CS 412 Intro. to Data Mining

Chapter 4. Data Warehousing and On-line Analytical Processing

Arindam Banerjee, Computer Science, UIUC, Fall 2021

Chapter 4: Data Warehousing and On-line Analytical Processing

- Data Warehouse: Basic Concepts 
- Data Warehouse Modeling: Data Cube and OLAP
- Summary

What is a Data Warehouse?

- Defined in many different ways, but not rigorously
  - Support decision
  - Maintained Separately
  - Information processing
- "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

- Help make decisions
  - A **simple** and **concise** view (modeling and analysis)
  - Not details (transaction processing)
  - Organizing around major subjects, such as customer, product, sales
  - Excluding data that are not useful in the decision support process

Data Warehouse—Integrated

- Integrating Multiple, heterogeneous sources
  - Ex. relational databases, flat files, on-line transaction records
- Consistency
  - Data cleaning and data integration techniques are applied
  - Ex. Hotel price: differences in currency, tax, breakfast covered, and parking
  - When data is moved to the warehouse, it is converted

Data Warehouse—Time Variant

Data Warehouse	Operational Database
Long time horizon (e.g., past 5-10 years)	current value data
Contains an element of time, explicitly or implicitly	data may or may not contain “time element”

Data Warehouse—Nonvolatile

- Independence – A physically separate store
- Static – No data management (updates, transaction processing, recovery, and concurrency control mechanisms)
- Requires only two operations in data accessing:
  - *initial loading of data* and *access of data*

Why a Separate Data Warehouse?

- 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
- Note: There are more and more systems which perform OLAP (online analytical processing) analysis directly on relational databases

OLTP vs. OLAP

- OLTP: Online transactional processing
  - DBMS operations
  - Query and transactional processing
- OLAP: Online analytical processing
  - Data warehouse operations
  - Drilling, slicing, dicing, etc.

	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

OLTP vs. OLAP

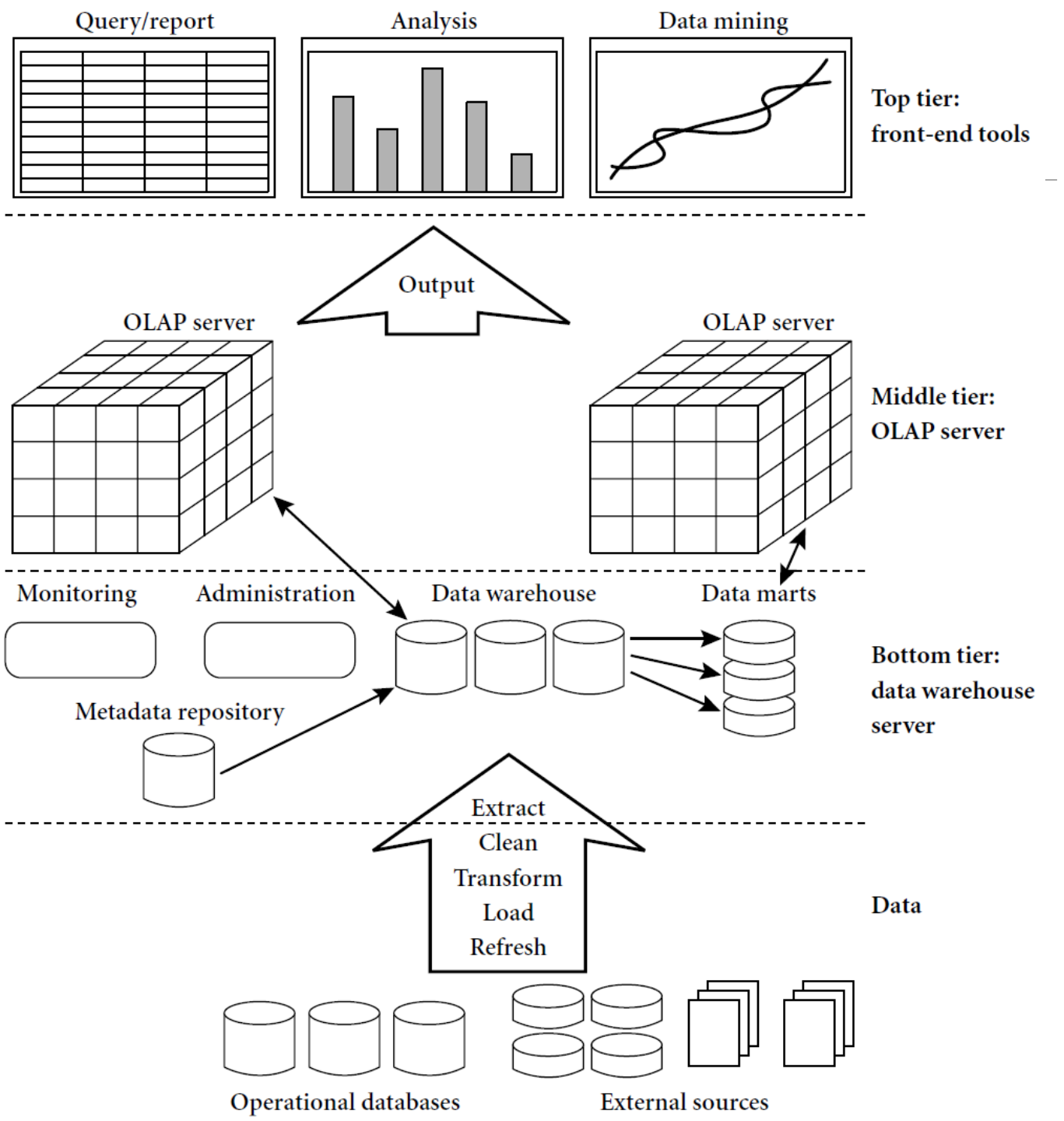
Table 4.1: Comparison of OLTP and OLAP Systems

Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements decision support
DB design	ER-based, application-oriented	star/snowflake, subject-oriented
Data	current, guaranteed up-to-date	historic, accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	GB to high-order GB	$\geq$ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Note: Table is partially based on Chaudhuri and Dayal [CD97].

Data Warehouse: A Multi-Tiered Architecture

- Top Tier: Front-End Tools
- Middle Tier: OLAP Server
- Bottom Tier: Data Warehouse Server
- Data



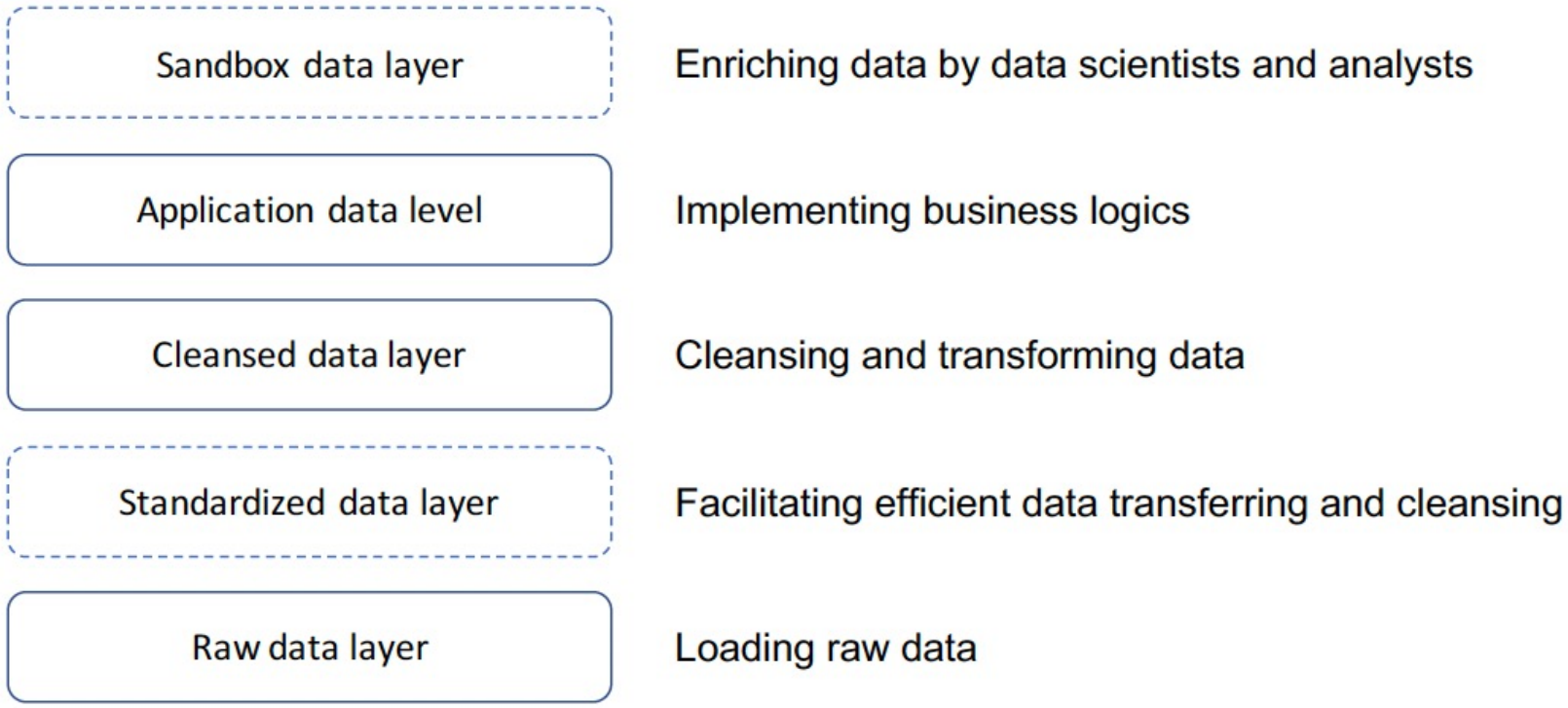
Extraction, Transformation, and Loading (ETL)

