DATA MINING AND DATA WAREHOUSING

(BCSDM515)

By,
Anusha K S

Assistant Professor, Dept. of CSE
VVCE, Mysuru.

MODULE 1

Data Warehousing: Basic Concepts, What is a Data Warehouse? Difference between Operational database Systems and Data Warehouses, Why have a Separate data Warehouse? Data Warehousing: A Multitiered Architecture, Extraction, Transformation and loading, metadata Repository.

Data Warehouse Modeling: Data Cube and OLAP, Data Cube: A multidimensional data model, Stars, Snowflakes and Fact constellations: Schemas for multidimensional Data models, Dimensions: The role of concept Hierarchies, Measures: Their Categorization and computation, Typical OLAP Operations.

SLT: Data Warehouse Models: Enterprise Warehouse, Data Mart and Virtual Warehouse

1. What is data…?

2. What is information…?

3. What is Database…?

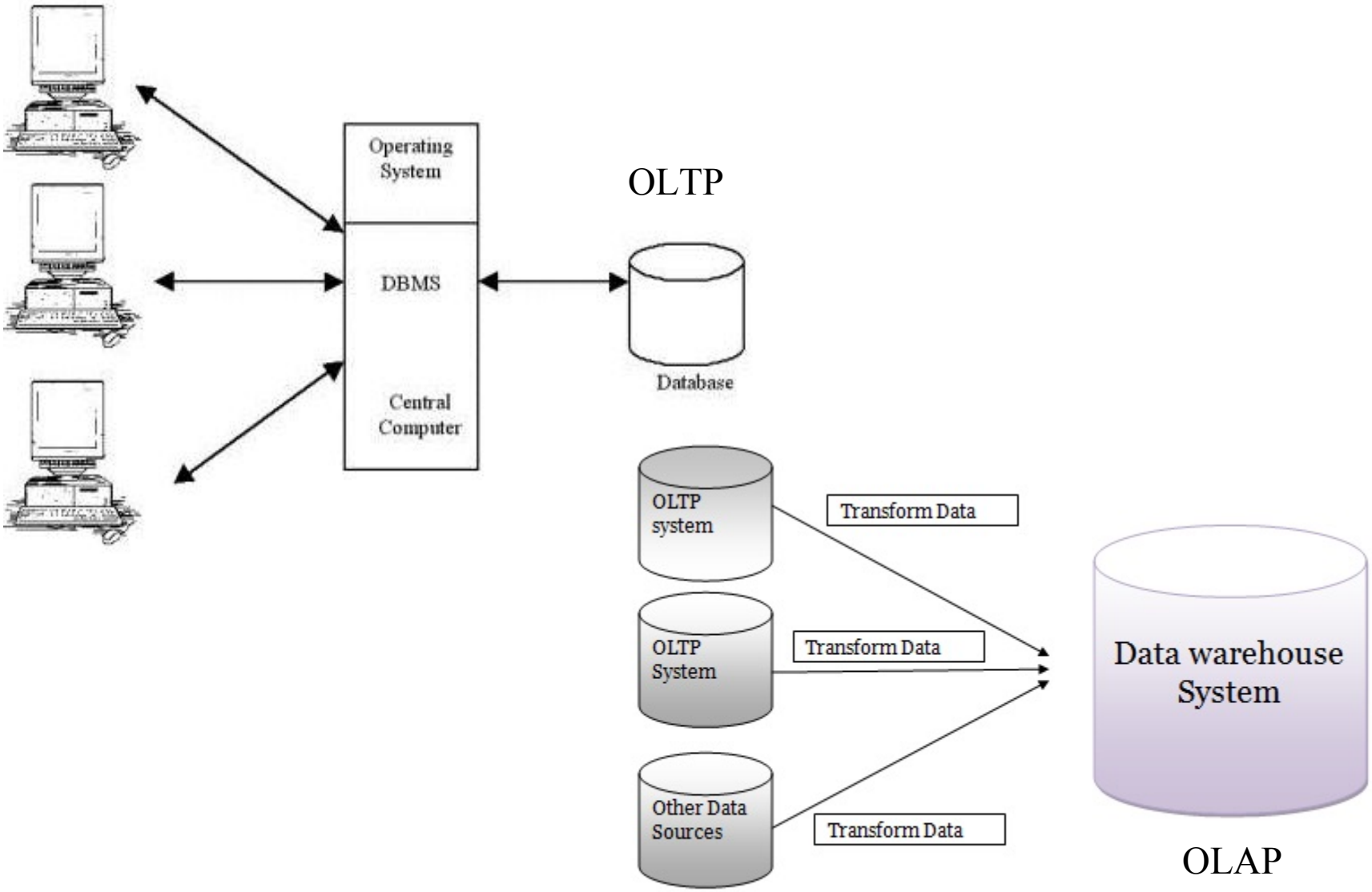


Figure: Data warehouse system

DATA WAREHOUSE

- A data warehouse refers to a data repository that is maintained separately from an organization operational database

- Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions

Key Features of Data Warehouse (According to William H. Inmon)

- Subject-oriented
- Integrated
- Time-variant
- Non-volatile

Differences between Operational Database Systems and Data Warehouses

The major distinguishing features of OLTP and OLAP are summarized as follows

- Users and system orientation

OLTP- customer-oriented OLAP – market-oriented

- Data contents

OLTP- current data OLAP – historic data

- Database design

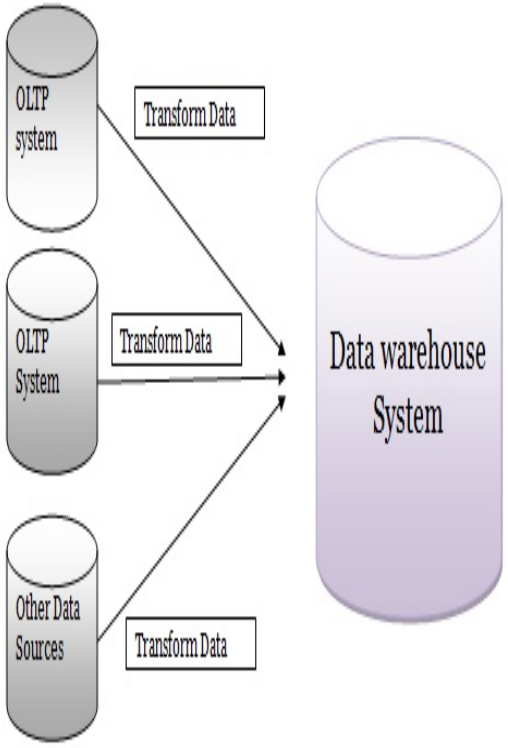
OLTP- ER model OLAP – star/Snowflake models

- View

OLTP- focus on current data OLAP – info originated form diff. sys

- Access patterns

OLTP- short atomic transaction OLAP – mostly read only operation



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 maintainedover 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	≥ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Data Warehousing: A Multitiered Architecture

