Overview of Data Warehousing and OLAP

Chapter 28

Chapter Outline

- What is a data warehouse (DW)
- Conceptual structure of DW
- Why separate DW
- Data modeling for DW
- Online Analytical Processing (OLAP)
- Building A Data Warehouse
- Warehouse vs. Data Views

- What is Data Warehouse?

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization’s operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”—W. H. Inmon
- Data warehousing:
  - The process of constructing and using data warehouses

-- Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

-- Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

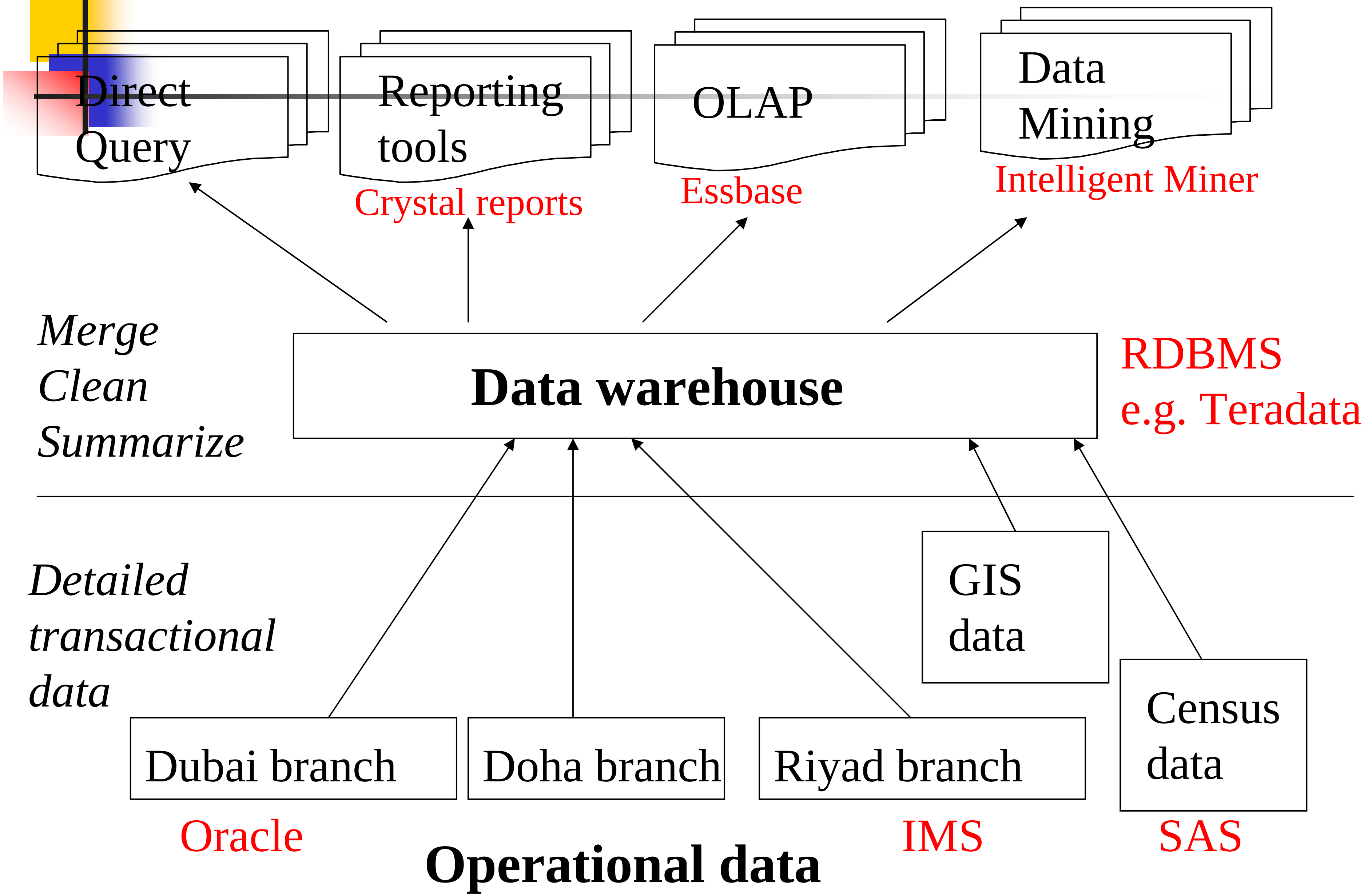
-- Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain “time element”.

-- Data Warehouse—Non-Volatile

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *Initial loading of data* and *access of data*.

Decision support tools



Direct Query

Reporting tools

Crystal reports

OLAP

Essbase

Data Mining

Intelligent Miner

Merge
Clean
Summarize

Data warehouse

RDBMS
e.g. Teradata

Detailed transactional data

GIS data

Census data

SAS

Dubai branch

Oracle

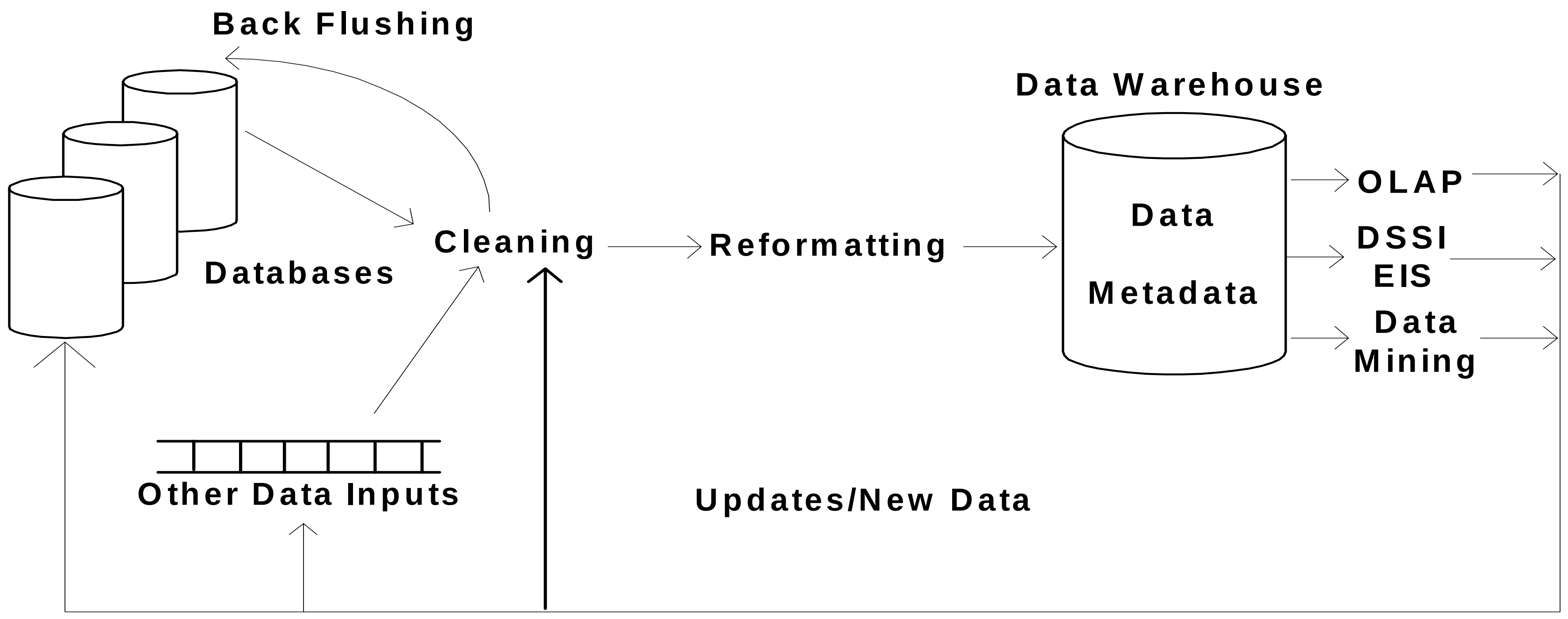
Doha branch

Riyad branch

IMS

Operational data

- Conceptual Structure of Data Warehouse



- Why Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for Online Transaction Processing (OLTP): access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for Online Analytical Processing (OLAP): complex OLAP queries, multidimensional view, consolidation.
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled

- Data Modeling for Data Warehouses ...