- **Data extraction**
  - get data from multiple, heterogeneous, and external sources
- **Data cleaning**
  - detect errors in the data and rectify them when possible
- **Data transformation**
  - convert data from legacy or host format to warehouse format
- **Load**
  - sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions
- **Refresh**
  - propagate the updates from the data sources to the warehouse

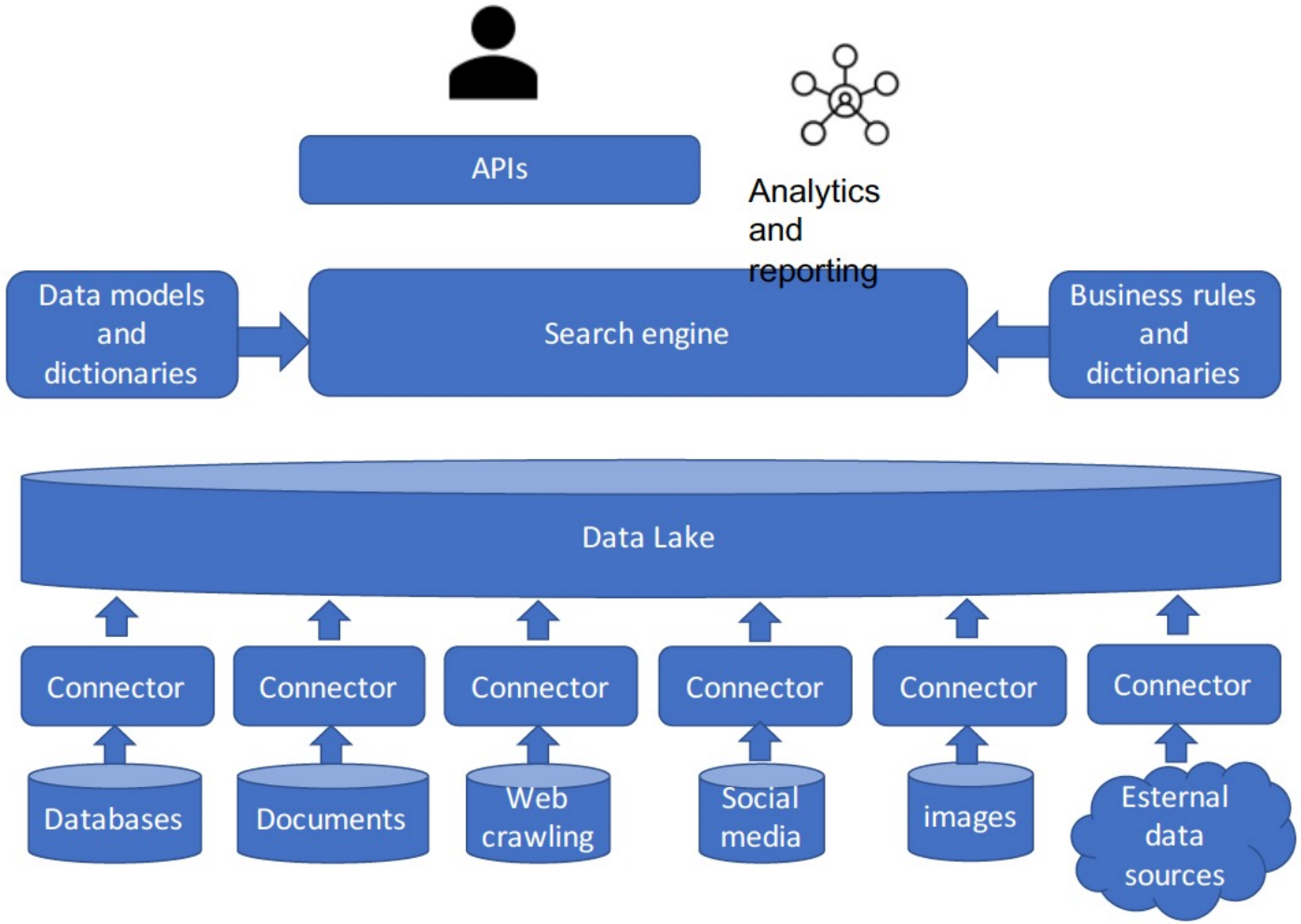
Three Data Warehouse Models

- **Enterprise warehouse -** Specially designed for the entire organization
- **Data Mart**
  - Specific, selected groups
  - Independent vs. dependent (directly from warehouse) data mart
- **Virtual warehouse**
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized
- **Data Lakes**
  - Single repo of all enterprise data in natural (possibly different) format
  - Base for all data related tasks, for all users, not structured like a warehouse
  - Does not need the design and development time like a warehouse

Layers of Data Storage in Data Lakes



Conceptual Architecture of Data Lakes



Metadata Repository

- **Meta data** is data about data. It stores:
  - Description of structure (schema, etc.)
  - Operational meta-data
    - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
  - The algorithms used for summarization
  - The mapping from operational environment to the data warehouse
  - Data related to system performance
    - warehouse schema, view and derived data definitions
  - Business data
    - business terms and definitions, ownership of data, charging policies

Chapter 4: Data Warehousing and On-line Analytical Processing

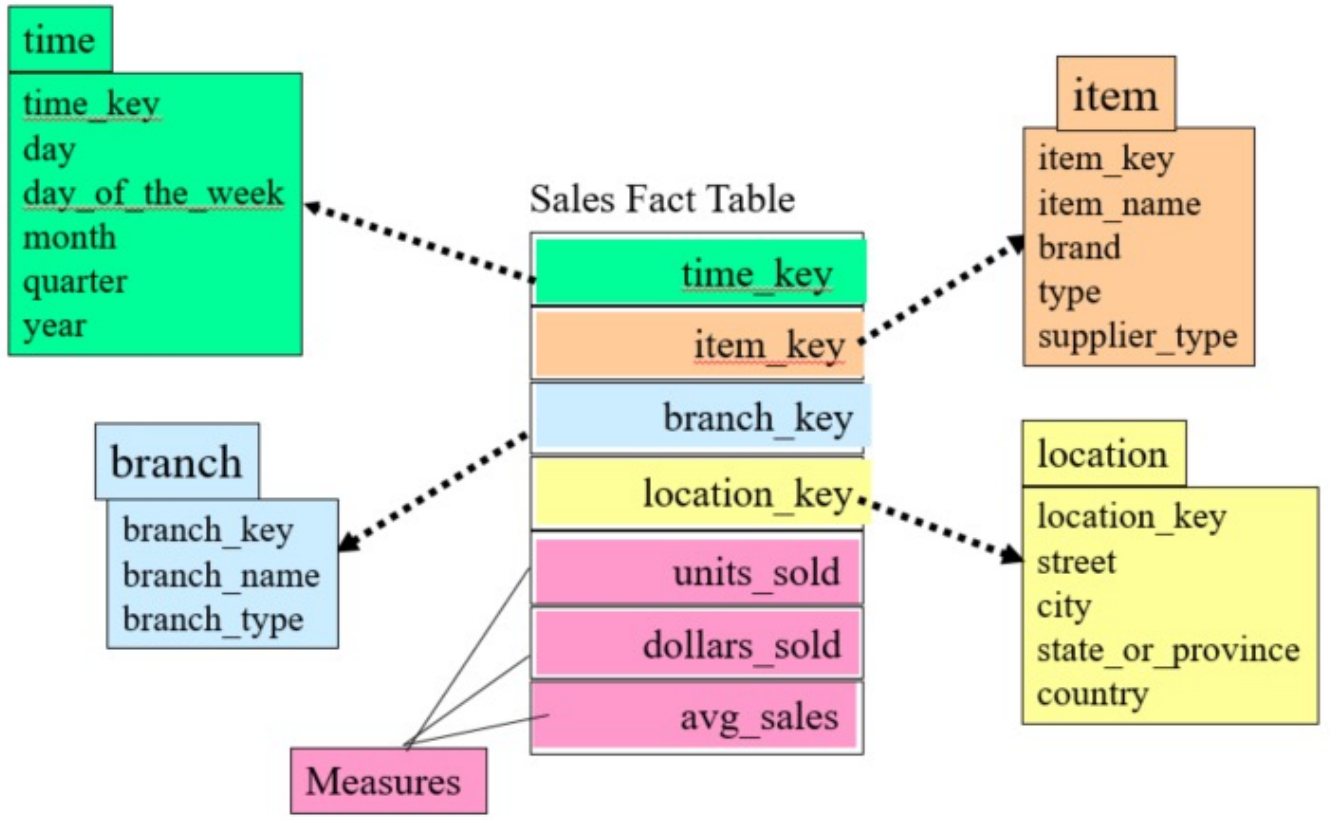
- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Summary

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
- Main function is to provide summarizations of the data
  - E.g., summarize the units or dollars sold at a particular store over a particular time period
- Can compute summarizations online (as they are requested)
  - Can be very slow
- Better to precalculate some summarizations

Design of Data Warehouses

- **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
- Different schema exist
  - Star
  - Snowflake
  - Fact constellation



Example: 2-D View of Sales Data

Table 4.2: 2-D View of Sales Data According to *time* and *item*
location = *"Vancouver"*

	item *(type)*			
time *(quarter)*	*home entertainment*	*computer*	*phone*	*security*
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

Note: The sales are from branches located in the city of Vancouver. The measure displayed is *dollars_sold* (in thousands).

