Data Warehousing

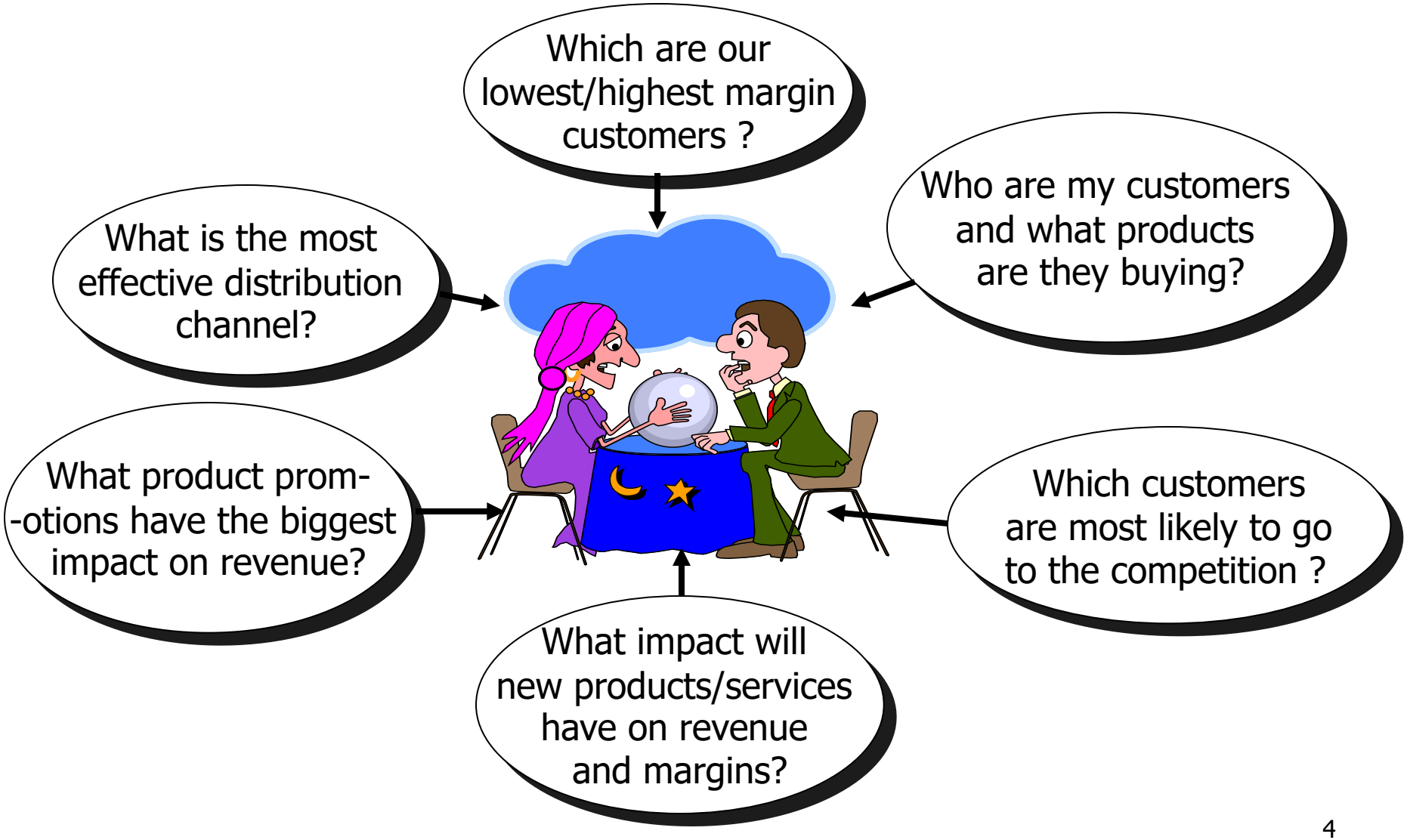
- Basic Concepts
  - Why? What? and Problems
  - Characteristics of DW
  - DW vs Operational DBMS
- DW Modeling
  - Data Cube and DW Schemas
  - OLAP
- Summary

Basic Concepts

Data, Data everywhere, yet ...

- I can't find the data I need
  - data is scattered over the network
  - many versions, subtle differences
- I can't get the data I need
  - need an expert to get the data
- I can't understand the data I found
  - available data poorly documented
- I can't use the data I found
  - results are unexpected
  - data needs to be transformed from one form to other

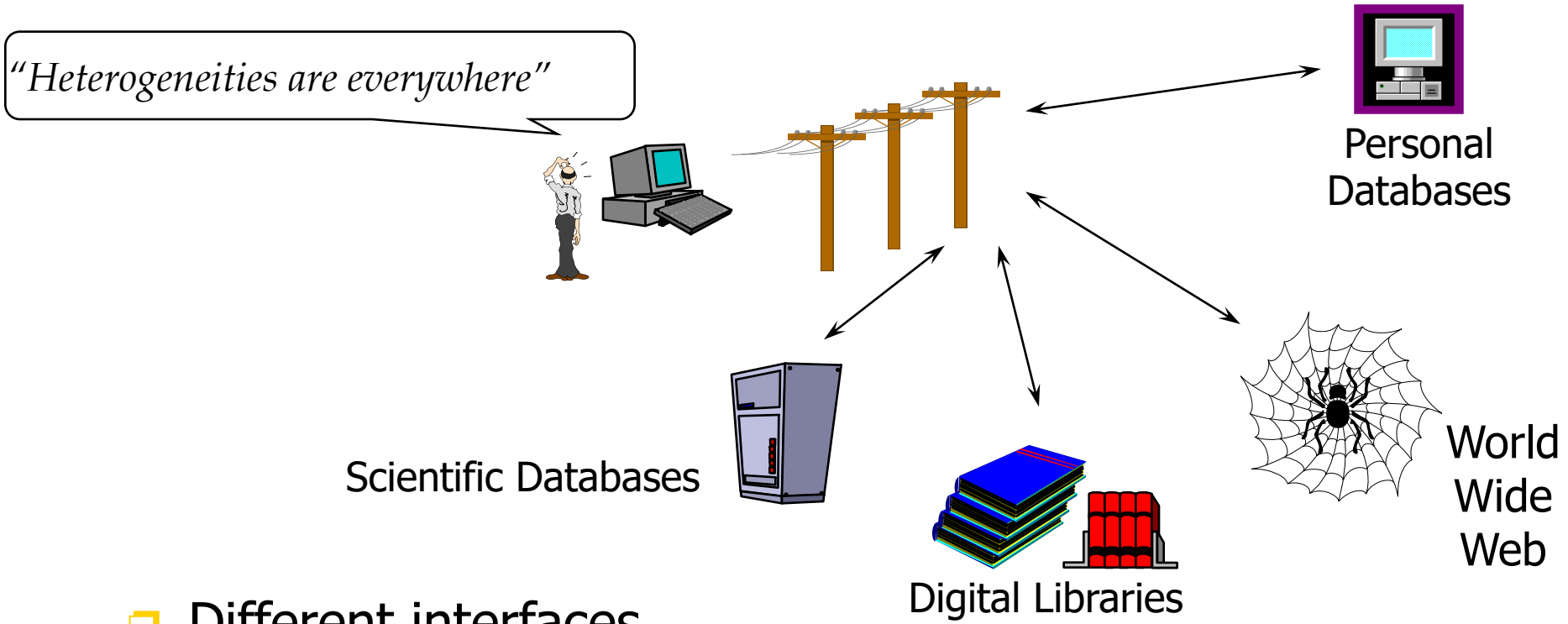
Why Data Warehousing?



Why Do We Need Data Warehouses?

- Consolidation of information resources
- Improved query performance
- Separate search and decision support functions from the operational systems
- Foundation for
  - data mining,
  - data visualization,
  - advanced reporting and
  - OLAP (On-Line Analytical Processing) tools

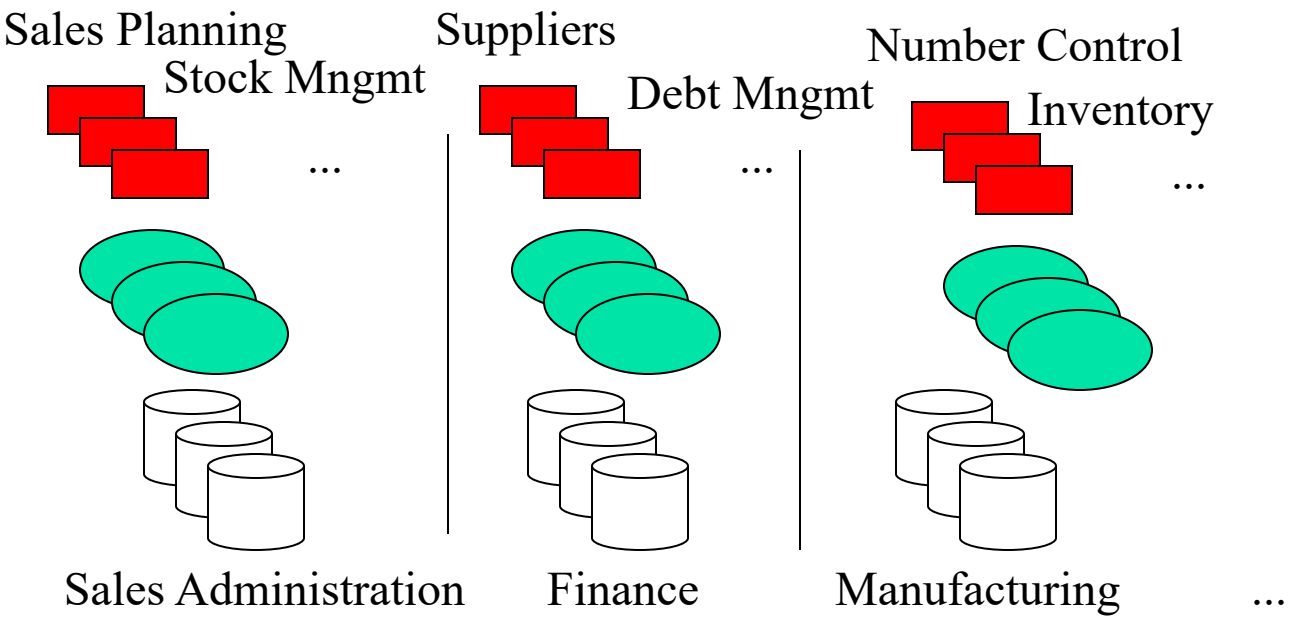
Root of the Problem: Heterogeneous Information Sources



- Different interfaces
- Different data representations
- Duplicate and inconsistent information

Additional Problem:
Data Management in Large Enterprises

- Vertical fragmentation of informational systems
- Result of application (user)-driven development of operational systems



What is a 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

What is a Data Warehouse Used for?

