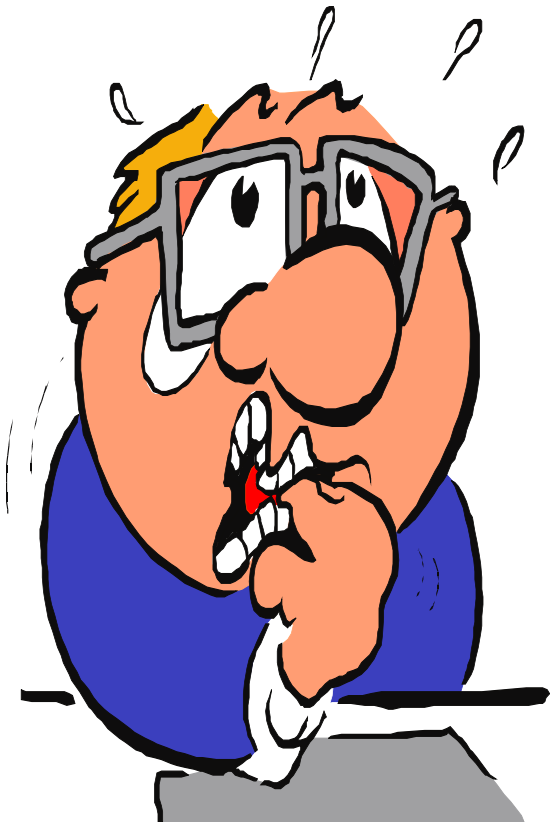
OLAP

-By Ujwala Bharambe

II. On-Line Analytical Processing (OLAP)

Making Decision Support Possible

Limitations of SQL



"A Freshman in Business needs a Ph.D. in SQL"

-- Ralph Kimball

Typical OLAP Queries

- Write a multi-table join to compare sales for each product line YTD this year vs. last year.
- Repeat the above process to find the top 5 product contributors to margin.
- Repeat the above process to find the sales of a product line to new vs. existing customers.
- Repeat the above process to find the customers that have had negative sales growth.

What Is OLAP?

- Online Analytical Processing - coined by EF Codd in 1994 paper contracted by Arbor Software*
- Generally synonymous with earlier terms such as Decisions Support, Business Intelligence, Executive Information System
- OLAP = Multidimensional Database
- MOLAP: Multidimensional OLAP (Arbor Essbase, Oracle Express)
- ROLAP: Relational OLAP (Informix MetaCube, Microstrategy DSS Agent)

* Reference: http://www.arborsoft.com/essbase/wht_ppr/coddTOC.html

Strengths of OLAP

- It is a powerful visualization paradigm
- It provides fast, interactive response times
- It is good for analyzing time series
- It can be useful to find some clusters and outliers
- Many vendors offer OLAP tools

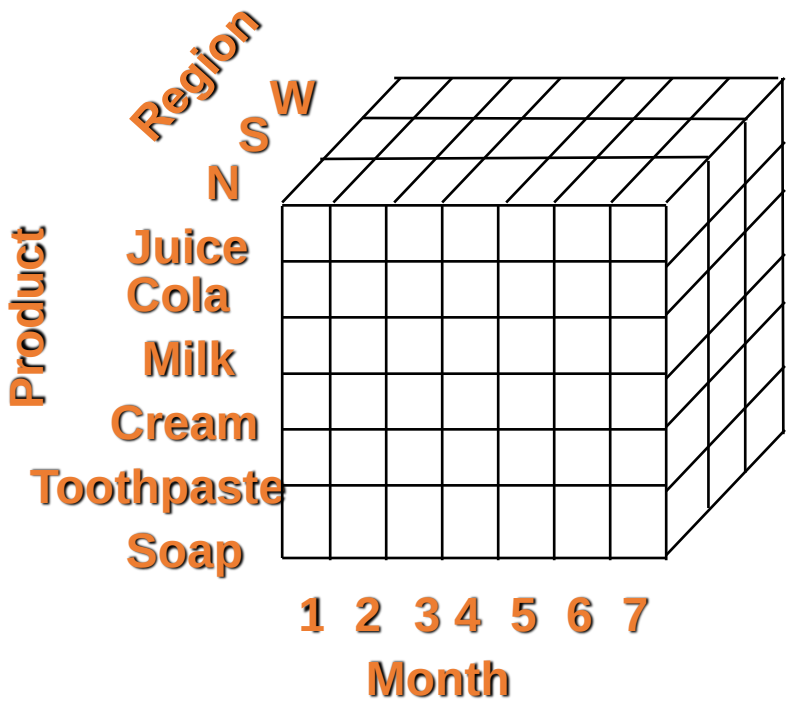
OLAP Is FASMI

- Fast
- Analysis
- Shared
- Multidimensional
- Information

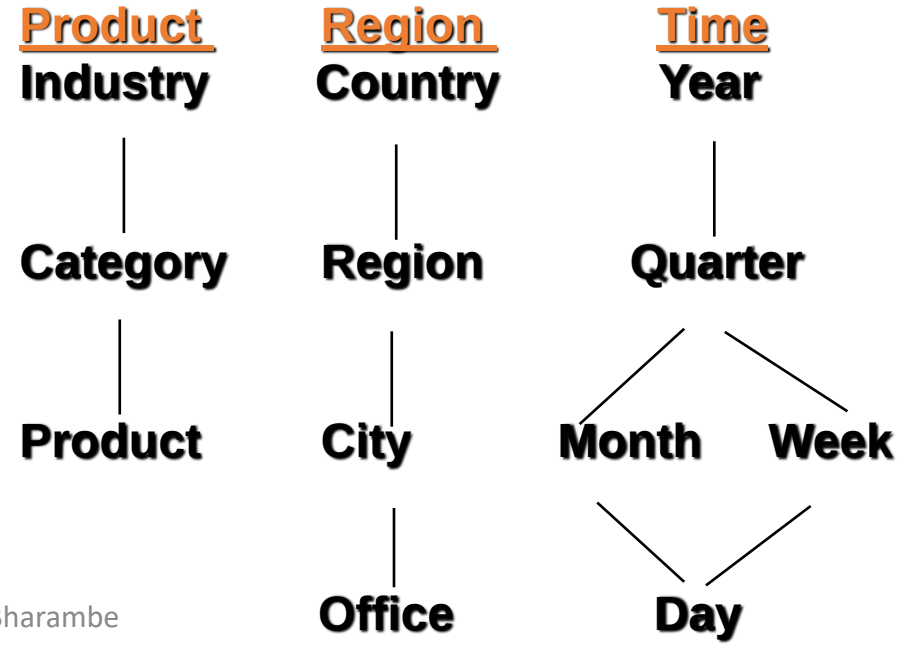
Nigel Pendse, Richard Creath - The OLAP Report

Multi-dimensional Data

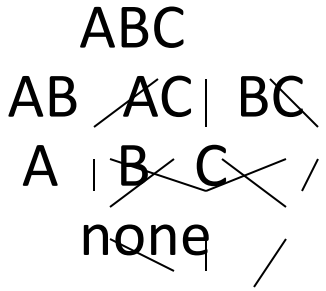
- "Hey…I sold $100M worth of goods"