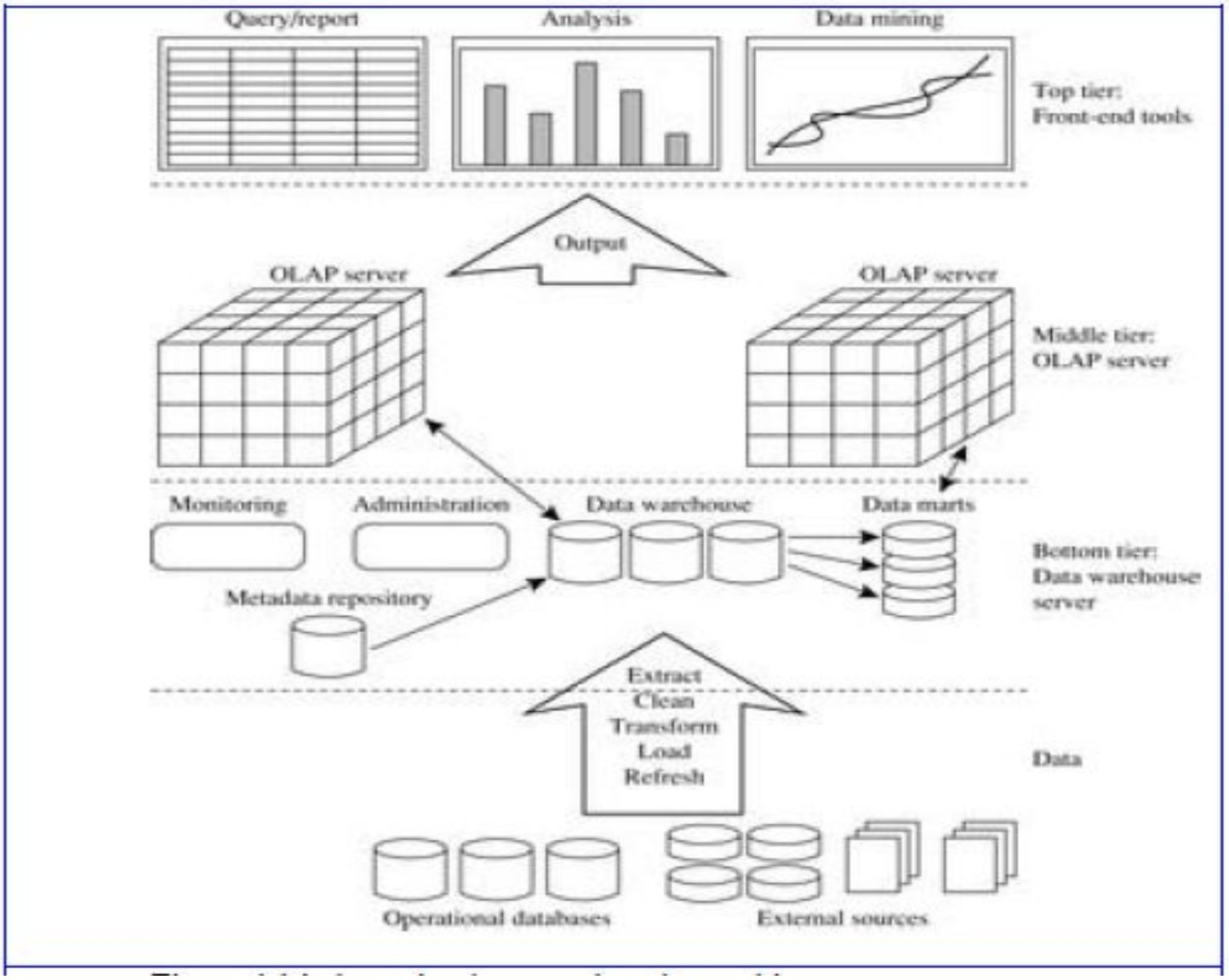


Figure: Three-tire Data warehouse Architecture

Data Warehouse Models

1. Enterprise Warehouse

2. Data Mart
   - Independent
   - Dependent

3. Virtual Warehouse

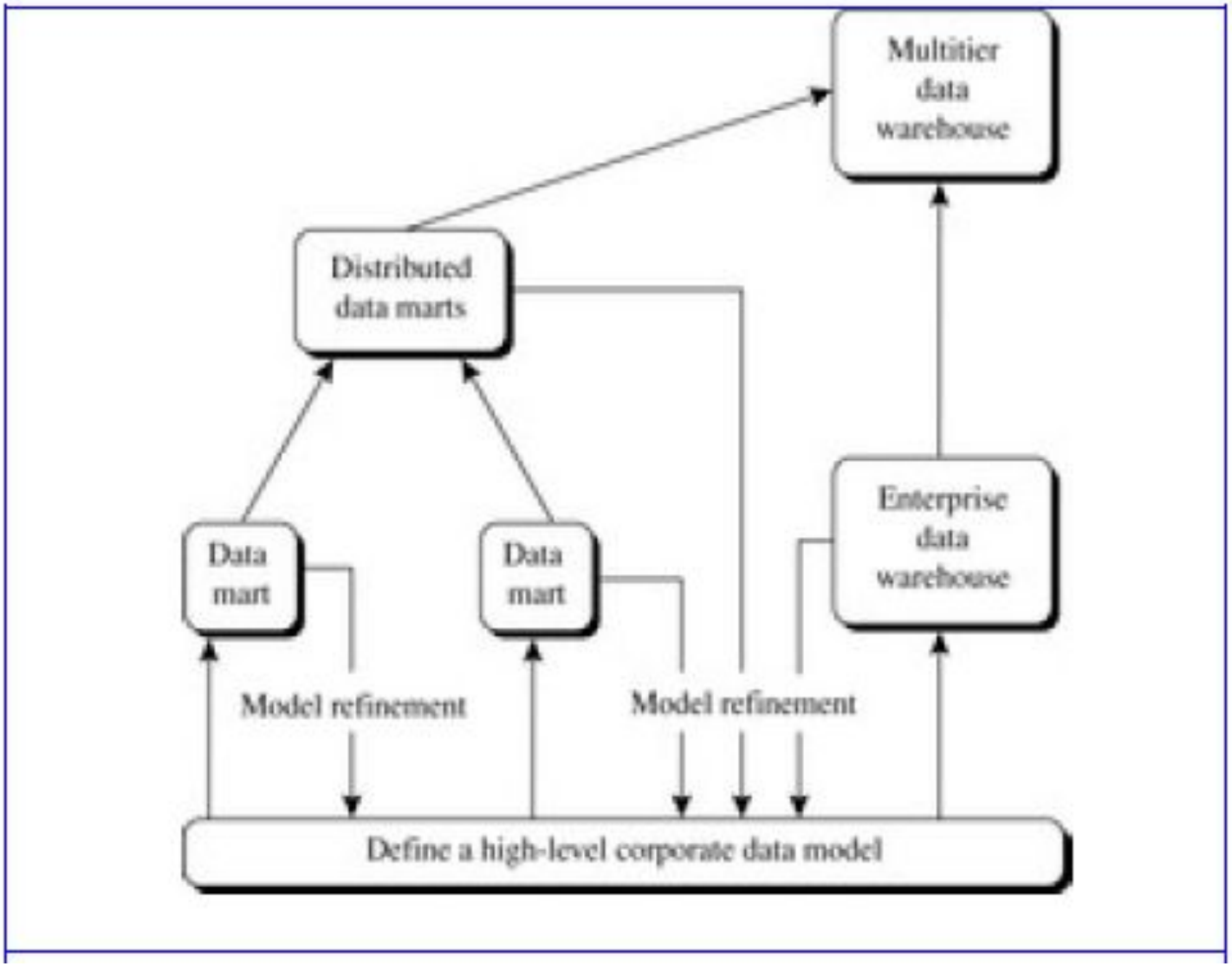


Figure : A recommended approach for data warehouse development

Data warehouse systems use back-end tools and utilities to populate and refresh their data

Data extraction

Data cleaning

Data transformation

Load

Refresh

Metadata Repository

- Metadata are data about data.
- When used in a data warehouse, metadata are the data that define warehouse objects
- A data warehouse contains different levels of summarization, of which metadata is one.
- A metadata repository should contain the following:

1. *A description of the data warehouse structure*
2. *Operational metadata*
3. *The algorithms used for summarization*
4. *Mapping from the operational environment to the data warehouse*
5. *Data related to system performance*
6. *Business metadata*

Data Cube: A Multidimensional Data Model

- A data cube allows data to be modeled and viewed **in multiple** dimensions
- It is defined by **dimensions** and **facts**
- **Dimensions** are the perspectives or entities with respect to which an organization wants to keep records
- **Facts** are numeric measures
- **Measures** are stored in fact table.
- In Data Warehousing the data cube is **n-dimensional**

2-D and 3-D View

Table 4.2 2-D View of Sales Data for *AllElectronics* 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).

