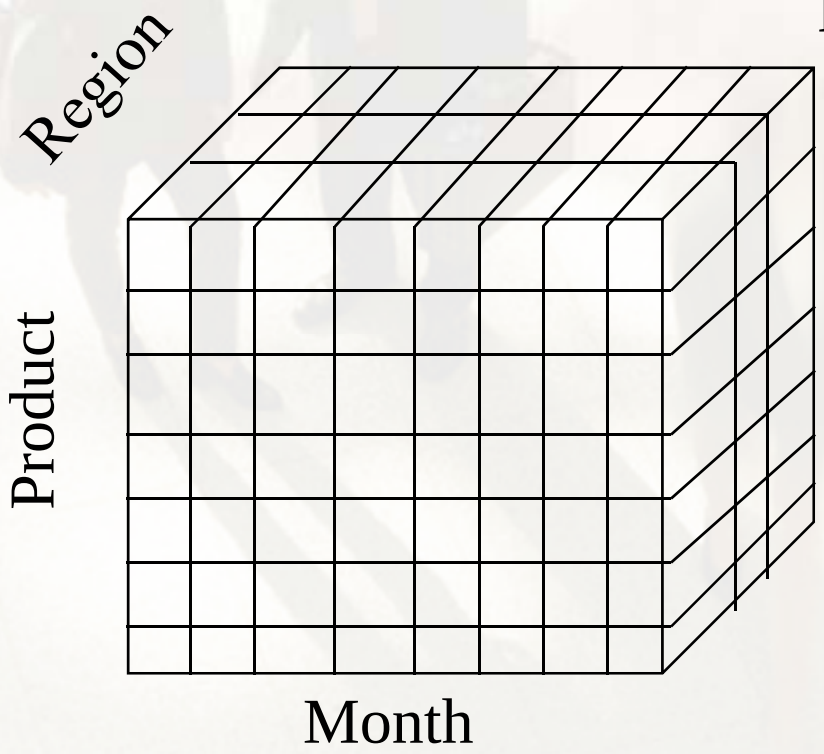
Multi-dimensional cube:



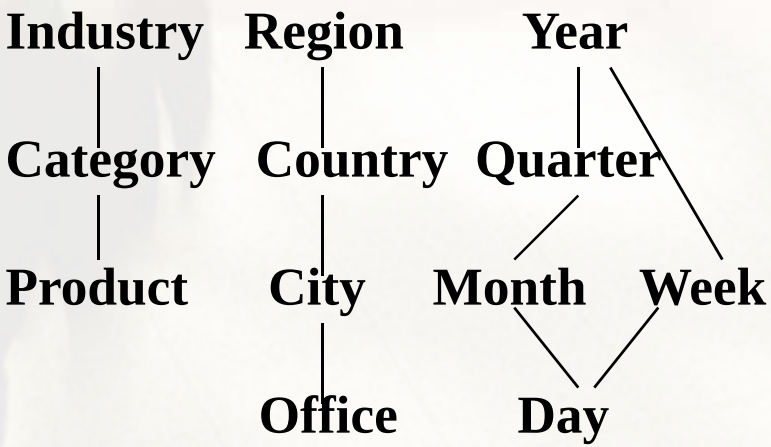
dimensions = 3

Multidimensional Cubes with Hierarchies

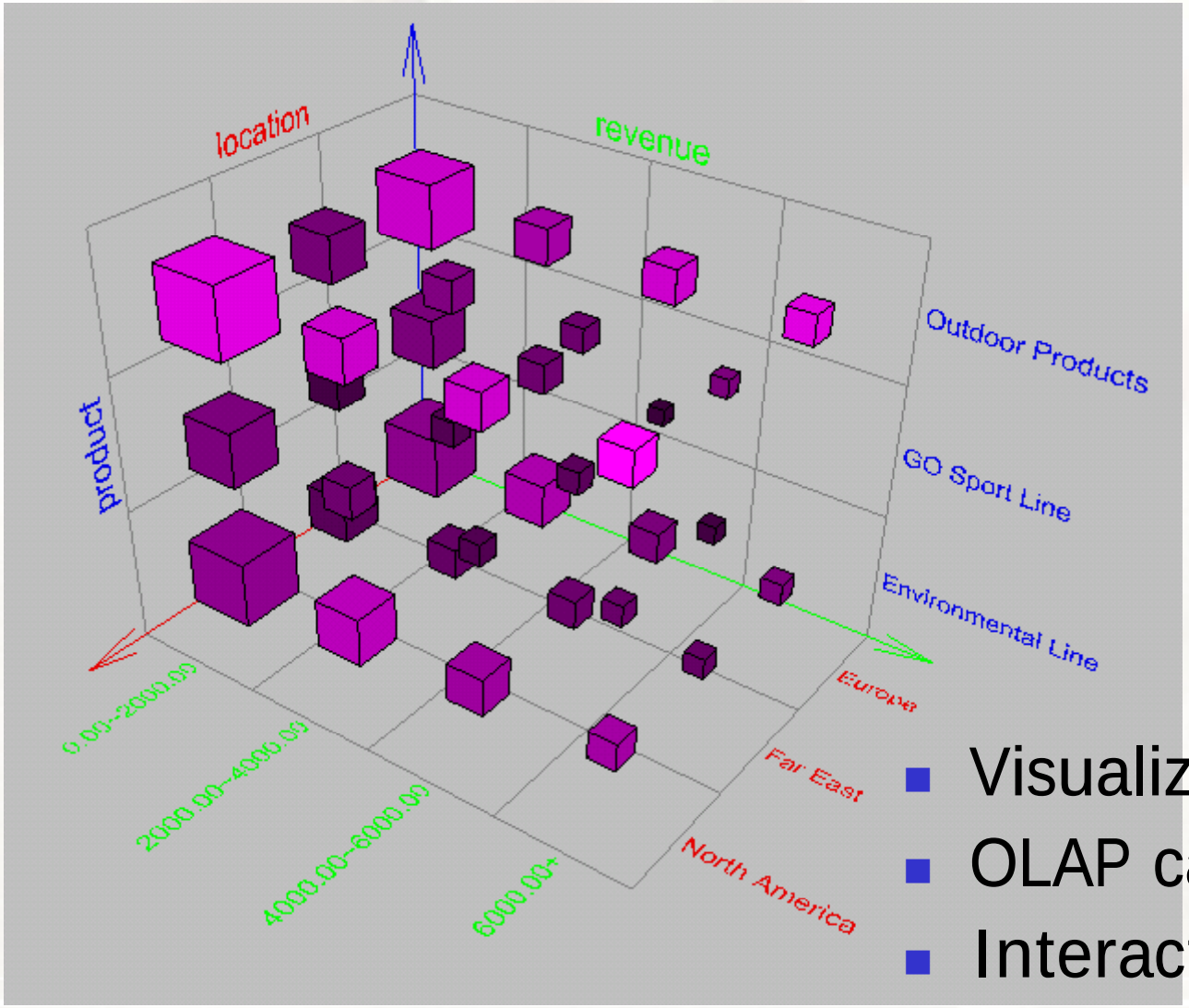
- Sales volume as a function of product, month, and region



