# What is Data Warehouse? (1)

- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site
- A data warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise need to make strategic decisions

# What is Data Warehouse? (2)

TRS

- Data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitate effective data generalization and data mining
- Many other data mining functions, such as association, classification, prediction, and clustering, can be integrated with OLAP operations to enhance interactive mining of knowledge at multiple levels of abstraction

# What is Data Warehouse? (3)

- A decision support database that is maintained separately from the organization's operational database
- *"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process* [Inm96]."—W. H. Inmon

# Data Warehouse Framework

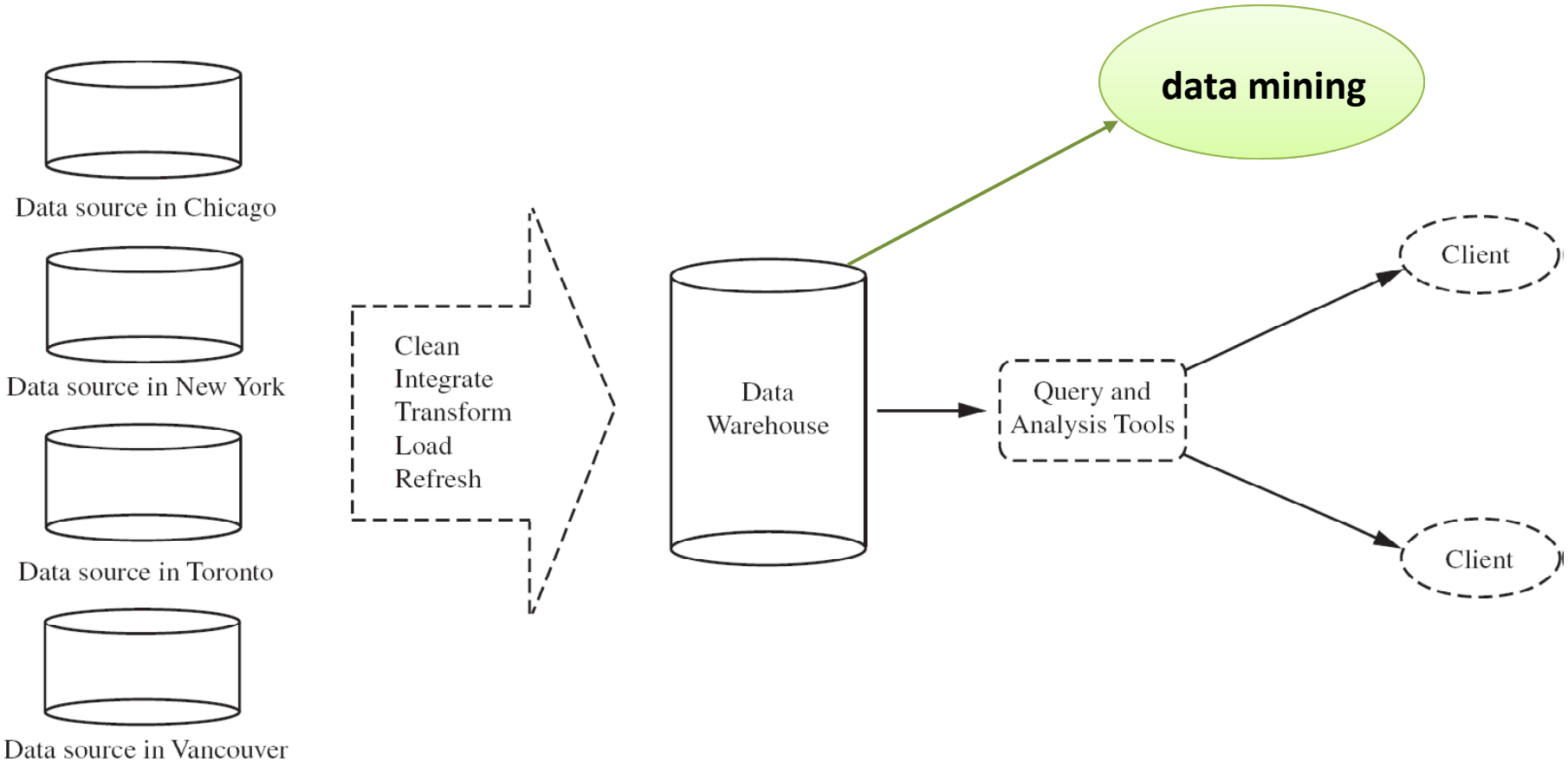


**Figure 1.7** Typical framework of a data warehouse for *AllElectronics*

# Data Warehouse is *Subject-Oriented*

- Organized around major subjects, such as customer, product, sales, etc.
- 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 is *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.

# Data Warehouse is *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

# Data Warehouse is *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*

# OLTP vs. OLAP

| 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 | historical; 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 | 100 MB to GB | 100 GB to TB |
| Priority | high performance, high availability | high flexibility, end-user autonomy |
| Metric | transaction throughput | query throughput, response time |

**Table 3.1** Comparison between OLTP and OLAP

# Why Separate is Data Warehouse Needed? (1)

- *Why not perform on-line analytical processing directly on operational databases instead of spending additional time and resources to construct a separate data warehouse?*

# Why Separate is Data Warehouse Needed? (2)

- High performance for both systems
  - DBMS— tuned for OLTP: searching for particular records, indexing, hashing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation (summarization and aggregation)

# Topics

- Definition of data warehouse
- Multidimensional data model
- Data warehouse architecture
- From data warehousing to data mining

# From Tables and Spreadsheets to Data Cubes

- A data warehouse is based on a multidimensional data model
- This model views data in the form of a data cube