- In many organizations, we want a central "store" of all of our entities, concepts, metadata, and historical information
  - For doing data validation, complex mining, analysis, prediction, etc.
- One of the "modern" uses of the data warehouse is not only to support **analytics** but to serve as a reference to all of the entities in the organization
  - A cleaned, validated repository of what we know
    … which can be linked to by data sources
    … which may help with data cleaning
    … and which may be the basis of **data governance** (processes by which data is created and modified in a systematic way, e.g., to comply with gov't regulations)

What is a Data Warehouse Used for?

- Knowledge discovery
  - Making consolidated reports
  - Finding relationships and correlations
  - Data mining
  - Examples
    - Banks identifying credit risks
    - Insurance companies searching for fraud
- OLAP (Online Analytical Processing)
  - It contrasts with OLTP (on-line transaction processing) used to deal with the everyday running of one aspect of an enterprise.
  - OLTP systems are usually designed independently of each other and it is difficult for them to share information.

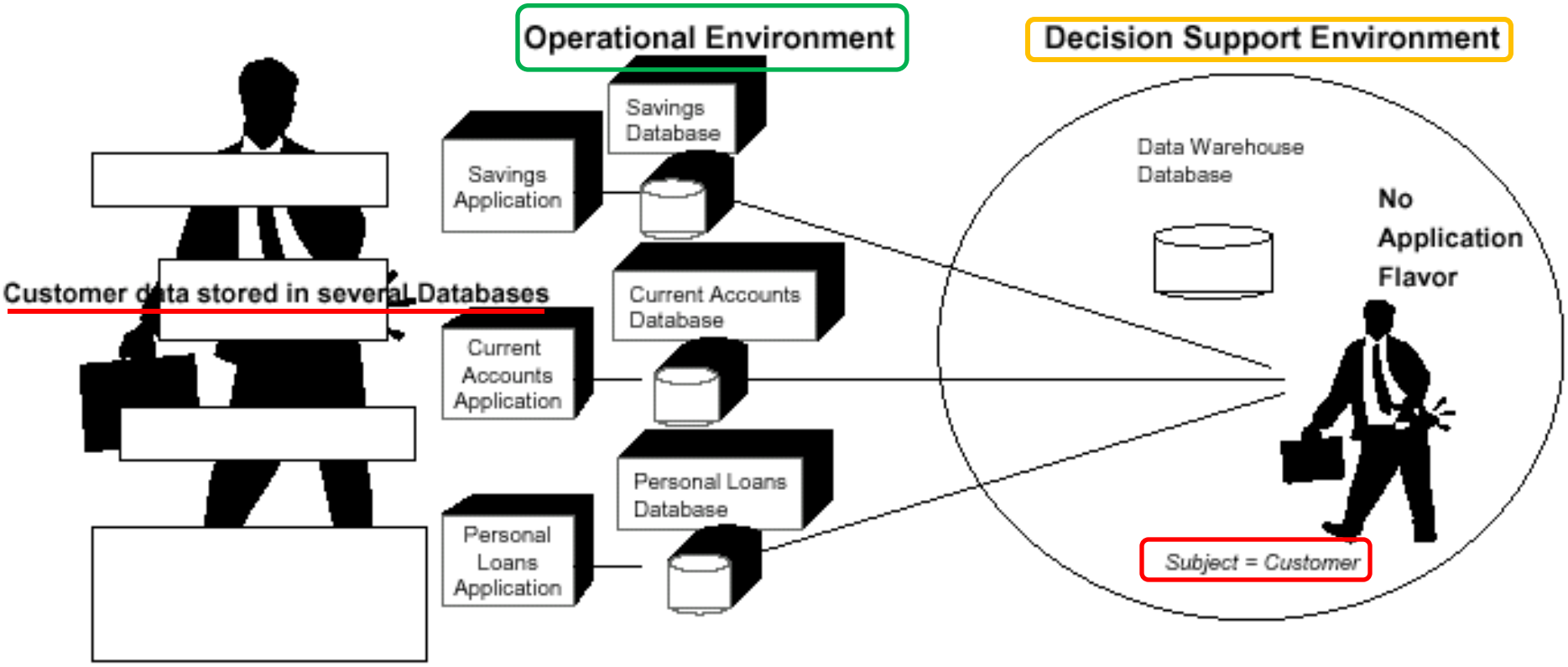
Characteristics of Data Warehouse: Subject-Oriented

- 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

Characteristics of 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.

Characteristics of Data Warehouse: Integrated (Cont.)



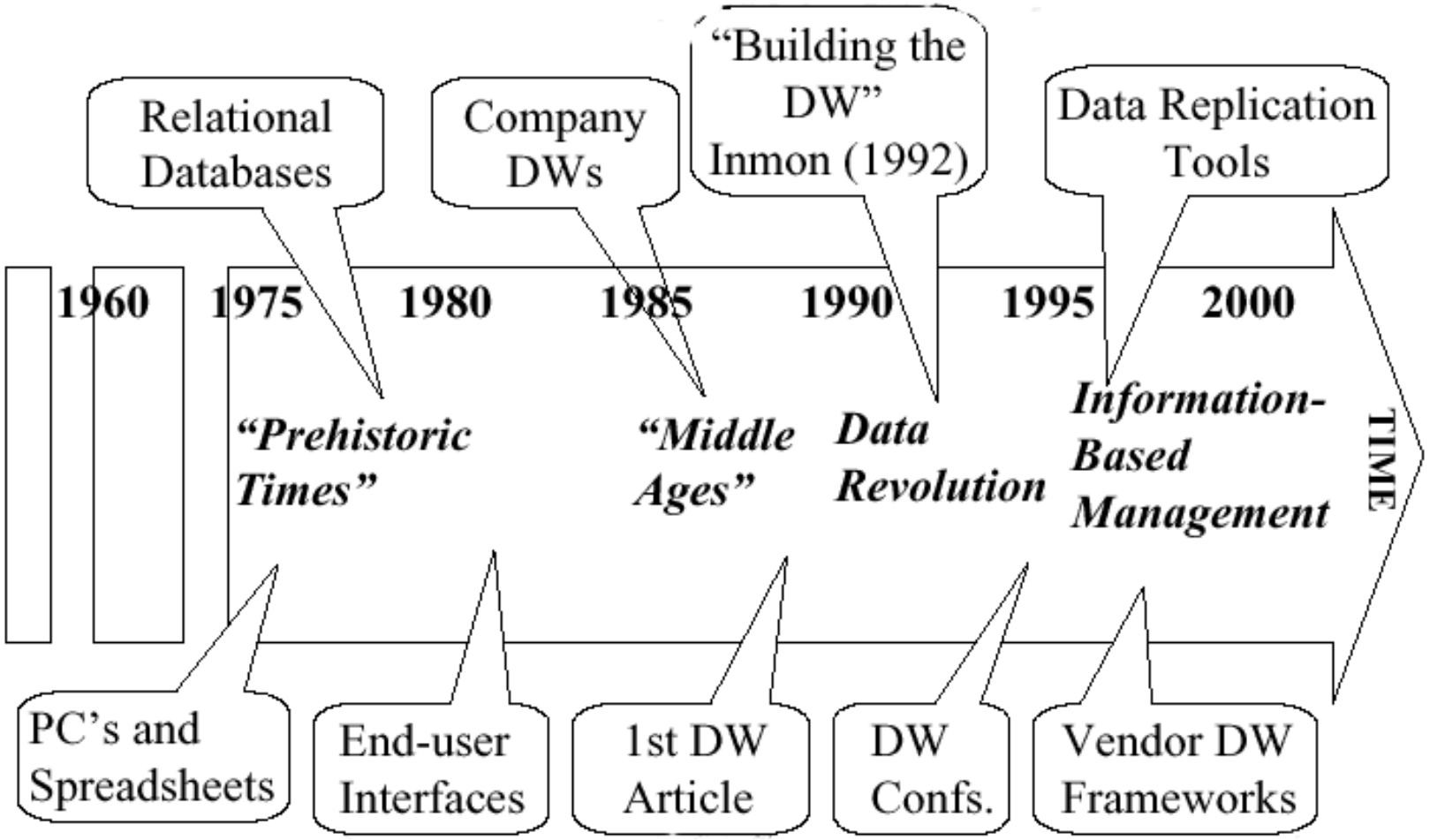
Characteristics of Data Warehouse: Time Variant

- 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"