Dimensions: Product, Region, Time
Hierarchical summarization paths



Data Cube Lattice

- Cube lattice
  - 

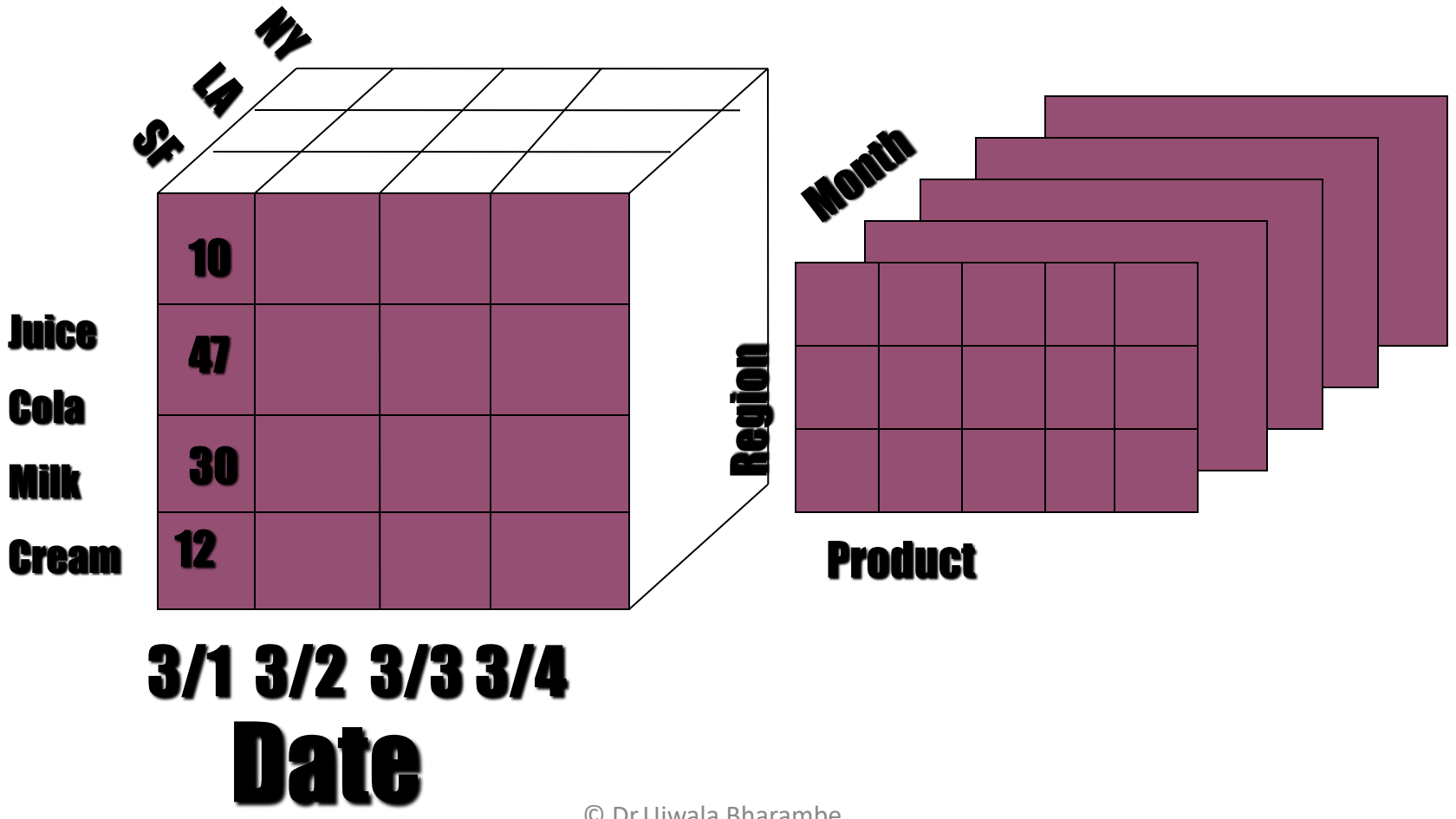
- Can materialize some groupbys, compute others on demand
- Question: which groupbys to materialze?
- Question: what indices to create
- Question: how to organize data (chunks, etc)

Visualizing Neighbors is simpler

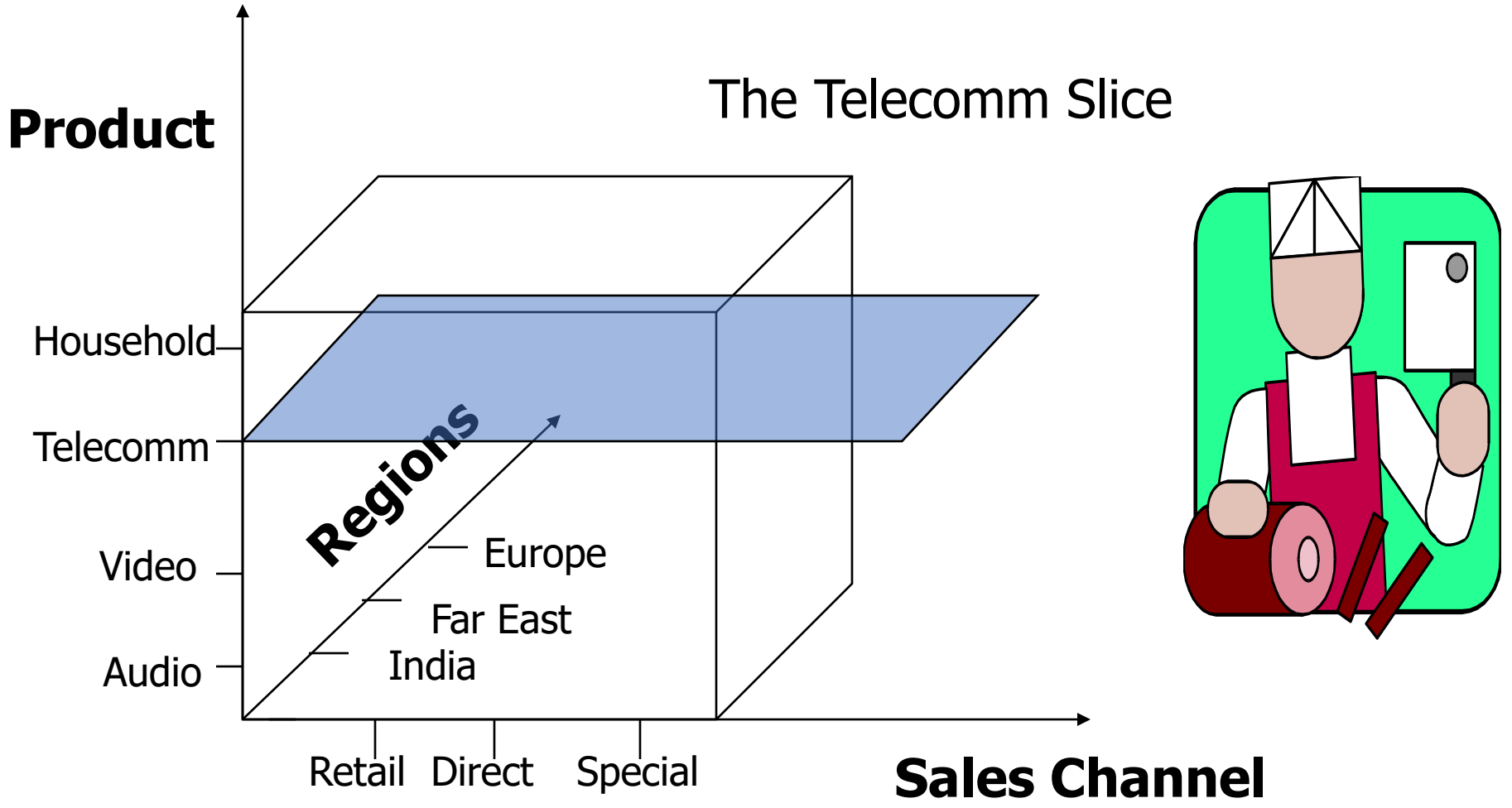
	1	2	3	4	5	6	7	8
Apr								
May								
Jun								
Jul								
Aug								
Sep								
Oct								
Nov								
Dec								
Jan								
Feb								
Mar								