Dimensions: Product, Location, Time
Hierarchical summarization paths



Browsing a Data Cube



- Visualization
- OLAP capabilities
- Interactive manipulation

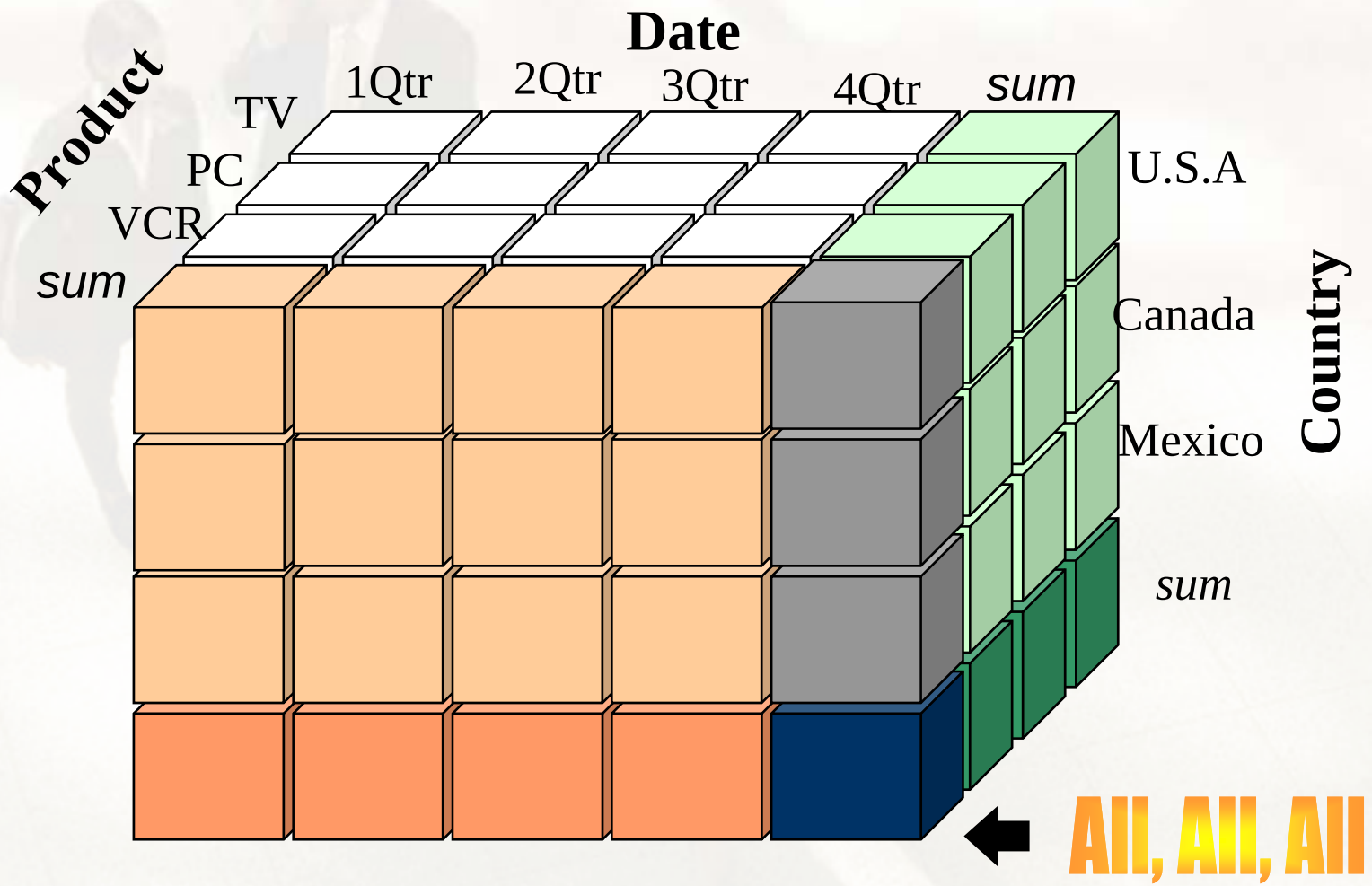
Cube aggregations

- The data cube summarizes the measure with respect to a set of n dimensions and provides summarizations for all subsets of them

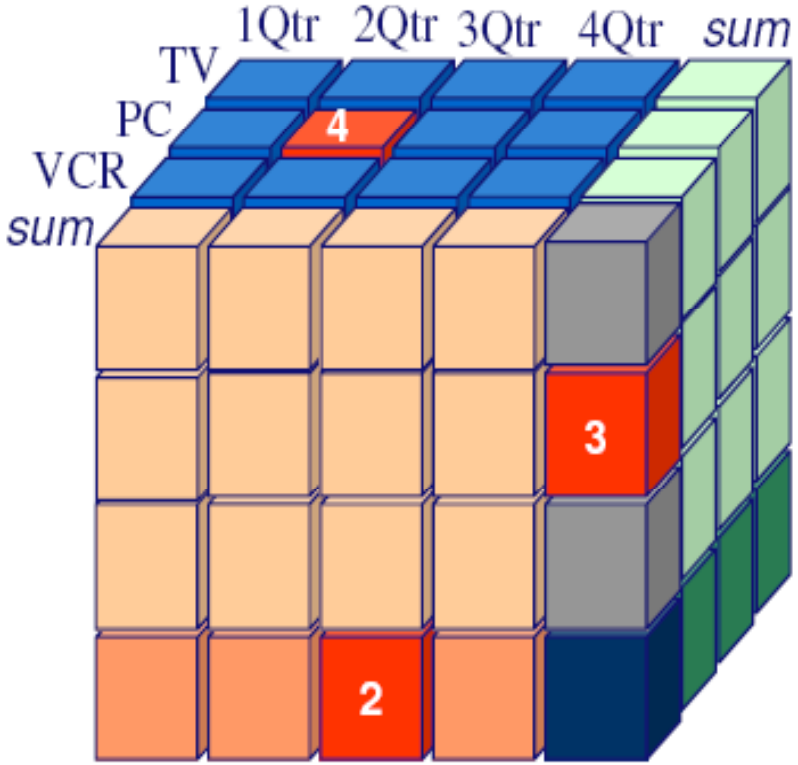
product \ year	1999	2000	2001	2002	ALL
chairs	25	37	89	21	172
tables	10	30	0	45	85
desks	56	84	9	35	184
shelves	19	20	0	71	110
boards	5	16	11	15	47
ALL	115	187	109	187	598

Data cube

Ex: 3-d data cube aggregations



Ex: 3-d data cube aggregations



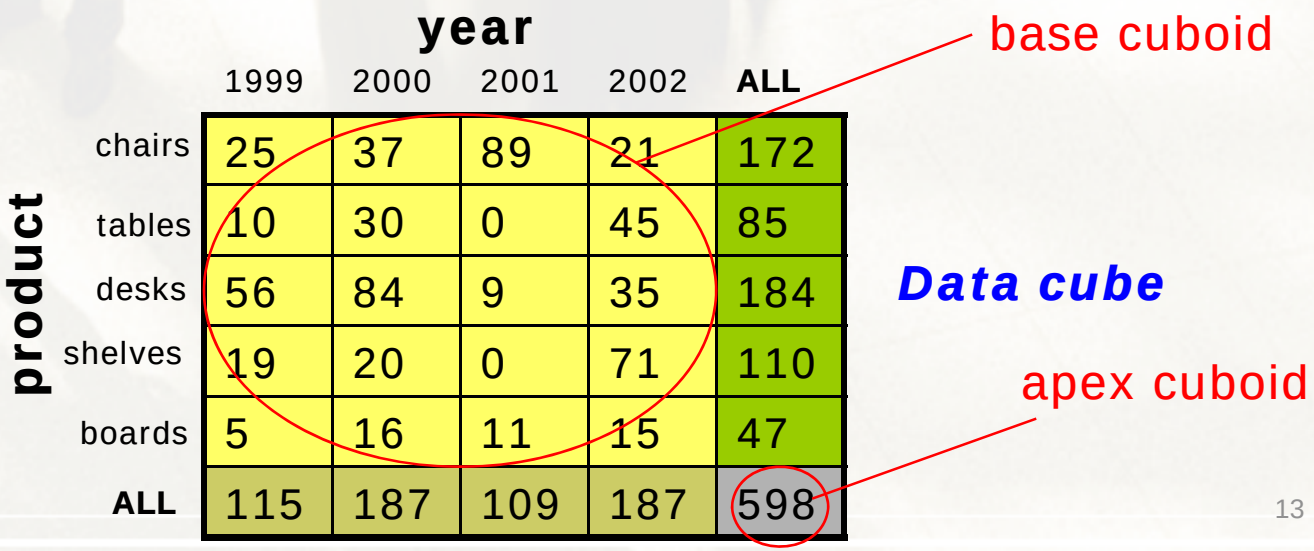
2) Total sales in 3Qtr?