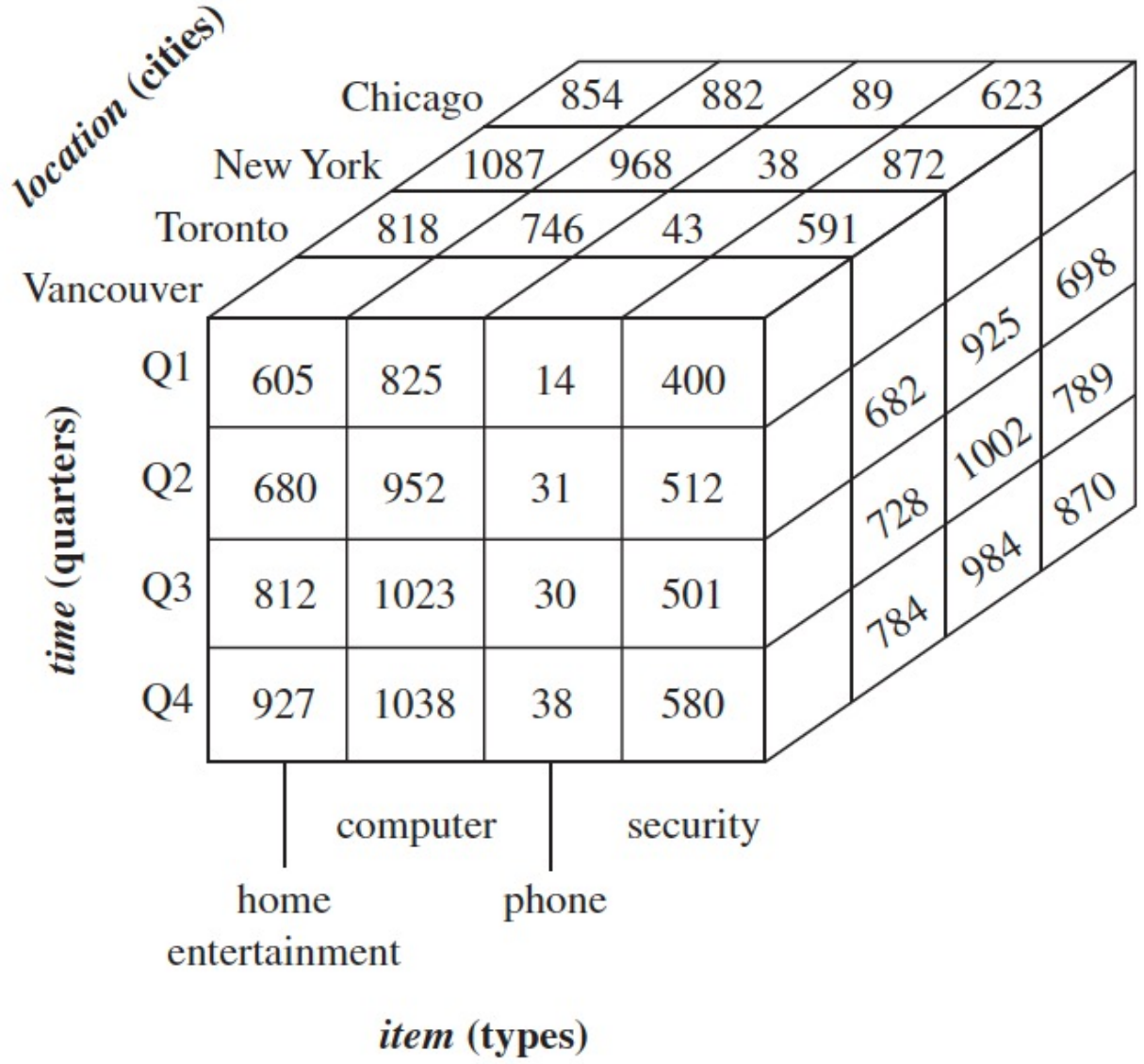
Example: 3-D View of Sales Data

Table 4.3: 3-D View of Sales Data According to *time*, *item*, and *location*

	location = ''Chicago''				*location* = ''New York''				*location* = ''Toronto''				*location* = ''Vancouver''			
	item				*item*				*item*				*item*			
	home				*home*				*home*				*home*			
time	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

Note: The measure displayed is *dollars_sold* (in thousands).

Example: 3-D View of Sales Data



Example: 4-D View of Sales Data

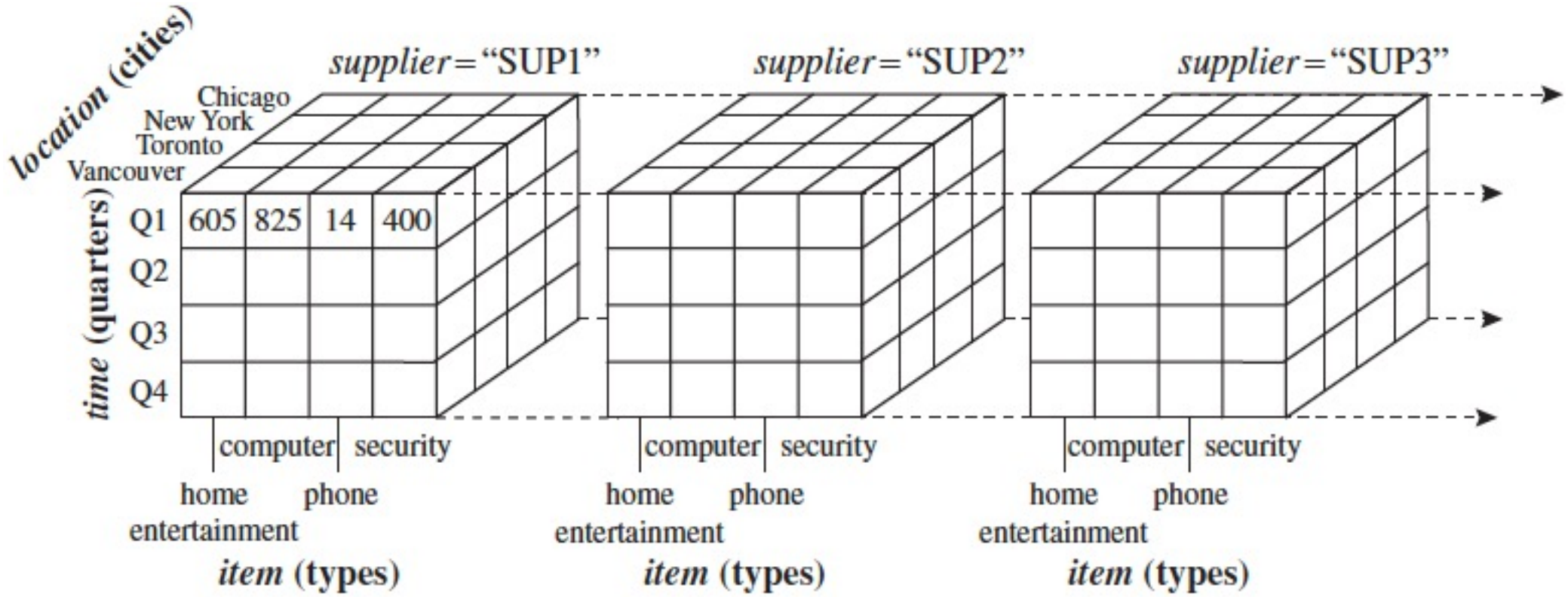
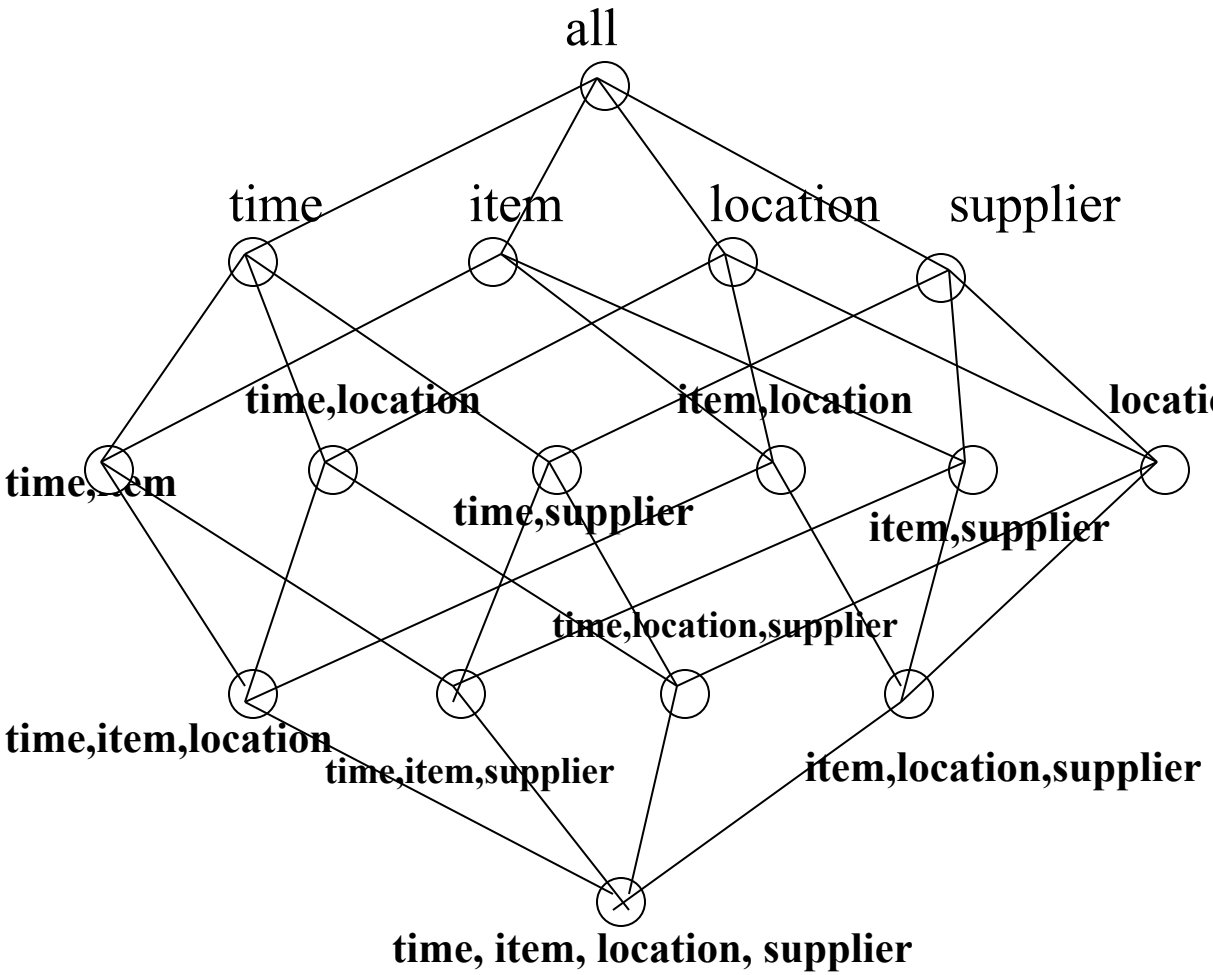