Month	Store	Sales
Apr	1	
Apr	2	
Apr	3	
Apr	4	
Apr	5	
Apr	6	
Apr	7	
Apr	8	
May	1	
May	2	
May	3	
May	4	
May	5	
May	6	
May	7	
May	8	
Jun	1	
Jun	2	

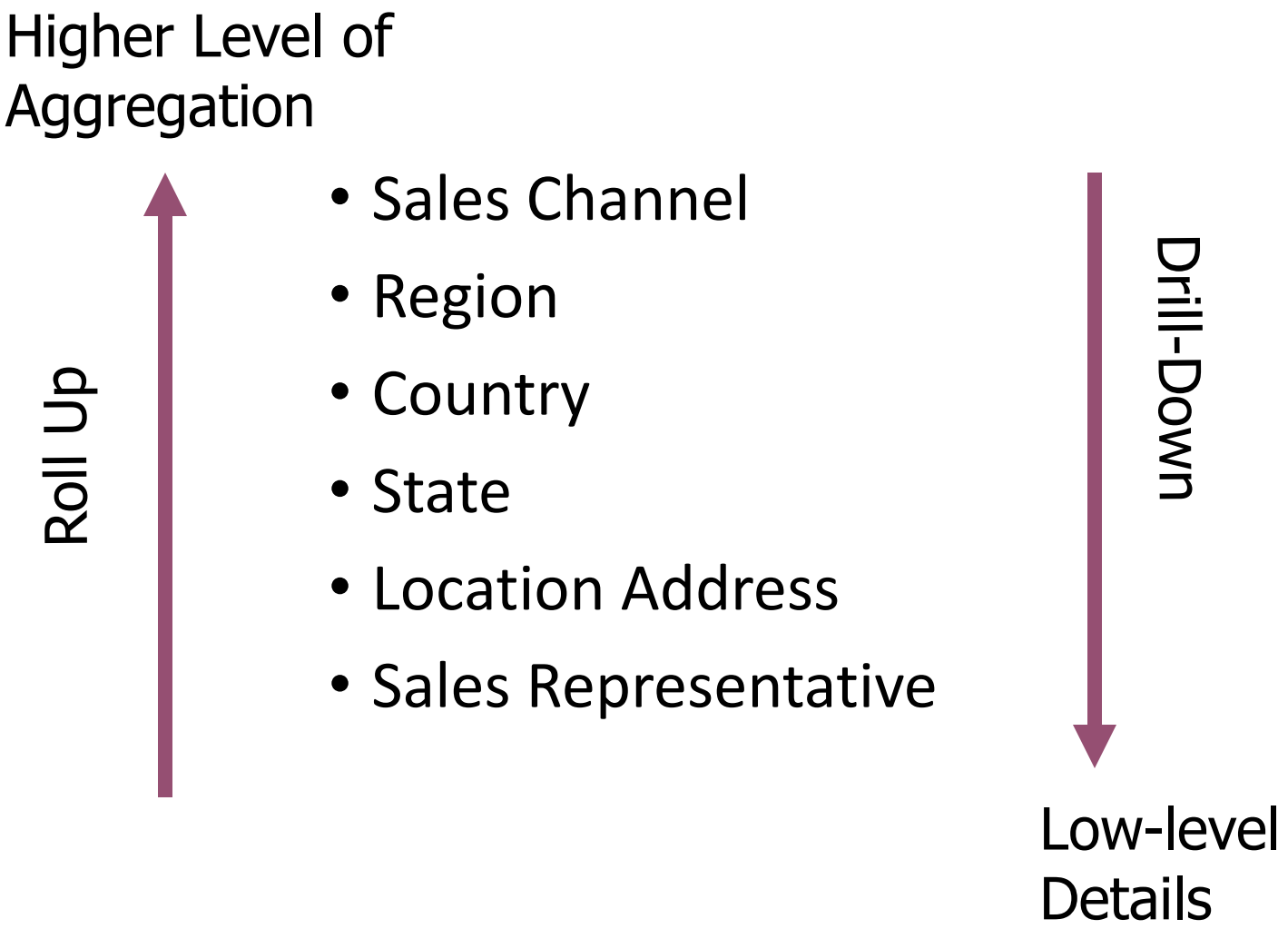
A Visual Operation: Pivot (Rotate)



“Slicing and Dicing”