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

	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

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

3-D Data Cube

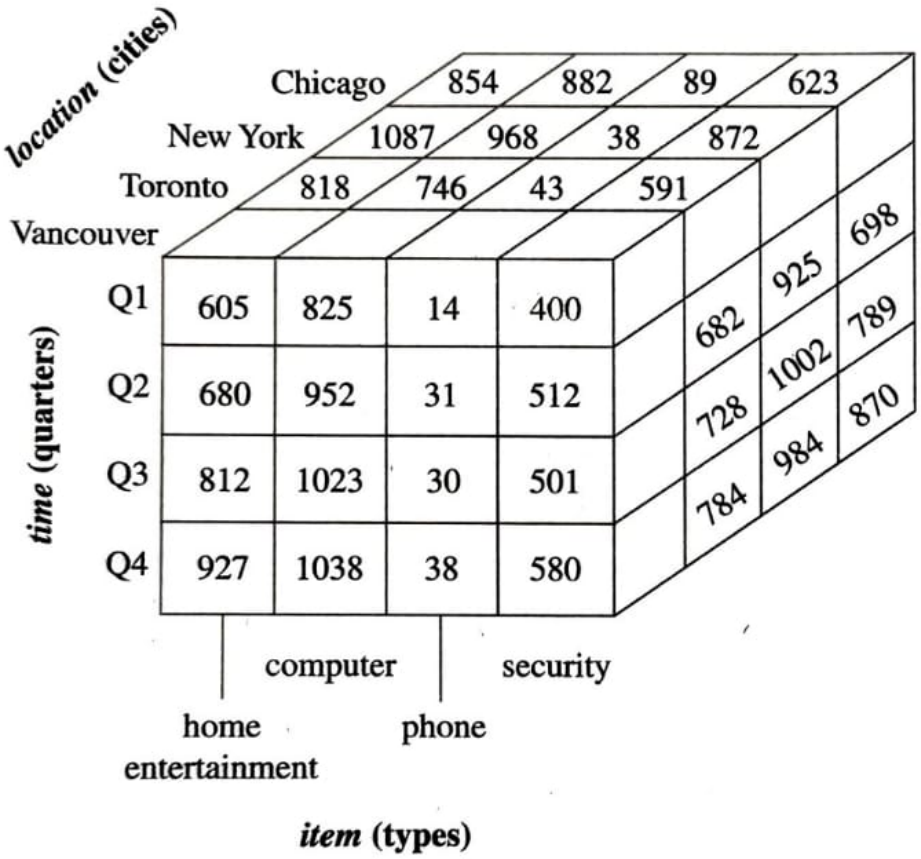


Figure 4.3 A 3-D data cube representation of the data in Table 4.3, according to *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

4-D Data Cube

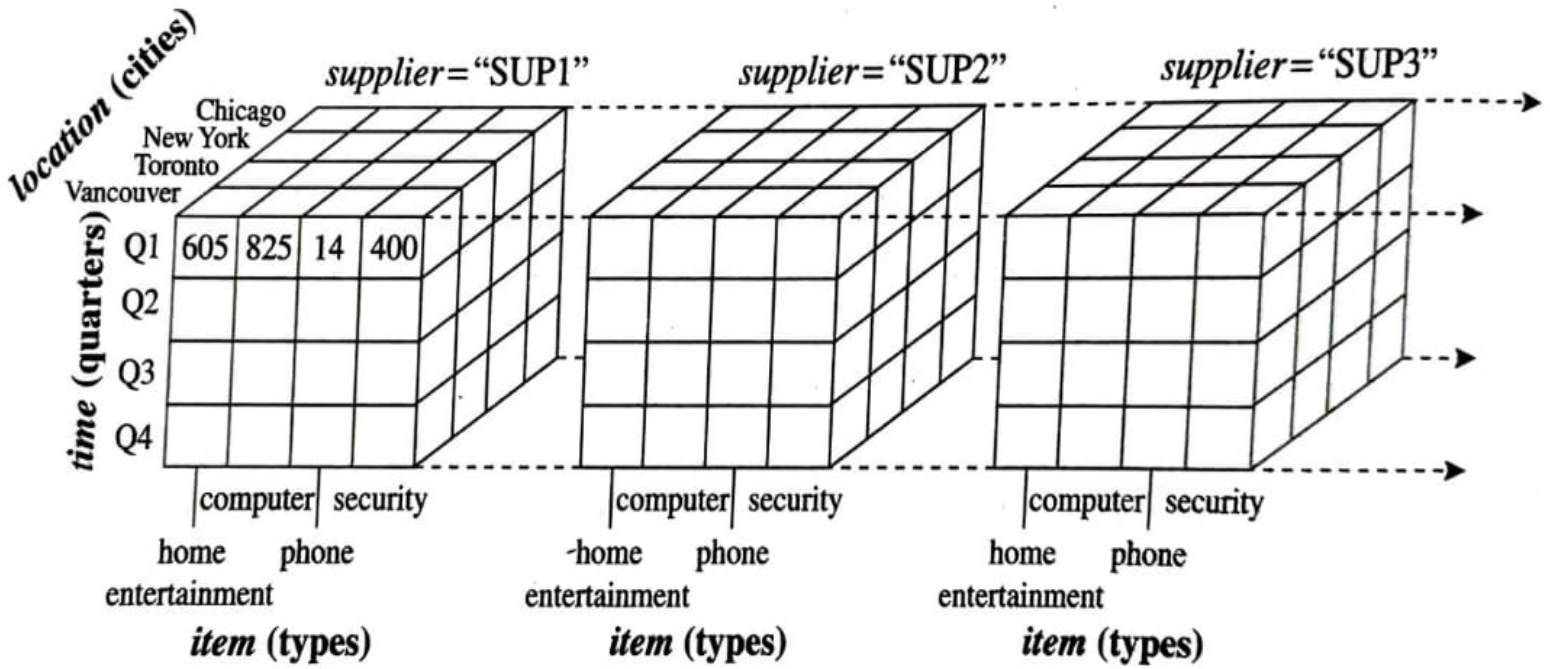


Figure 4.4 A 4-D data cube representation of sales data, according to *time*, *item*, *location*, and *supplier*. The measure displayed is *dollars_sold* (in thousands). For improved readability, only some of the cube values are shown.