Figure 4.4: A 4-D data cube representation of sales data, according to *time*, *item*, *location*, and *supplier*. The measure displayed is *dollars_sold* (in thousands). For improved readability, only some of the cube values are shown.

Data Cube: A Lattice of Cuboids



0-D (*apex*) cuboid

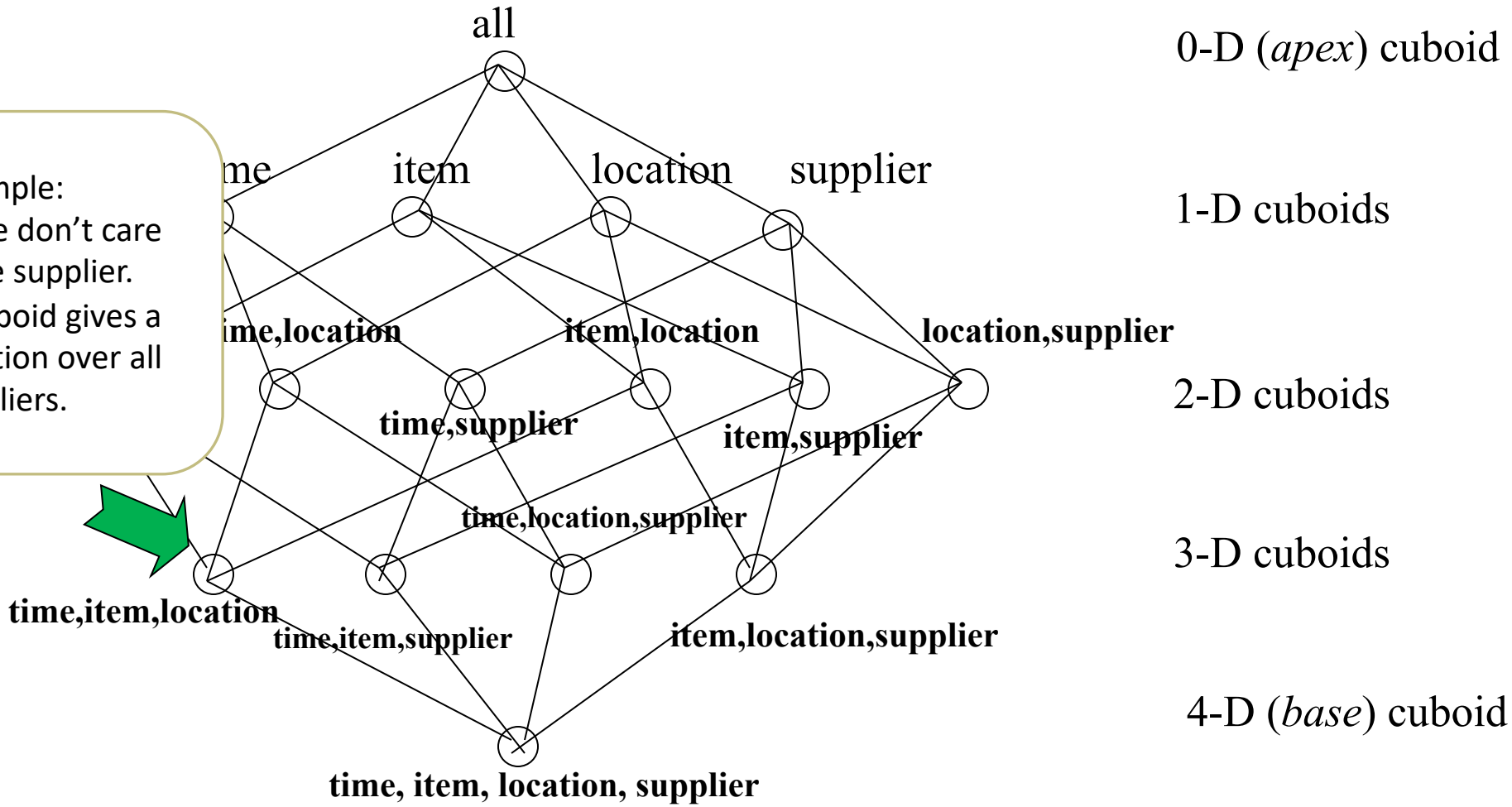
1-D cuboids

2-D cuboids

3-D cuboids

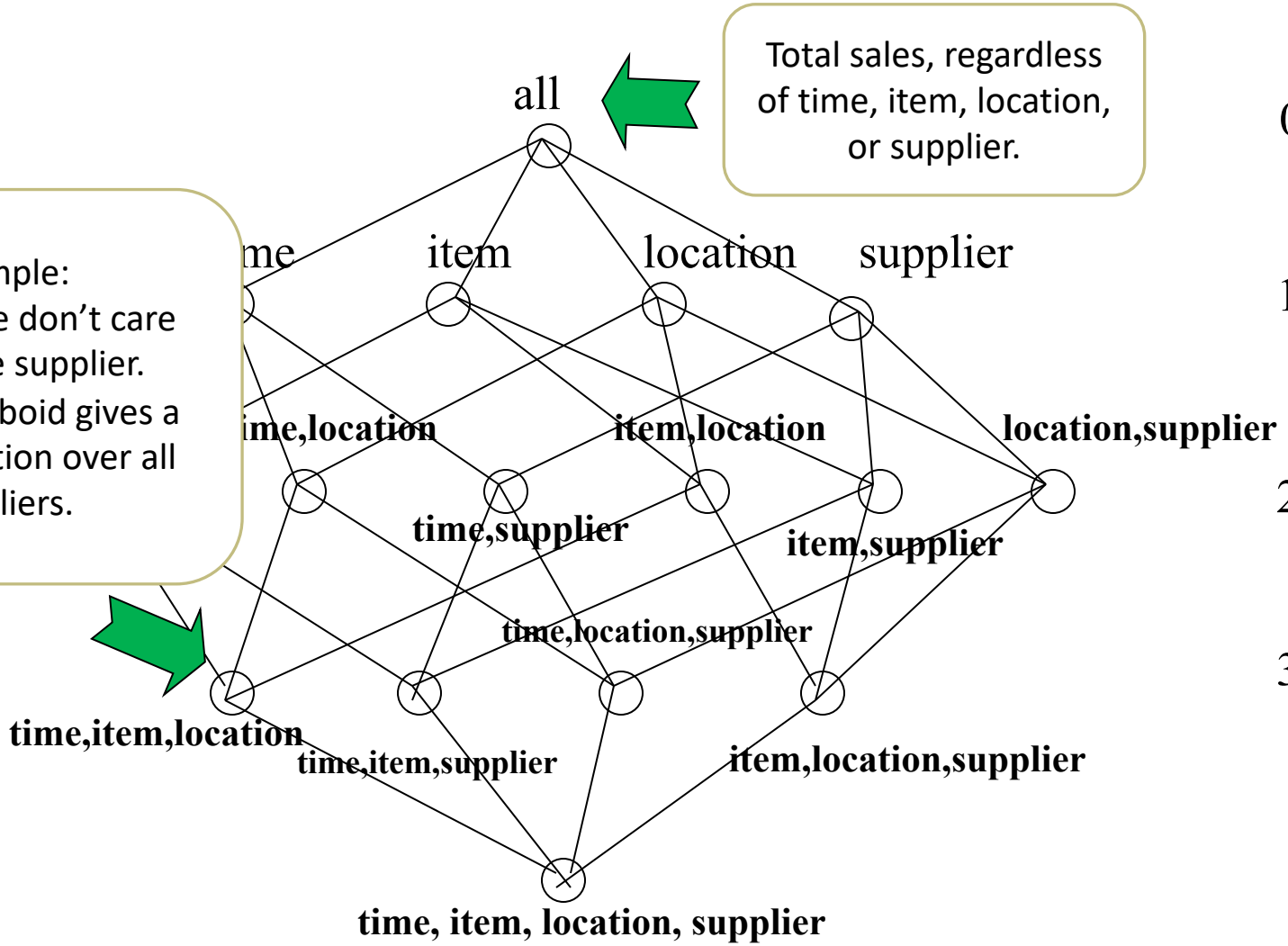
4-D (*base*) cuboid

Data Cube: A Lattice of Cuboids



Example:
Suppose we don't care about the supplier.
This 3-D cuboid gives a summarization over all suppliers.