3) #products sold in France this year

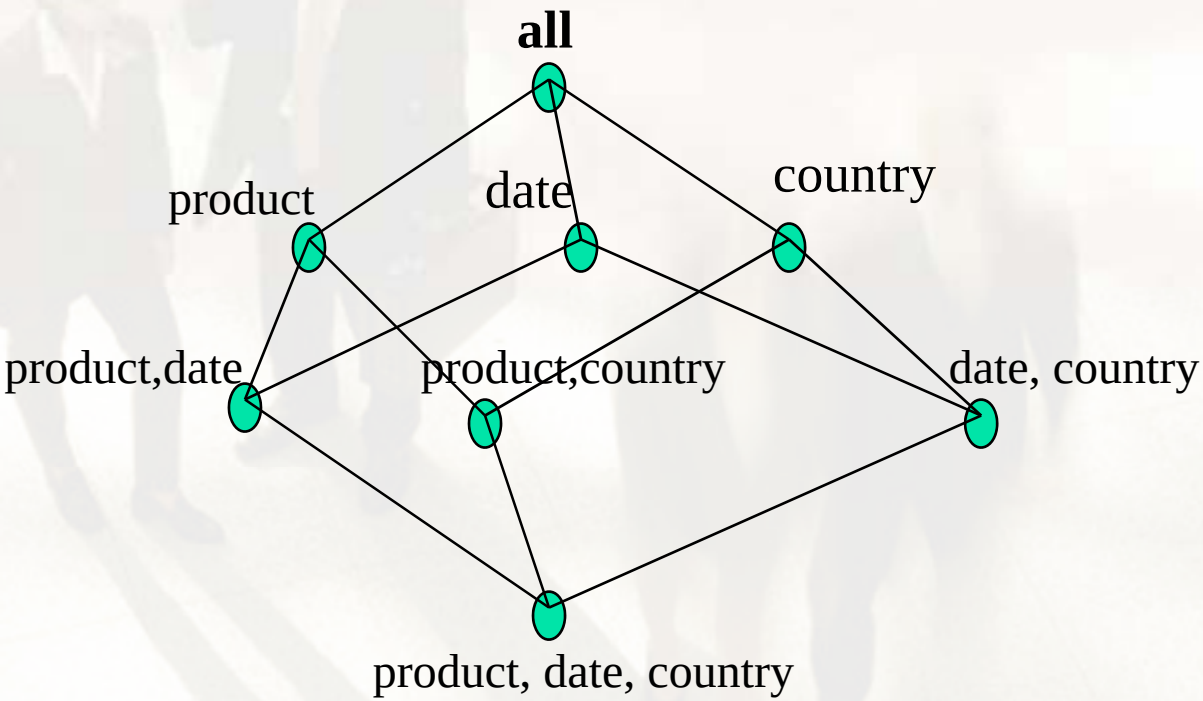
4) #PC's in Ireland in 2Qtr?

Cube as set of cuboids

- In data warehousing literature, the most detailed part of the cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.



Cuboids Corresponding to the Cube



0-D(apex) cuboid

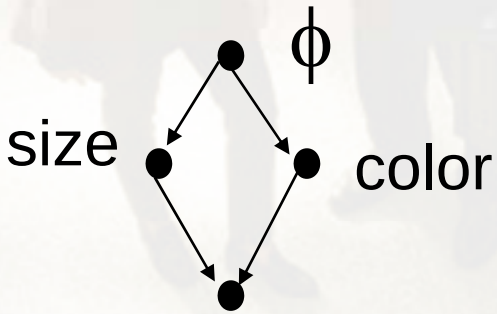
1-D cuboids

2-D cuboids

3-D(base) cuboid

Ex: Cube and cuboids

‘color’, ‘size’: DIMENSIONS
‘count’: MEASURE

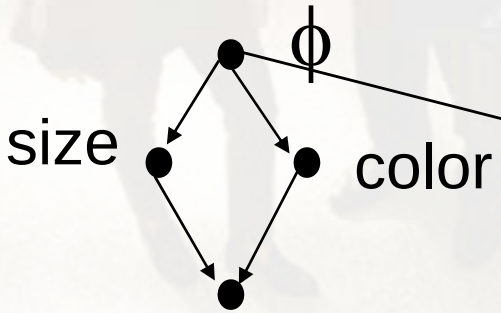


color; size

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

Ex: Cube and cuboids

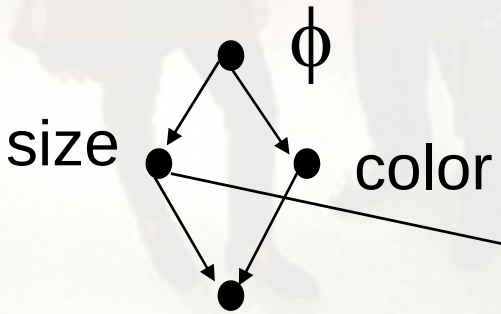
'color', 'size': DIMENSIONS
'count': MEASURE



C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

Ex: Cube and cuboids

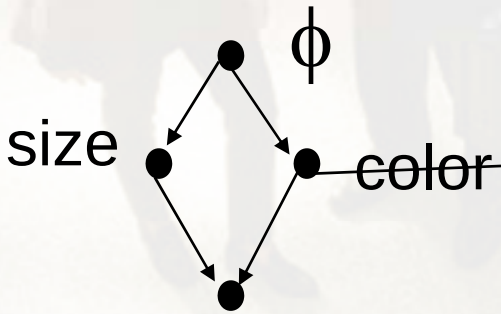
'color', 'size': DIMENSIONS
'count': MEASURE



C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

Ex: Cube and cuboids

'color', 'size': DIMENSIONS
'count': MEASURE



color; size

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

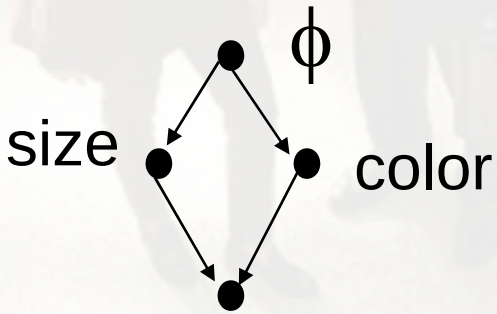
Ex: Cube and cuboids

'color', 'size': DIMENSIONS
'count': MEASURE

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

Ex: Cube and cuboids

'color', 'size': DIMENSIONS
'count': MEASURE



color; size

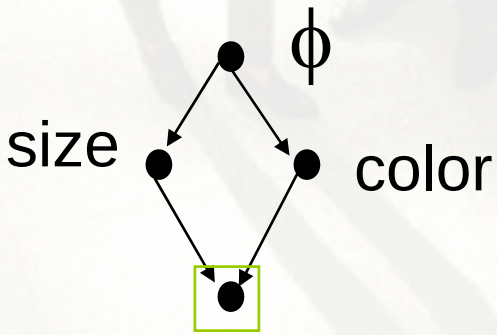
C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

DataCube

Typical OLAP Operations

- Roll up (drill-up): summarize data
  - *by climbing up hierarchy or by dimension reduction*
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- Slice and dice:
  - *project and select*
- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes.*
- Other operations
  - *drill across: involving (across) more than one fact table*
  - *drill through: through the bottom level of the cube to its back-end relational tables (using SQL)*

Example of operations on a Datacube



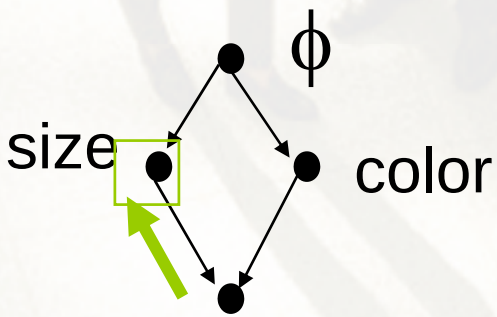
color; size

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

Roll-up

Roll-up:

- In this example we reduce one dimension
- It is possible to climb up one hierarchy
  - Example (product, city) → (product, country)