Lattice of Cuboids

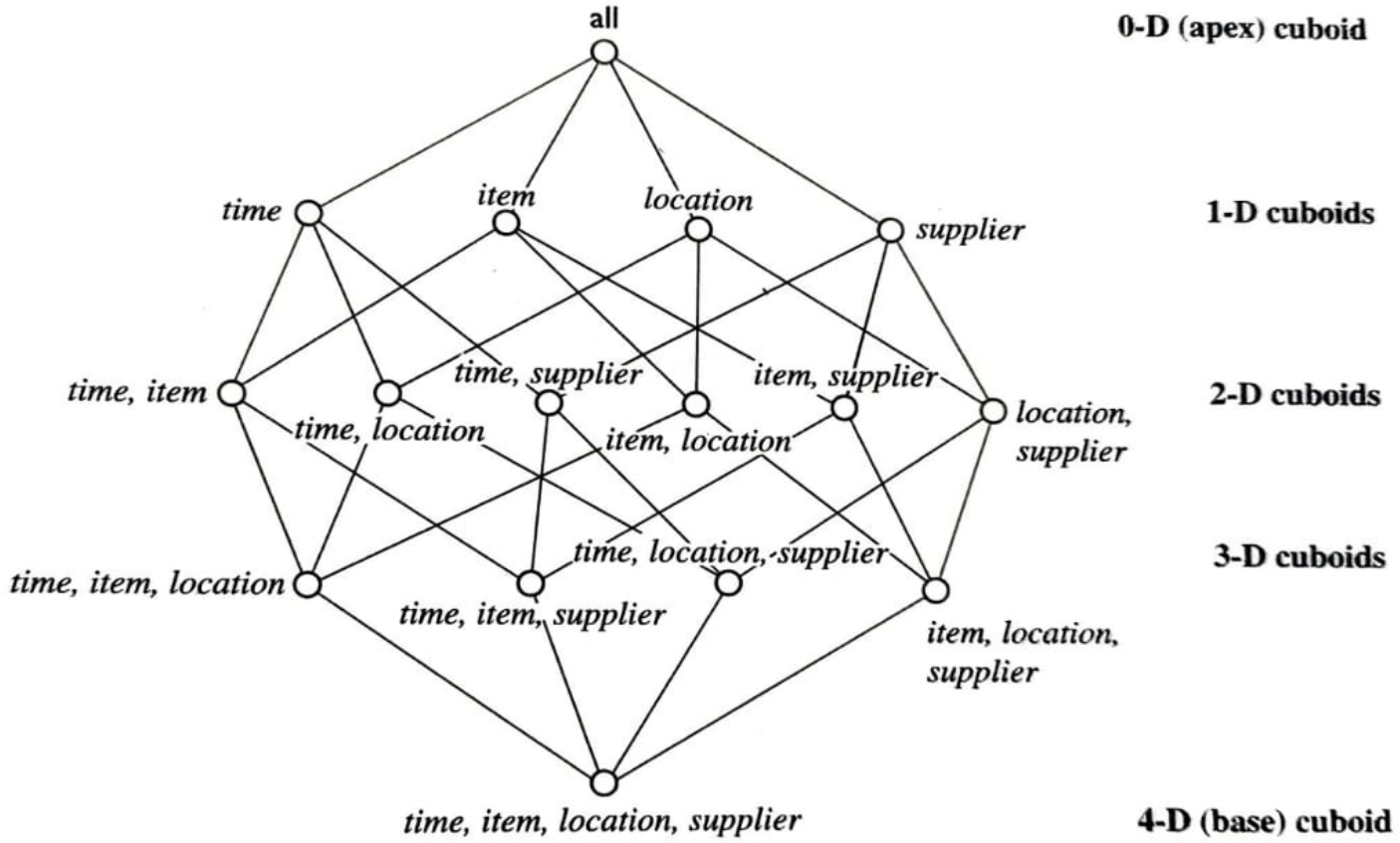
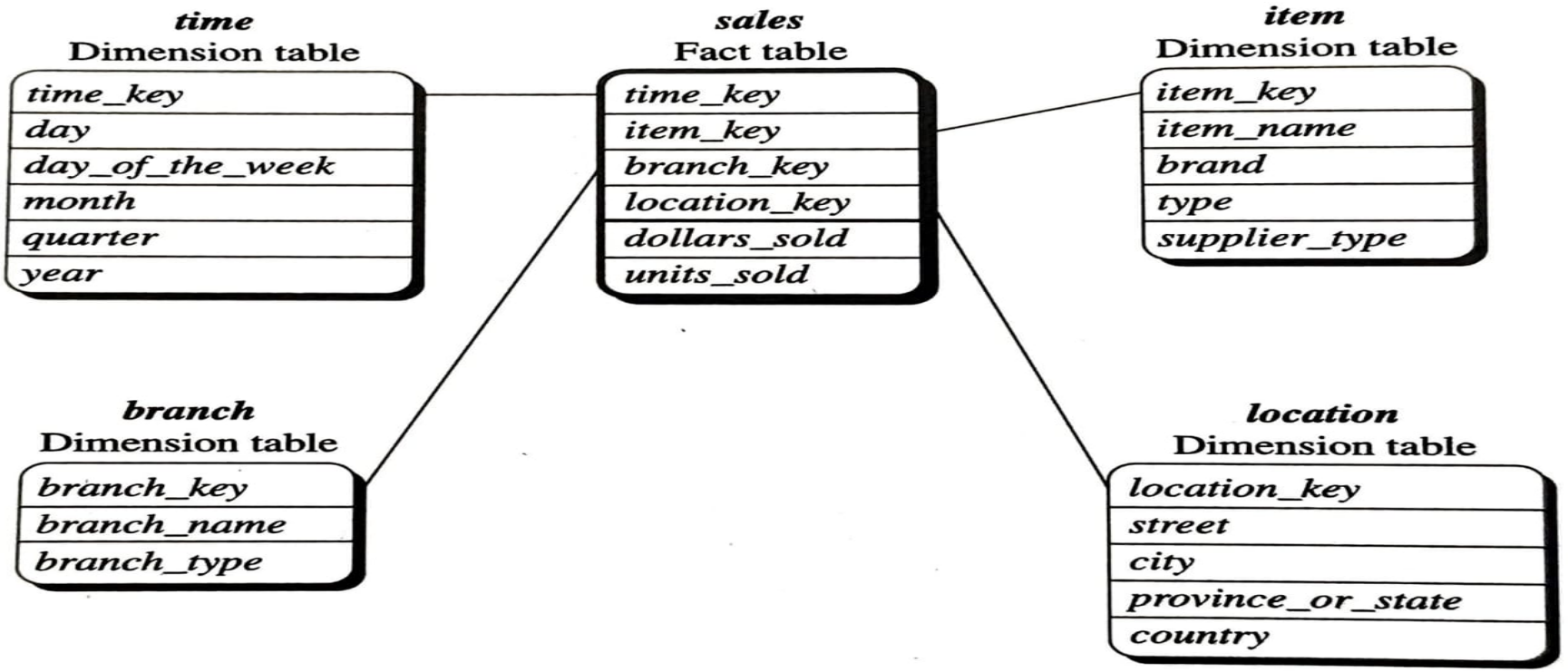


Figure 4.5 Lattice of cuboids, making up a 4-D data cube for *time*, *item*, *location*, and *supplier*. Each cuboid represents a different degree of summarization.

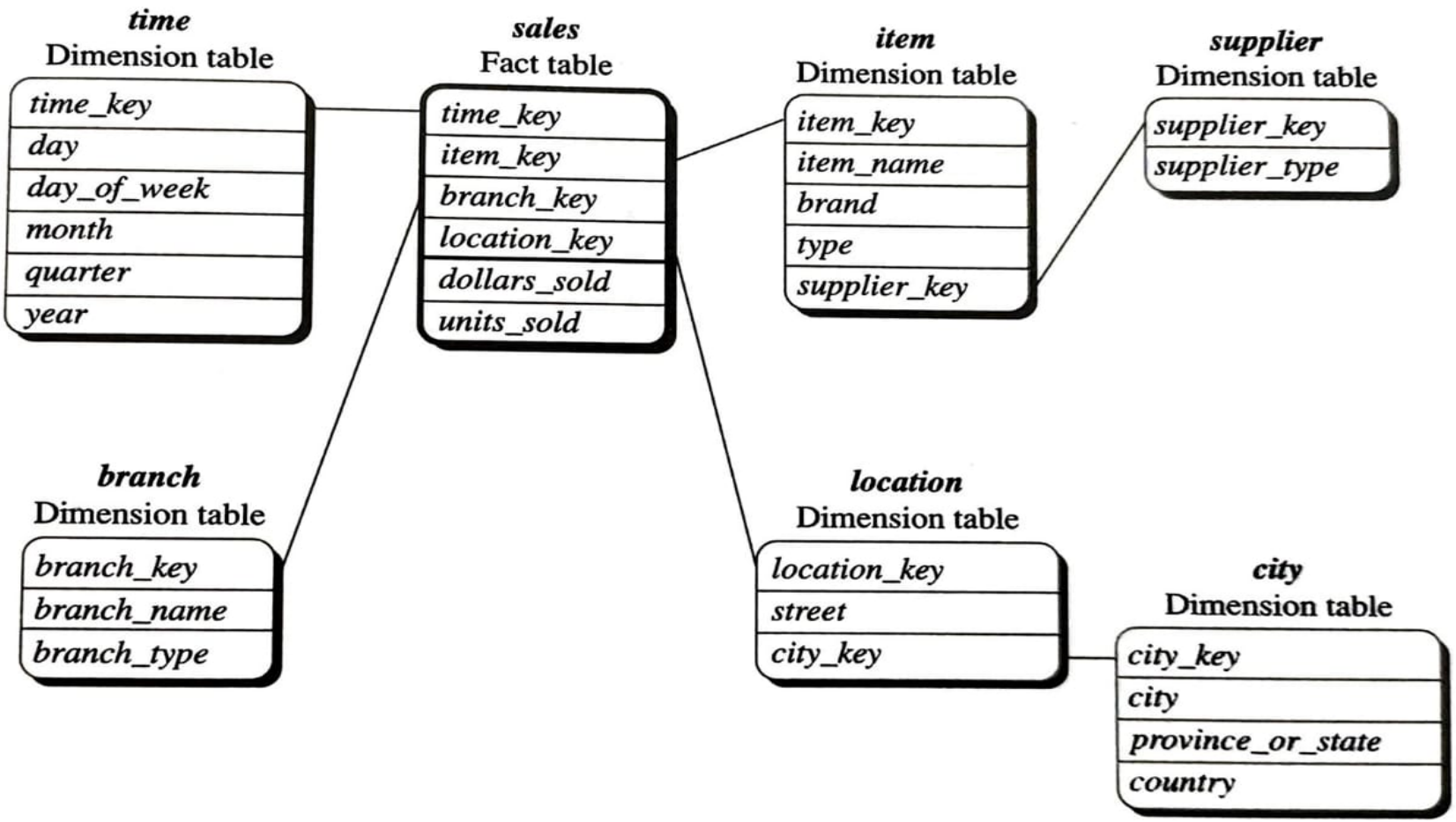
Stars, Snowflakes, and Fact Constellations:
Schemas for Multidimensional Data Models