Data Cube: A Lattice of Cuboids



Total sales, regardless of time, item, location, or supplier.

Example:
Suppose we don't care about the supplier.
This 3-D cuboid gives a summarization over all suppliers.

0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

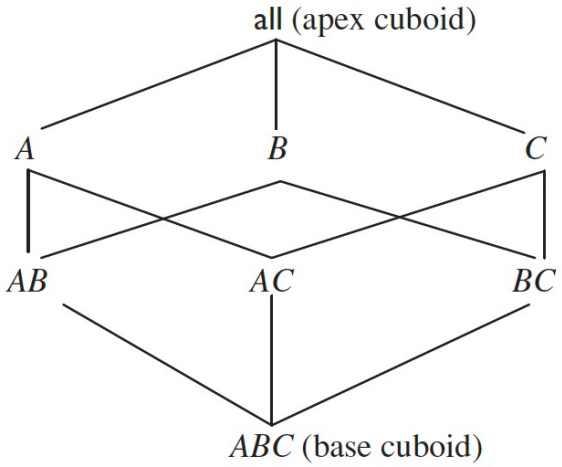
4-D (*base*) cuboid

Calculating Number of Cuboids

- Consider dimensions as binary numbers
- Example: 4 dimensions
  - Each is either in the cuboid, or not in the cuboid
  - (, , ,) ← choice of 0 or 1 for each element of vector
  - Sum up for each position: $2^3 + 2^2 + 2^1 + 2^0 + 1$ (0-d cuboid) = 2^4
- In general, 2^d cuboids (d = number of dimensions)

Calculating Number of Cuboids

- Tuple in a cuboid is a cell
  - Base cuboid is a base cell
  - Otherwise, aggregate cell
- Ancestor and descendant relationship
  - (Jan,*,*,2800) is ancestor of (Jan, Chicago,*, 2800)
- Each dimension may have concept hierarchies
  - Li for dimension i



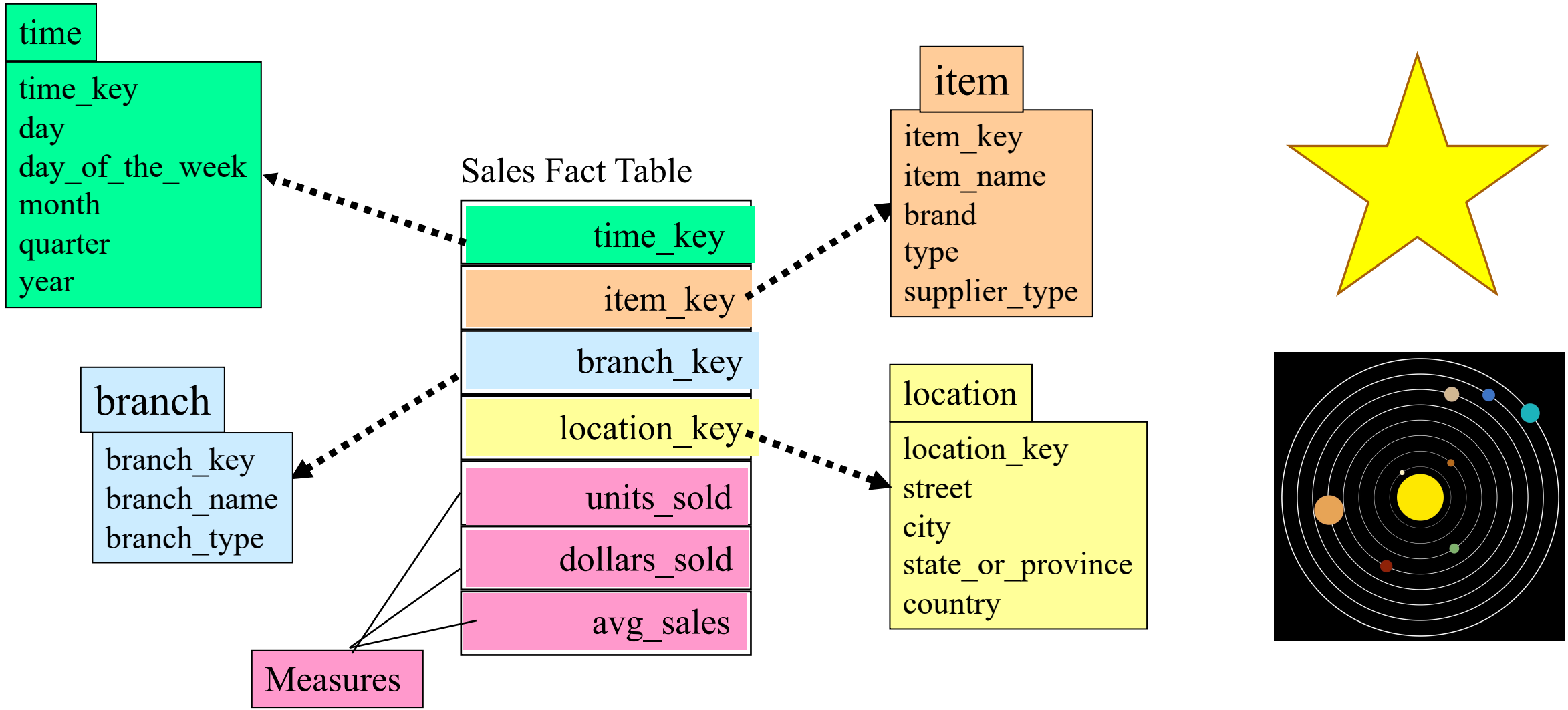
$$Total\ cubiods = \prod_{i=1}^{n}(L_i + 1)$$

Conceptual Modeling of Data Warehouses

- Modeling data warehouses: dimensions & measures
  - Star schema
  - Snowflake schema
  - Fact constellations