- Example of Two- Dimensional vs. Multi- Dimensional

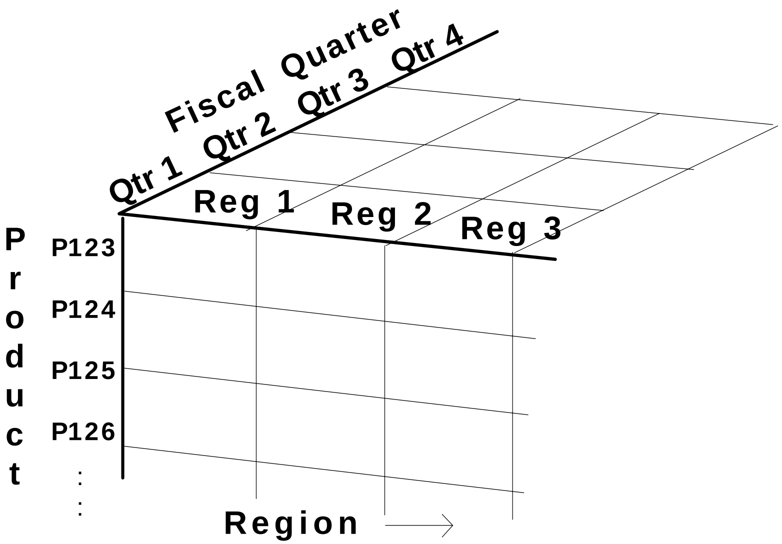
Two Dimensional Model

REGION

PRODUCT

	REG1	REG2	REG3
P123			
P124			
P125			
P126			
: :			

Three dimensional data cube



… - Data Modeling for Data Warehouses

- Advantages of a multi-dimensional model
  - Multi-dimensional models lend themselves readily to hierarchical views in what is known as roll-up display and drill-down display.
  - The data can be directly queried in any combination of dimensions, bypassing complex database queries.

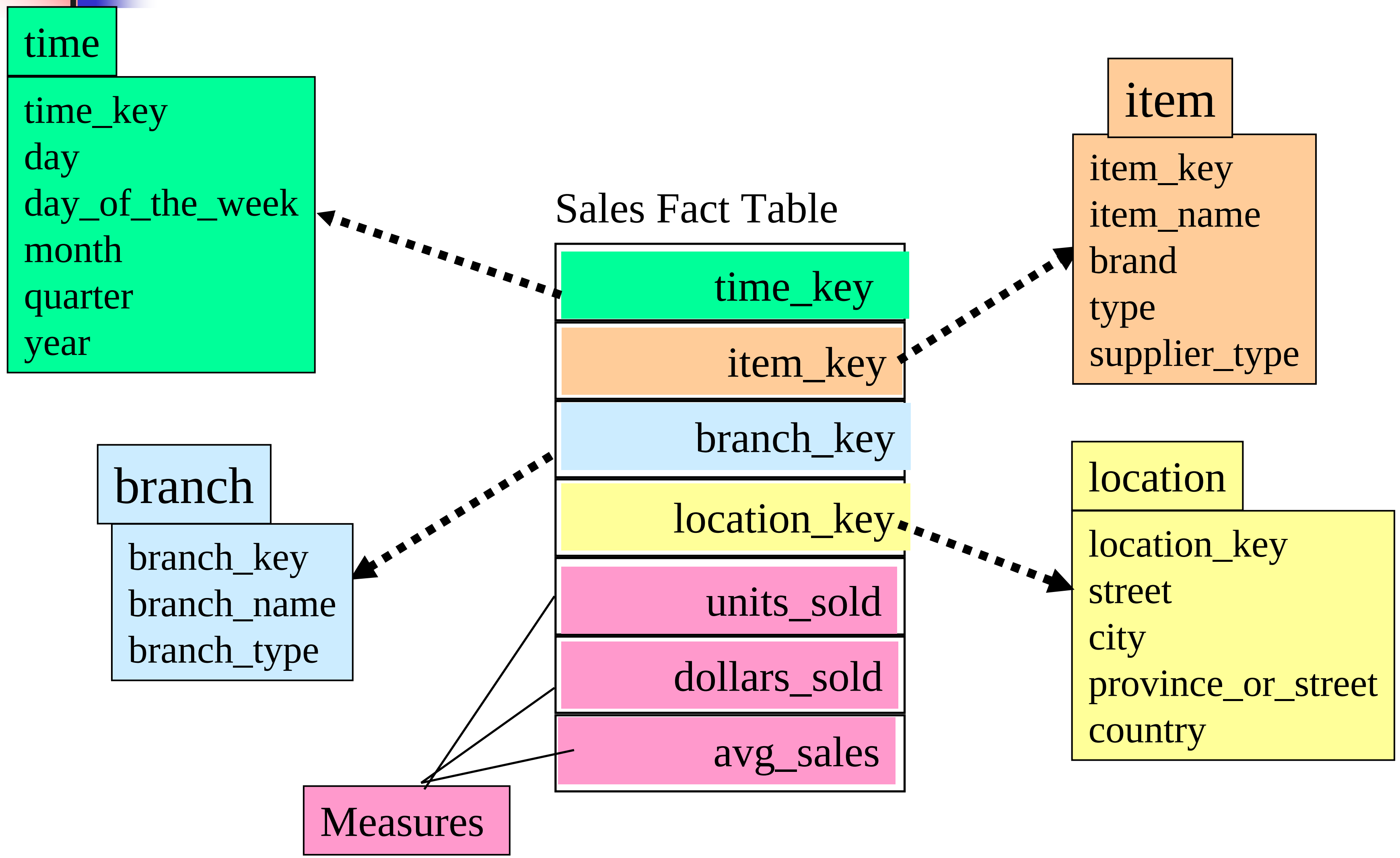
-- Multi-dimensional Schemas ...

- Multi-dimensional schemas are specified using:
  - **Dimension table** – It consists of tuples of attributes of the dimension.
  - **Fact table** – Each tuple is a recorded fact. This fact contains some measured or observed variable (s) and identifies it with pointers to dimension tables. The fact table contains the data, and the dimensions to identify each tuple in the data

... -- Multi-dimensional Schemas ...

- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

--- Example of Star Schema



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item

item_key
item_name
brand
type
supplier_type

branch

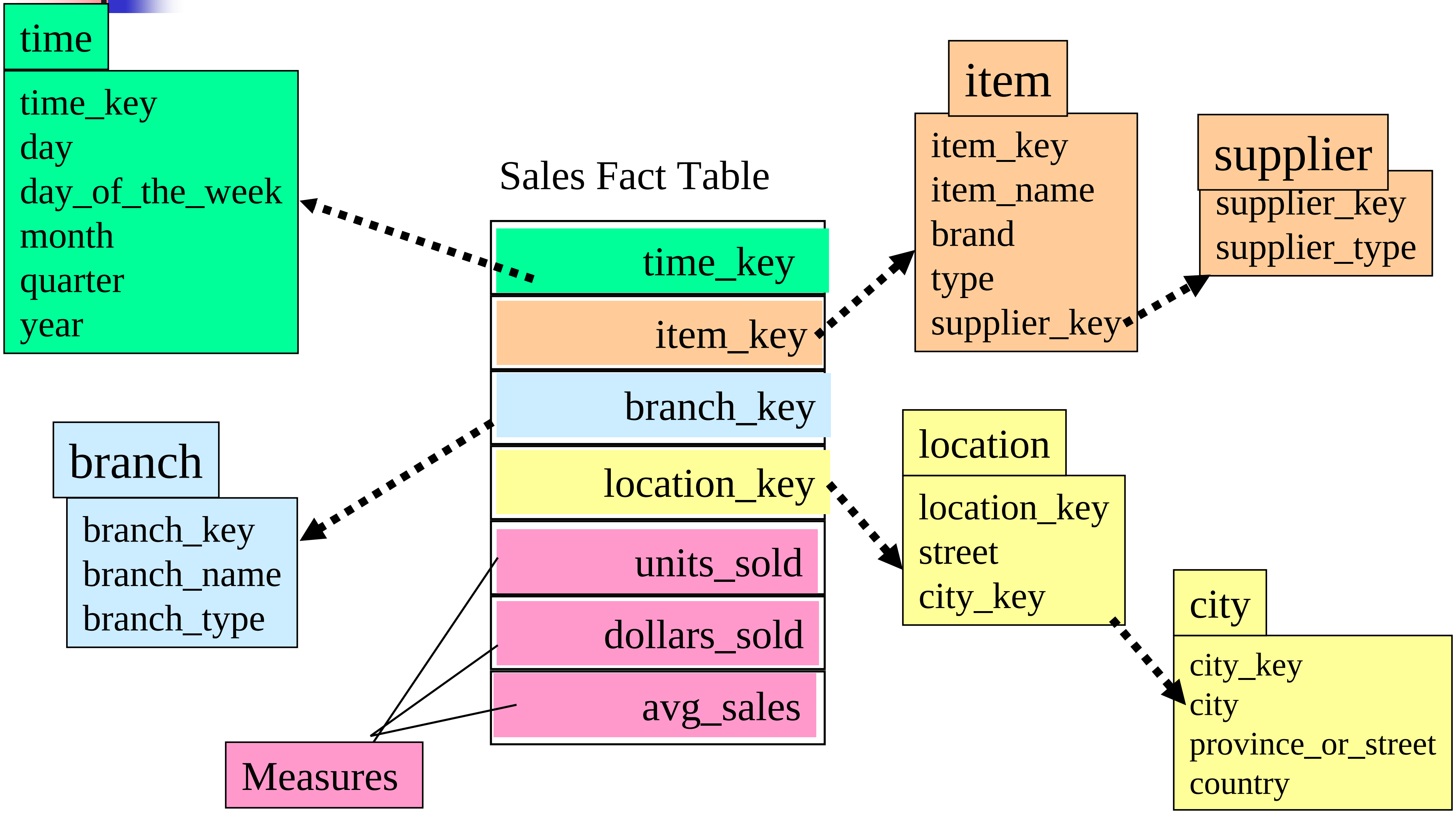
branch_key
branch_name
branch_type

location

location_key
street
city
province_or_street
country

Measures

--- Example of Snowflake Schema



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

Measures

item

item_key
item_name
brand
type
supplier_key

supplier

supplier_key
supplier_type

branch

branch_key
branch_name
branch_type

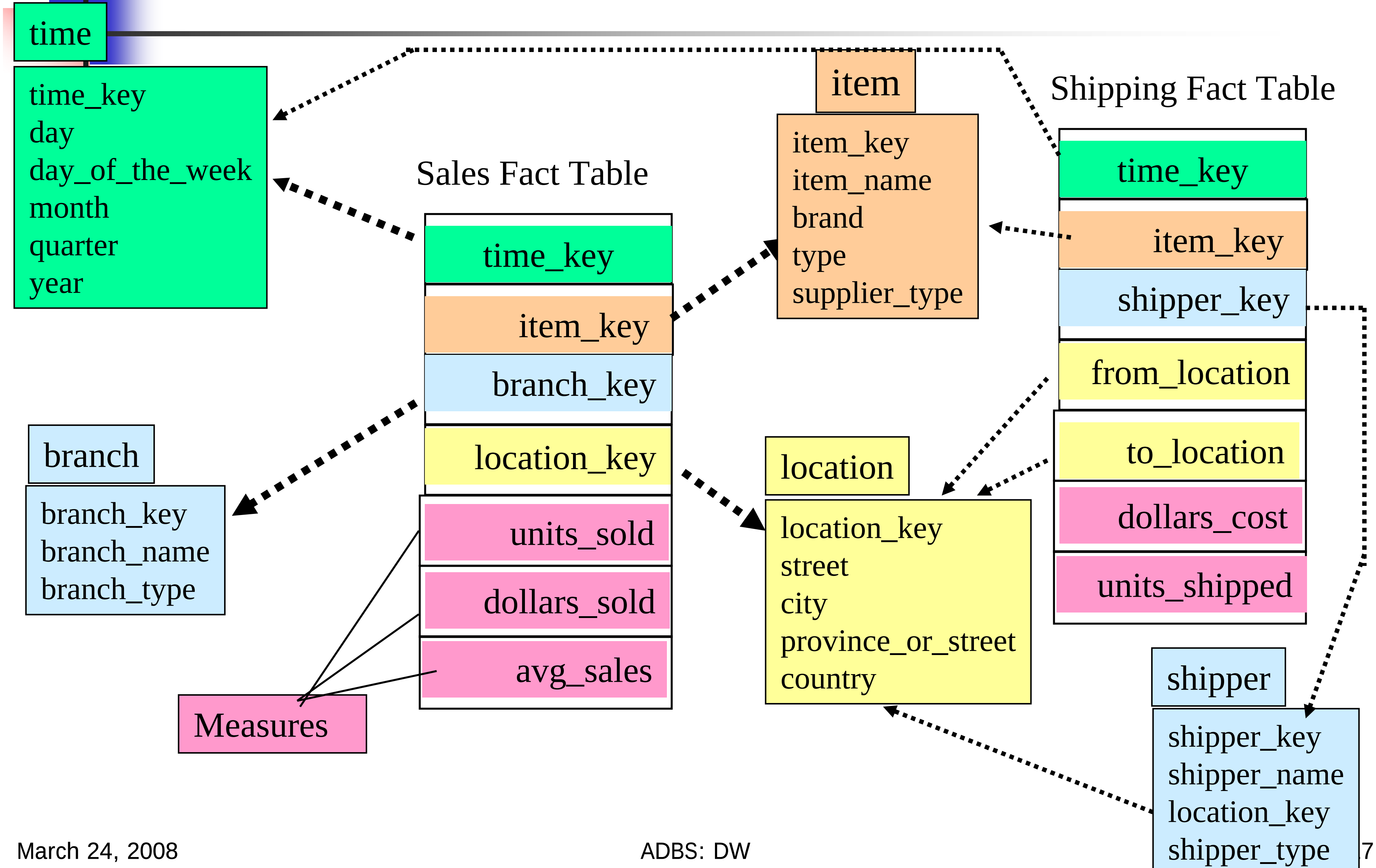
location

location_key
street
city_key

city

city_key
city
province_or_street
country

--- Example of Fact Constellation



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

Measures

item

item_key
item_name
brand
type
supplier_type

Shipping Fact Table

time_key
item_key
shipper_key
from_location
to_location
dollars_cost
units_shipped