Characteristics of Data Warehouse: Nonvolatile

- 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*

History of Data Warehouse



Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

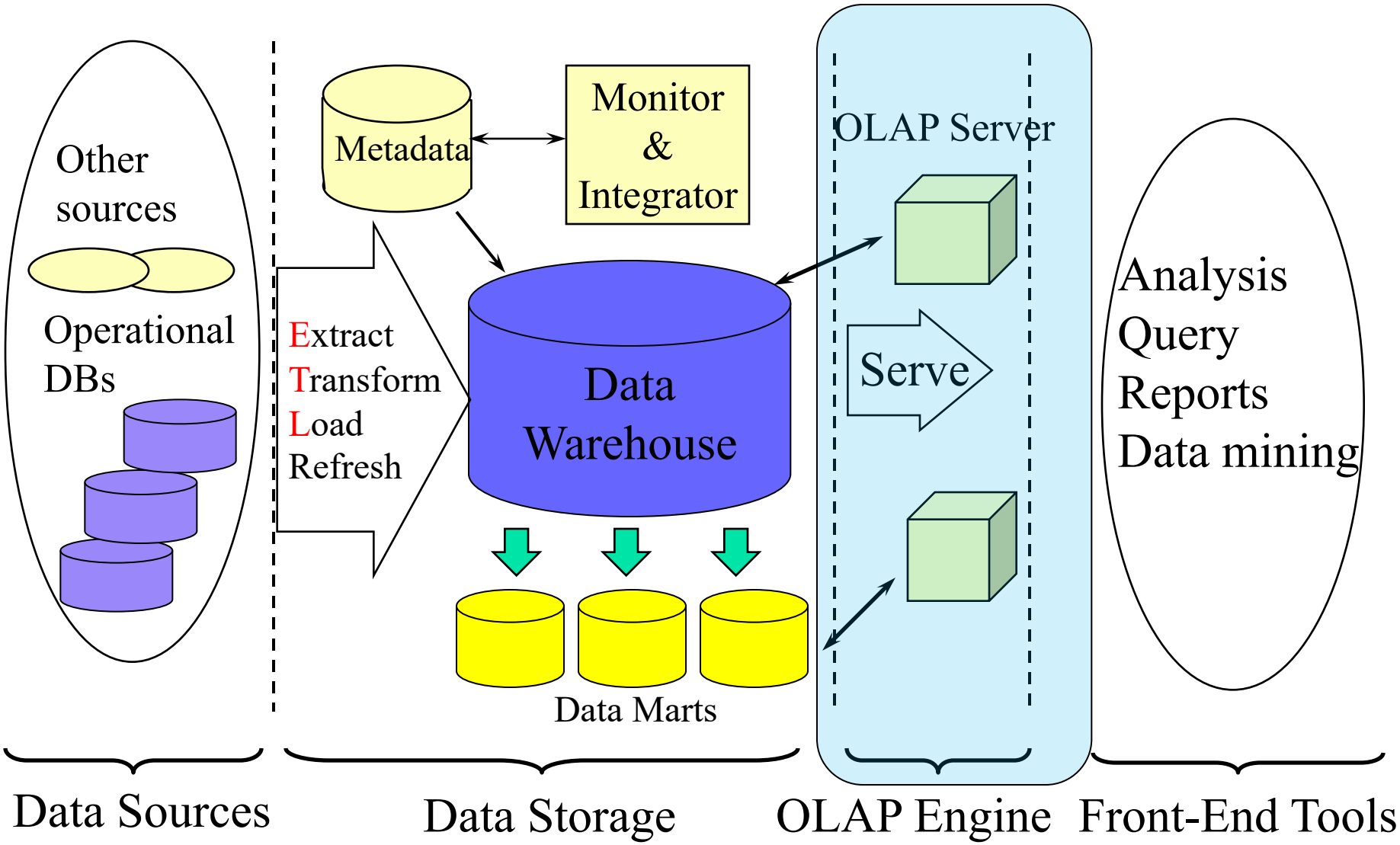
OLTP vs. OLAP

	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

Why a Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for 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: Decision support 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 analysis directly on relational databases

Data Warehouse's Multi-Tiered Architecture



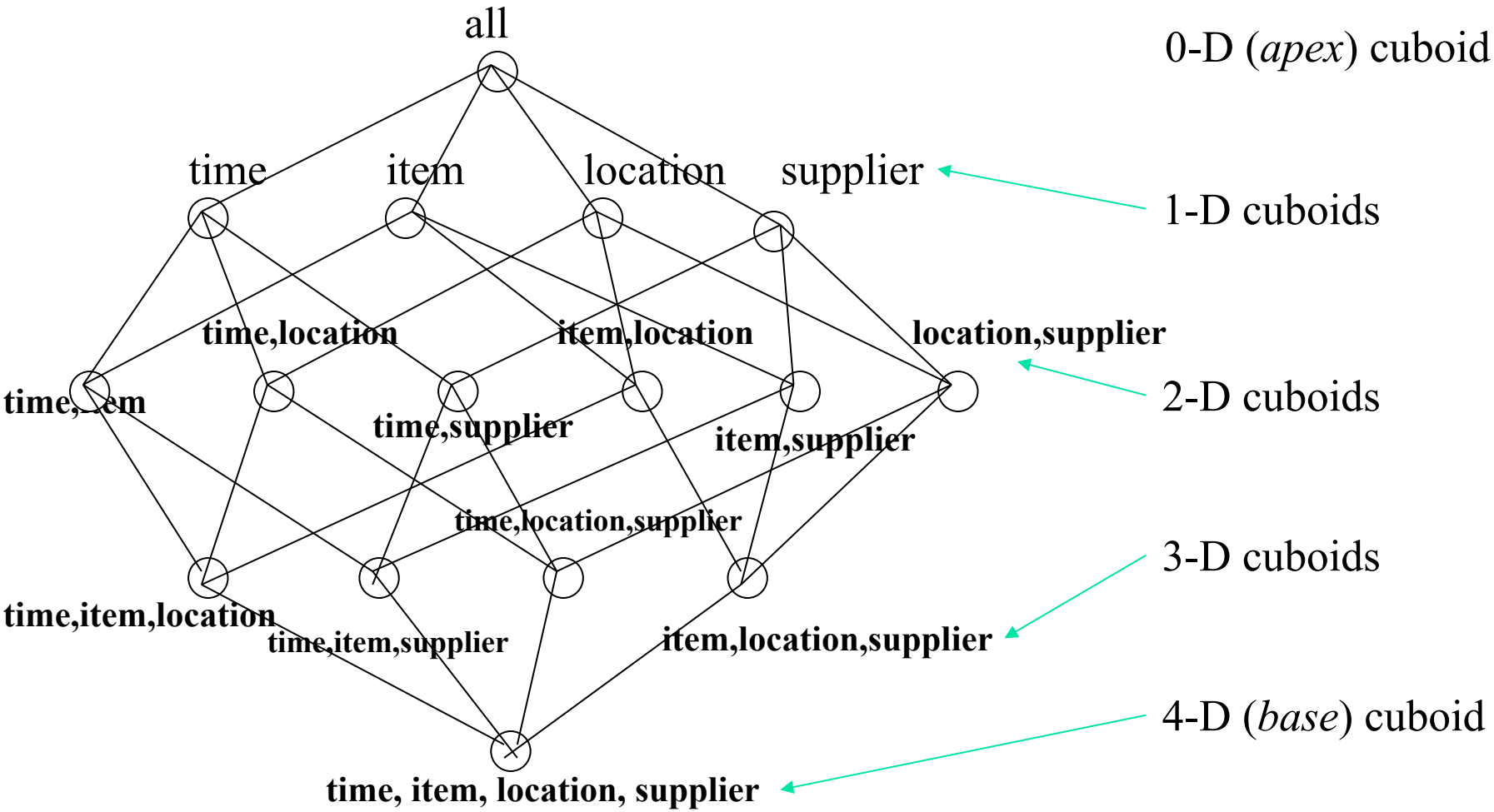
DW Modeling-

Data Cube and OLAP

From Tables and Spreadsheets to Data Cubes

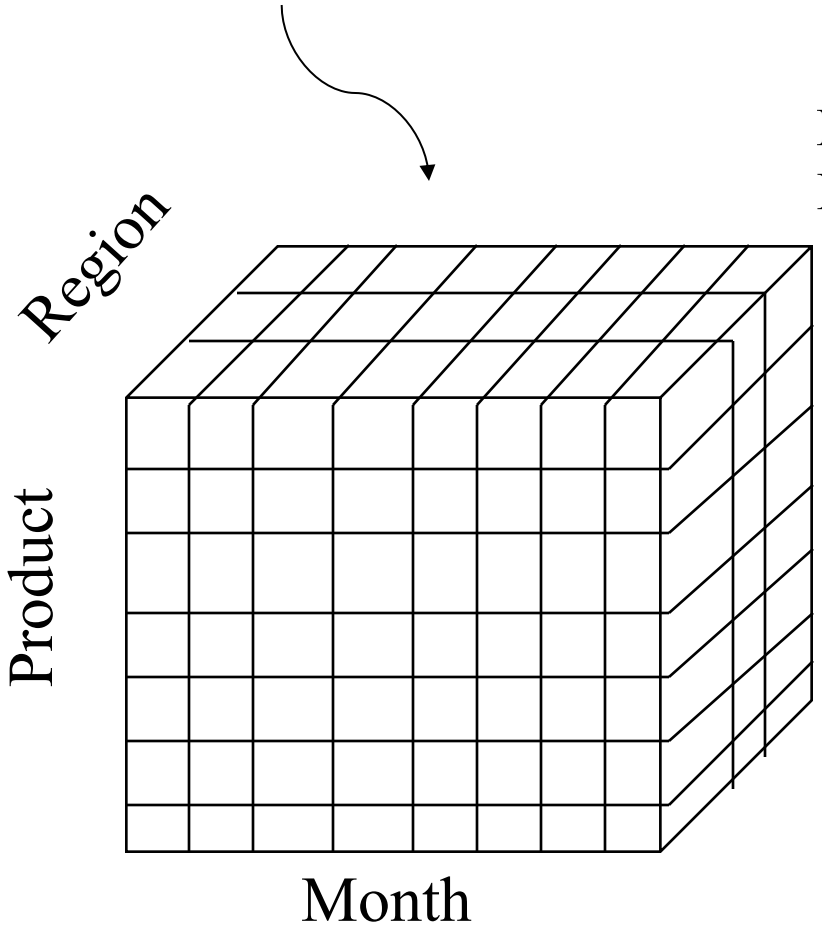
- A **data warehouse** is based on a multidimensional data model which views data in the form of a data cube
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as item (item_name, brand, type), or time(day, week, month, quarter, year)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