Roll-up and Drill Down

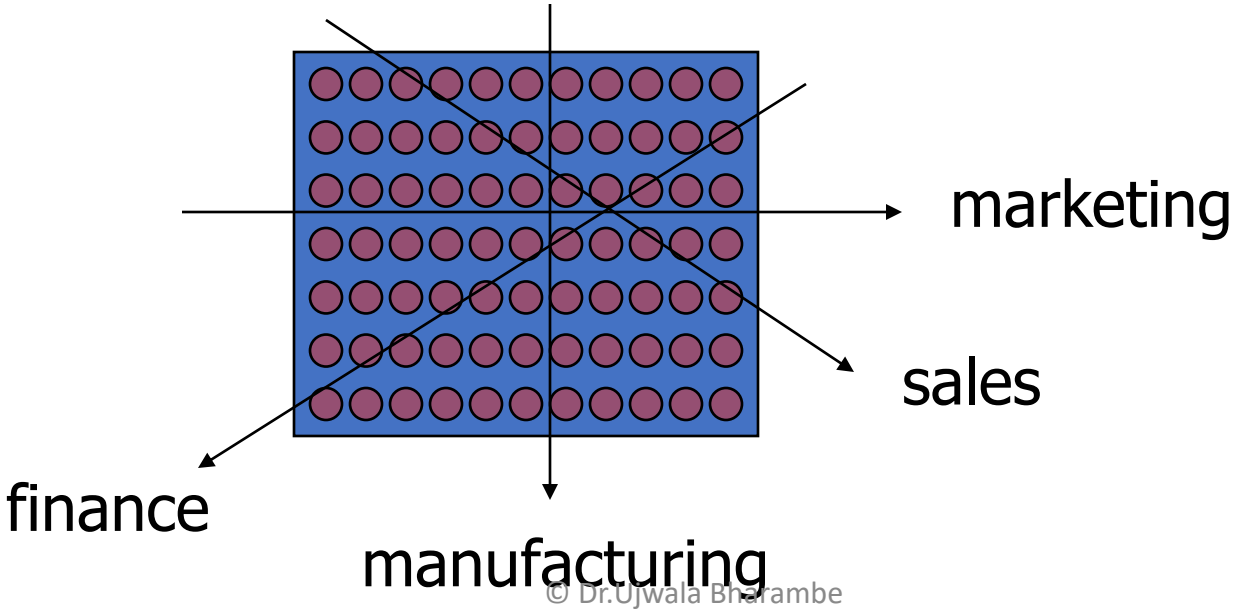


Nature of OLAP Analysis

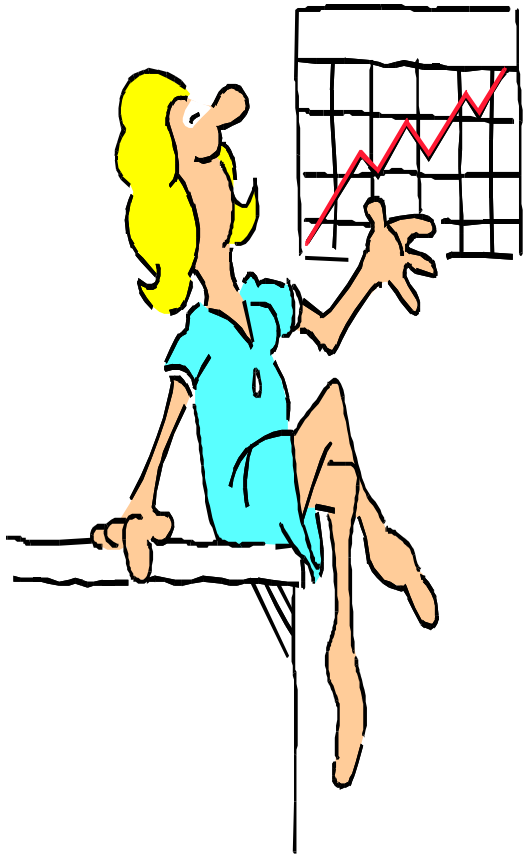
- Aggregation -- (total sales, percent-to-total)
- Comparison -- Budget vs. Expenses
- Ranking -- Top 10, quartile analysis
- Access to detailed and aggregate data
- Complex criteria specification
- Visualization

Organizationally Structured Data

- Different Departments look at the same detailed data in different ways. Without the detailed, organizationally structured data as a foundation, there is no reconcilability of data



Multidimensional Spreadsheets



- Analysts need spreadsheets that support
  - pivot tables (cross-tabs)
  - drill-down and roll-up
  - slice and dice
  - sort
  - selections
  - derived attributes
- Popular in retail domain

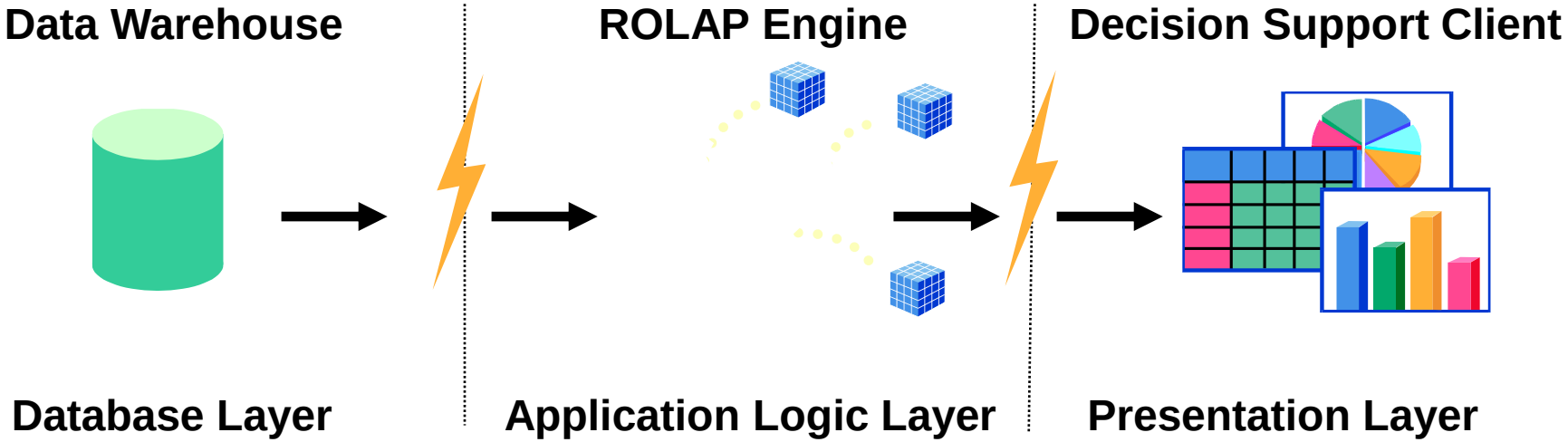
OLAP - Data Cube

- Idea: analysts need to group data in many different ways
  - eg. Sales(region, product, prodtype, prodstyle, date, saleamount)
  - saleamount is a measure attribute, rest are dimension attributes
  - groupby every subset of the other attributes
    - materialize (precompute and store) groupbys to give online response
  - Also: hierarchies on attributes: date -> weekday,
    date -> month -> quarter -> year

SQL Extensions

- Front-end tools require
  - Extended Family of Aggregate Functions
    - rank, median, mode
  - Reporting Features
    - running totals, cumulative totals
  - Results of multiple group by
    - total sales by month and total sales by product
  - Data Cube