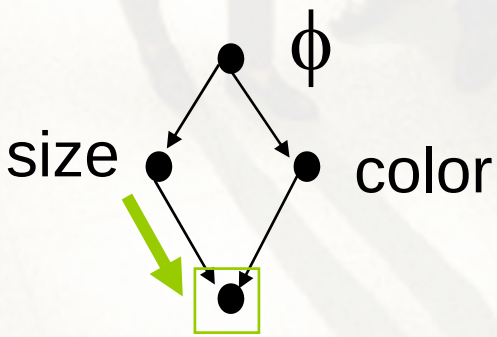


C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

Drill-down

Drill-down

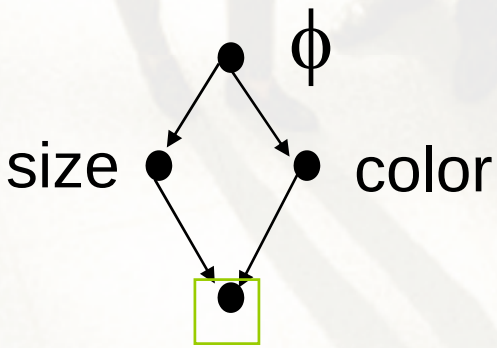
- In this example we add one dimension
- It is possible to climb down one hierarchy
  - Example (product, year) → (product, month)



C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

Slice

Slice: Perform a selection on one dimension

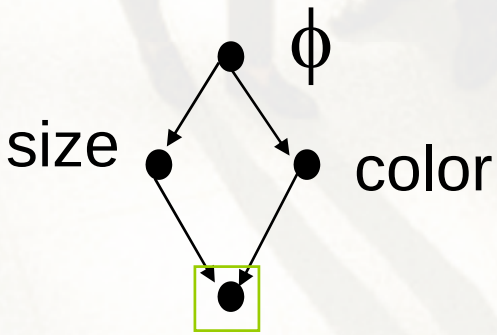


color; size

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

Dice

Dice: Perform a selection on two or more dimensions

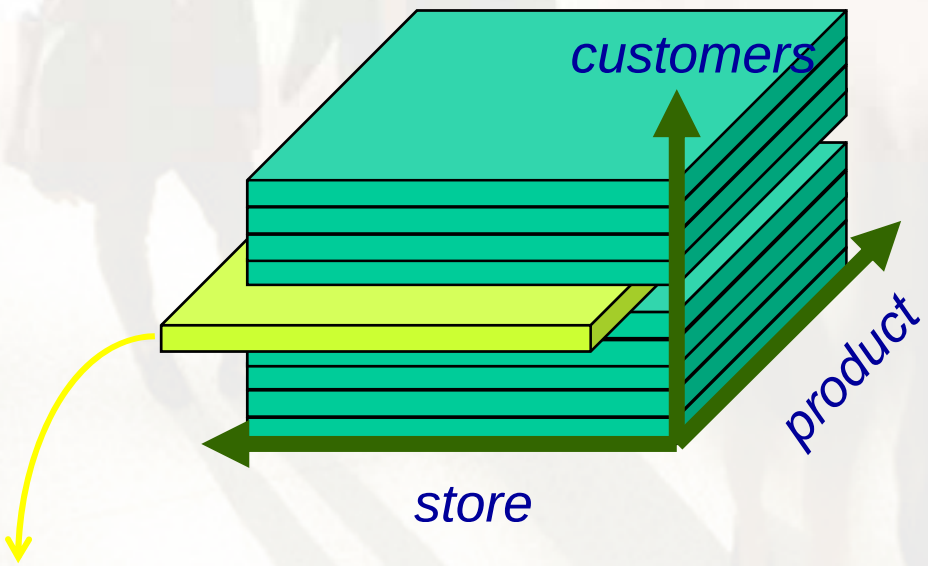


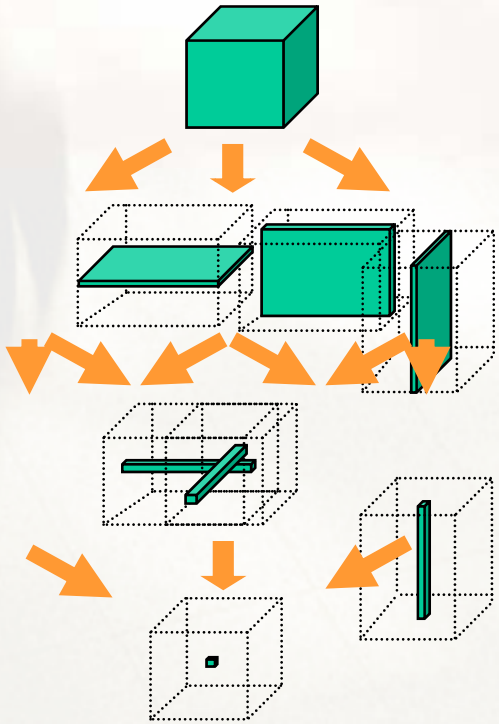
color; size

C / S	S	M	L	TOT
Red	20	3	5	28
Blue	3	3	8	14
Gray	0	0	5	5
TOT	23	6	18	47

Slice/Dice

Easy terms compared to
Select-Where in SQL

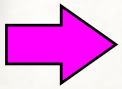




Background: SQL Aggregates

- Add up amounts for day 1
- In SQL: SELECT sum(amt) FROM SALE
  WHERE date = 1

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

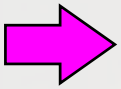


81

Background: SQL Aggregates

- Add up amounts by day
- In SQL: SELECT date, sum(amt) FROM SALE
  GROUP BY date

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

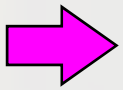


ans	date	sum
	1	81
	2	48

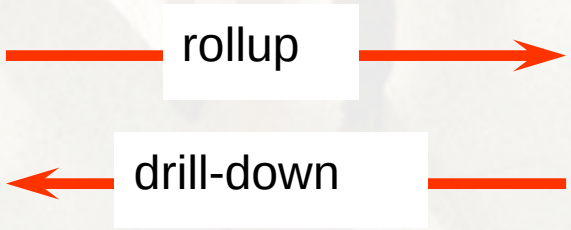
Ex: Roll-up vs. Drill-down

- Add up amounts by day, product
- In SQL: SELECT date, sum(amt) FROM SALE GROUP BY date, prodId