- A data cube allows data to be modeled and viewed in multiple dimensions

# From Tables and Spreadsheets to Data Cubes (1)

- A data cube is defined by facts and dimensions
  - Facts are data which data warehouse focus on
    - Fact tables contain numeric measures (such as dollars_sold) and keys to each of the related dimension tables
  - Dimensions are perspectives with respect to fact
    - Dimension tables describe the dimension with attributes. For example, item (item_name, brand, type), or time(day, week, month, quarter, year)

**Figure 1.6.** Fragments of relations from a relational database for *AllElectronics*

*customer*

| *cust_ID* | *name* | *address* | *age* | *income* | *credit_info* | *category* | ... |
|---|---|---|---|---|---|---|---|
| C1 | Smith, Sandy | 1223 Lake Ave., Chicago, IL | 31 | $78000 | 1 | 3 | ... |
| ... | ... | ... | ... | ... | ... | ... | ... |

*item*

| *item_ID* | *name* | *brand* | *category* | *type* | *price* | *place_made* | *supplier* | *cost* |
|---|---|---|---|---|---|---|---|---|
| I3 | hi-res-TV | Toshiba | high resolution | TV | $988.00 | Japan | NikoX | $600.00 |
| I8 | Laptop | Dell | laptop | computer | $1369.00 | USA | Dell | $983.00 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

*employee*

| *empl_ID* | *name* | *category* | *group* | *salary* | *commission* |
|---|---|---|---|---|---|
| E55 | Jones, Jane | home entertainment | manager | $118,000 | 2% |
| ... | ... | ... | ... | ... | ... |

*branch*

| *branch_ID* | *name* | *address* |
|---|---|---|
| B1 | City Square | 396 Michigan Ave., Chicago, IL |
| ... | ... | ... |

*purchases*

| *trans_ID* | *cust_ID* | *empl_ID* | *date* | *time* | *method_paid* | *amount* |
|---|---|---|---|---|---|---|
| T100 | C1 | E55 | 03/21/2005 | 15:45 | Visa | $1357.00 |
| ... | ... | ... | ... | ... | ... | ... |

*items_sold*

| *trans_ID* | *item_ID* | *qty* |
|---|---|---|
| T100 | I3 | 1 |
| T100 | I8 | 2 |
| ... | ... | ... |

*works_at*

| *empl_ID* | *branch_ID* |
|---|---|
| E55 | B1 |
| ... | ... |

# From Tables and Spreadsheets to Data Cubes (2)

| *time* (quarter) | *location* = "Vancouver" | | | |
|---|---|---|---|---|
| | *item* (type) | | | |
| | 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 |

dimensions

Facts (numerical measures)

**Table 3.2** A 2-D view of sales data for *AllElectronics* according to the dimensions *time* and *item*, where the sales are from branches located in the city of Vancouver. The measure displayed is *dollar_sold* (in thousands).