Cube: A Lattice of Cuboids

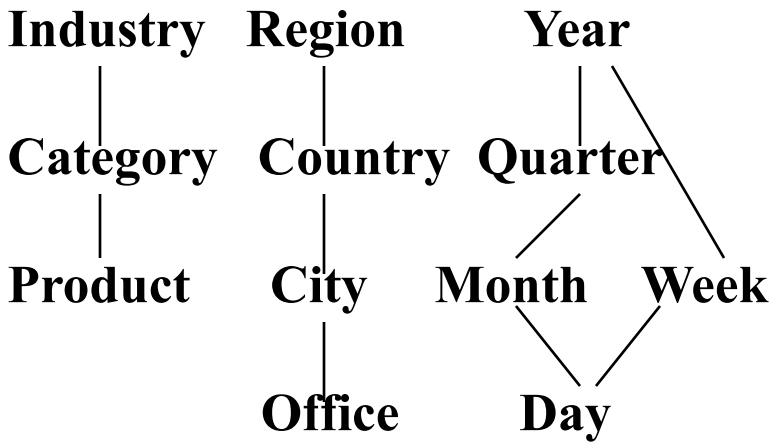


Multidimensional Data Modeling

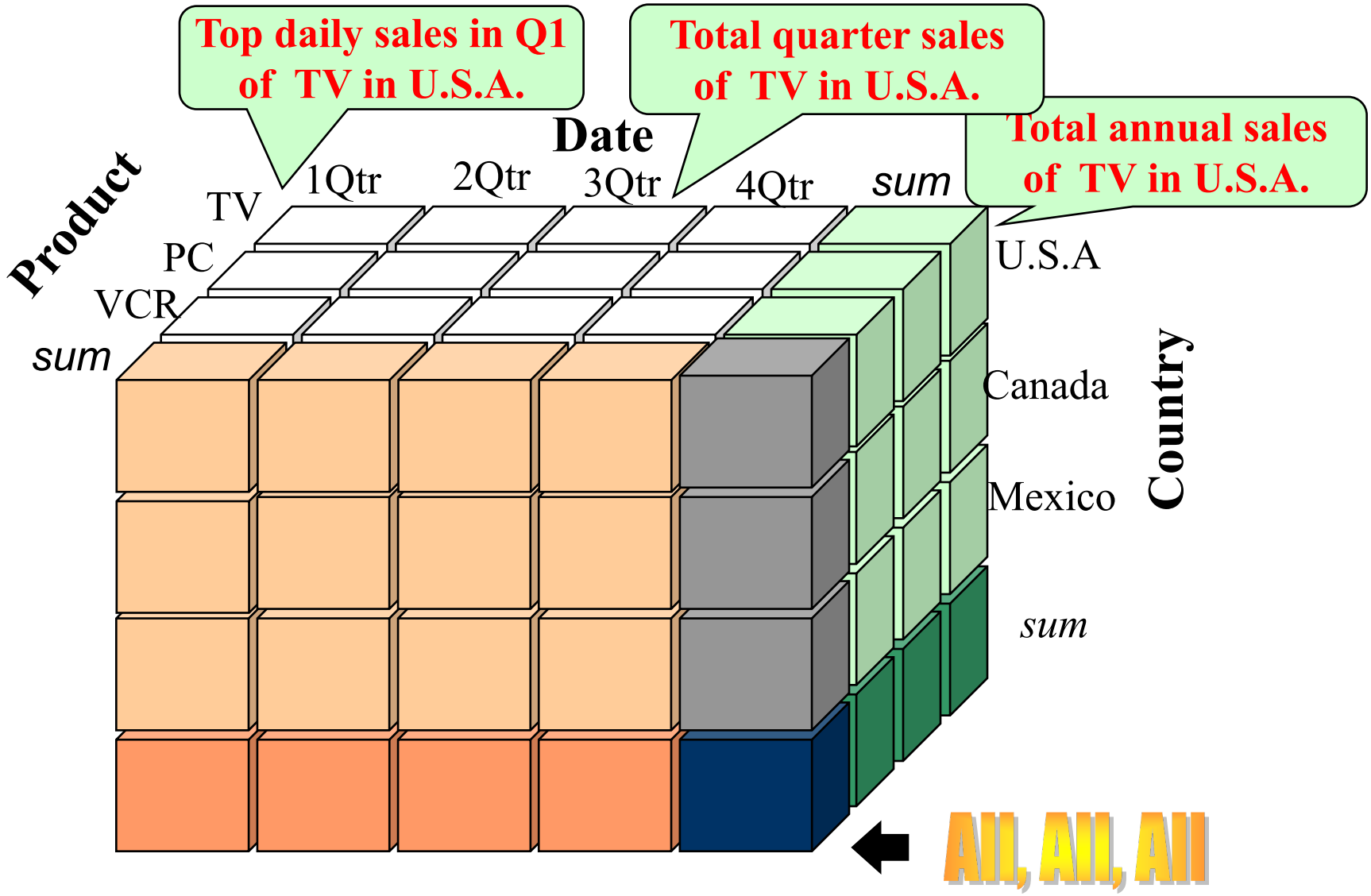
- Sales volume as a function of product, month, and region



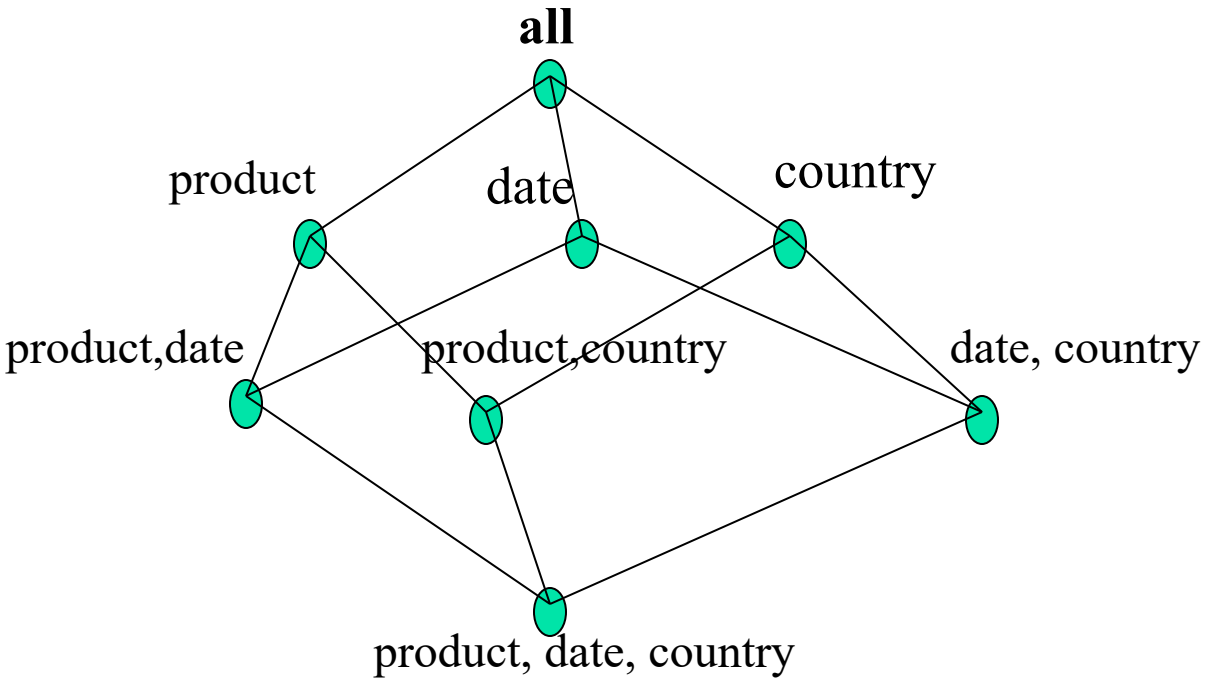
Dimensions: *Product, Location, Time*
Hierarchical summarization paths



A Hypothetical Data Cube



Cuboids Corresponding to the Cube



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D (*base*) cuboid

More data cube example:
3-D Data Cube

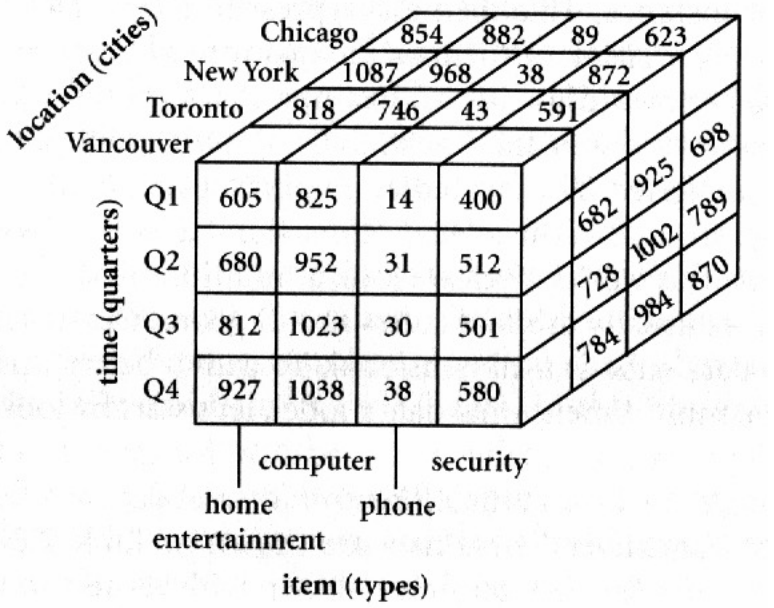
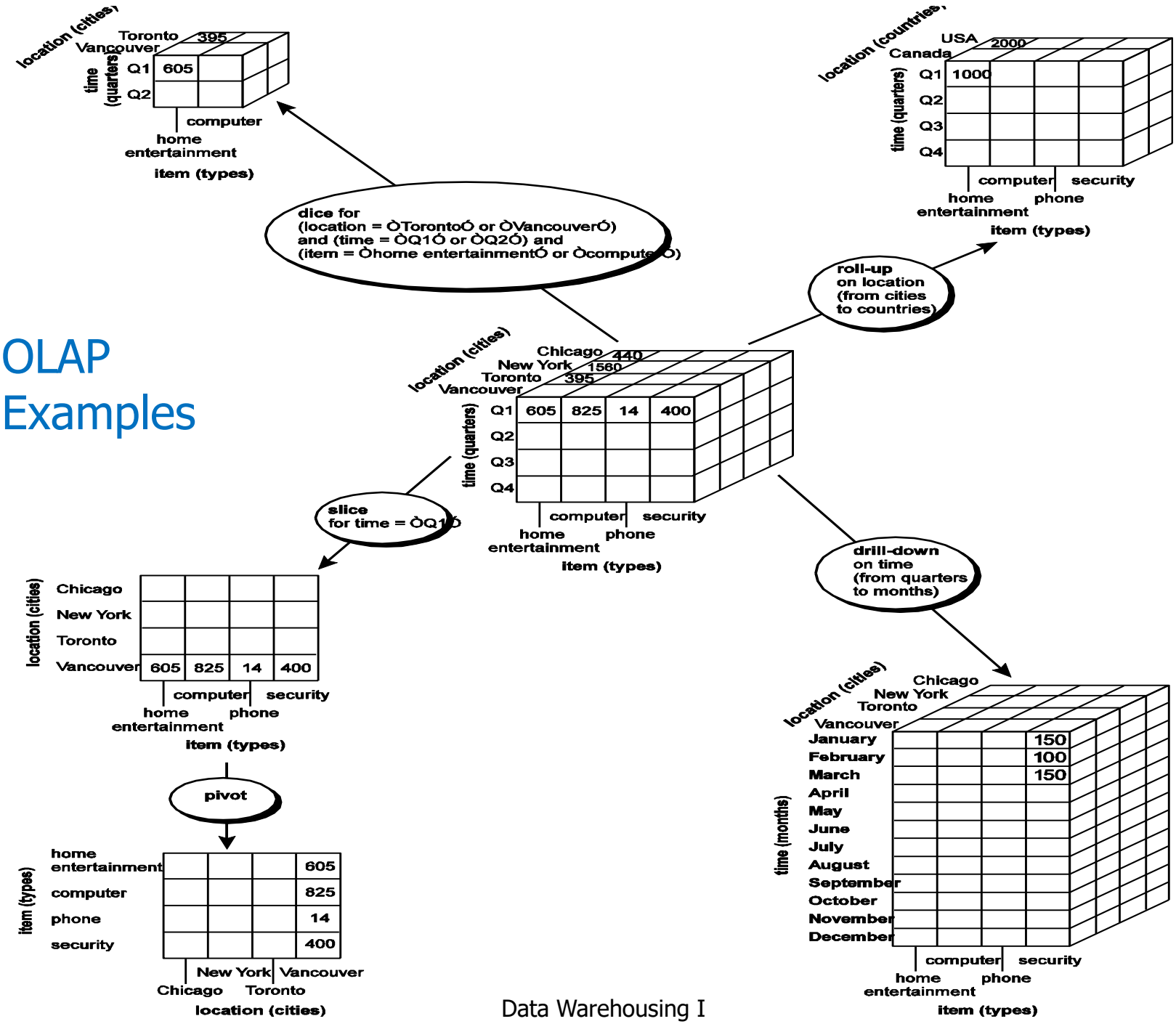