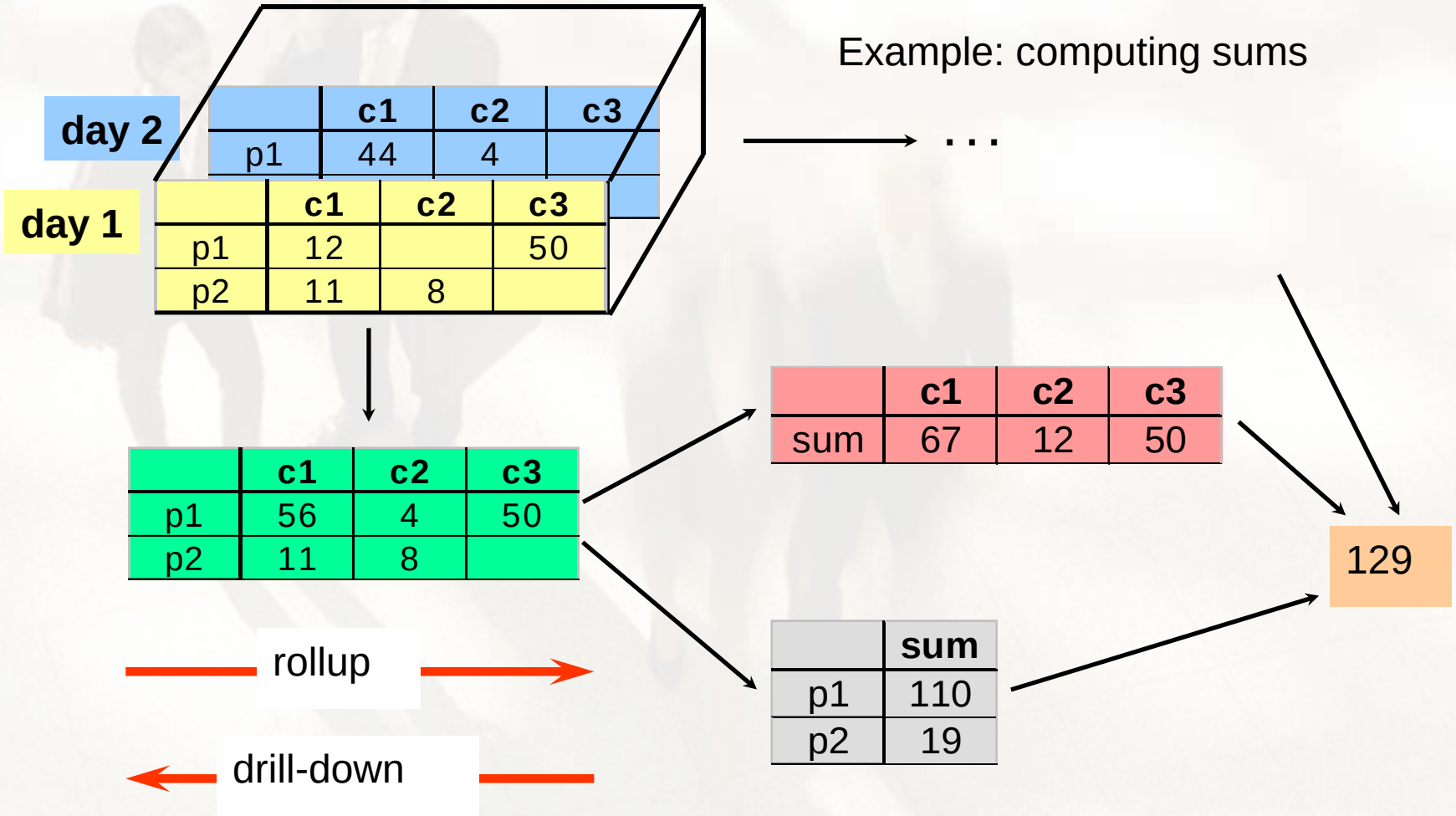
sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4



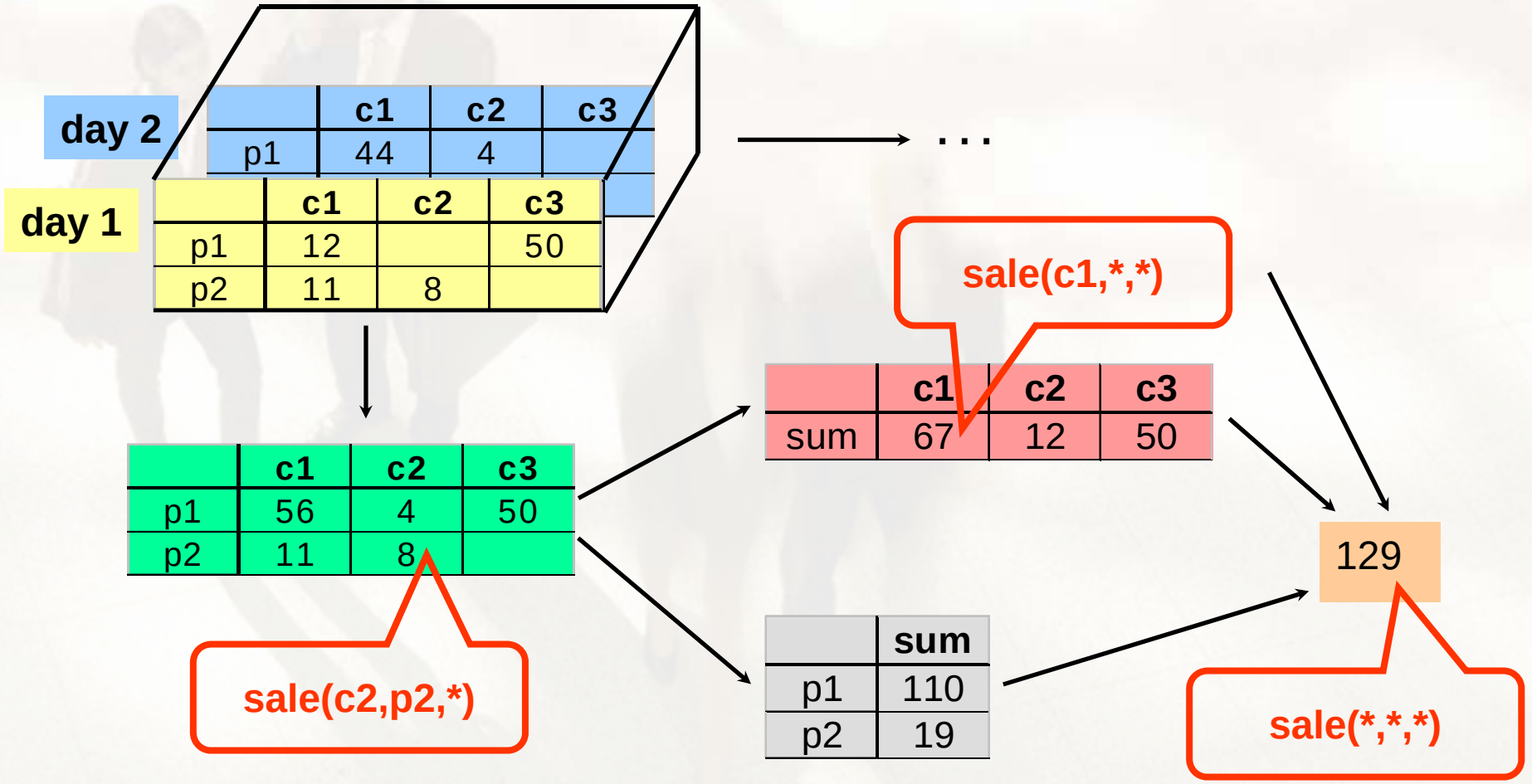
sale	prodId	date	amt
	p1	1	62
	p2	1	19
	p1	2	48



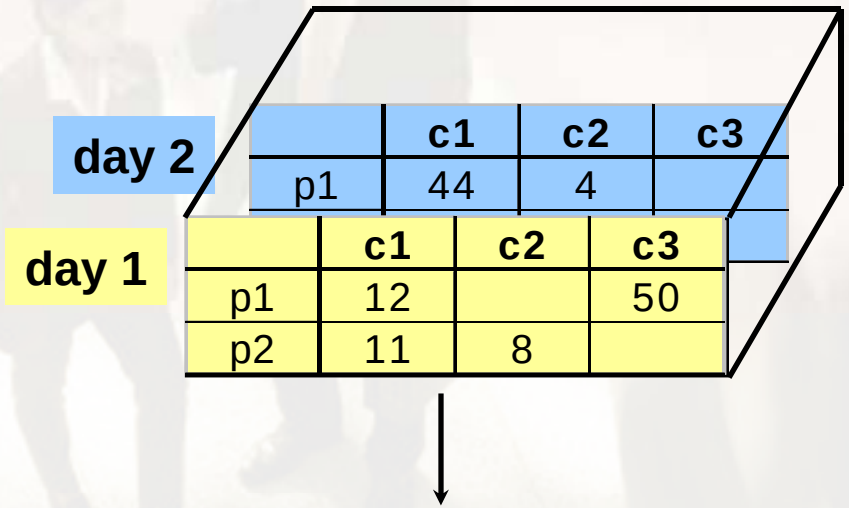
Ex: Roll-up vs. Drill-down



Cube Operators



Don’t forget!
Aggregation Using Hierarchies



	region A	region B
p1	56	54
p2	11	8

customer

region

country

(customer c1 in Region A;
customers c2, c3 in Region B)

The Data Cube Operator (Gray et al)

Cross-Tabulation (products/store)

Sales		Product				
		1	2	3	4	ALL
Store	1	454	-	-	925	1379
	2	468	800	-	-	1268
	3	296	-	240	-	536
	4	652	-	540	745	1937
	ALL	1870	800	780	1670	5120

Sub-totals per store

Total sales

Sub-totals per product

4 Group-bys here:
(store,product)
(store)
(product)
()

Need to write 4 queries!!!