# From Tables and Spreadsheets to Data Cubes (3)

| | ***location*** = "Chicago" | | | | ***location*** = "New York" | | | | ***location*** = "Toronto" | | | | ***location*** = "Vancouver" | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***time*** | ***item*** | | | | ***item*** | | | | ***item*** | | | | ***item*** | | | |
| | 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 |

**Table 3.3** A 3-D view of sales data for *AllElectronics* according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollar_sold* (in thousands).

# From Tables and Spreadsheets to Data Cubes (4)

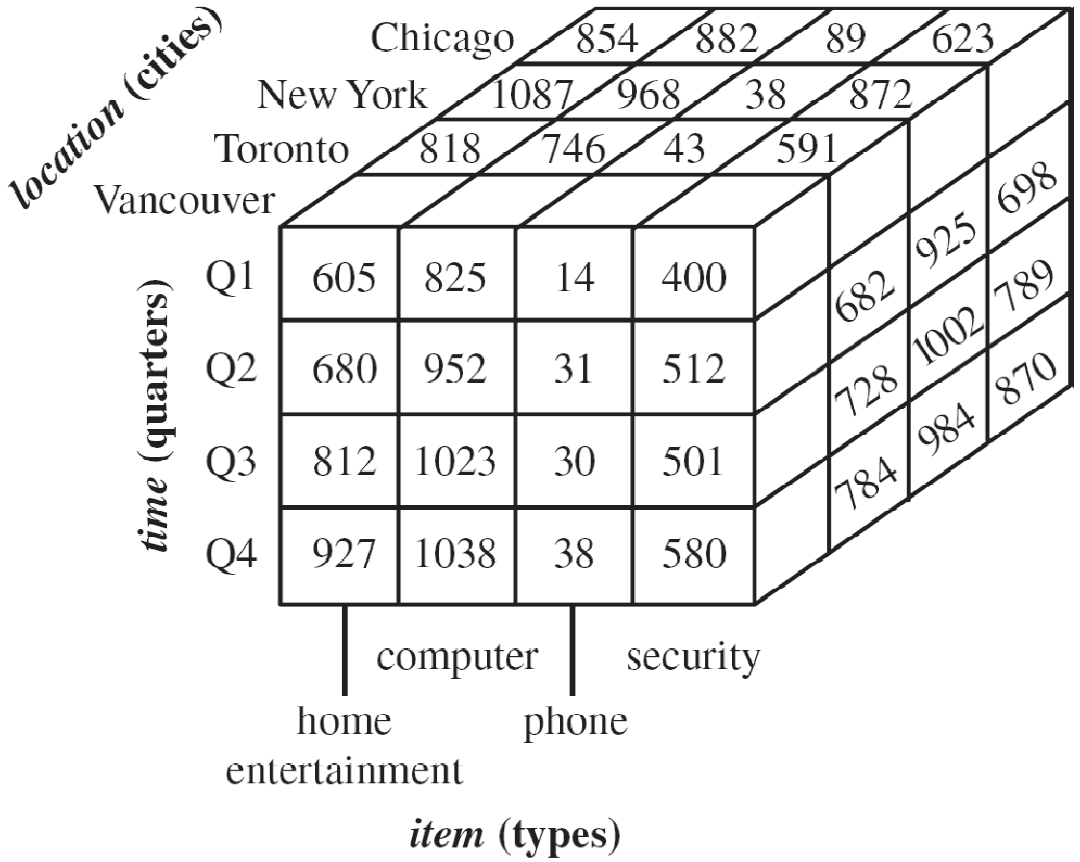


**Figure 3.1** A 3-D data cube representation of the data in Table 3.3, according to the dimensions *time, item*, and *location*. The measure displayed is *dollar_sold* (in thousands).