Table 2.3 A 3-D view of sales data for *AllElectronics*, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

	location = *"Chicago"*				***location*** = *"New York"*				***location*** = *"Toronto"*				***location*** = *"Vancouver"*			
	item				***item***				***item***				***item***			
time	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home 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

OLAP Examples



More data cube example: 4-D Data Cube

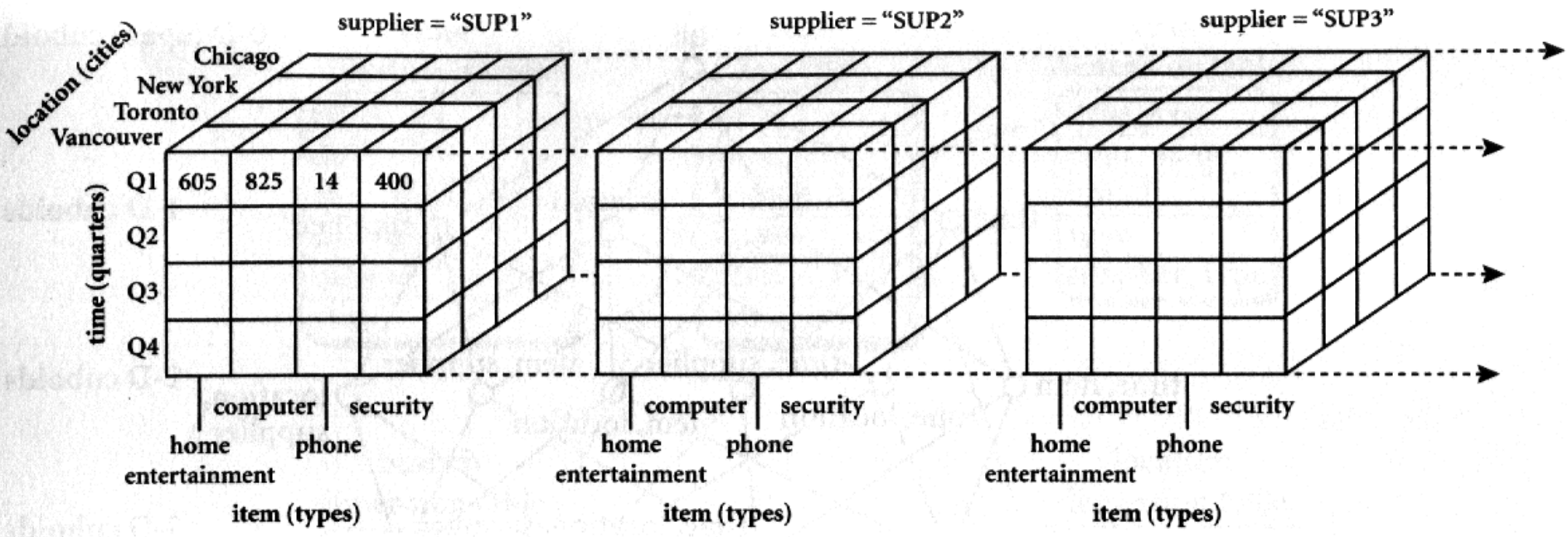


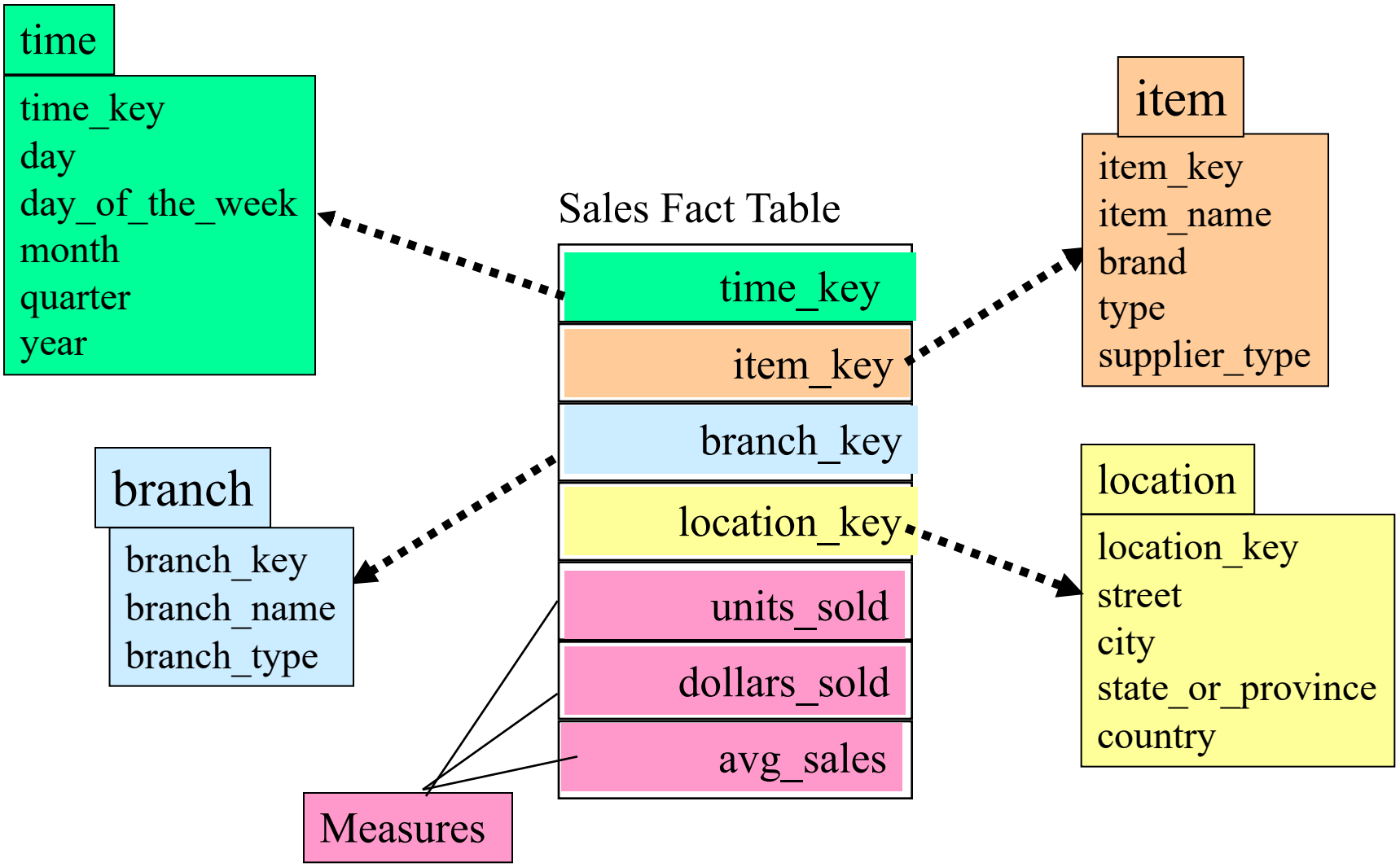
Figure 2.2 A 4-D data cube representation of sales data, according to the dimensions *time*, *item*, *location*, and *supplier*. The measure displayed is *dollars_sold* (in thousands). For improved readability, only some of the cube values are shown.