Stars Schema



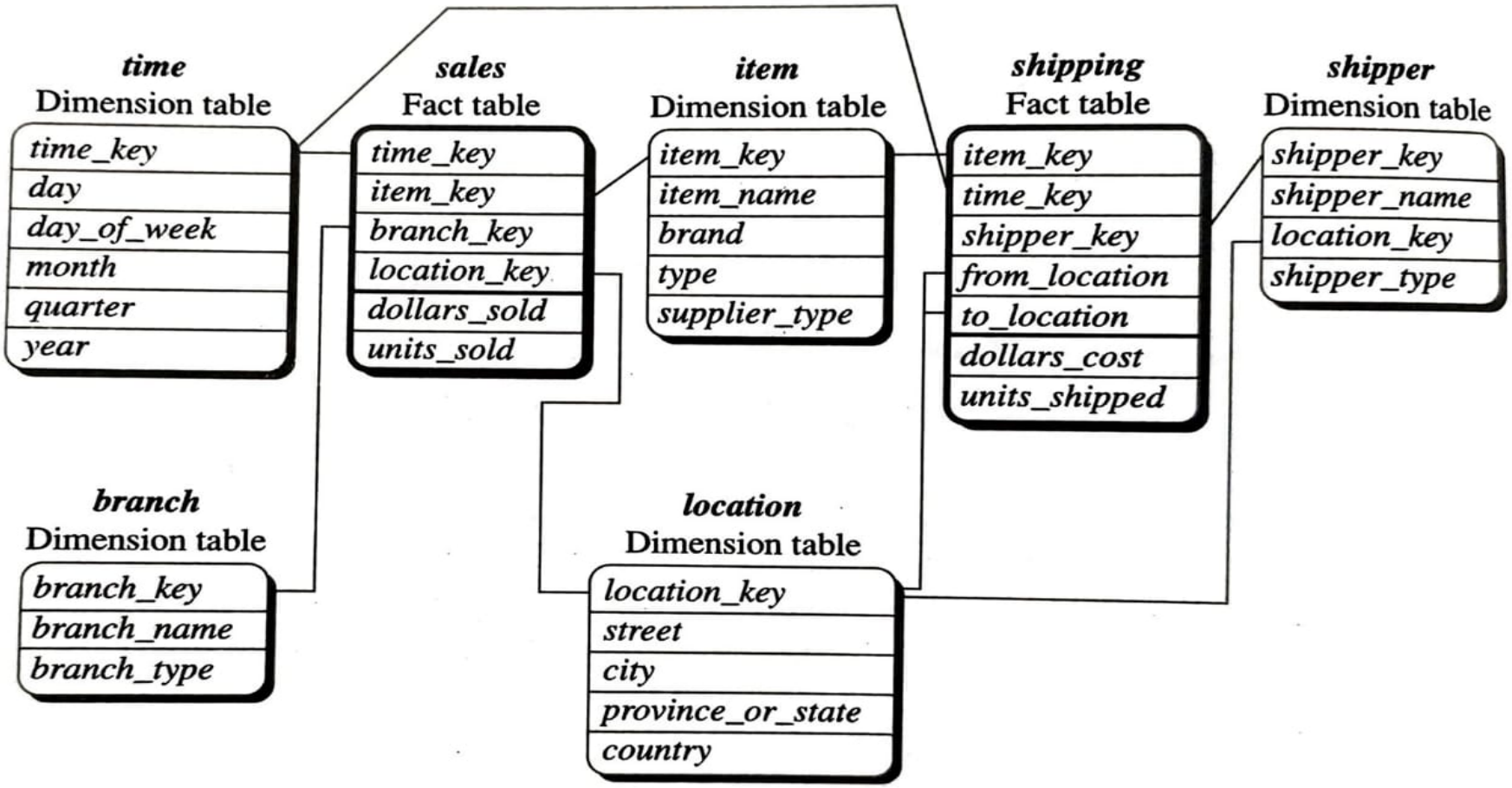
Star schema of *sales* data warehouse.

Snowflakes Schema



Snowflake schema of a *sales* data warehouse.

Fact Constellations Schema

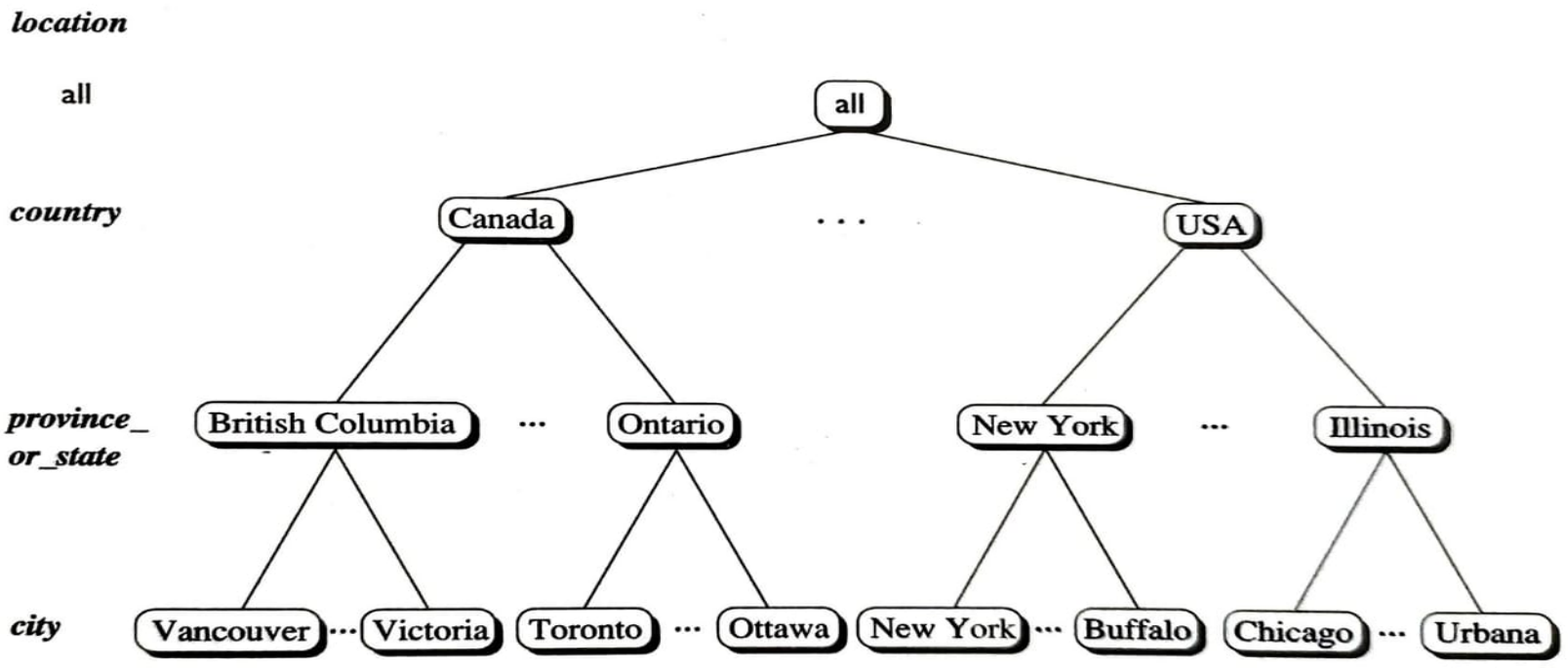


Fact constellation schema of a sales and shipping data warehouse.