# From Tables and Spreadsheets to Data Cubes (5)

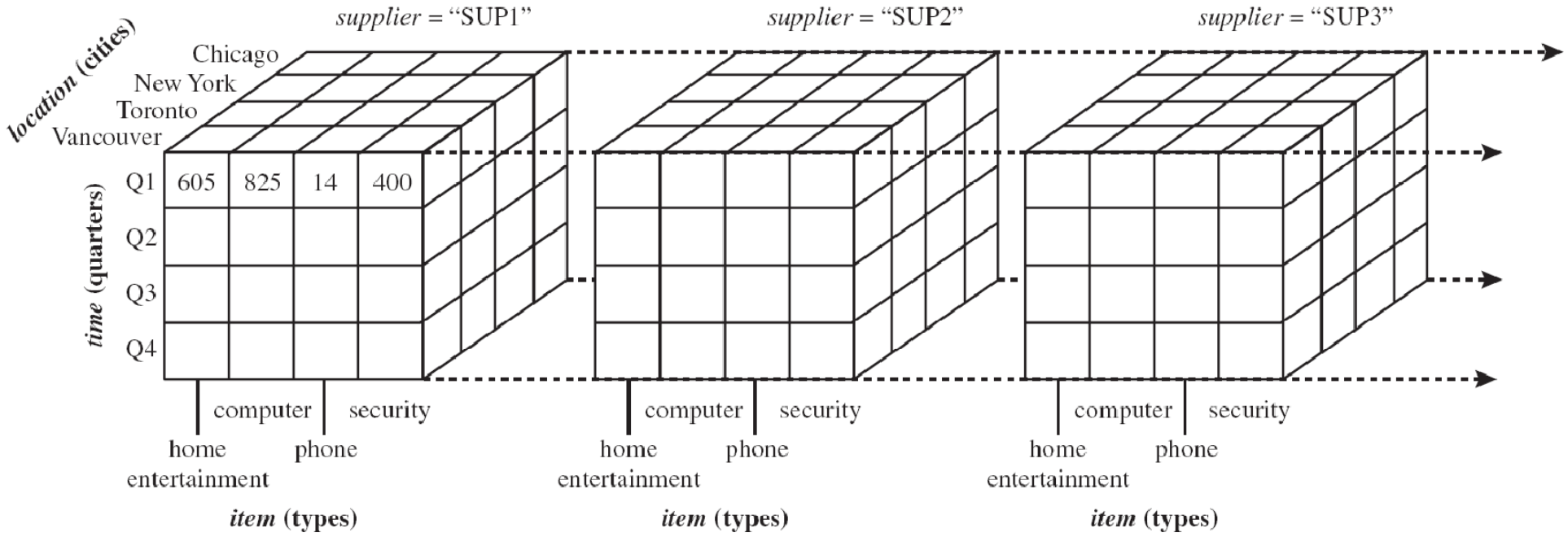


**Figure 3.2** A 4-D data cube representation, according to the dimensions *time, item*, *location,* and *supplier*. The measure displayed is *dollar_sold* (in thousands).

# Cuboid

- A data cube is a lattice of cuboids
- The total number of cuboids
- The apex cuboid
- The base cuboid