The Data Cube Operator (Gray et al)

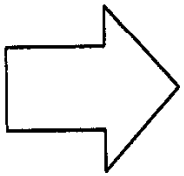
Sales		Product				
		1	2	3	4	ALL
Store	1	454	-	-	925	1379
	2	468	800	-	-	1268
	3	296	-	240	-	536
	4	652	-	540	745	1937
	ALL	1870	800	780	1670	5120

```
SELECT LOCATION.store, SALES.product_key, SUM (amount)
FROM SALES, LOCATION
WHERE SALES.location_key=LOCATION.location_key
CUBE BY SALES.product_key, LOCATION.store
```

Store	Product_key	sum(amout)
1	1	454
1	4	925
2	1	468
2	2	800
3	1	296
3	3	240
4	1	625
4	3	240
4	4	745
1	ALL	1379
1	ALL	1268
1	ALL	536
1	ALL	1937
ALL	1	1870
ALL	2	800
ALL	3	780
ALL	4	1670
ALL	ALL	5120

Ex: 3-d cube operator

Model	Year	Color	Sales
Chevy	1990	Red	5
Chevy	1990	Blue	87
Ford	1990	Green	64
Ford	1990	Blue	99
Ford	1991	Red	8
Ford	1991	Blue	7



Model	Year	Color	Sales
Chevy	1990	Blue	87
Chevy	1990	Red	5
Chevy	1990	ALL	92
Chevy	ALL	Blue	87
Chevy	ALL	Red	5
Chevy	ALL	ALL	92
Ford	1990	Blue	99
Ford	1990	Green	64
Ford	1990	ALL	163
Ford	1991	Blue	7
Ford	1991	Red	8
Ford	1991	ALL	15
Ford	ALL	Blue	106
Ford	ALL	Green	64
Ford	ALL	Red	8
ALL	1990	Blue	186
ALL	1990	Green	64
ALL	1991	Blue	7
ALL	1991	Red	8
Ford	ALL	ALL	178
ALL	1990	ALL	255
ALL	1991	ALL	15
ALL	ALL	Blue	193
ALL	ALL	Green	64
ALL	ALL	Red	13
ALL	ALL	ALL	270

Conceptual vs. Actual

- The “cube” is a logical way of visualizing the data in an OLAP setting
- Not how the data is actually represented on disk
- Two opposite ways of storing data:
  - ROLAP: Relational OLAP
  - MOLAP: Multidimensional OLAP

OLAP Server Architectures

- Relational OLAP (ROLAP)
  - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware to support missing pieces
  - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
  - greater scalability
- Multidimensional OLAP (MOLAP)
  - Array-based multidimensional storage engine (sparse matrix techniques)
  - fast indexing to pre-computed summarized data
- Hybrid OLAP (HOLAP)
  - User flexibility, e.g., low level: relational, high-level: array

Points about MOLAP

- Pre-calculating or pre-consolidating transactional data improves speed.

BUT

Fully pre-consolidating incoming data, MDDs require an enormous amount of overhead both in processing time and in storage. An input file of 200MB can easily expand to 5GB

MDDs are great candidates for the <50GB department data marts.

- Rolling up and Drilling down through aggregate data.

ROLAP vs. MOLAP

- Performance:
  - How fast will the system appear to the end-user?
  - MDD server vendors believe this is a key point in their favor.

- Data volume and scalability:
  - While MDD servers can handle up to 100GB of storage, RDBMS servers can handle hundreds of gigabytes and terabytes.

ROLAP vs. MOLAP

Benefits		MOLAP	ROLAP
User Benefits	Multidimensional View	✓	✓
	Excellent Performance	✓	
	Analytical Flexibility	✓	
	Real-time Data Access		✓
	High Data Capacity		✓
MIS Benefits	Leverages Data Warehouse		✓
	Easy Development	✓	
	Low Structure Maintenance		✓
	Low Aggregate Maintenance	✓	

HOLAP

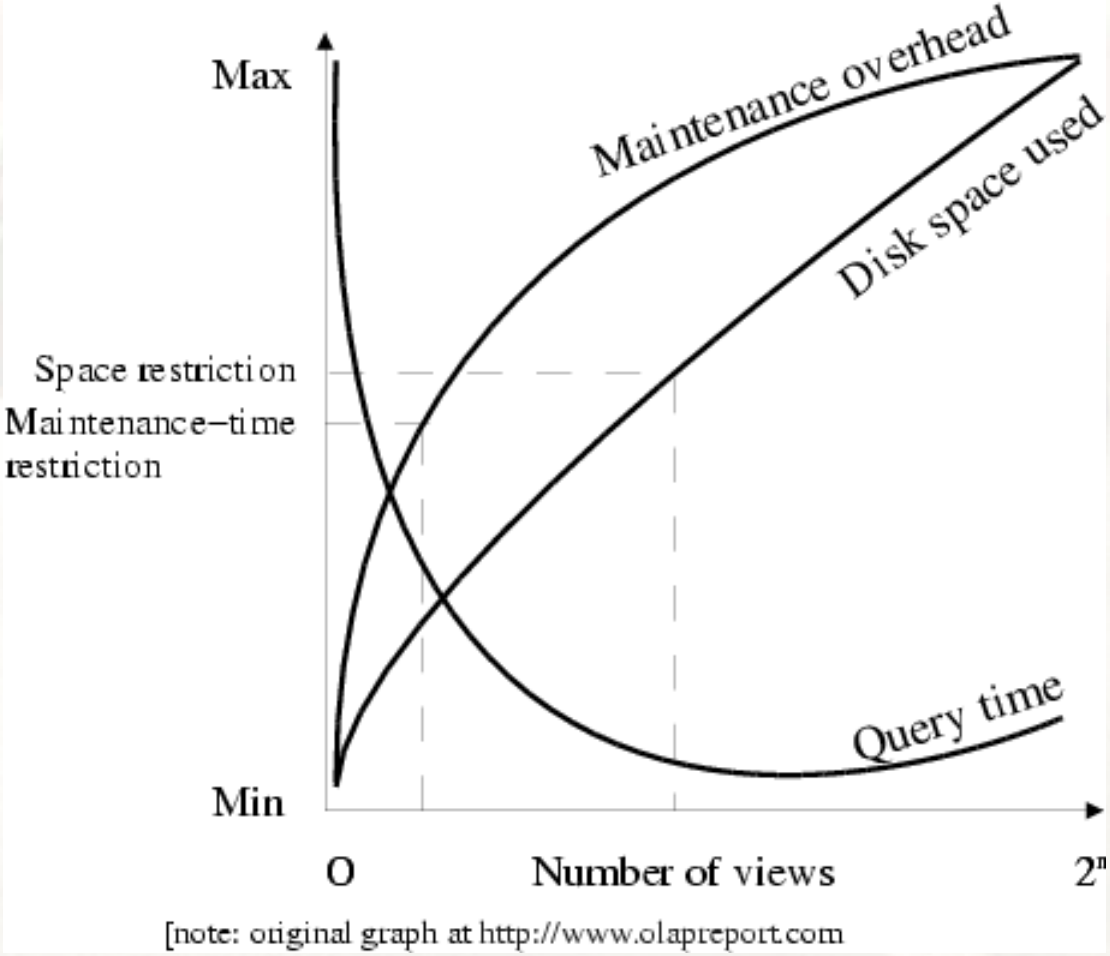
- Best of both worlds
- Storing part of data in RDBMS
- Storing rest data in MDBMS
- User access via MOLAP tools

Efficient Data Cube Computation

- Data cube can be viewed as a lattice of cuboids
  - The bottom-most cuboid is the base cuboid
  - The top-most cuboid (apex) contains only one cell
  - How many cuboids in an n-dimensional cube? **2^n**

- Materialization of data cube
  - Materialize every (cuboid) (full materialization), none (no materialization), or some (partial materialization)
  - Selection of which cuboids to materialize
    - Based on size, sharing, access frequency, etc.