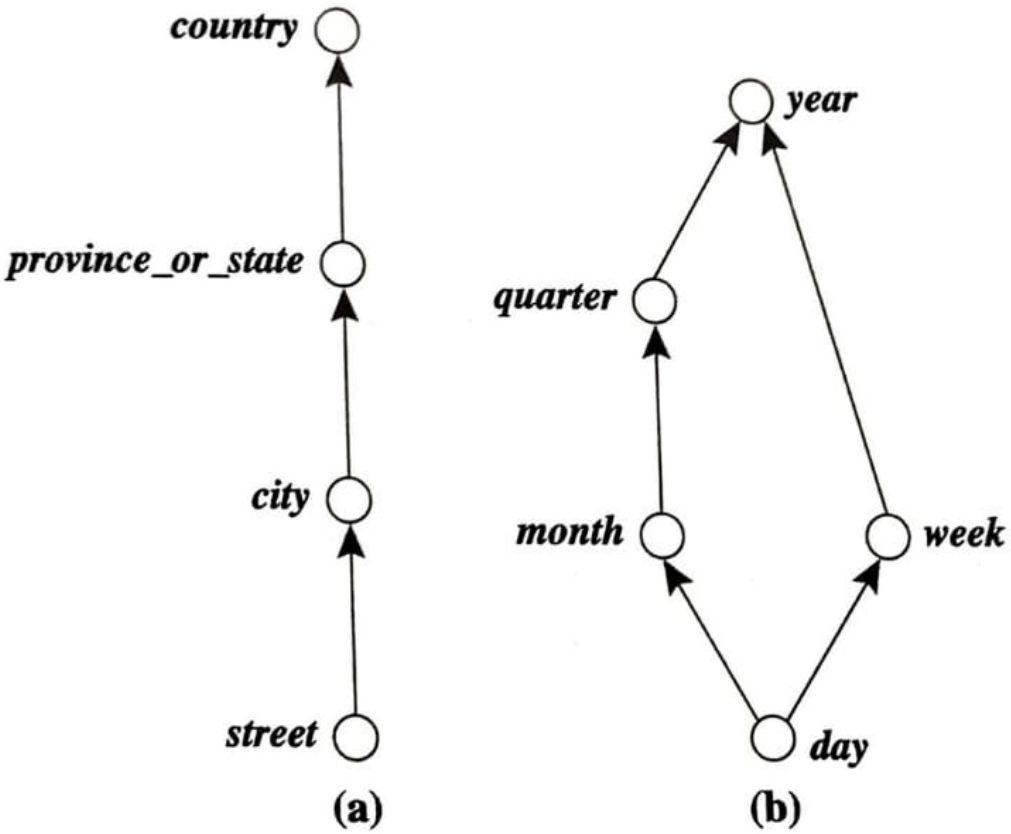
Dimensions: The Role of Concept Hierarchies

- A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts

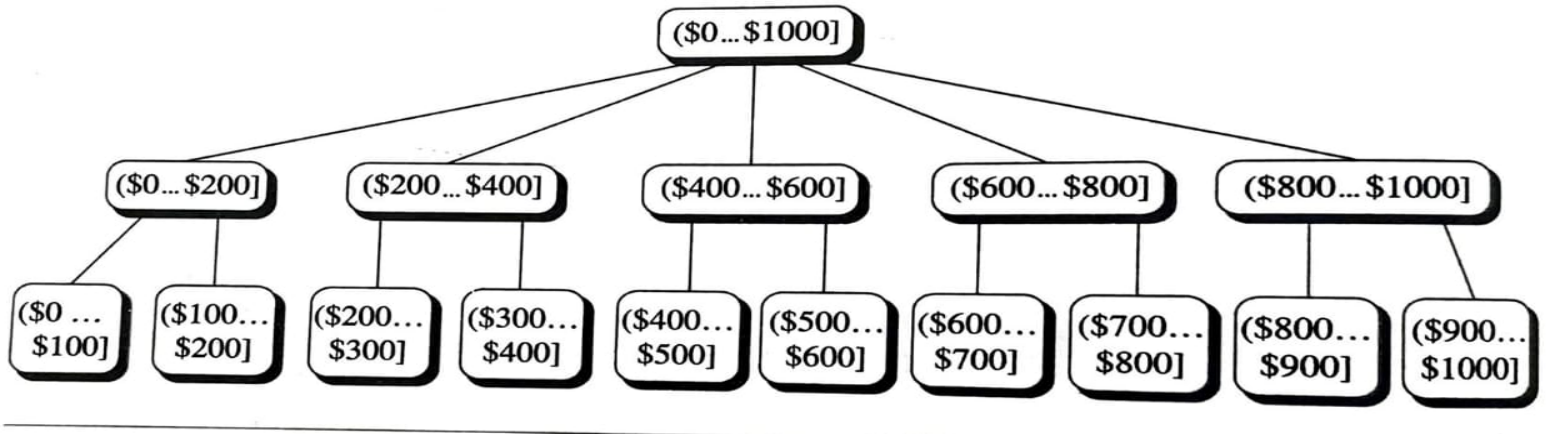
- Example Concept hierarchy for the dimension ***location***.



A concept hierarchy for *location*. Due to space limitations, not all of the hierarchy nodes are shown, indicated by ellipses between nodes.



Hierarchical and lattice structures of attributes in warehouse dimensions: (a) a hierarchy for *location* and (b) a lattice for *time*.



A concept hierarchy for *price*.

- Concept hierarchies may also be defined by discretizing or grouping values for a given dimension or attribute, resulting in a **"Set-Grouping hierarchy"**
- For dimension "prize" with interval range **$x (exclusive) to $y (inclusive)**
- There may be more than one concept hierarchy for a given attribute or Dimension based on different user view points
- Concept hierarchies may be provided manually by system users, domain experts or knowledge engineers or may be automatically generated based on statistical analysis of the data distribution.

Measures: Their Categorization and Computation

- A data cube measure is a numeric function that can be evaluated at each point in the data cube space

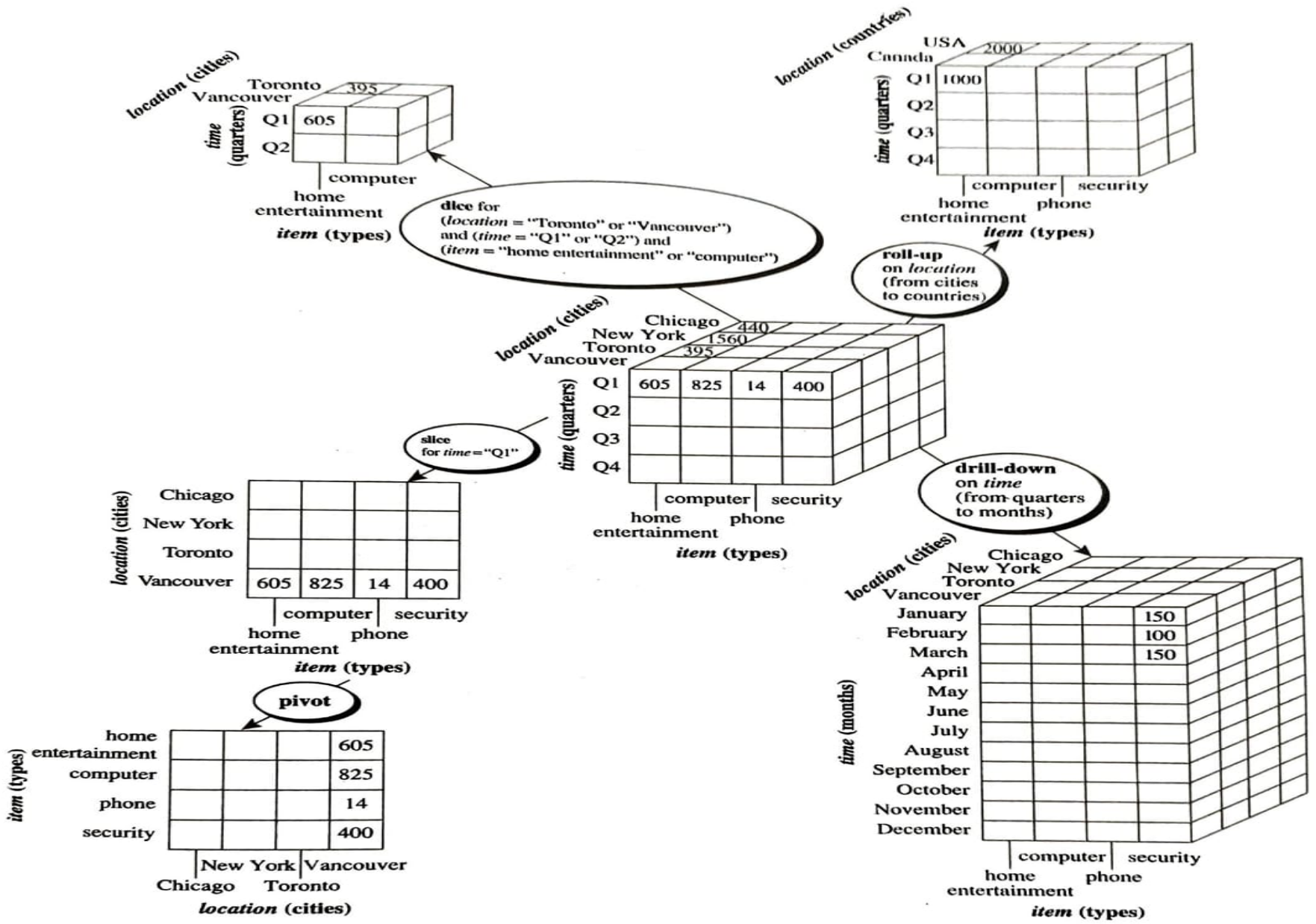
- A measure value is computed for a given point by aggregating the data corresponding to the respective dimension–value pairs defining the given point.