Relational OLAP: 3 Tier DSS

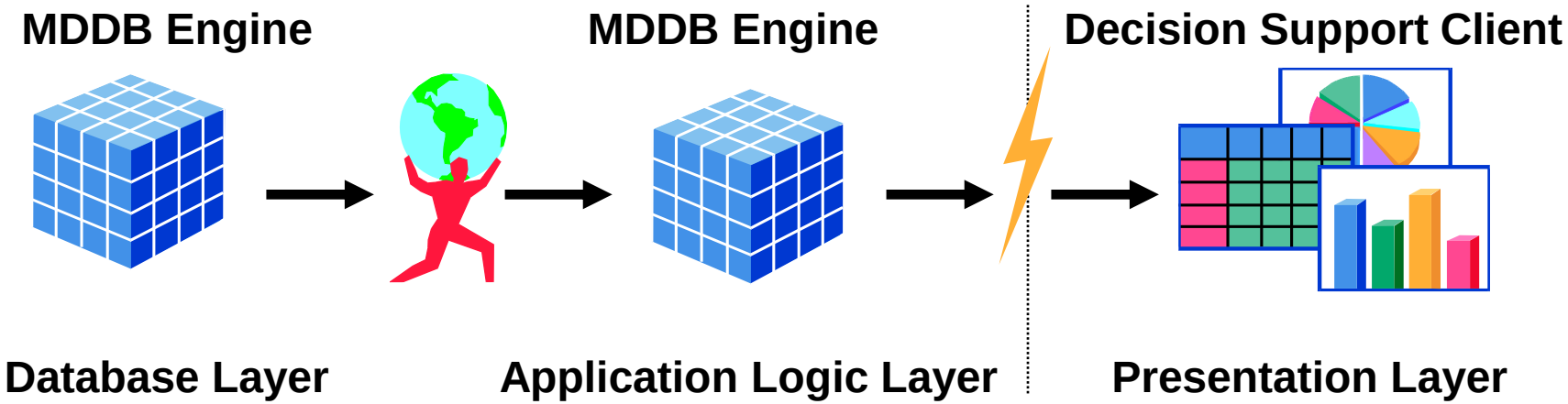


Store atomic data in industry standard RDBMS.

Generate SQL execution plans in the ROLAP engine to obtain OLAP functionality.

Obtain multi-dimensional reports from the DSS Client.

MD-OLAP: 2 Tier DSS

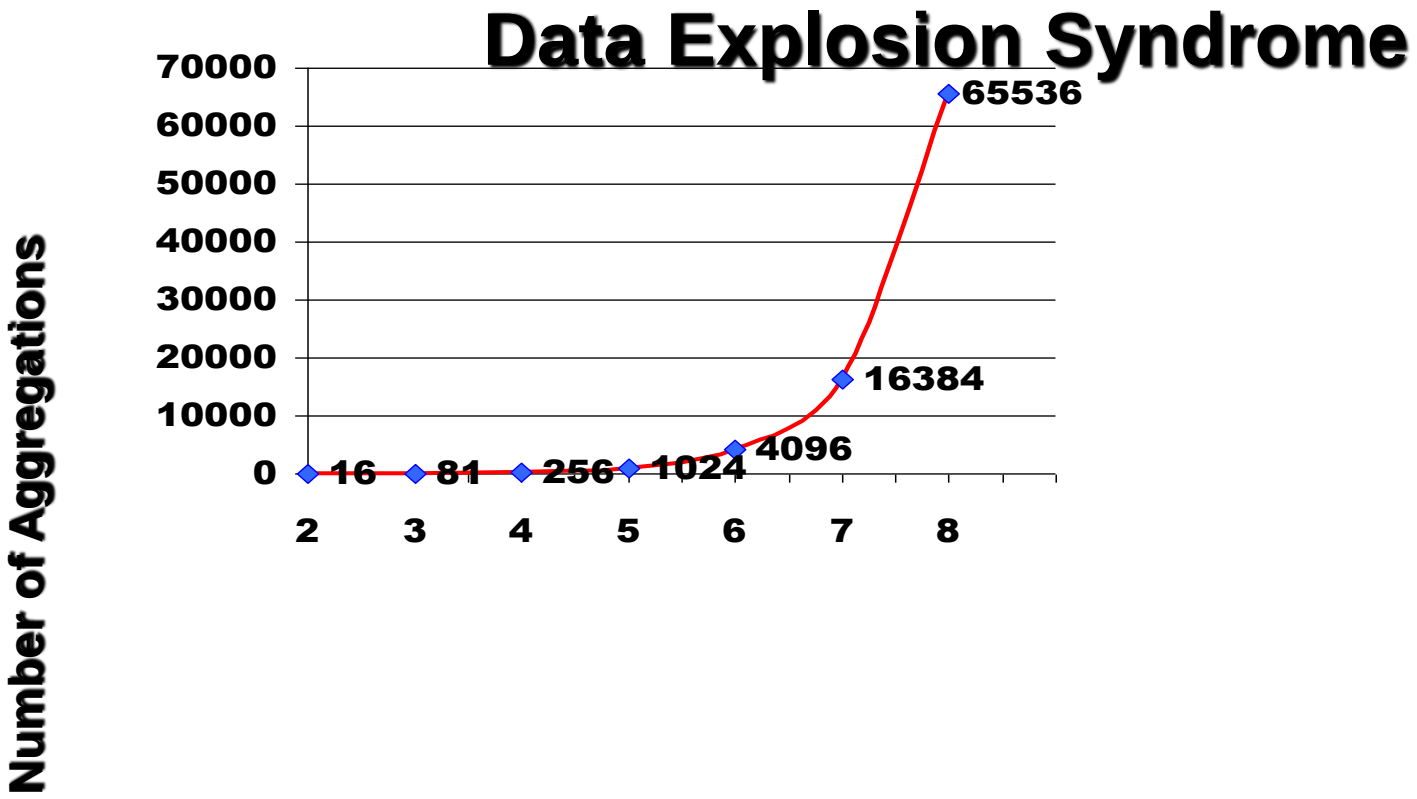


Store atomic data in a proprietary data structure (MDDB), pre-calculate as many outcomes as possible, obtain OLAP functionality via proprietary algorithms running against this data.

Obtain multi-dimensional reports from the DSS Client.