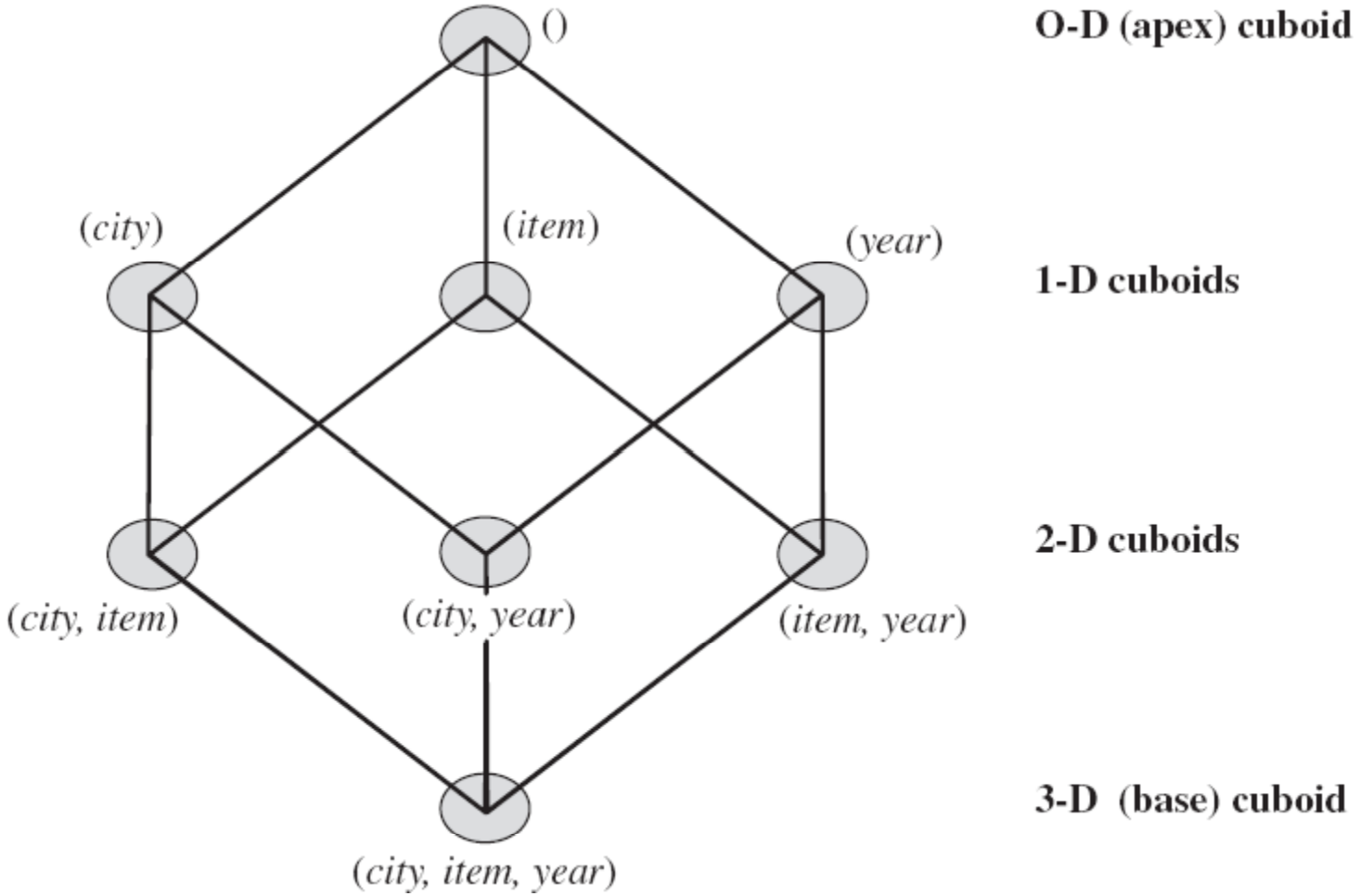


**Figure 3.14** Lattice of cuboids, making up a 3-D data cube. Each cuboid represents a different group-by. The base cuboid contains the three dimensions city, item, and year.

# The Curse of Dimensionality

- How many cuboids are there in a n-dimensional data cube?
- How many cuboids are there in a n-dimensional data cube and each dimension (i) has the number of level, ($L_i$)?