branch

branch_key
branch_name
branch_type

location

location_key
street
city
province_or_street
country

shipper

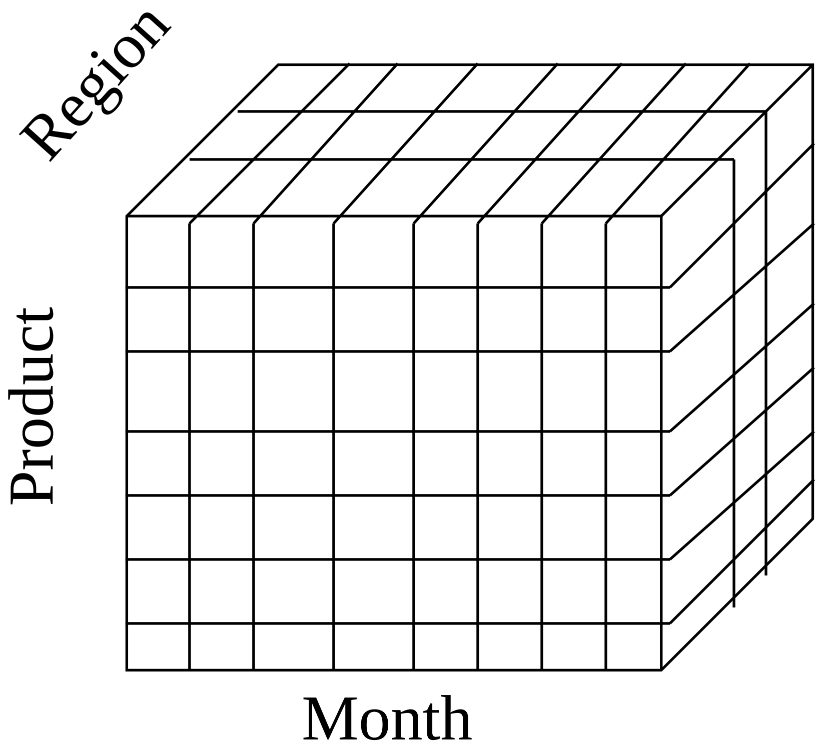
shipper_key
shipper_name
location_key
shipper_type

- OLAP

- *Fast, interactive answers to large aggregate queries*.
- Multidimensional model: *dimensions* with *hierarchies*
  - Dim 1: Bank location:
    - branch-->city-->state
  - Dim 2: Customer:
    - sub profession --> profession
  - Dim 3: Time:
    - month --> quarter --> year
- *Measures*: loan amount, #transactions, balance

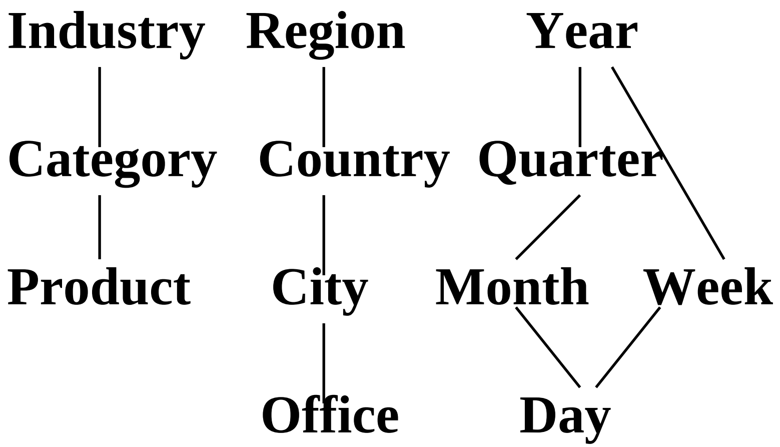
-- Multidimensional Data

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



Dimensions: Product, Location, Time

Hierarchical summarization paths



Industry
Category
Product

Region
Country
City
Office

Year
Quarter
Month
Week
Day

-- Data Cubes ...

Fact relation

sale	Product	Client	Amt
	p1	c1	12
	p2	c1	11
	p1	c3	50
	p2	c2	8

Two-dimensional cube

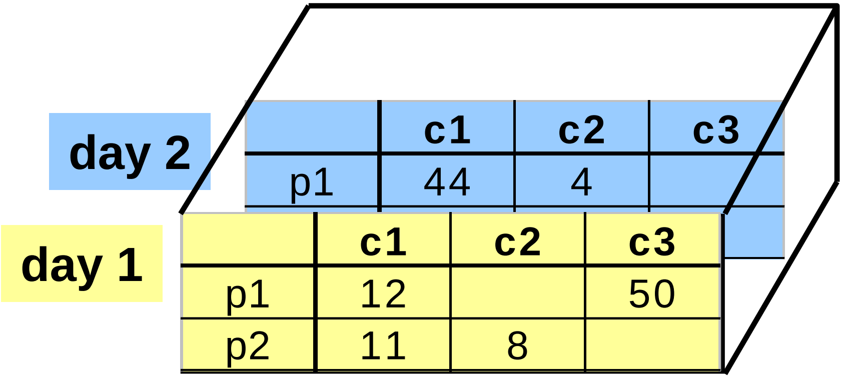
	c1	c2	c3
p1	12		50
p2	11	8	

... -- Data Cubes ...