Typical OLAP Problems

Data Explosion



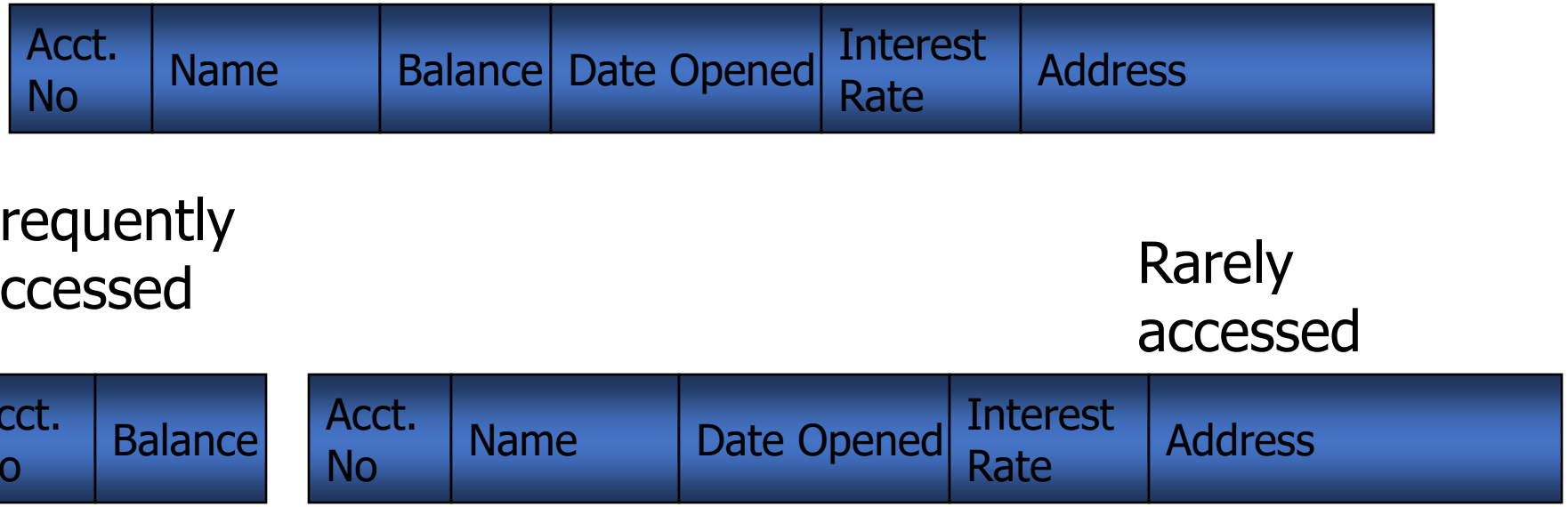
Granularity in Warehouse

- Can not answer some questions with summarized data
  - Did Anand call Seshadri last month? Not possible to answer if total duration of calls by Anand over a month is only maintained and individual call details are not.
- Detailed data too voluminous

Granularity in Warehouse

- Tradeoff is to have dual level of granularity
  - Store summary data on disks
    - 95% of DSS processing done against this data
  - Store detail on tapes
    - 5% of DSS processing against this data

Vertical Partitioning



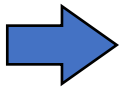
Derived Data

- Introduction of **derived** (calculated data) may often help
- Have seen this in the context of dual levels of granularity
- Can keep **auxiliary views** and indexes to speed up query processing

Aggregates

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

sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4

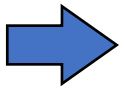


81

Aggregates

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

sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4

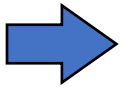


ans	date	sum
	1	81
	2	48

Another Example

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

sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4



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

Aggregates

- Operators: sum, count, max, min, median, ave
- “Having” clause
- Using dimension hierarchy
  - average by region (within store)
  - maximum by month (within date)

Data Cube

Fact table view:

sale	prodId	storeId	amt
	p1	s1	12
	p2	s1	11
	p1	s3	50
	p2	s2	8

Multi-dimensional cube:

	s1	s2	s3
p1	12		50
p2	11	8	

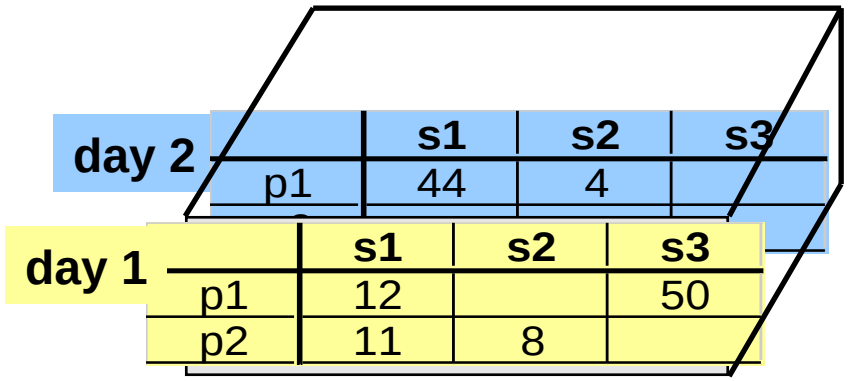
dimensions = 2

3-D Cube

Fact table view:

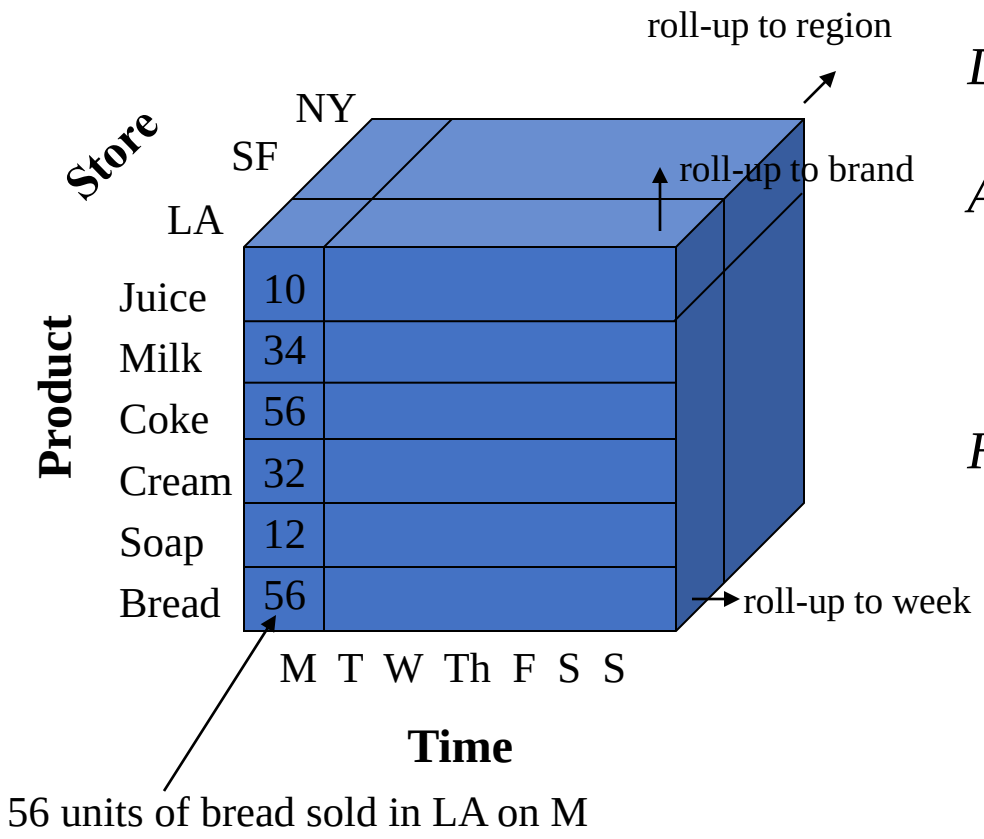
sale	prodId	storeId	date	amt
	p1	s1	1	12
	p2	s1	1	11
	p1	s3	1	50
	p2	s2	1	8
	p1	s1	2	44
	p1	s2	2	4

Multi-dimensional cube:



dimensions = 3

Example



Dimensions:

Time, Product, Store

Attributes:

Product (upc, price, …)