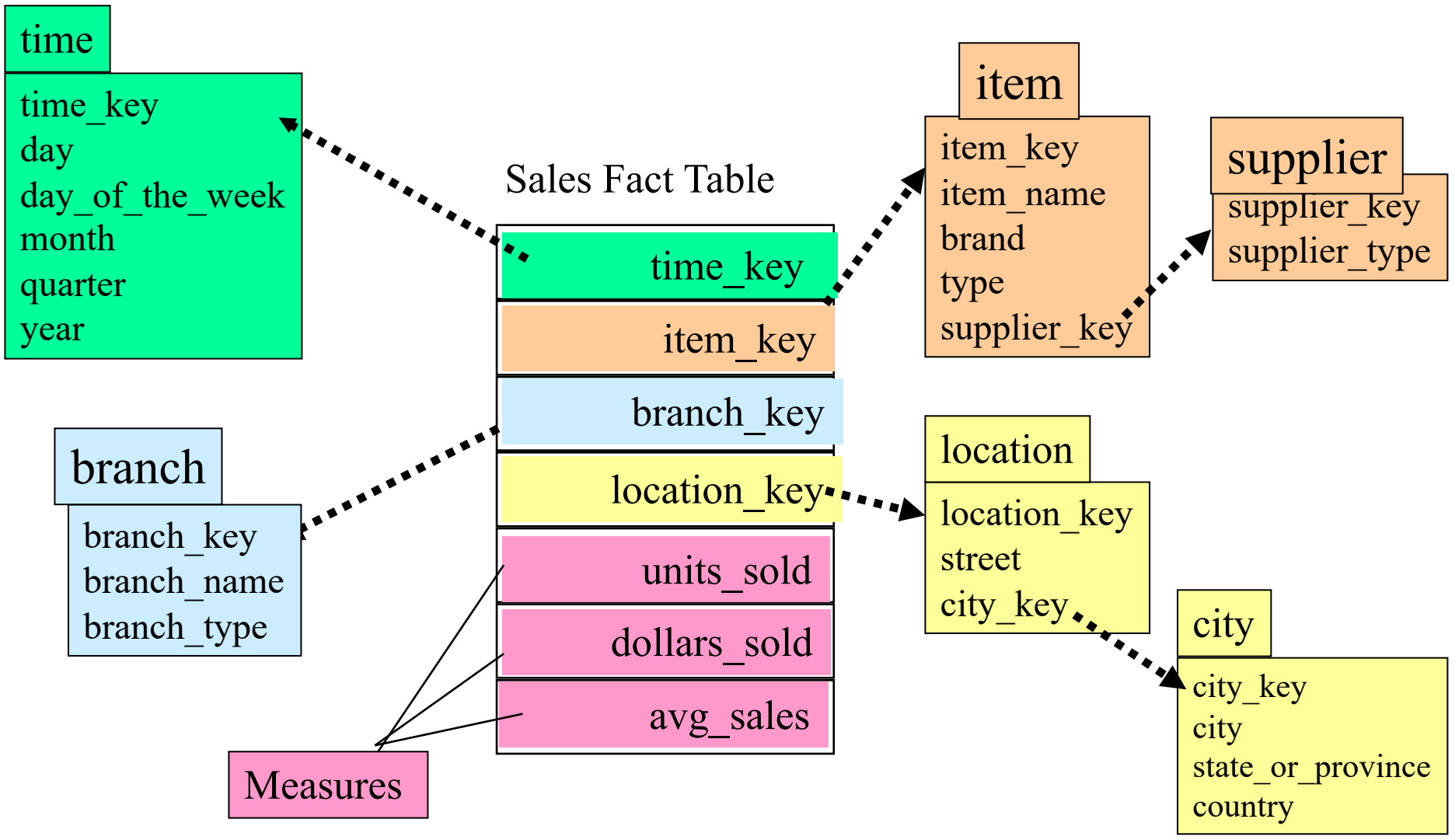
Star Schema: An Example

A fact table in the middle connected to a set of dimension tables



Snowflake Schema: An Example

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 Constellation: An Example

Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

time
time_key
day
day_of_the_week
month
quarter
year

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

Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

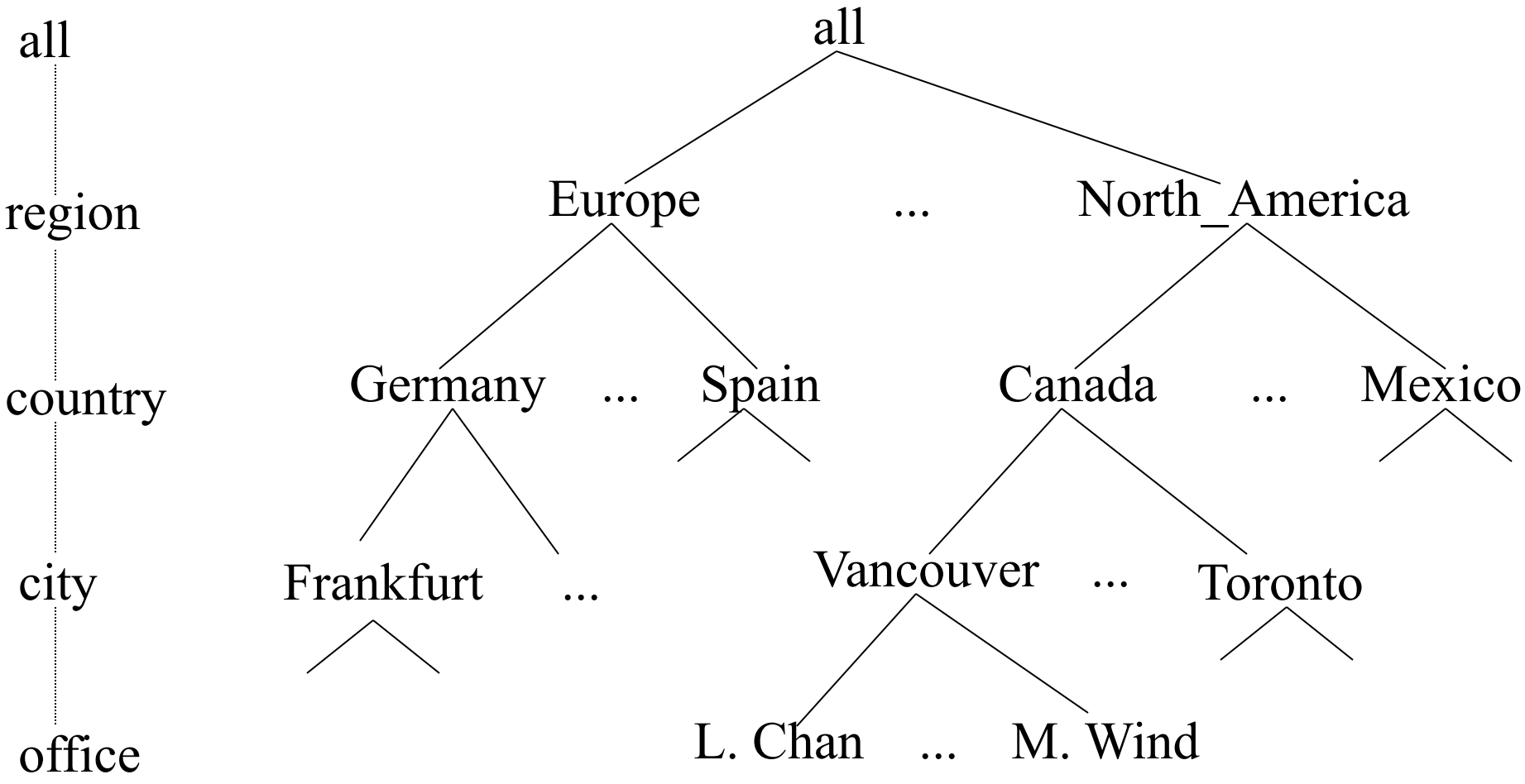
branch
branch_key
branch_name
branch_type

location
location_key
street
city
province_or_state
country

shipper
shipper_key
shipper_name
location_key
shipper_type

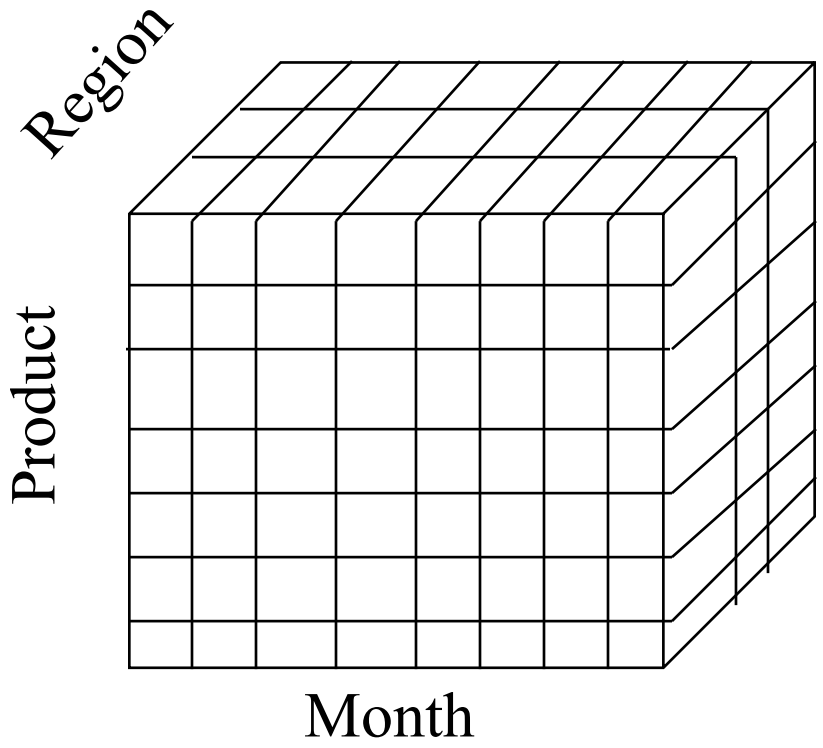
Measures

A Concept Hierarchy for a Dimension (location)