Fact relation

sale	Product	Client	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

3-dimensional cube



day 2

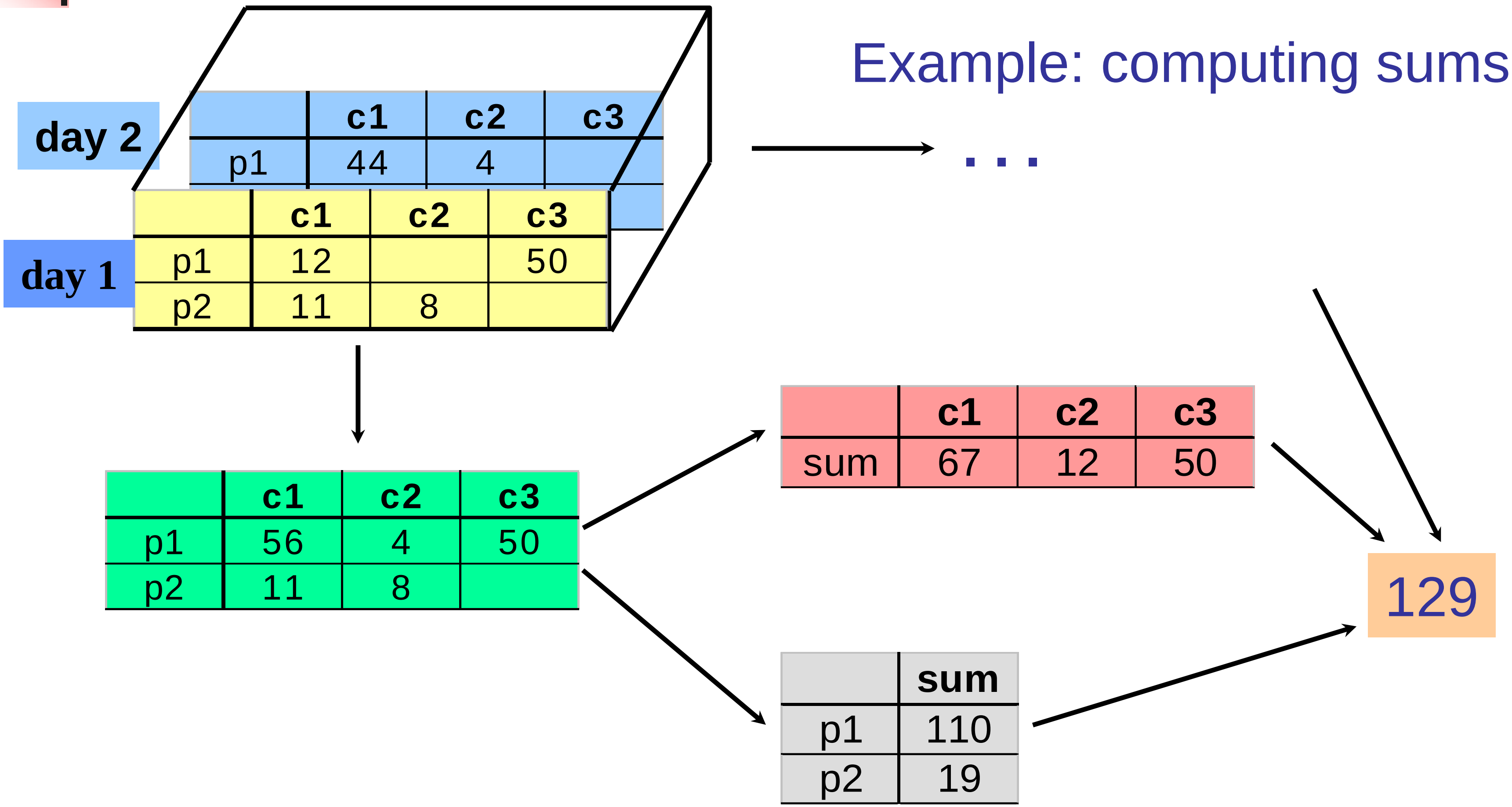
	c1	c2	c3
p1	44	4	

day 1

	c1	c2	c3
p1	12		50
p2	11	8	

... -- Data Cubes ...

Example: computing sums



day 2

	c1	c2	c3
p1	44	4	

day 1

	c1	c2	c3
p1	12		50
p2	11	8	

. . .

	c1	c2	c3
p1	56	4	50
p2	11	8	

	c1	c2	c3
sum	67	12	50

	sum
p1	110
p2	19

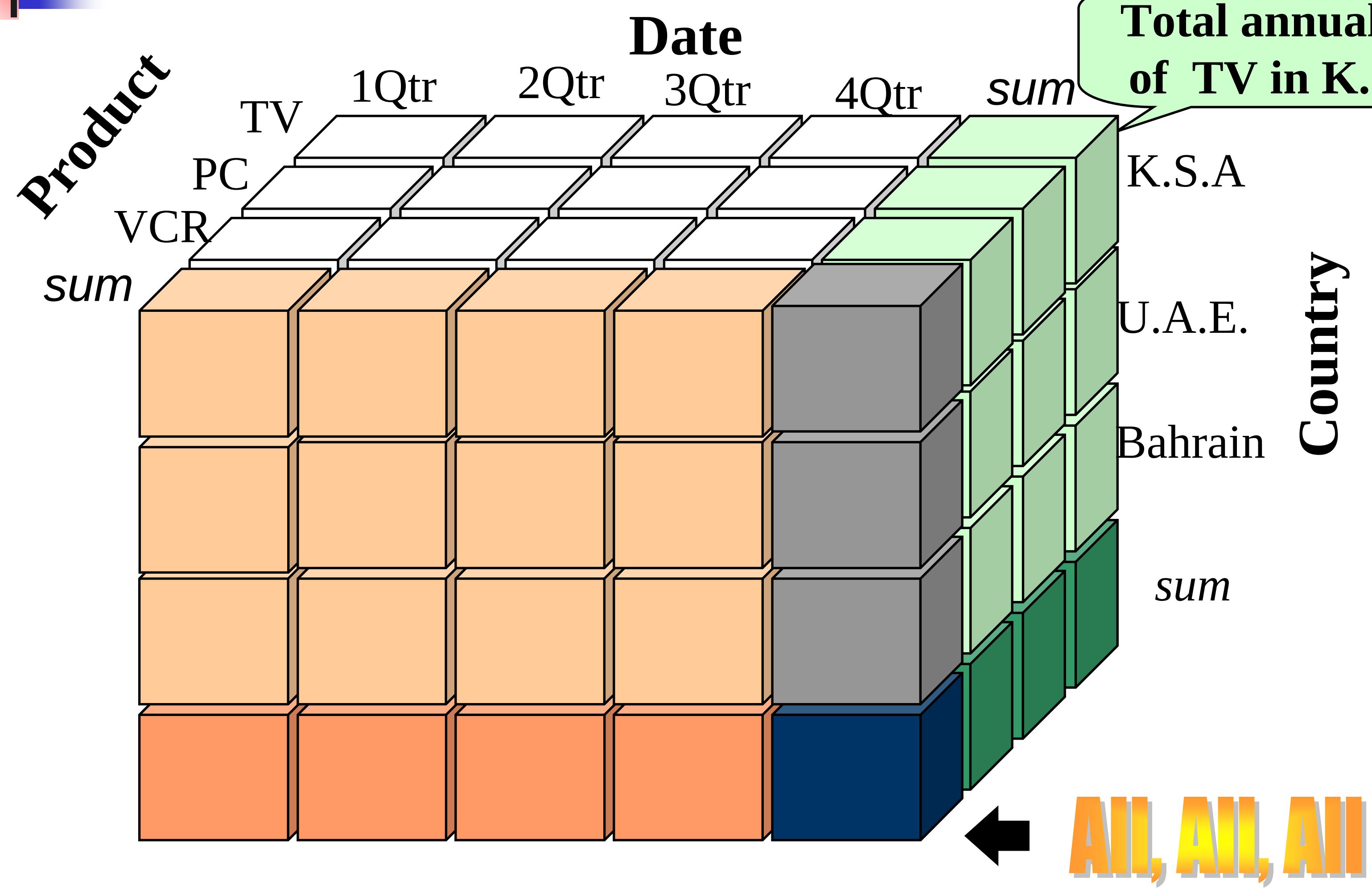
129

... -- Data Cubes ...