A link

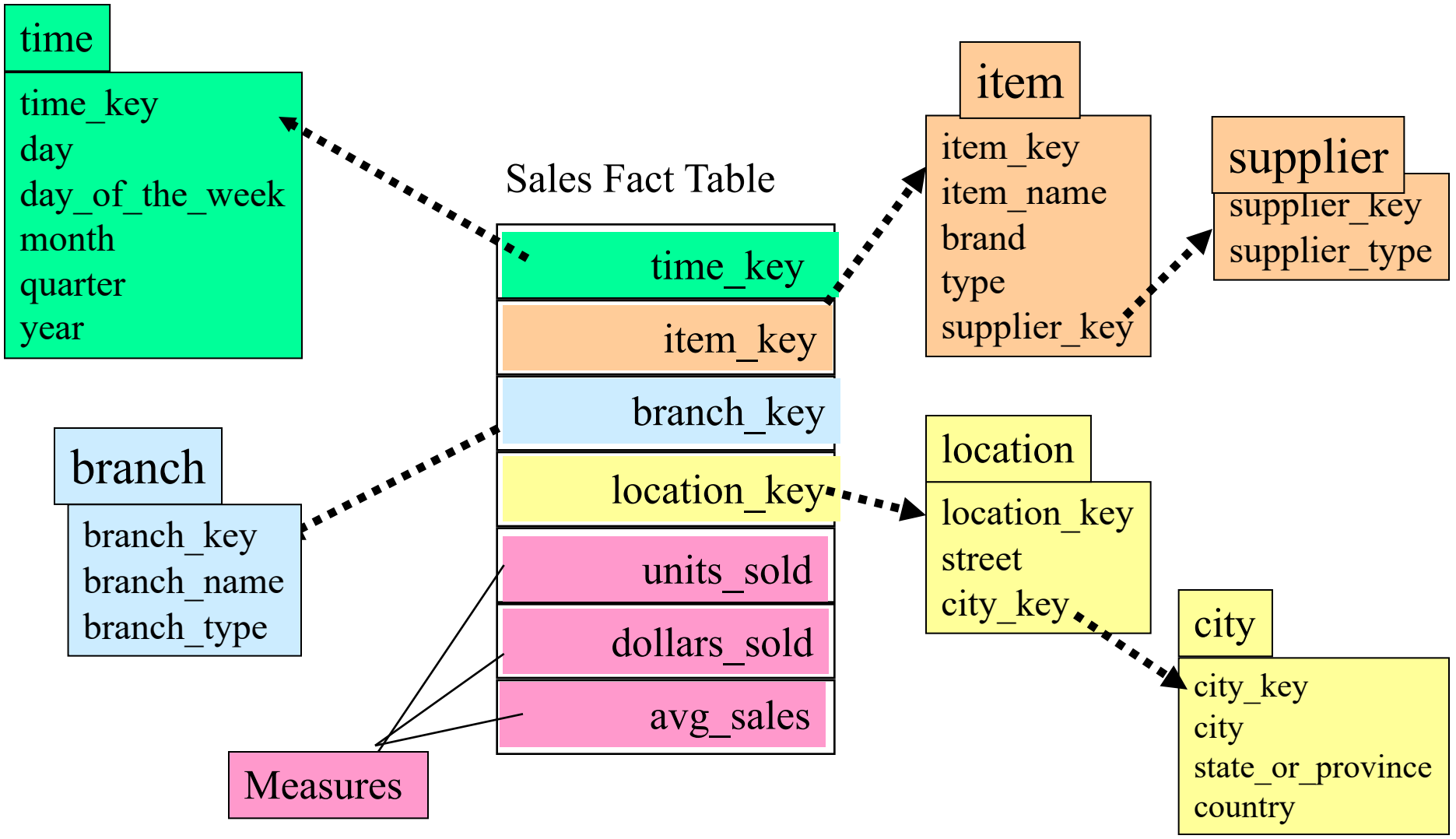
Conceptual Modeling of Data Warehouses

- Modeling data warehouses: **dimensions** & measures
- DW Schema
  - 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