- Measures can be organized into three categories—distributive, algebraic, and holistic—based on the kind of aggregate functions used

- **Distributive:** Distributive measure can be computed efficiently because of the way the computation can be partitioned
  Ex: sum(),count(), min(), max()
- **Algebraic:** A measure is algebraic if it is obtained by applying an algebraic aggregate function
  Ex- Avg() computed by sum()/count(), standard_deviation()
- **Holistic:** A measure is holistic if it is obtained by applying a holistic aggregate function
  Ex- median(), mode(), rank()

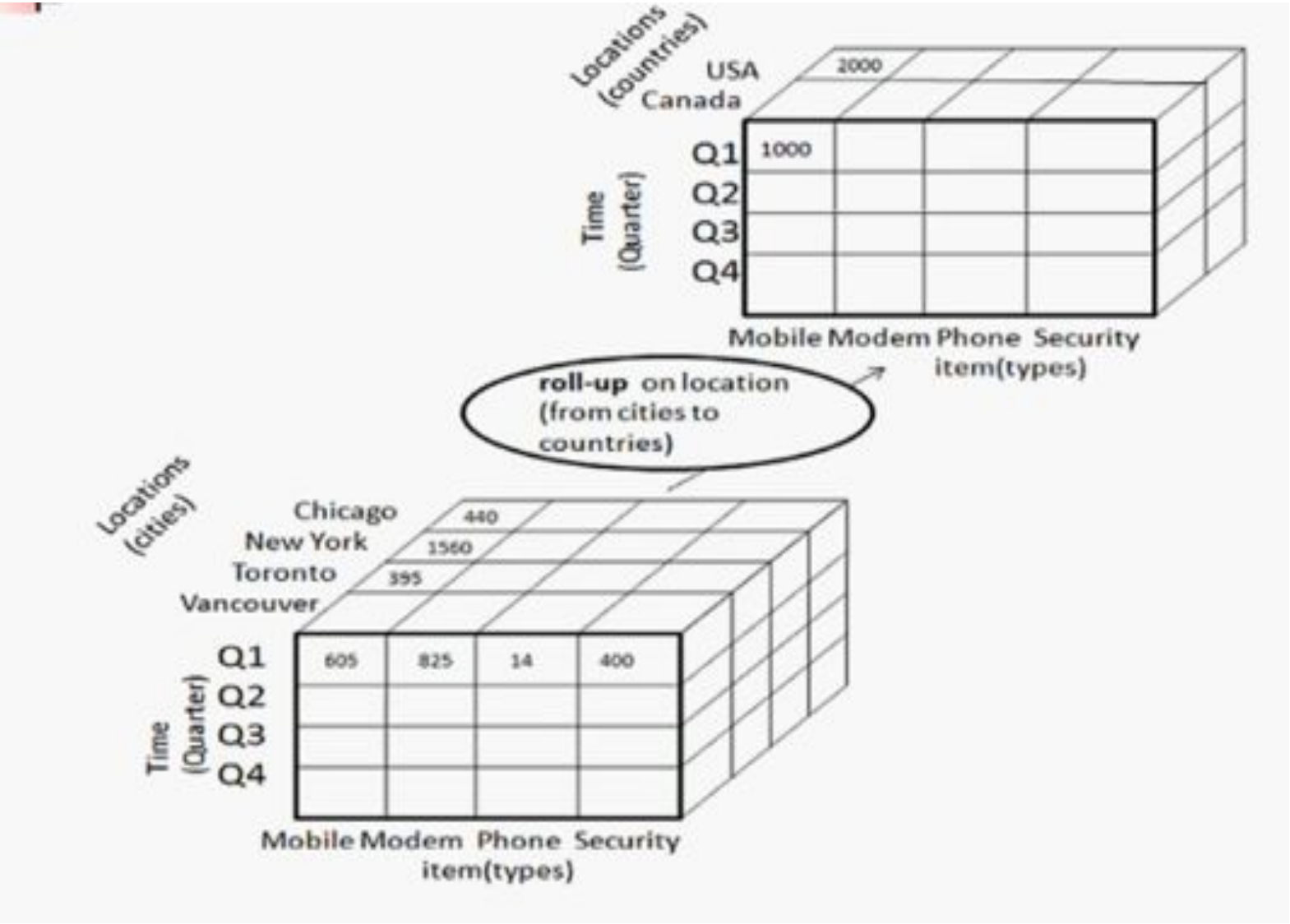
Typical OLAP Operations

- **Roll – Up**
- **Drill – Down**
- **Slice & Dice**
- **Pivot or Rotate**

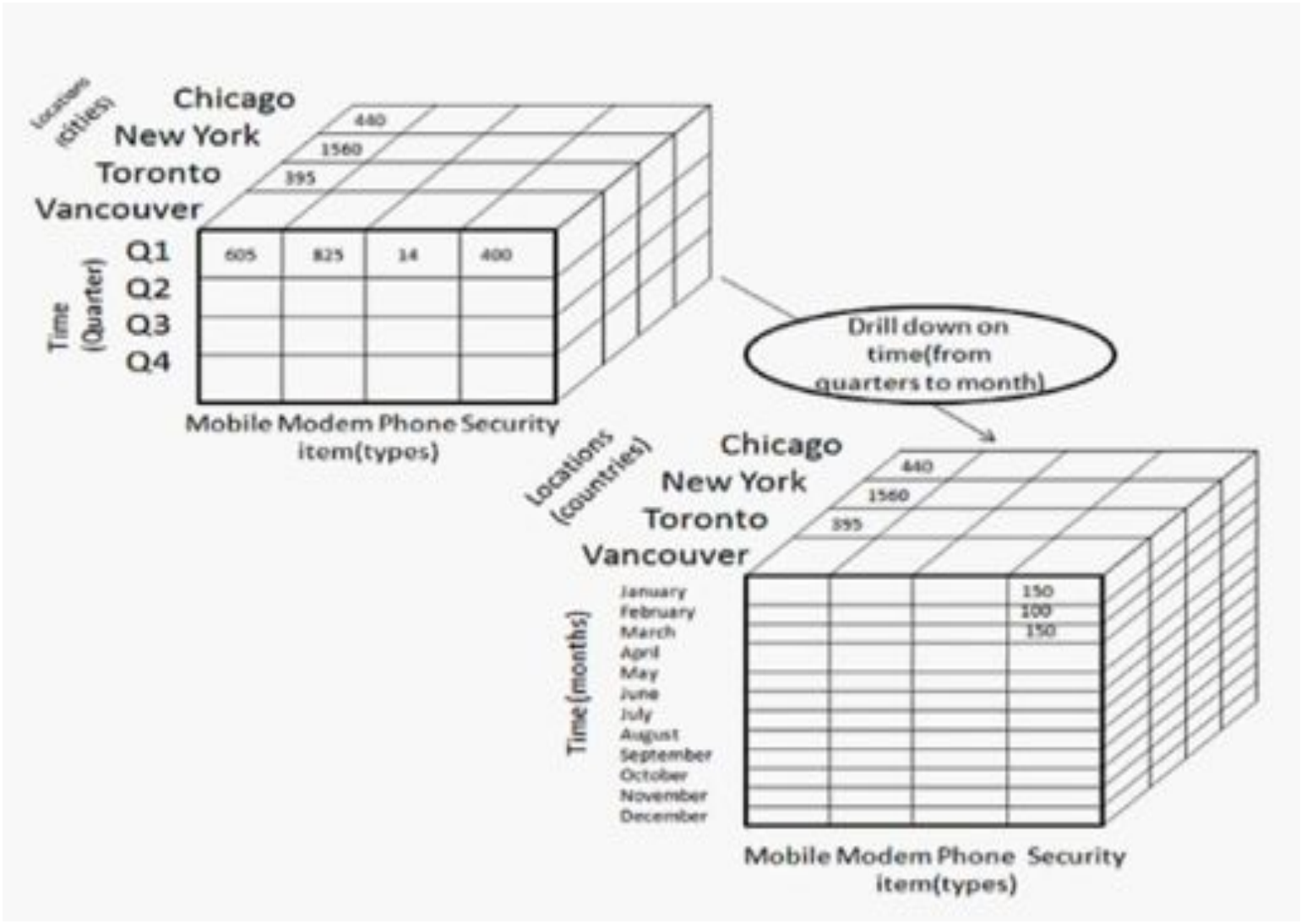


ure 4.12 Examples of typical OLAP operations on multidimensional data.