- In multidimensional data model together with measure values usually we store summarizing information (aggregates)

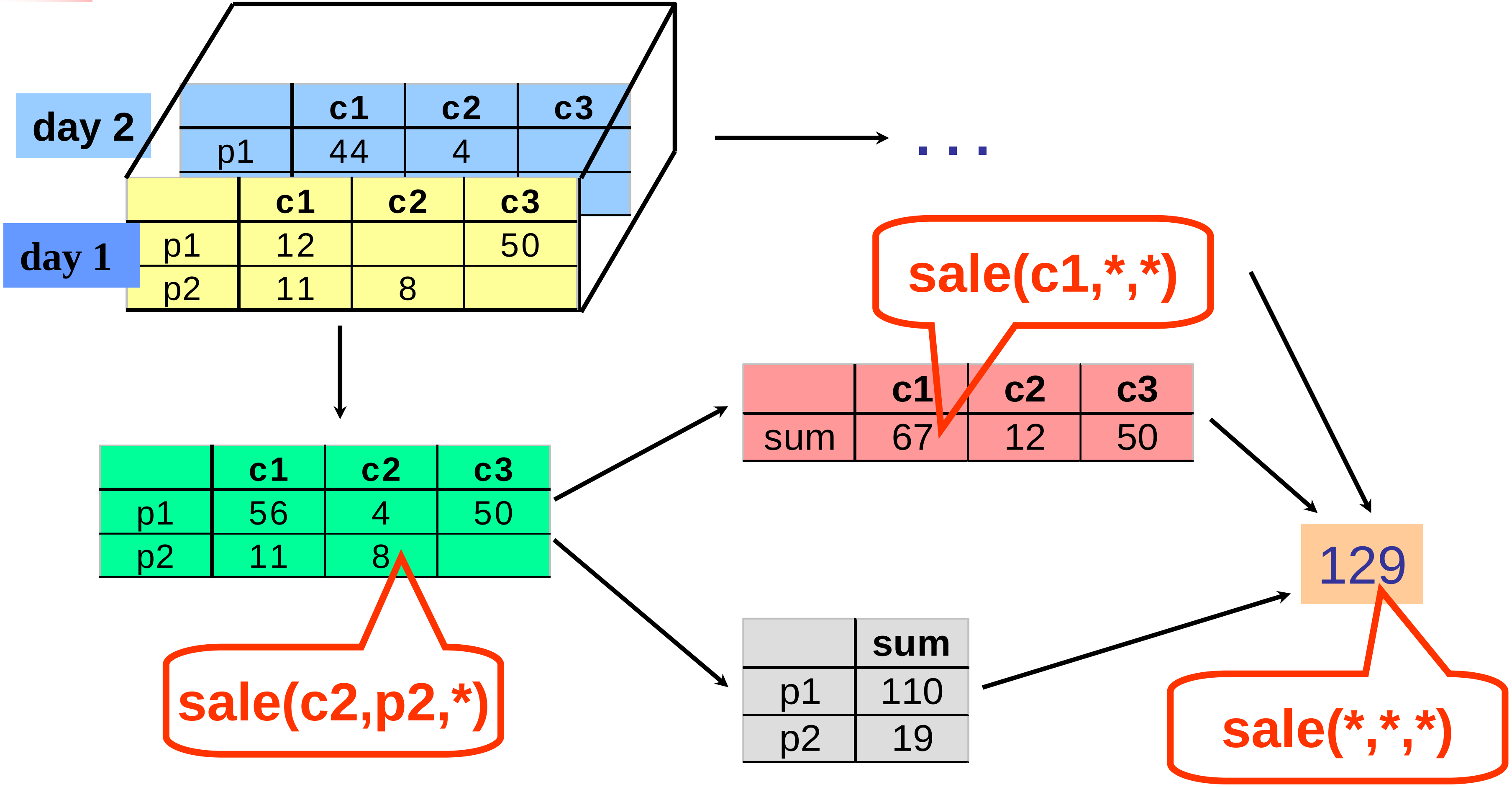
	c1	**c2**	**c3**	**Sum**
p1	56	4	50	110
p2	11	8		19
Sum	67	12	50	**129**

- A Sample Data Cube



Total annual sales of TV in K. S. A.

-- The Cube Operator ...



day 2

	c1	c2	c3
p1	44	4	

day 1

	c1	c2	c3
p1	12		50
p2	11	8	

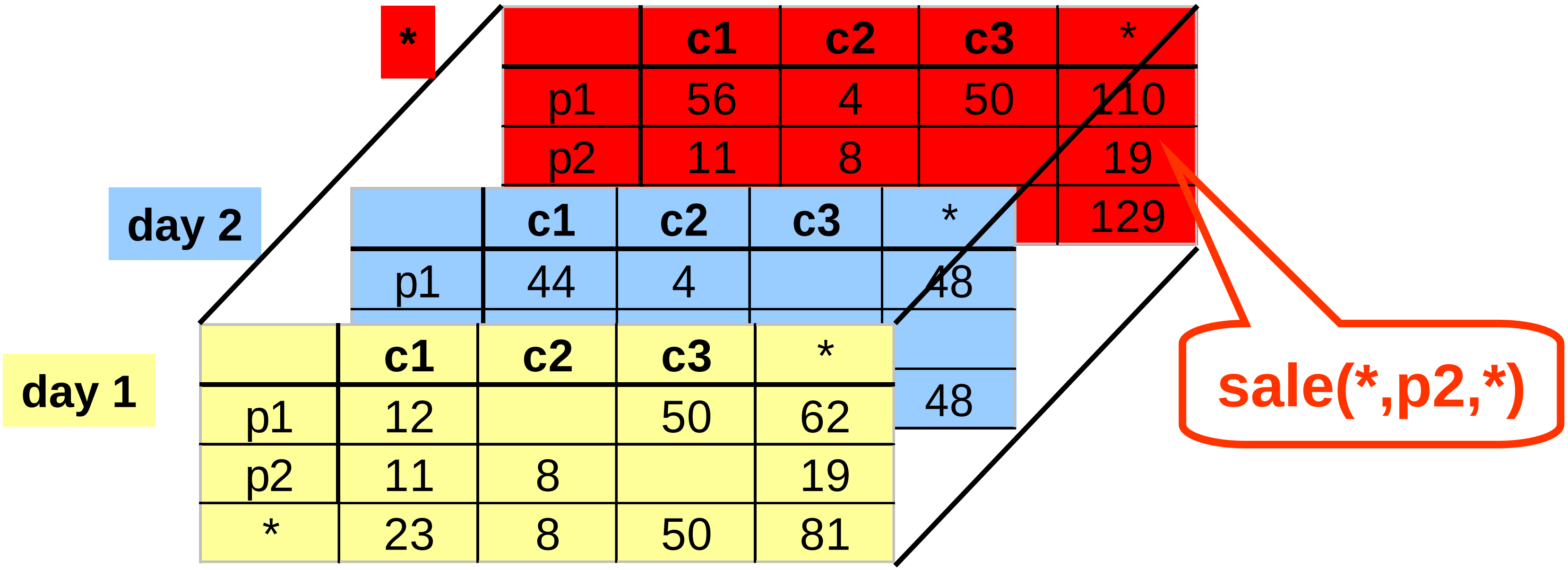
	c1	c2	c3
p1	56	4	50
p2	11	8	

	c1	c2	c3
sum	67	12	50

	sum
p1	110
p2	19

129

... -- The Cube Operator ...