# Conceptual Modeling of Data Warehouses

- 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

# Star Schema

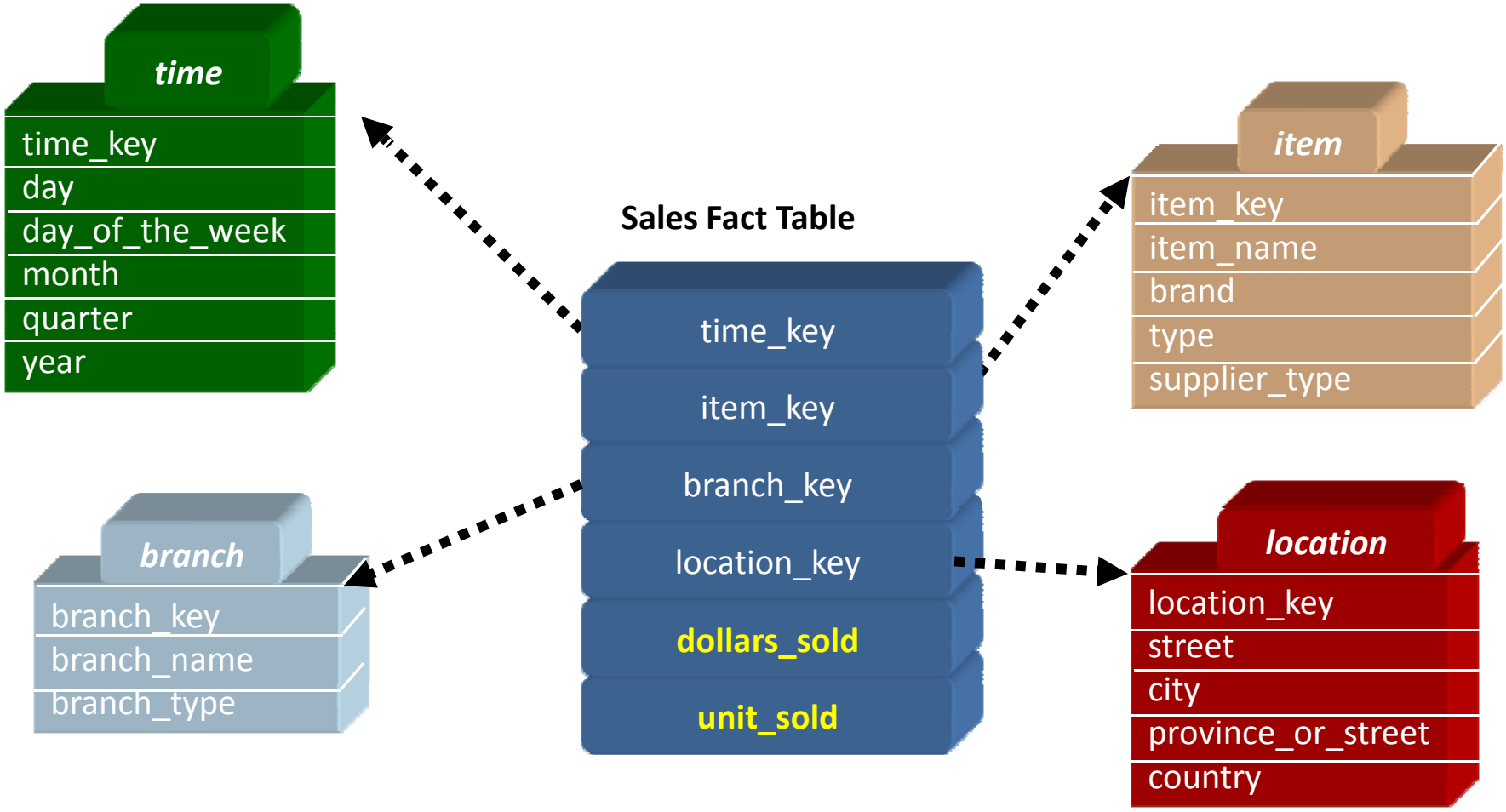