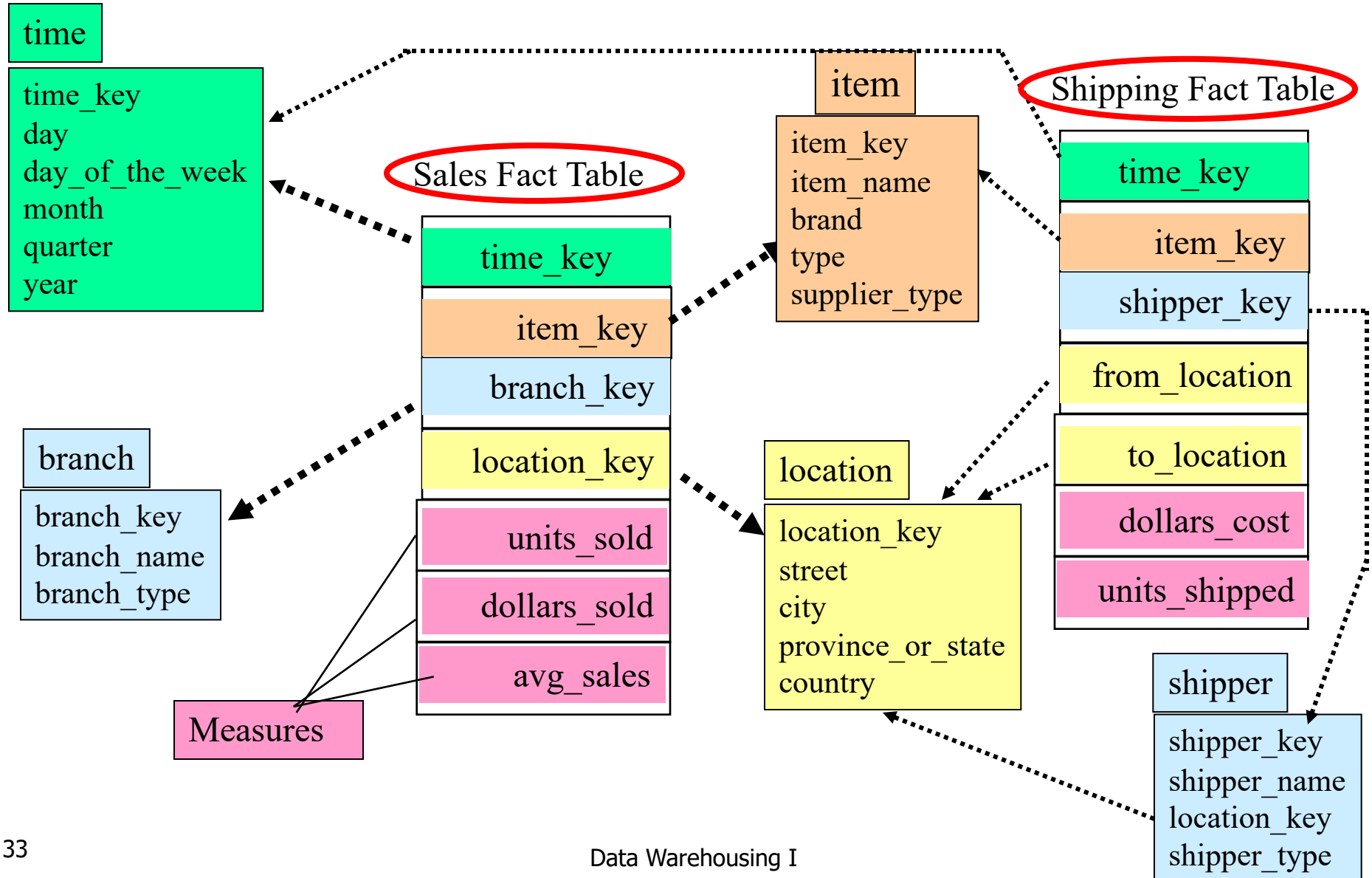
Example of Star Schema



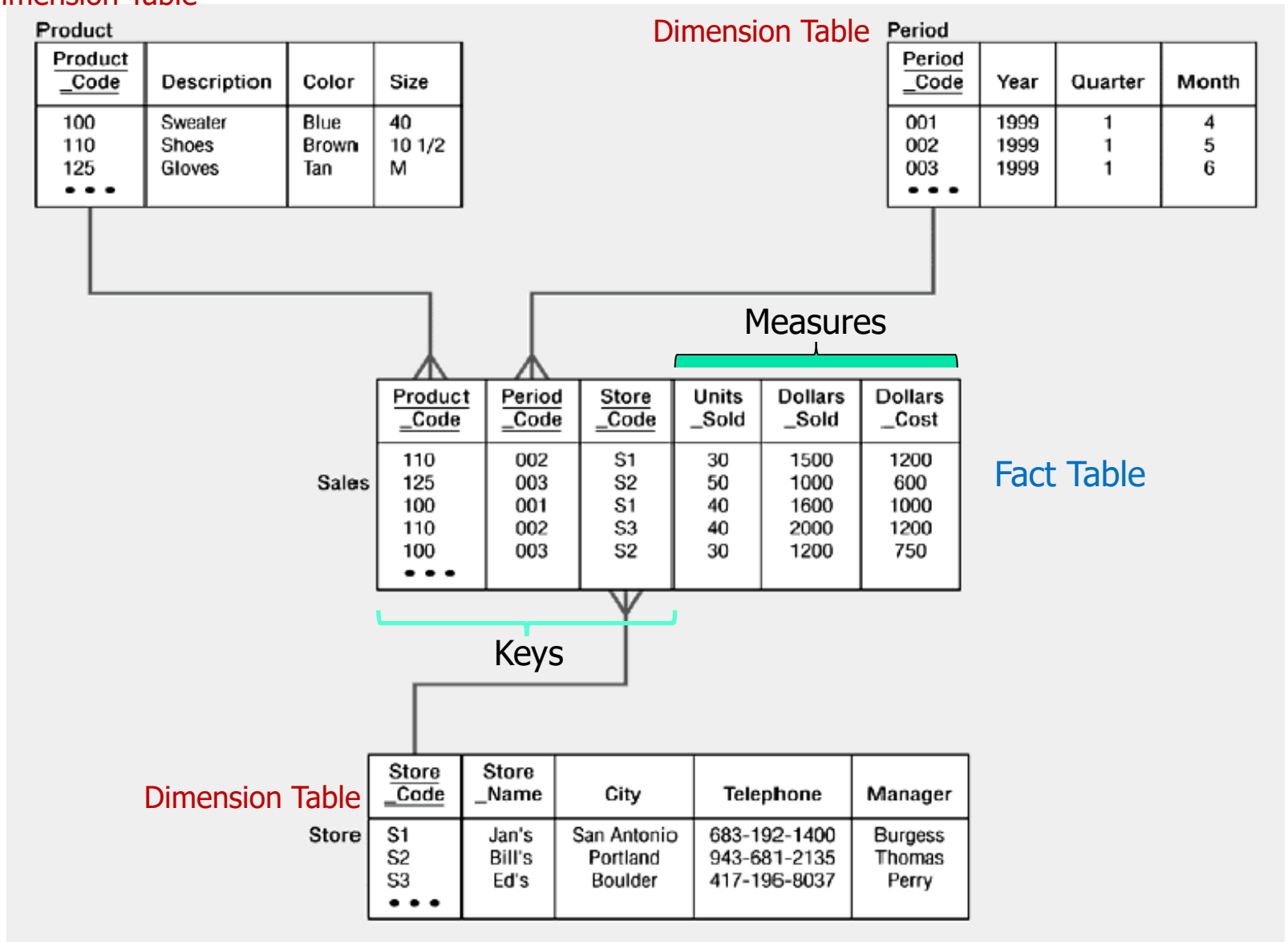
Example of Snowflake Schema



Example of Fact Constellation Schema



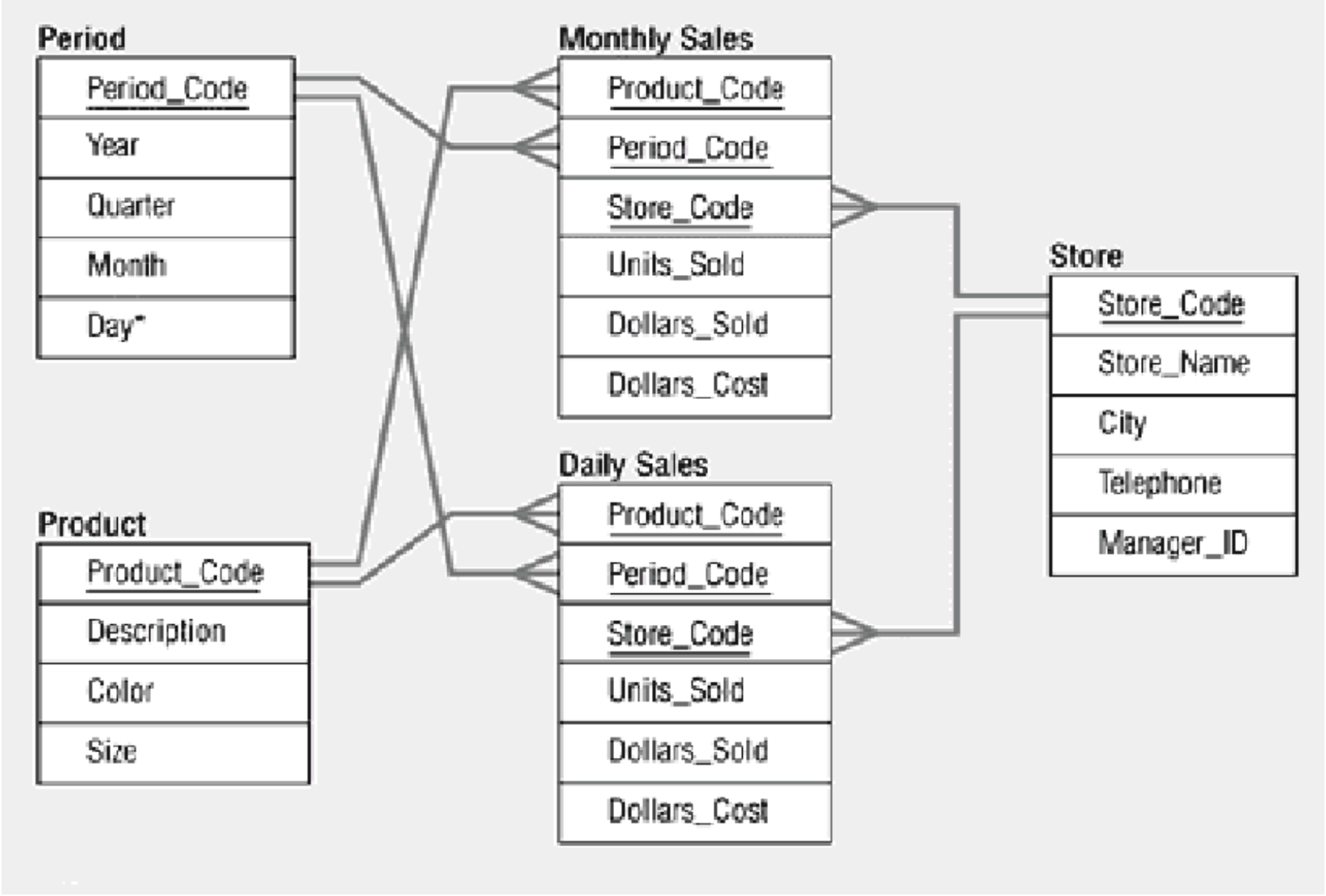
Example of Star Schema with Data



Multiple Fact Tables ⇒ Galaxy Schema

- For performance or other reasons, we can define multiple fact tables in a given star schema
  - e.g. various users require different levels of aggregation
- Performance can be improved by defining a different fact table for each level of aggregation (see the example in next slide)
- Designers of DW need decide whether increased storage requirements are justified by the prospective performance improvement

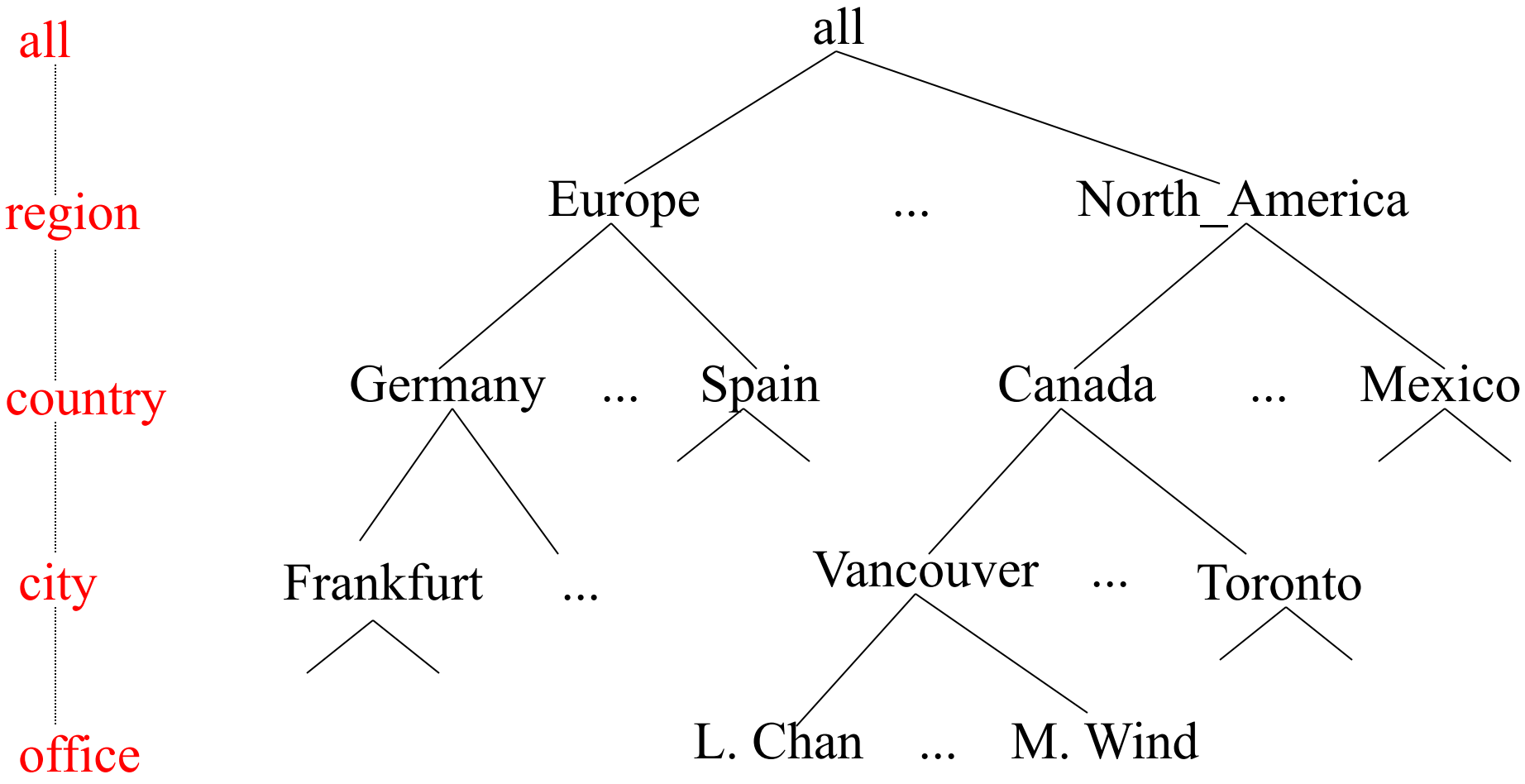
Star Schema with Two Fact Tables (Galaxy Schema)



Snowflake Schema

- Sometimes a dimension in a star schema forms a natural hierarchy
  - e.g. a dimension named Market has geographic hierarchy:
    - several markets within a state
    - several markets within a region
    - several markets within a country
- When a dimension participates in a hierarchy, the designer has two basic choices.
  - Include all the information for the hierarchy in a single table
    - i.e., a big flat table
  - normalize the tables
    - resulting in an expanded schema $\Rightarrow$ the *snowflake schema*!
- A snowflake schema is an expanded version of a star schema in which all of the tables are fully normalized.