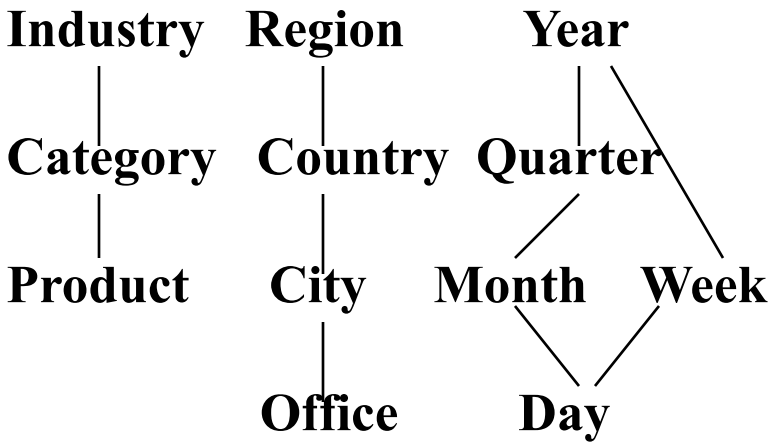


Multidimensional Data

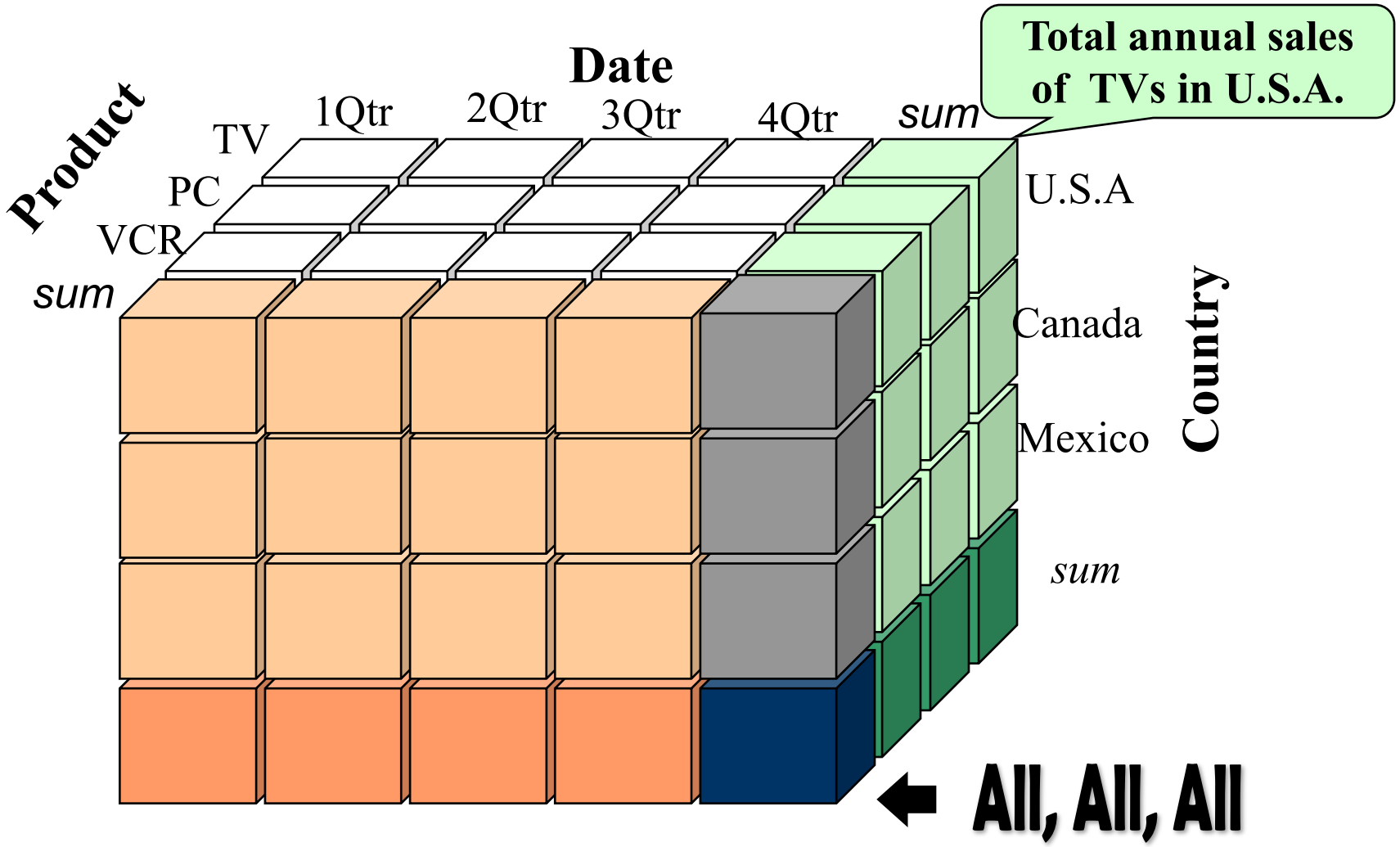
- Sales volume as a function of product, month, and region

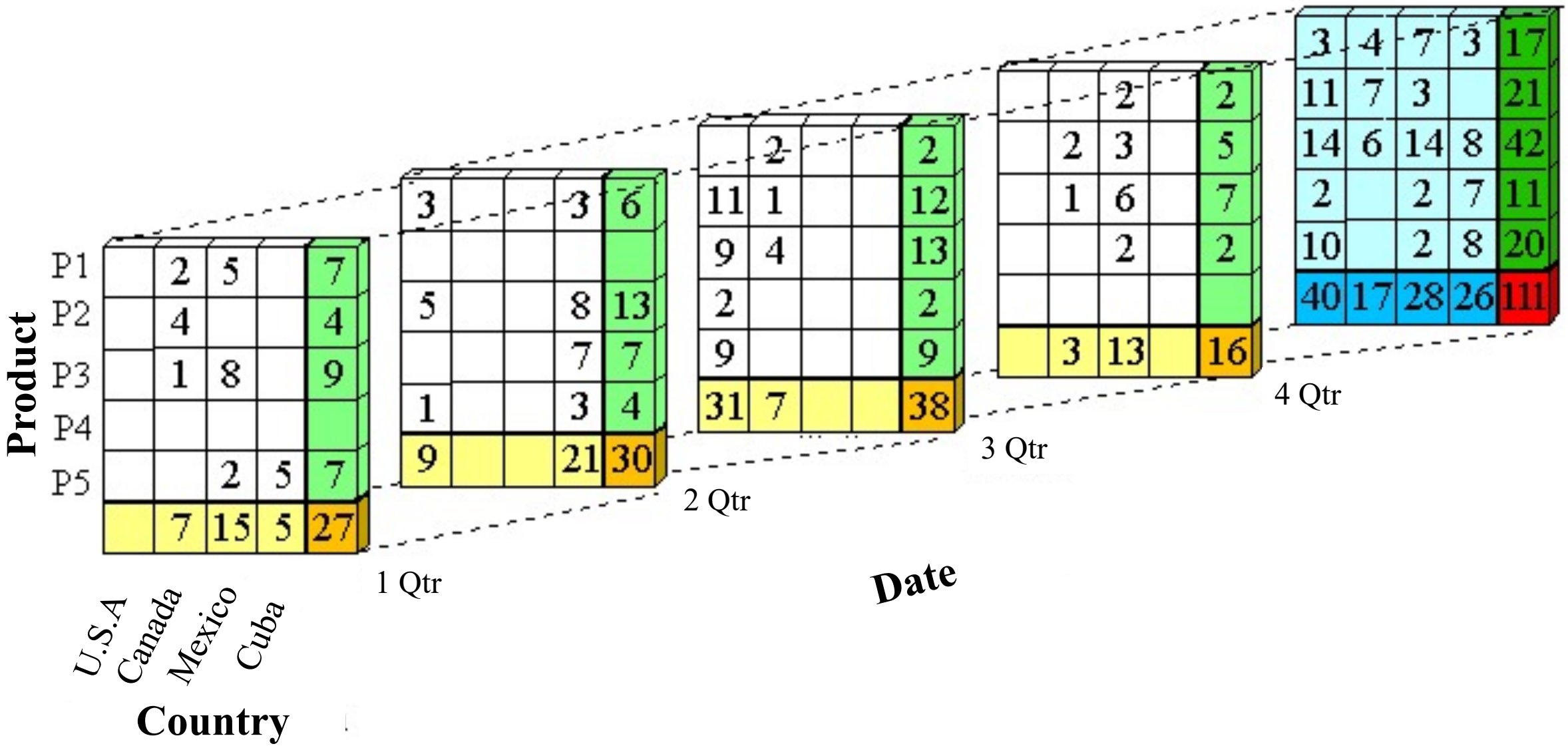


Dimensions: *Product, Location, Time*
Hierarchical summarization paths



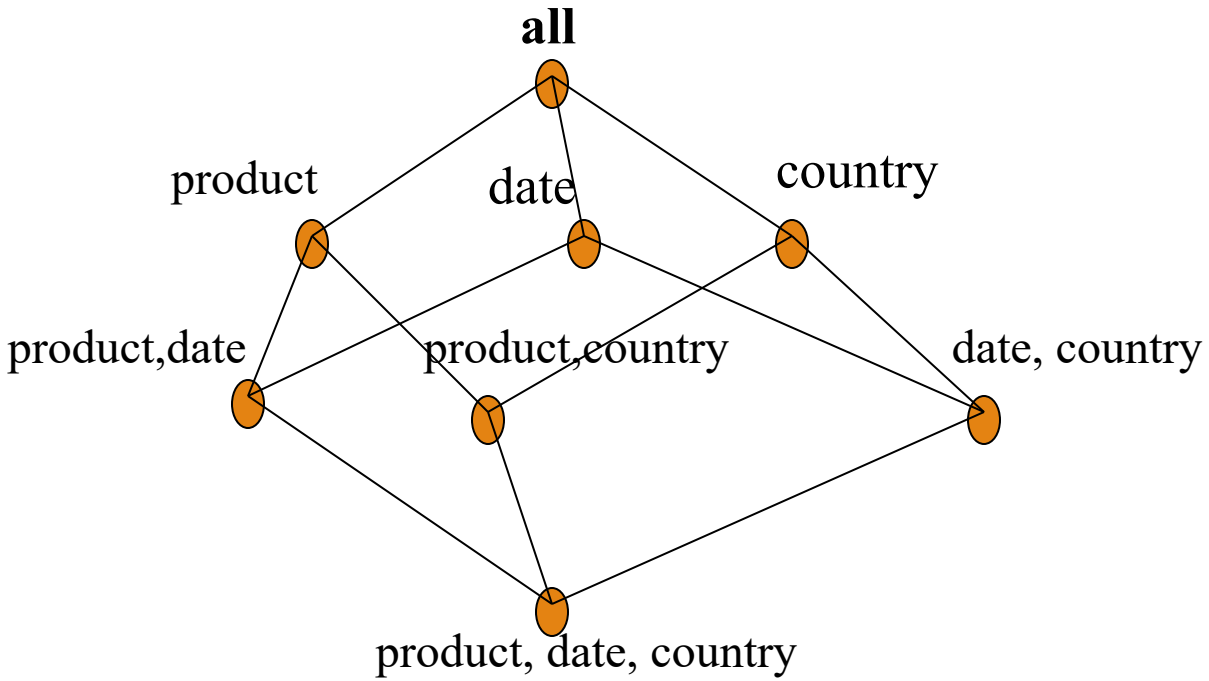
A Sample Data Cube




Product
P1
P2
P3
P4
P5
2 5 7
4 4
1 8 9
2 5 7
7 15 5 27
3 3 6
5 8 13
7 7
1 3 4
9 21 30
2 2
11 1 12
9 4 13
2 2
9 9
31 7 38
2 2
2 3 5
1 6 7
2 2
3 13 16
3 4 7 3 17
11 7 3 21
14 6 14 8 42
2 2 7 11
10 2 8 20
40 17 28 26 111
1 Qtr
2 Qtr
3 Qtr
4 Qtr
Date
U.S.A
Canada
Mexico
Cuba
Country