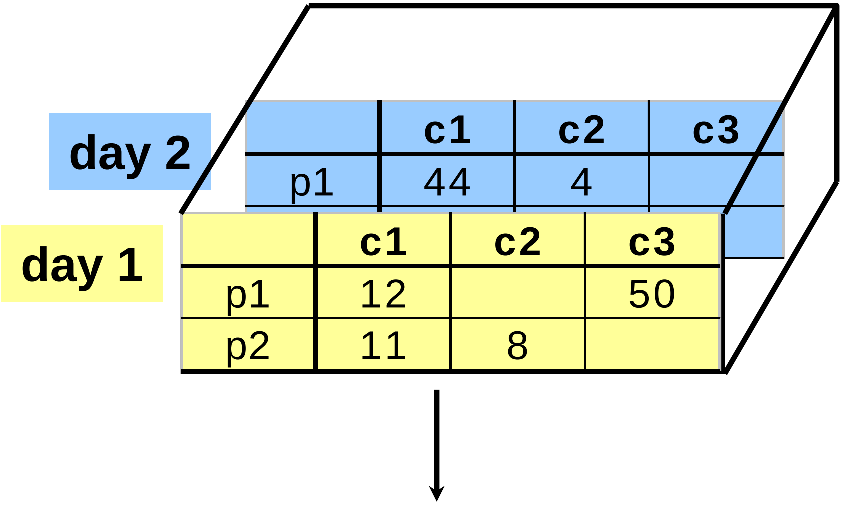


	c1	c2	c3	*
p1	56	4	50	110
p2	11	8		19
				129

	c1	c2	c3	*
p1	44	4		48
				48

	c1	c2	c3	*
p1	12		50	62
p2	11	8		19
*	23	8	50	81

-- Aggregation Using Hierarchies ...



day 2

	c1	c2	c3
p1	44	4	

day 1

	c1	c2	c3
p1	12		50
p2	11	8	

customer

region

country

	region A	region B	
p1	12	50	
p2	11	8	

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

-- OLAP Servers

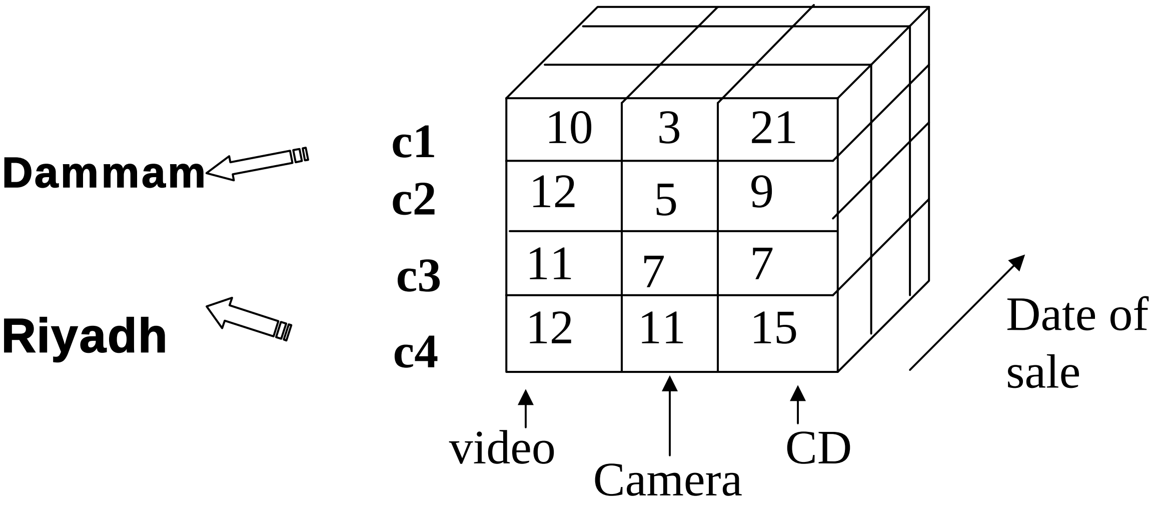
- Relational OLAP (ROLAP)
  - Extended relational DBMS that maps operations on multidimensional data to standard relations operations.
  - Store all information, including fact tables, as relations
- Multidimensional OLAP (MOLAP)
  - Special purpose server that directly implements multidimensional data and operations
  - Store multidimensional datasets as arrays.

- OLAP Queries

- 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)*

-- OLAP Queries: Roll Up

- Summarizes data along dimension.



	video	Camera	CD
c1	10	3	21
c2	12	5	9
c3	11	7	7
c4	12	11	15

client
|
city
|
region

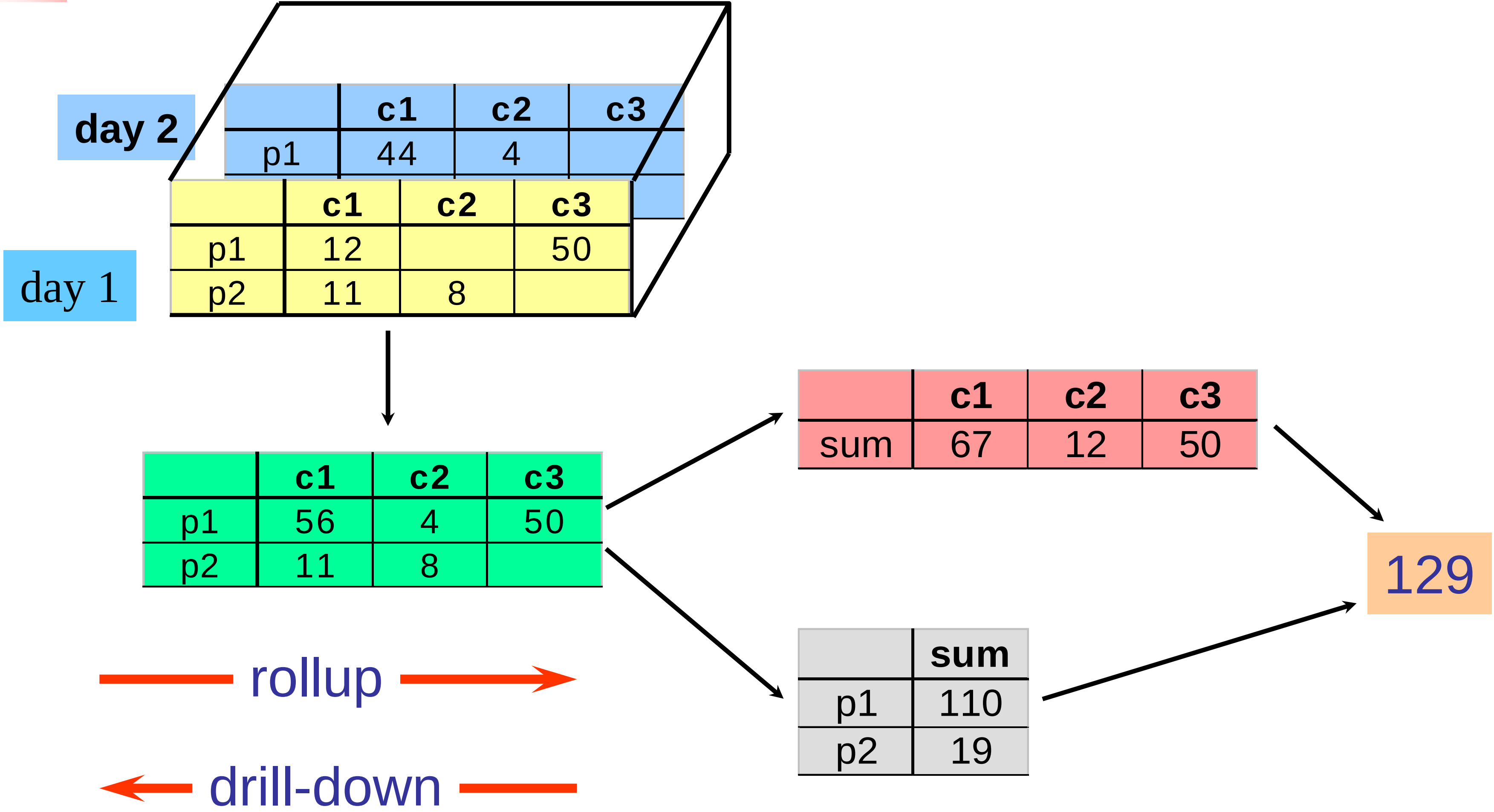
Roll up
aggregation with respect to city