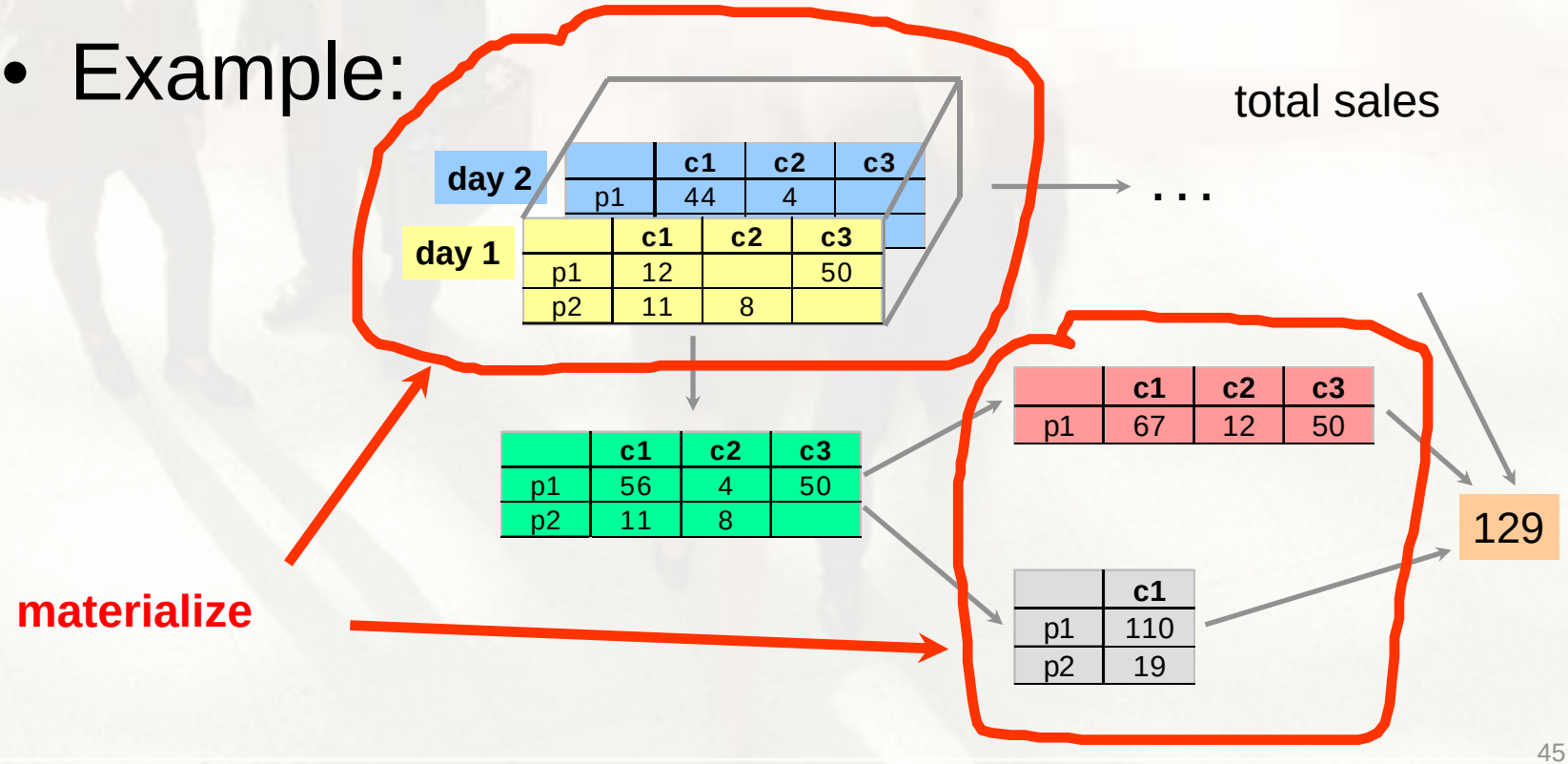
Partial materialization



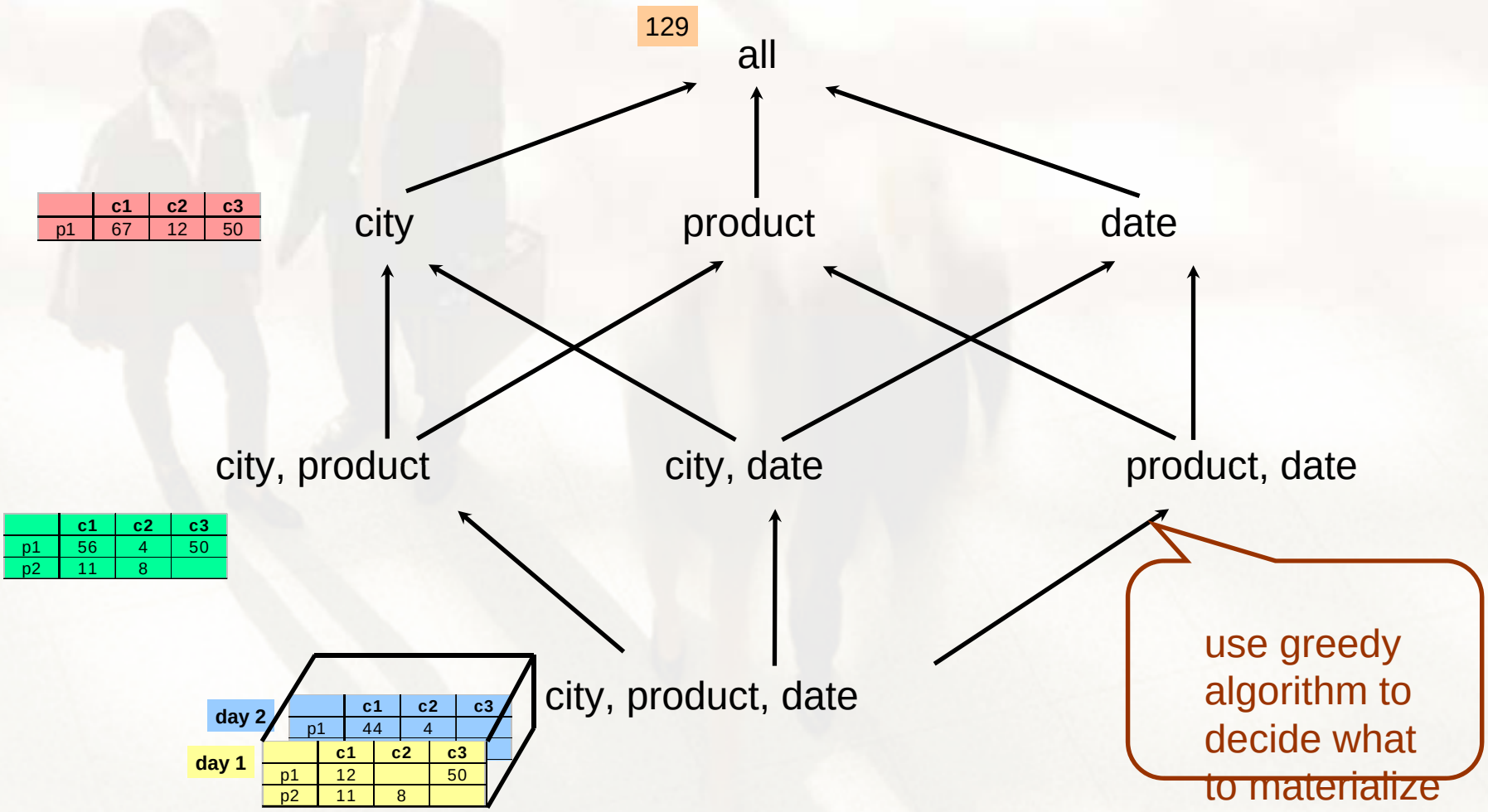
- 2^n views for *n* dimensions (no-hierarchies)
- Storage/update-time explosion
- More pre-computation doesn't mean better performance!!!!