Store …

…

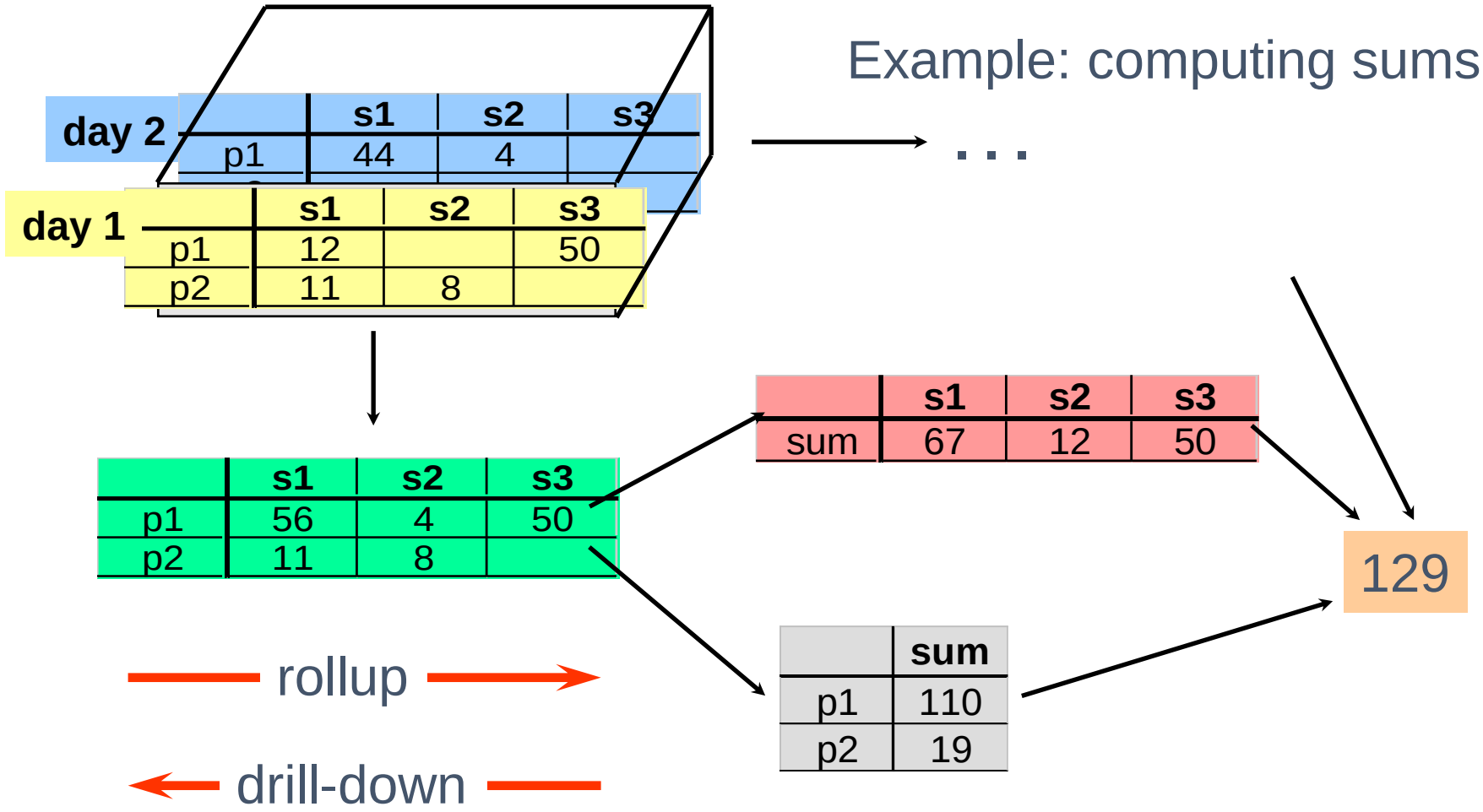
Hierarchies:

Product → Brand → …

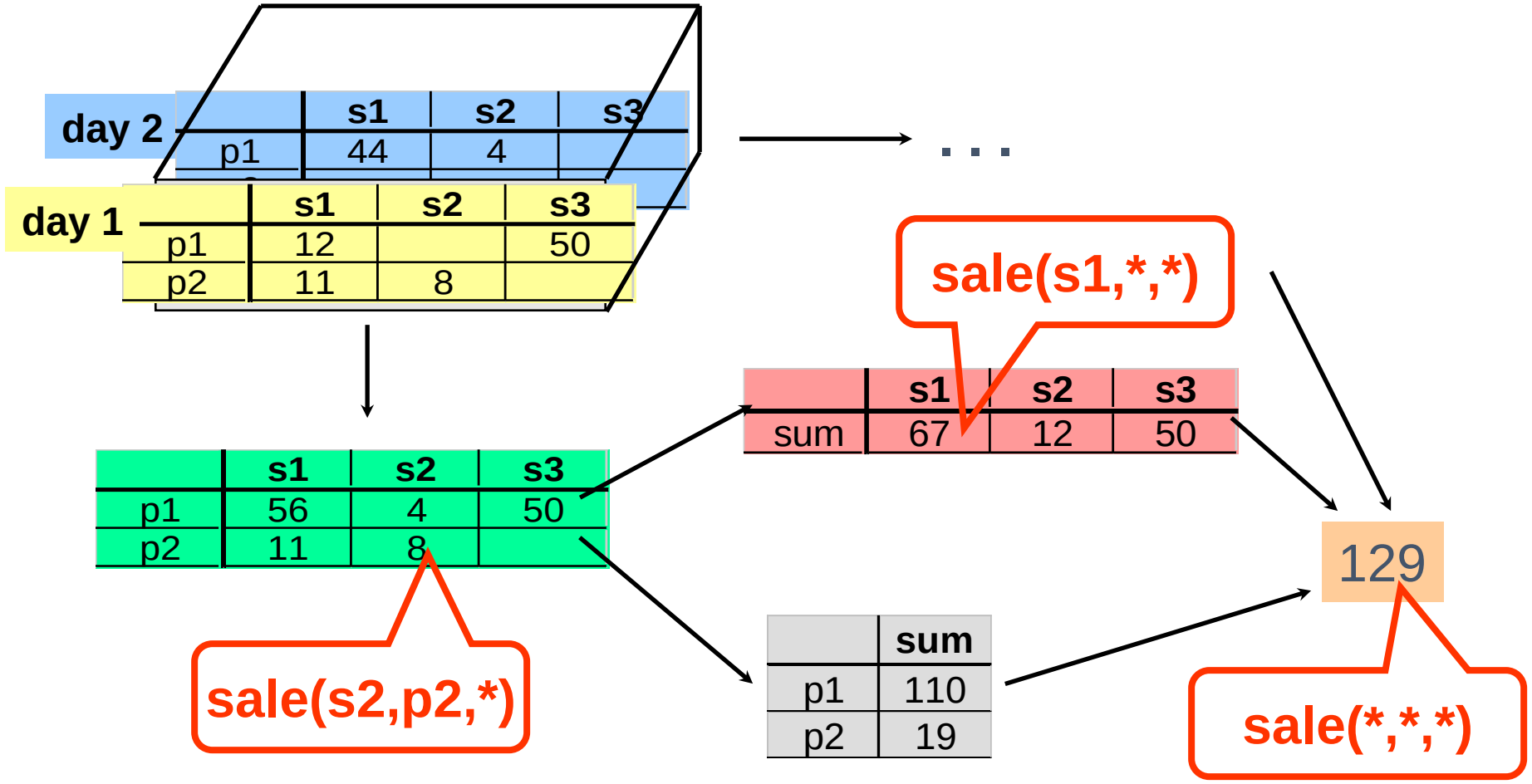
Day → Week → Quarter

Store → Region → Country

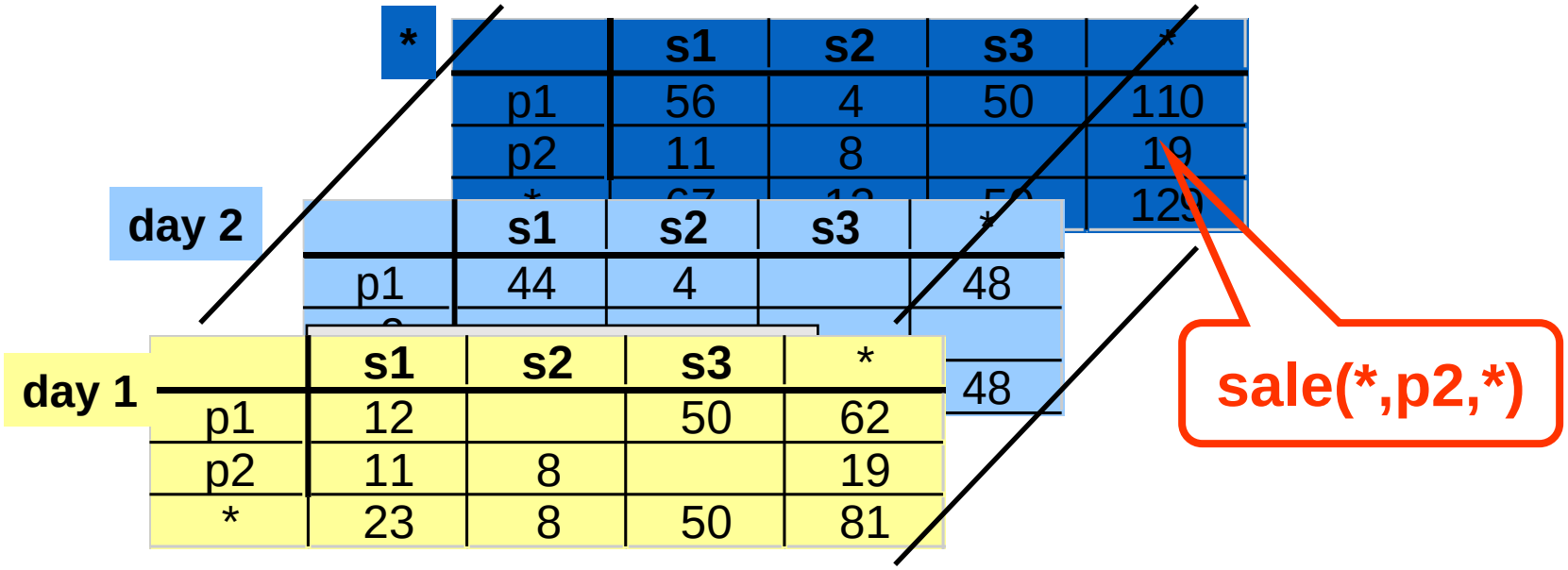
Cube Aggregation: Roll-up



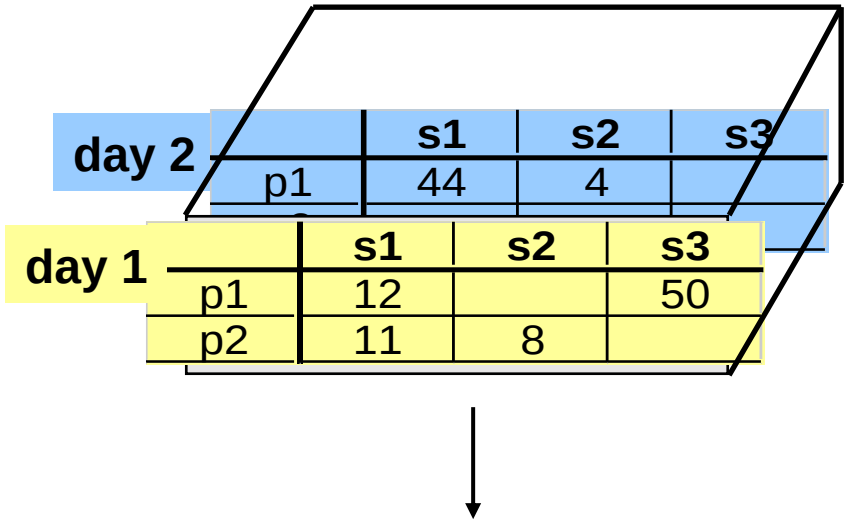
Cube Operators for Roll-up



Extended Cube



Aggregation Using Hierarchies

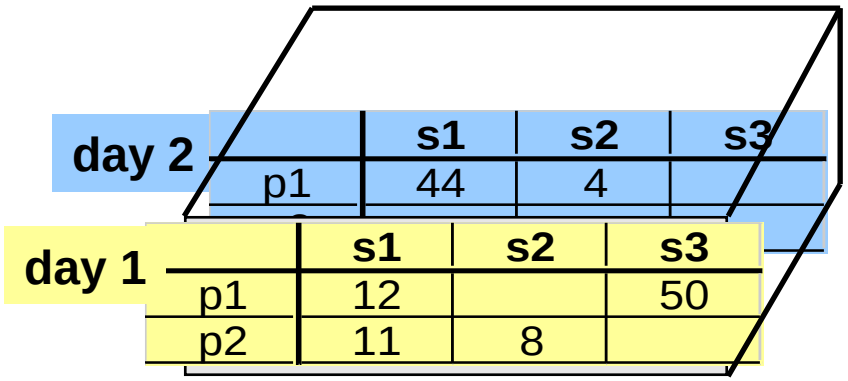


	region A	region B
p1	56	54
p2	11	8

store
region
country

(store s1 in Region A;
stores s2, s3 in Region B)

Slicing



TIME = day 1

	s1	s2	s3
p1	12		50
p2	11	8	

Slicing & Pivoting

	Sales ($ millions)			
	Products	Time		
		d1	d2	
Store s1	Electronics	$5.2		
	Toys	$1.9		
	Clothing	$2.3		
	Cosmetics	$1.1		
Store s2	Electronics	$8.9		
	Toys	$0.75		
	Clothing	$4.6		
	Cosmetics	$1.5		

	Sales ($ millions)			
	Products	d1		
		Store s1	Store s2	
Store s1	Electronics	$5.2	$8.9	
	Toys	$1.9	$0.75	
	Clothing	$2.3	$4.6	
	Cosmetics	$1.1	$1.5	
Store s2	Electronics			
	Toys			
	Clothing			

Summary of Operations

- Aggregation (roll-up)
  - aggregate (summarize) data to the next higher dimension element
  - e.g., total sales by city, year $\rightarrow$ total sales by region, year
- Navigation to detailed data (drill-down)
- Selection (slice) defines a subcube
  - e.g., sales where city ='Gainesville' and date = '1/15/90'
- Calculation and ranking
  - e.g., top 3% of cities by average income
- Visualization operations (e.g., Pivot)
- Time functions
  - e.g., time average