	Video	Camera	CD
Dammam	22	8	30
Riyadh	23	18	22

-- OLAP Queries: Drill Down ...

- Roll down, drill down: go from higher level summary to lower level summary or detailed data
  - For a particular product category, find the detailed sales data for each salesperson by date
  - Given total sales by state, we can ask for sales per city, or just sales by city for a selected state

... -- OLAP Queries: Drill down



day 2

	c1	c2	c3
p1	44	4	

day 1

	c1	c2	c3
p1	12		50
p2	11	8	

	c1	c2	c3
p1	56	4	50
p2	11	8	

	c1	c2	c3
sum	67	12	50

	sum
p1	110
p2	19

129

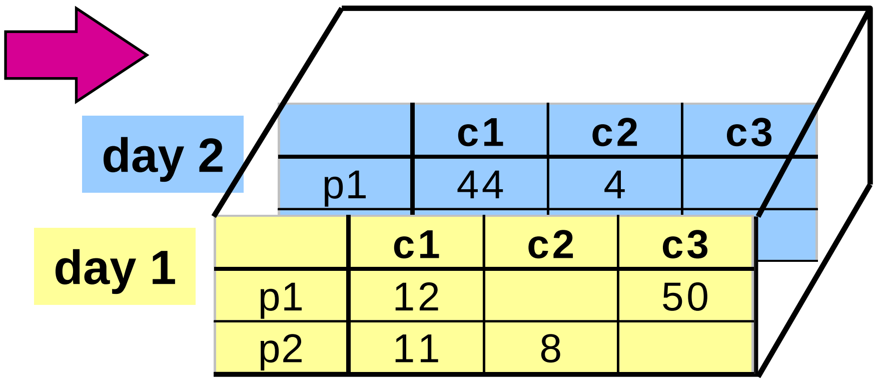
-- Other OLAP Queries ...

- Slice and dice: select and project
  - Sales of video in USA over the last 6 months
  - Slicing and dicing reduce the number of dimensions
- Pivot: reorient cube
  - The result of pivoting is called a cross-tabulation
  - If we pivot the Sales cube on the Client and Product dimensions, we obtain a table for each client for each product value

... Other OLAP Queries

- Pivoting can be combined with aggregation

sale	prodId	clientid	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



day 2

	c1	c2	c3
p1	44	4	

day 1

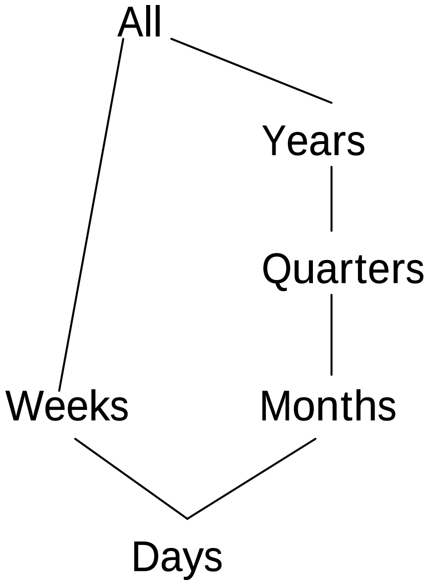
	c1	c2	c3
p1	12		50
p2	11	8	

	c1	c2	c3	Sum
1	23	8	50	81
2	44	4		48
Sum	67	12	50	129

	c1	c2	c3	Sum
p1	56	4	50	110
p2	11	8		19
Sum	67	12	50	129

-- Cube Implementations

- Data cubes are implemented by materialized views
- A materialized view is the result of some query, which we chose to store its output table in the database.
- For the data cube, the views we would choose to materialize will typically be aggregations of the full data cube.
- Lattice of views are created for performance reasons



Lattice

-- OLTP vs. OLAP

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

- Building A Data Warehouse ...

- The builders of Data warehouse should take a broad view of the anticipated use of the warehouse.
- The design should support ad-hoc querying
- An appropriate schema should be chosen that reflects the anticipated usage.

... - Building A Data Warehouse ...

- The Design of a Data Warehouse involves following steps
  - Acquisition of data for the warehouse.
  - Ensuring that Data Storage meets the query requirements efficiently.
  - Giving full consideration to the environment in which the data warehouse resides.

... - Building A Data Warehouse ...

- Acquisition of data for the warehouse ...
  - The data must be extracted from multiple, heterogeneous sources.
  - Data must be formatted for consistency within the warehouse.
  - The data must be cleaned to ensure validity.
    - Difficult to automate cleaning process.
    - Back flushing, upgrading the data with cleaned data.

... - Building A Data Warehouse ...

- ... Acquisition of data for the warehouse
  - The data must be fitted into the data model of the warehouse.
  - The data must be loaded into the warehouse.
    - Proper design for refresh policy should be considered.

… - Building A Data Warehouse …

- Storing the data according to the data model of the warehouse
- Creating and maintaining required data structures.
- Creating and maintaining appropriate access paths.
- Providing for time-variant data as new data are added.
- Supporting the updating of warehouse data.
- Refreshing the data.
- Purging data.

... - Building A Data Warehouse

- Usage projections
- The fit of the data model
- Characteristics of available resources
- Design of the metadata component
- Modular component design
- Design for manageability and change
- Considerations of distributed and parallel architecture
    - Distributed vs. federated warehouses.

- Warehouse vs. Data Views

- Views and data warehouses are alike in that they both have read-only extracts from the databases. However, data warehouses are different from views in the following ways:
  - Data Warehouses exist as persistent storage instead of being materialized on demand.
  - Data Warehouses are not usually relational, but rather multi-dimensional.
  - Data Warehouses can be indexed for optimization.
  - Data Warehouses provide specific support of functionality.
  - Data Warehouses deals huge volumes of data that is contained generally in more than one database.

- Difficulties of implementing Data Warehouses

- Lead time is huge in building a data warehouse. Potentially it takes years to build and efficiently maintain a data warehouse.
- Both quality and consistency of data are major concerns.
- Revising the usage projections regularly to meet the current requirements. The data warehouse should be designed to accommodate addition and attrition of data sources without major redesign
- Administration of data warehouse would require far broader skills than are needed for a traditional database.

END