What to Materialize?

- Store in warehouse results useful for common queries
- Example:



What to Materialize?



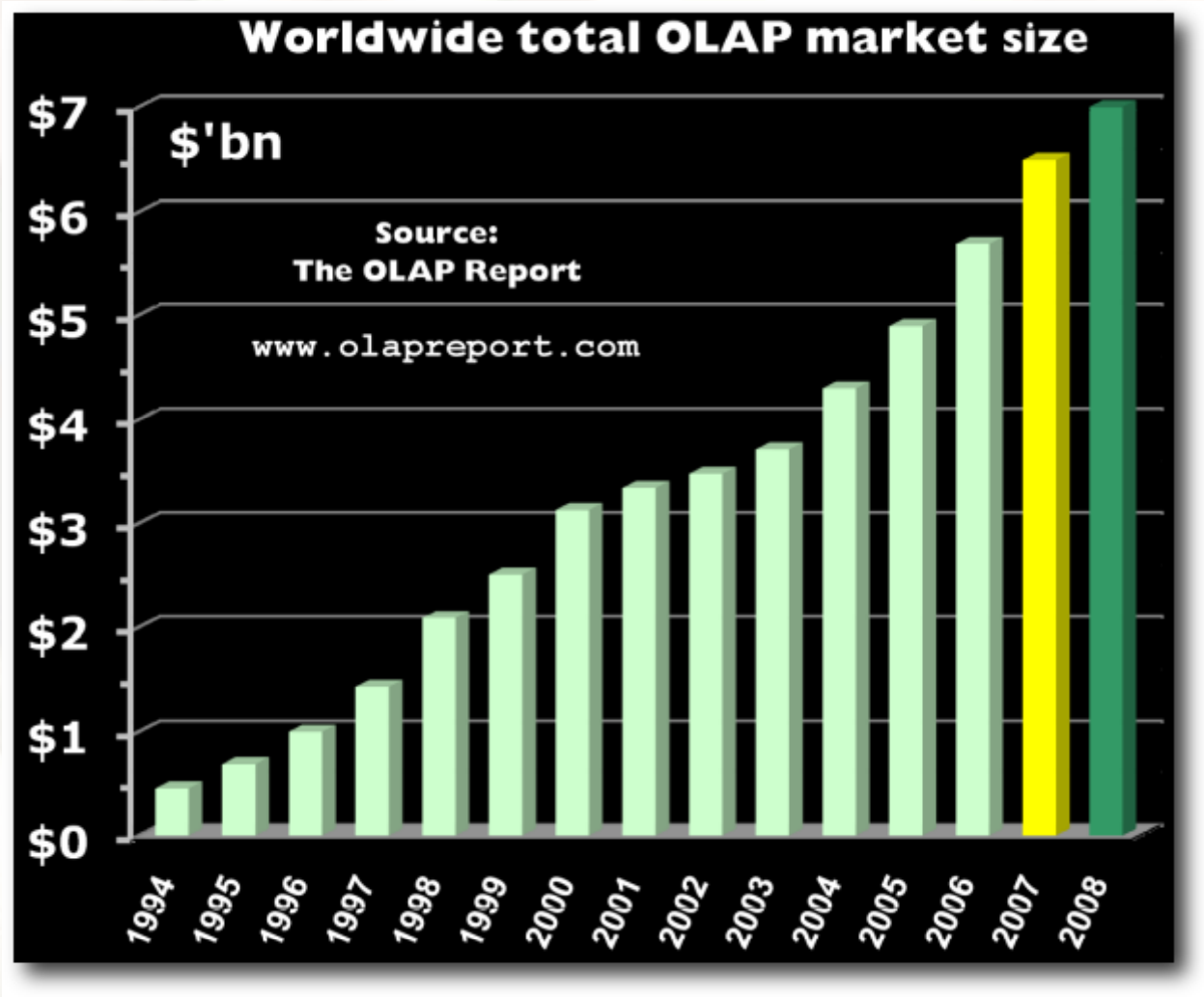
	c1	c2	c3
p1	67	12	50

	c1	c2	c3
p1	56	4	50
p2	11	8	

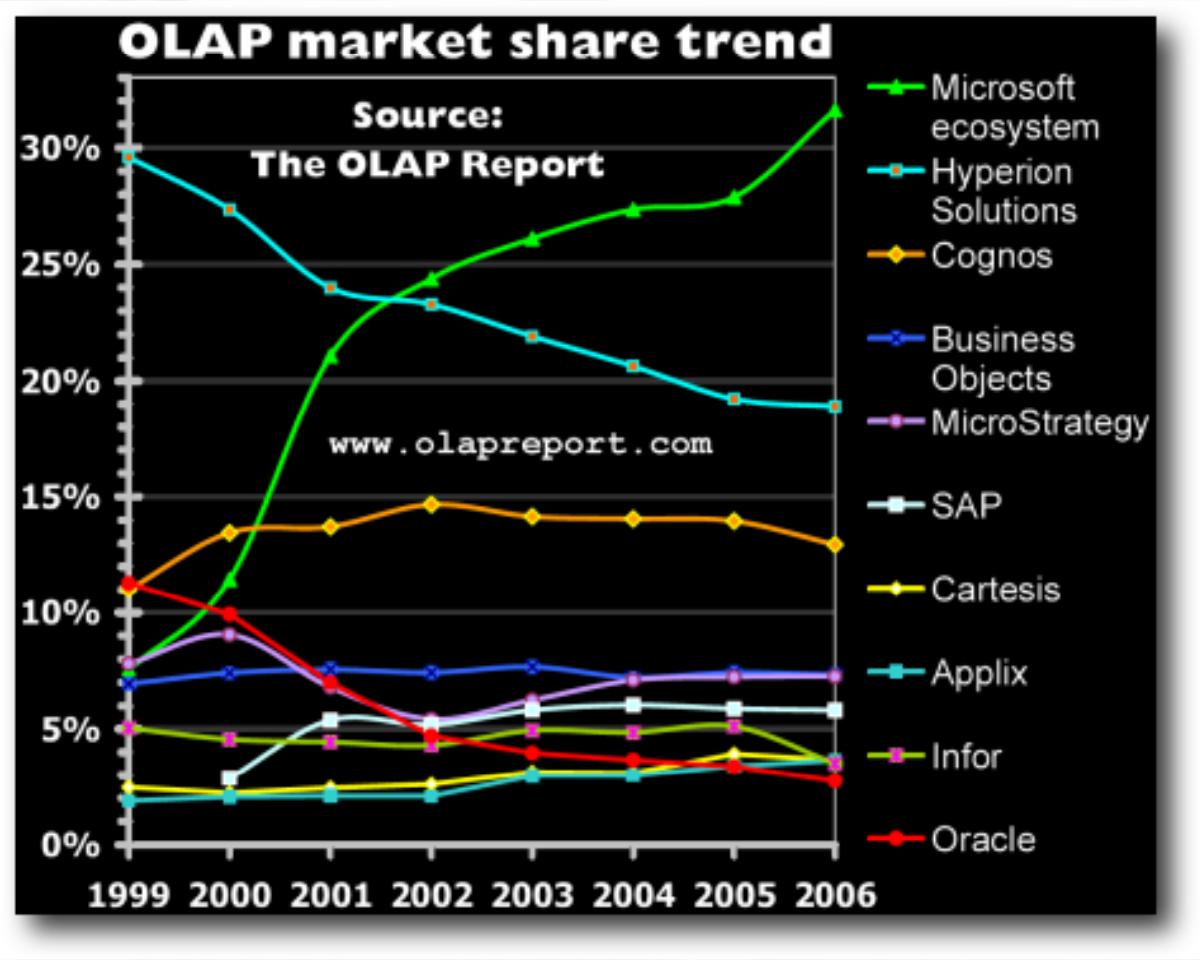
	c1	c2	c3
p1	44	4	

	c1	c2	c3
p1	12		50
p2	11	8	

The OLAP market



The OLAP Market



http://www.bi-verdict.com/fileadmin/FreeAnalyses/market.htm