36

Cuboids Corresponding to the Cube



0-D (*apex*) cuboid

1-D cuboids

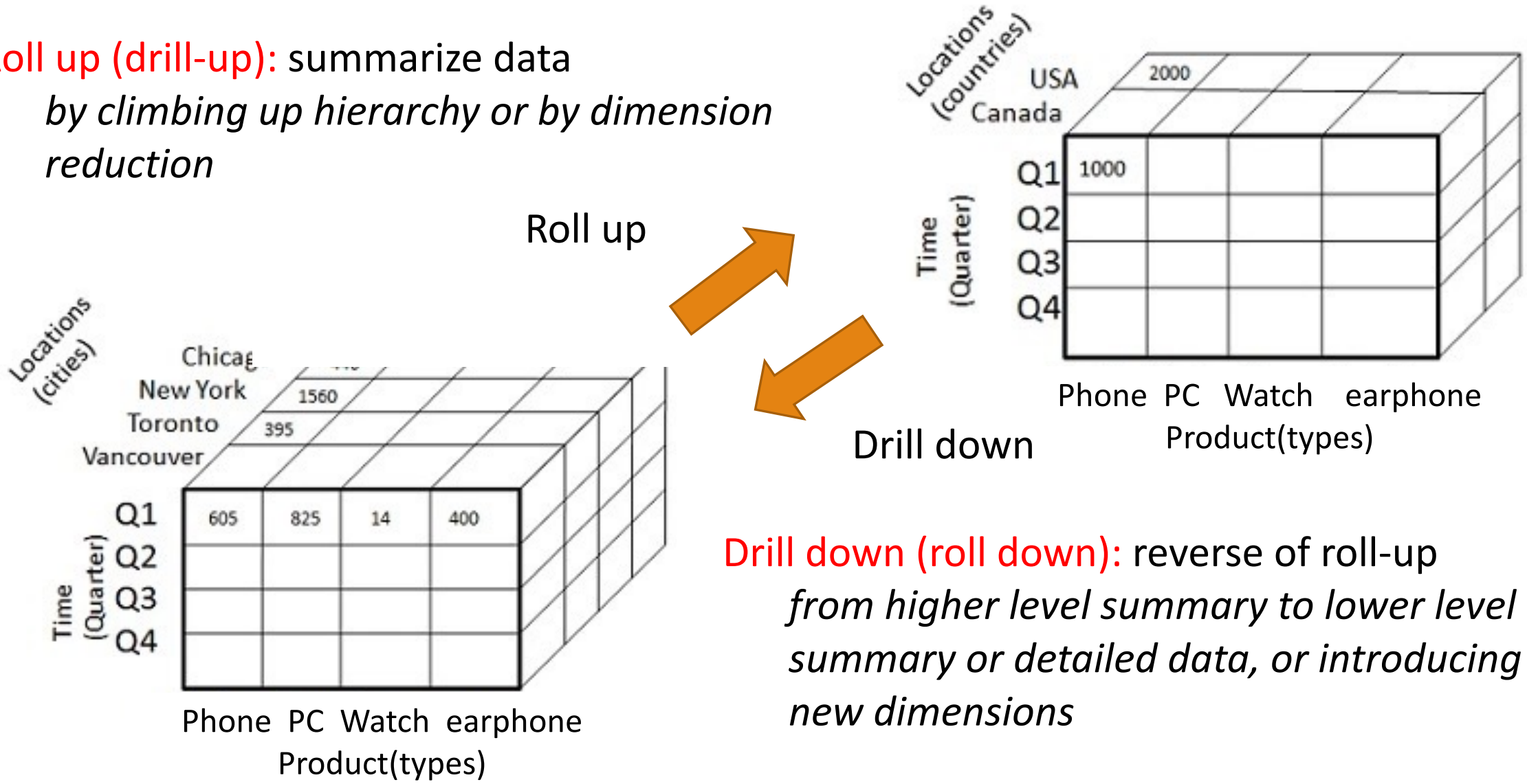
2-D cuboids

3-D (*base*) cuboid

How can we play with the Cube?

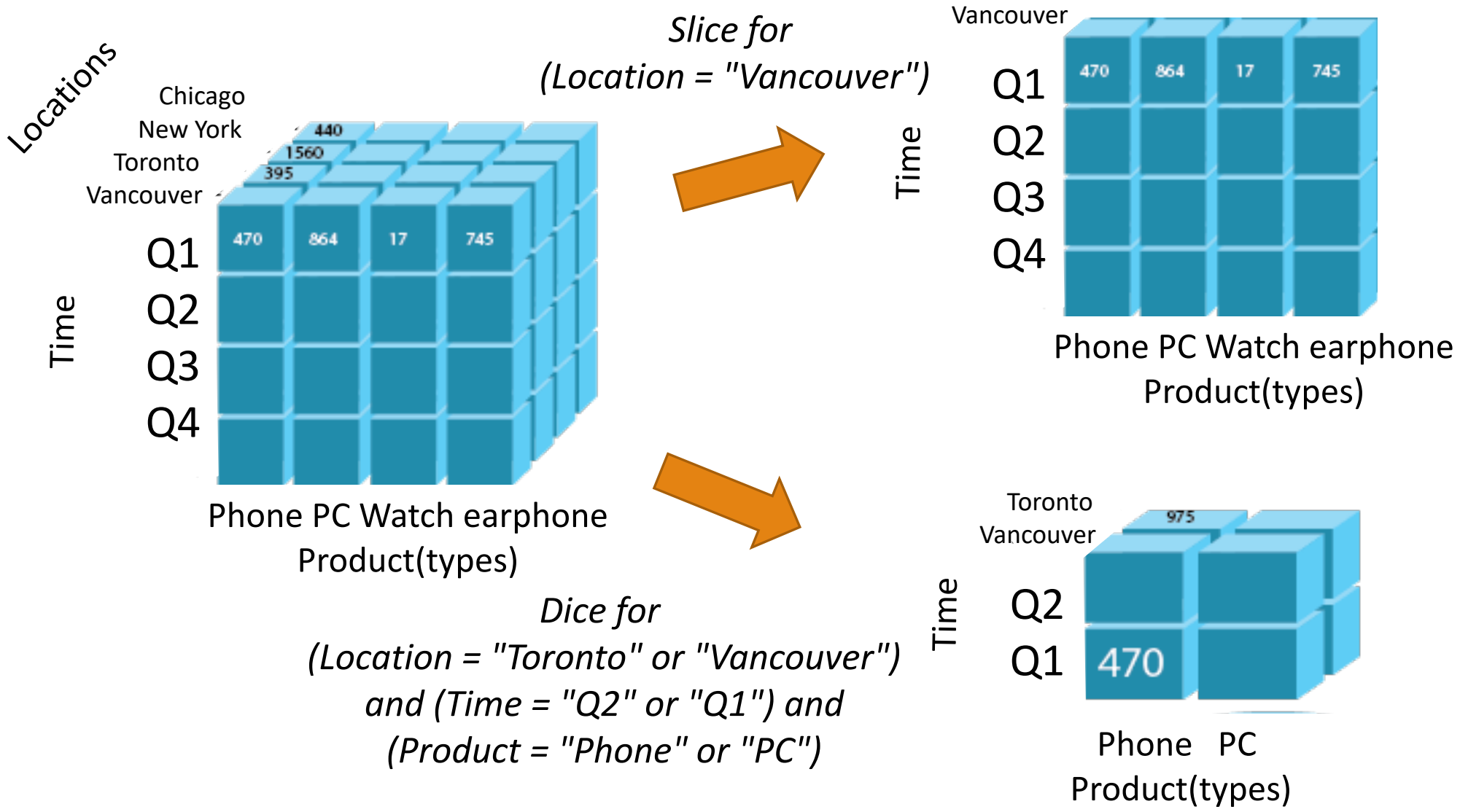
Roll up & Drill down

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

Dice and Slice



Chapter 4: Data Warehousing and On-line Analytical Processing

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Summary 

Summary

- Data warehousing: A multi-dimensional model of a data warehouse
  - A data cube consists of *dimensions* & *measures*
  - Star schema, snowflake schema, fact constellations
  - OLAP operations: drilling, rolling, slicing, dicing and pivoting
- Data Warehouse Architecture, Design, and Usage
  - Multi-tiered architecture
  - Business analysis design framework
  - Information processing, analytical processing, data mining

References (I)

- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB'96
- D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD'97
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE'97
- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.
- A. Gupta and I. S. Mumick. Materialized Views: Techniques, Implementations, and Applications. MIT Press, 1999
- J. Han. Towards on-line analytical mining in large databases. *SIGMOD Record*, 1998
- V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. SIGMOD'96

References (II)

- C. Imhoff, N. Galemmo, and J. G. Geiger. Mastering Data Warehouse Design: Relational and Dimensional Techniques. John Wiley, 2003
- W. H. Inmon. Building the Data Warehouse. John Wiley, 1996
- R. Kimball and M. Ross. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2ed. John Wiley, 2002
- P. O'Neil and D. Quass. Improved query performance with variant indexes. SIGMOD'97
- S. Sarawagi and M. Stonebraker. Efficient organization of large multidimensional arrays. ICDE'94
- P. Valduriez. Join indices. ACM Trans. Database Systems, 12:218-246, 1987.
- J. Widom. Research problems in data warehousing. CIKM'95.
- K. Wu, E. Otoo, and A. Shoshani, Optimal Bitmap Indices with Efficient Compression, ACM Trans. on Database Systems (TODS), 31(1), 2006, pp. 1-38.