**Figure 3.4** Star schema of a data warehouse for sales.

# Snowflake Schema

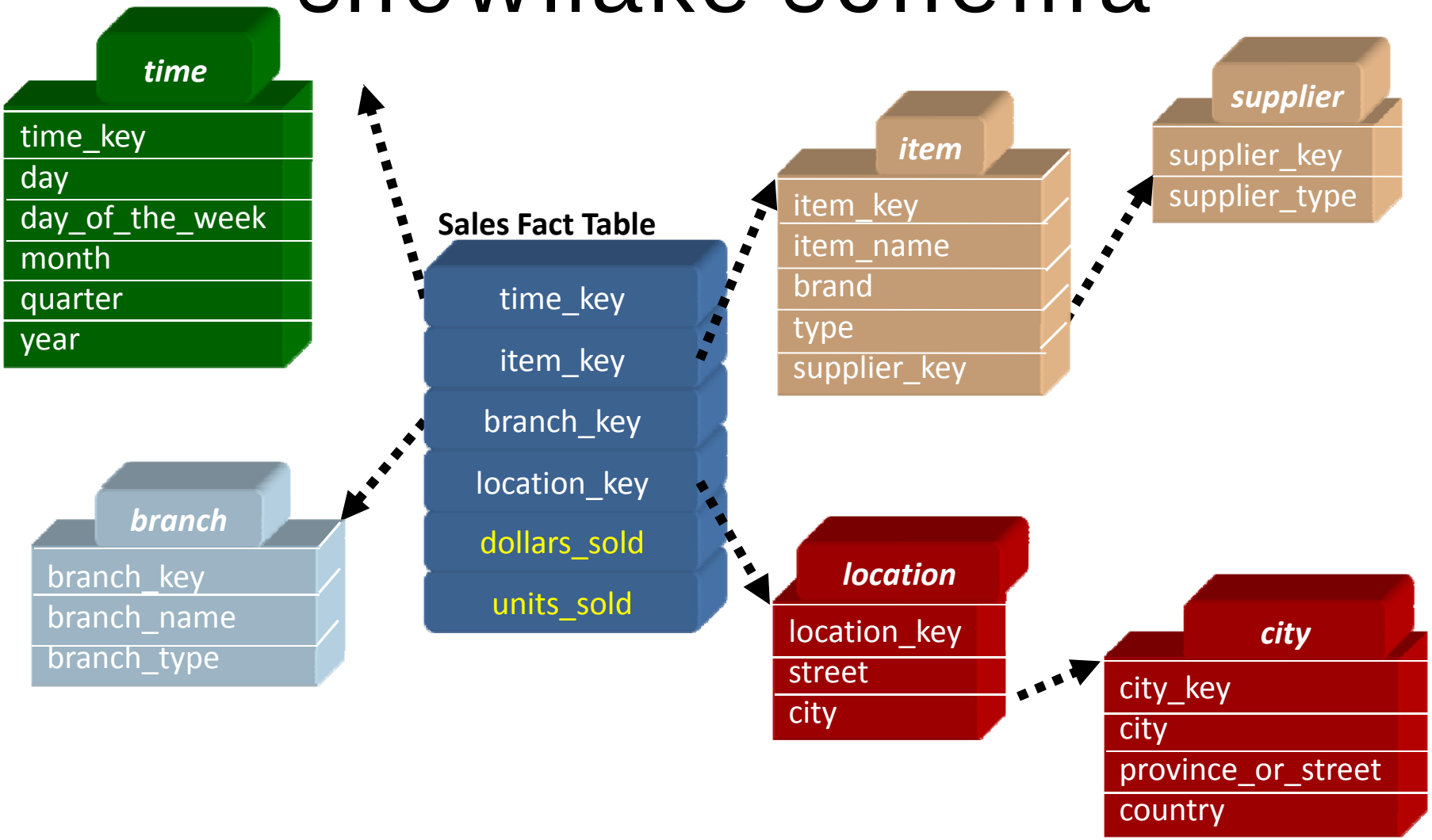


**Figure 3.4** Snowflake schema of a data warehouse for sales.