A Concept Hierarchy: Dimension (location)



Data Cube Measures: Three Categories

- Distributive: if the result derived by applying the function to *n* aggregate values is the same as that derived by applying the function on all the data without partitioning
  - E.g., count(), sum(), min(), max()

(total) units_sold

- Algebraic: if it can be computed by an algebraic function with *M* arguments (where *M* is a bounded integer), each of which is obtained by applying a distributive aggregate function
  - E.g., avg(), min_N(), standard_deviation()

(average) units_sold

- Holistic: if there is no constant bound on the storage size needed to describe a sub-aggregate.
  - E.g., median(), mode(), rank()

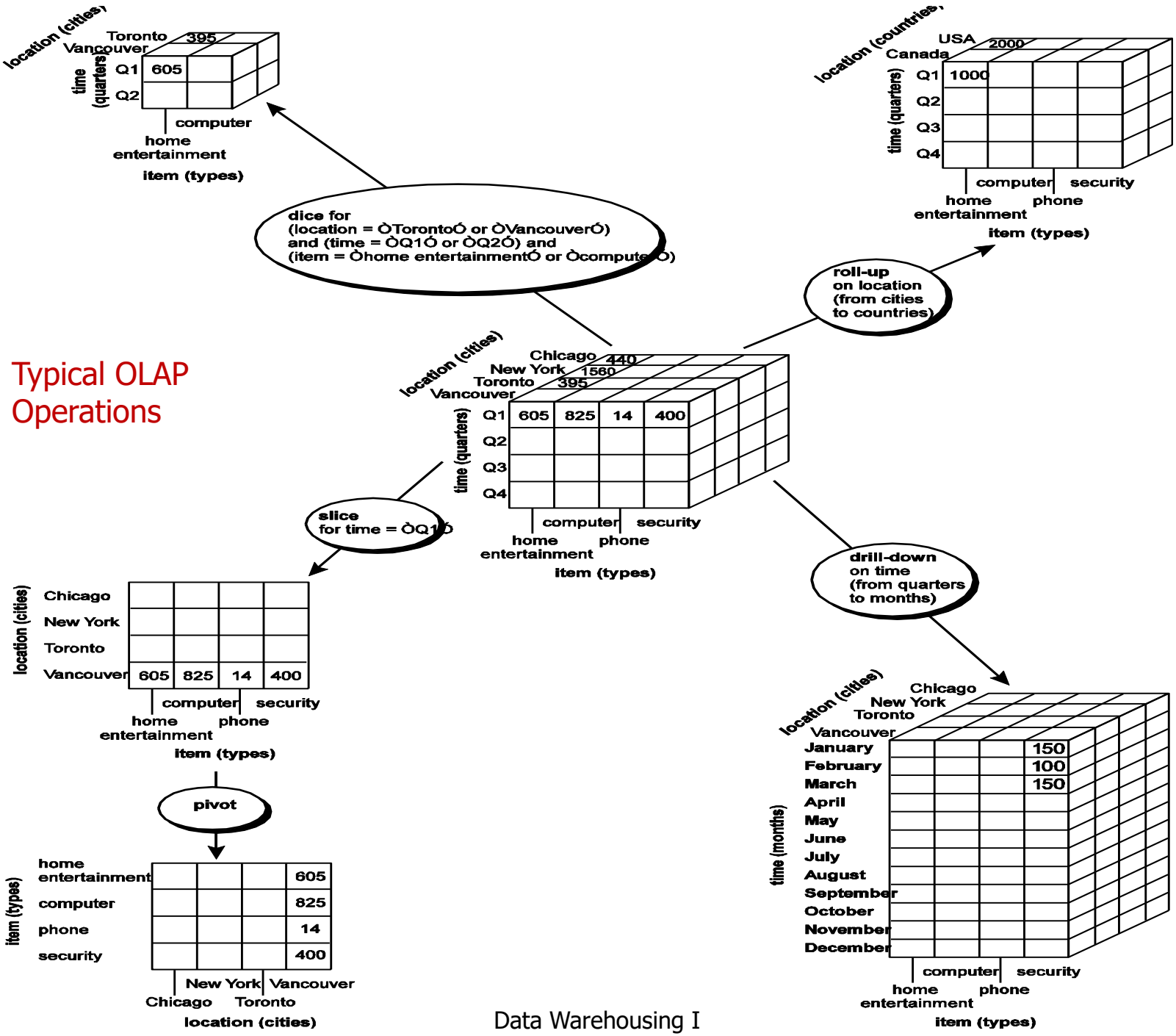
(median) units_sold

Online Analytical Processing (OLAP)

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)*

Typical OLAP Operations



Roll-up Operation

- Roll-up operation corresponds to taking the current aggregation level of fact values and doing a further aggregation on one (or more) of the dimensions
- That is equivalent to doing GROUP BY to this dimension(s) by using attribute hierarchy
- Roll-up operation can be understood as lowering the number of dimensions
- In this case, the measure is calculated without regard to dimensions to be omitted.

Drill Down Operation

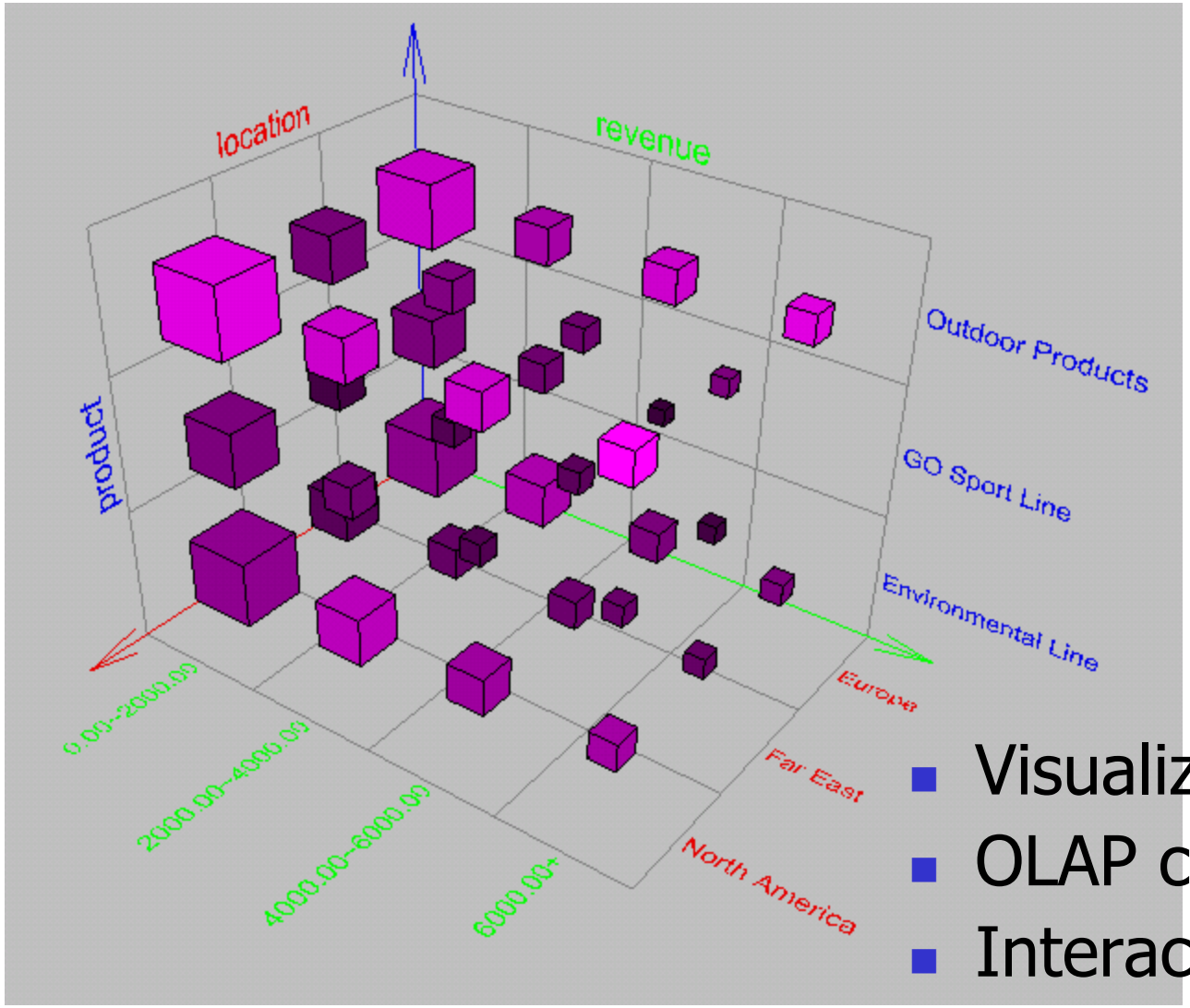
- Analyzing a set of data at a finer level of detail
  - e.g. a summary report for the total sales of three package sizes for a given brand of paper towels
  - Further breakdown of sales by color within each of these package sizes
- A drill-down presentation is equivalent to adding another column to the original report (a color column)
- Executing a drill-down may require that the OLAP tool "reach back" to the DW to obtain the detailed data necessary for the drill-down
- Some tools even permit the OLAP tool to reach back to the operational data if necessary for a given query

Slicing and Dicing a Cube

- Slicing the data cube to produce a simple 2-D table or view
  - e.g. A slice is for the product named shoes
  - other views developed by simple "drag and drop"
  - This type of operation is often called "slicing and dicing" the cube
- Slice-and-dice operations reduce the number of dimensions by taking a projection of facts on a subset of dimensions and for some selected values of dimensions that are being dropped.
- Closely related to slicing and dicing is data pivoting
  - This term refers to rotating the view for a particular data point, to obtain another perspective
  - The analyst could pivot this view to obtain the sale of shoes by store for the same month

Browsing a Data Cube

Back



- Visualization
- OLAP capabilities
- Interactive manipulation

Part I's Summary

- Data warehouse can be considered as a central "store" of all of our entities, concepts, metadata, and historical information
  - For doing data validation, complex mining, analysis, prediction, etc.
- Multi-dimensional modellinh 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
  - Basically, a kind of navigation through the data

Acknowledgement

- Slides/Materials of
  - J. Han et al.'s DM: Concepts and Techniques textbook
  - https://web2.utc.edu/~djy471/
- Photos from Internet

References

- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- A. Gupta and I. S. Mumick. Materialized Views: Techniques, Implementations, and Applications. MIT Press, 1999.