# Fact Constellation

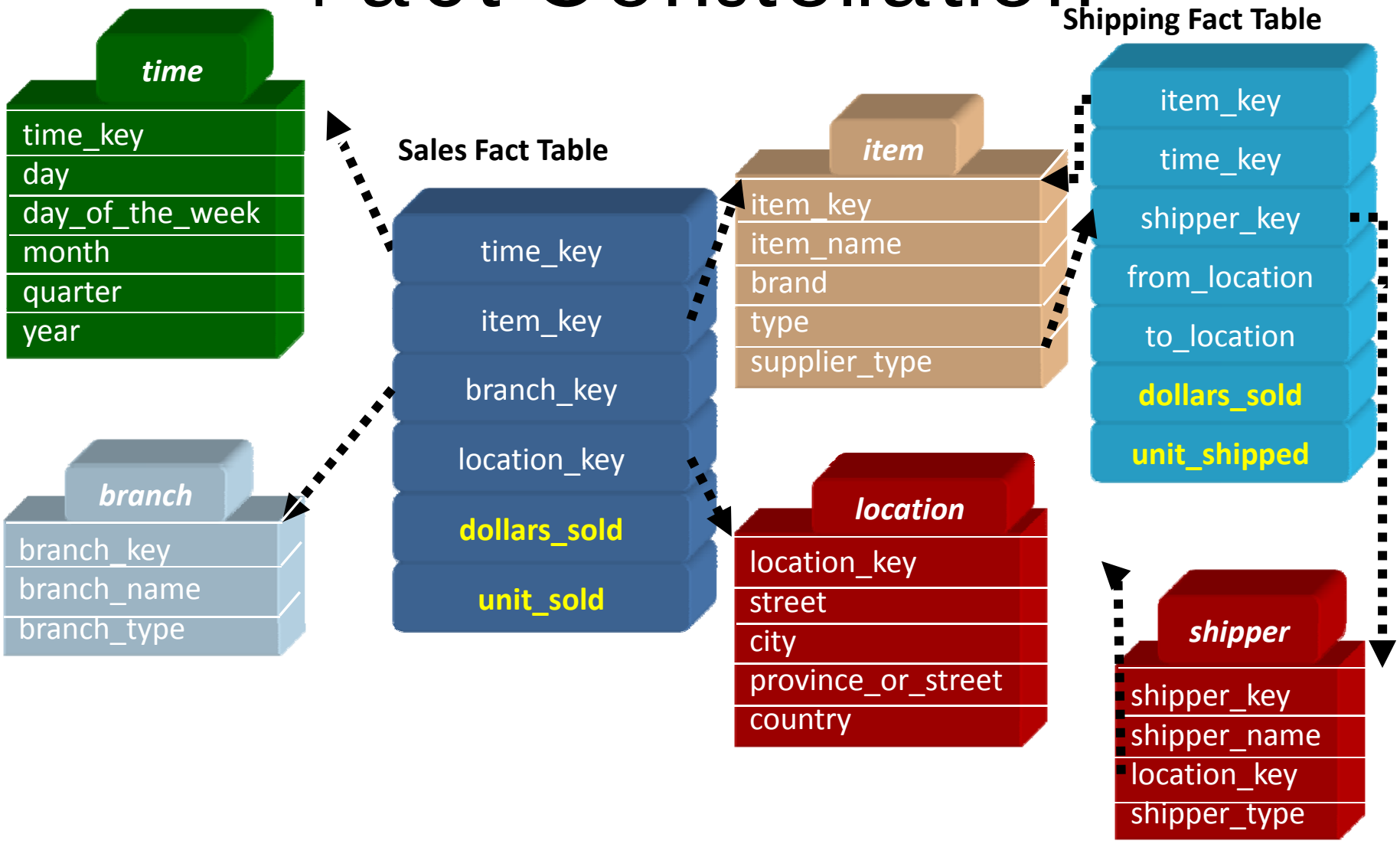


**Figure 3.5** Fact constellation schema of a data warehouse for sales and shipping.