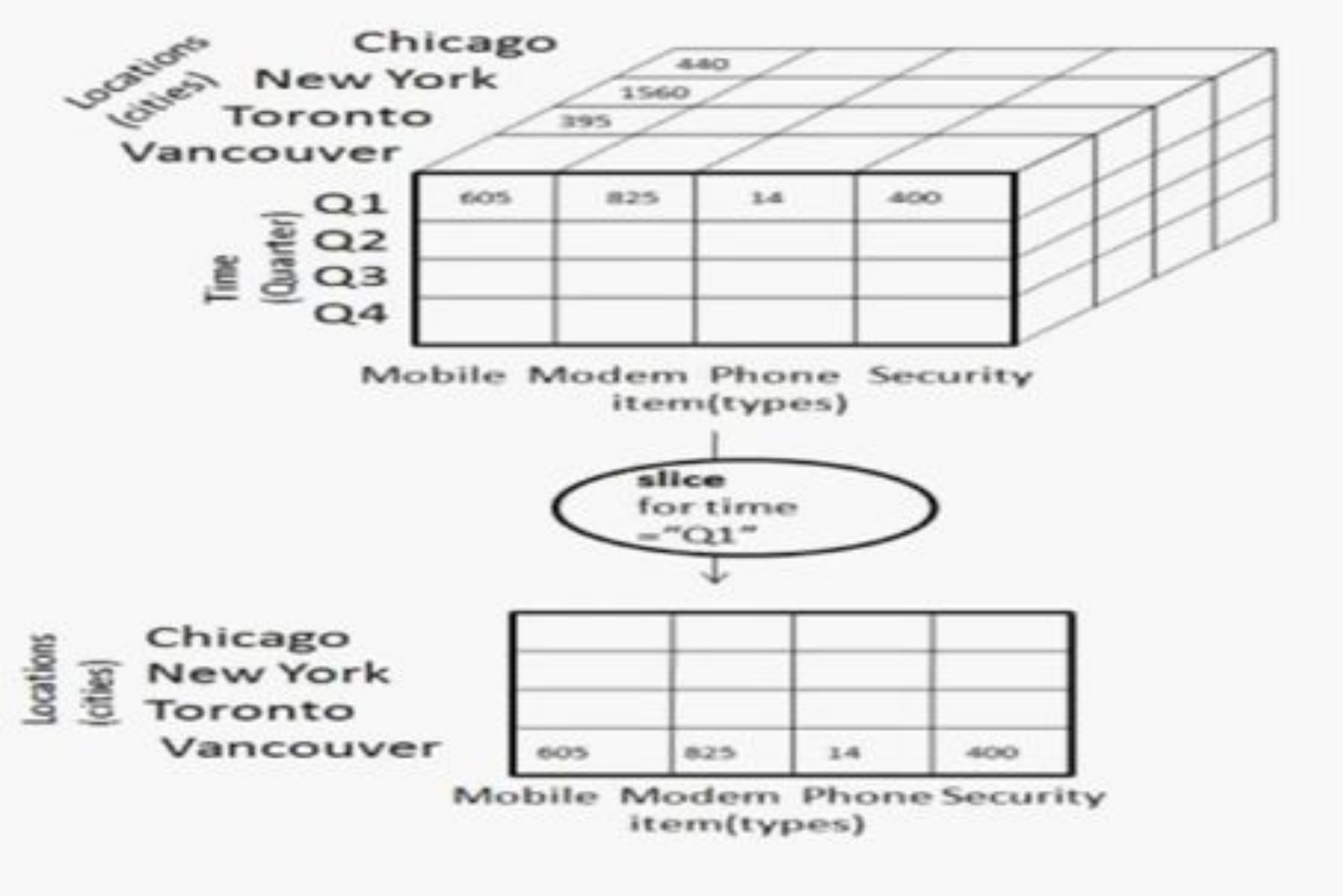
Roll – Up


Locations (countries)
USA
Canada
2000
Q1
Q2
Q3
Q4
1000
Time (Quarter)
Mobile Modem Phone Security
item(types)
roll-up on location (from cities to countries)
Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
Q2
Q3
Q4
605
825
14
400
Time (Quarter)
Mobile Modem Phone Security
item(types)

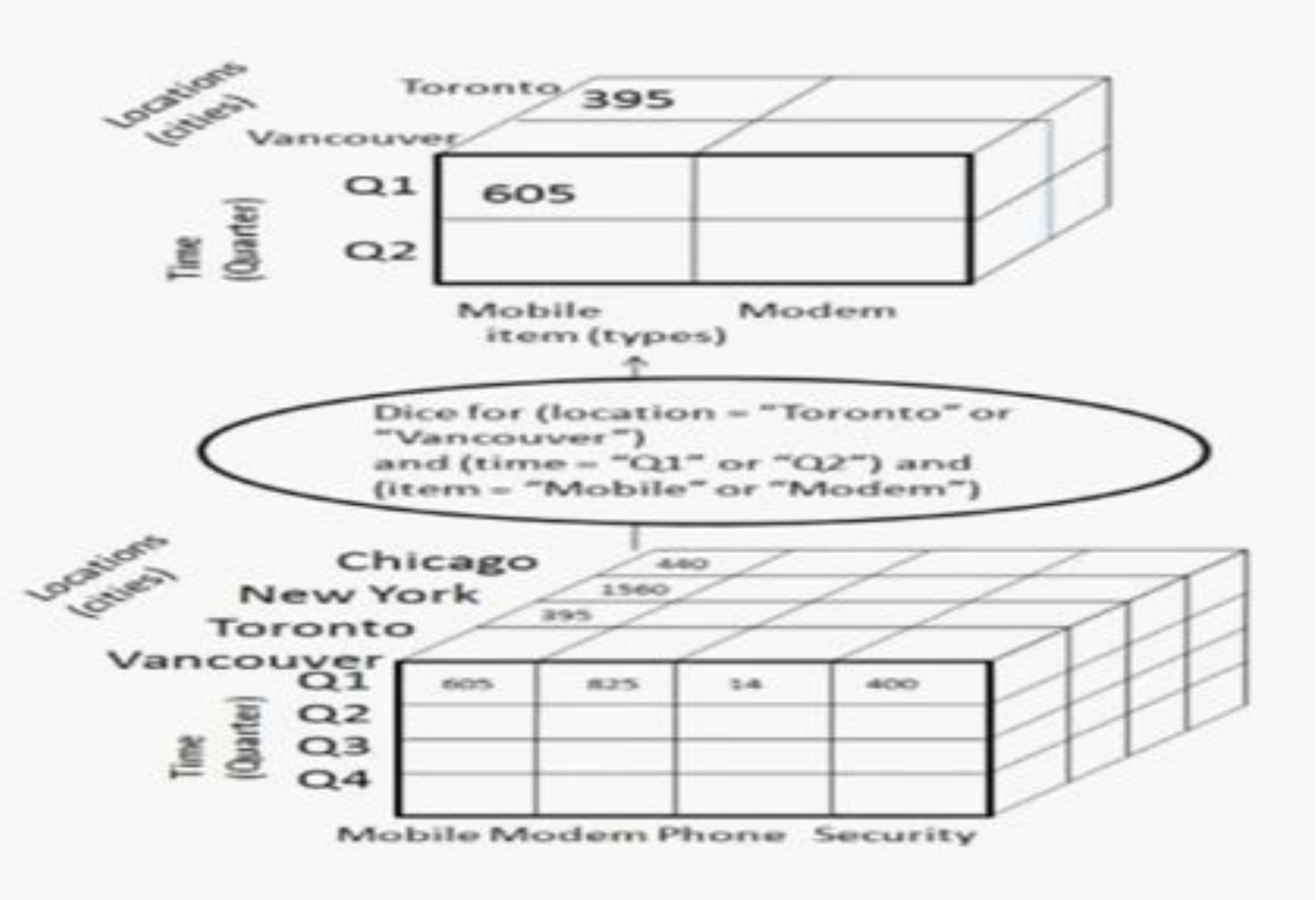
Drill – Down


Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
Mobile Modem Phone Security
item(types)
Drill down on time(from quarters to month)
Locations (countries)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time (months)
January
February
March
April
May
June
July
August
September
October
November
December
150
100
150
Mobile Modem Phone Security
item(types)

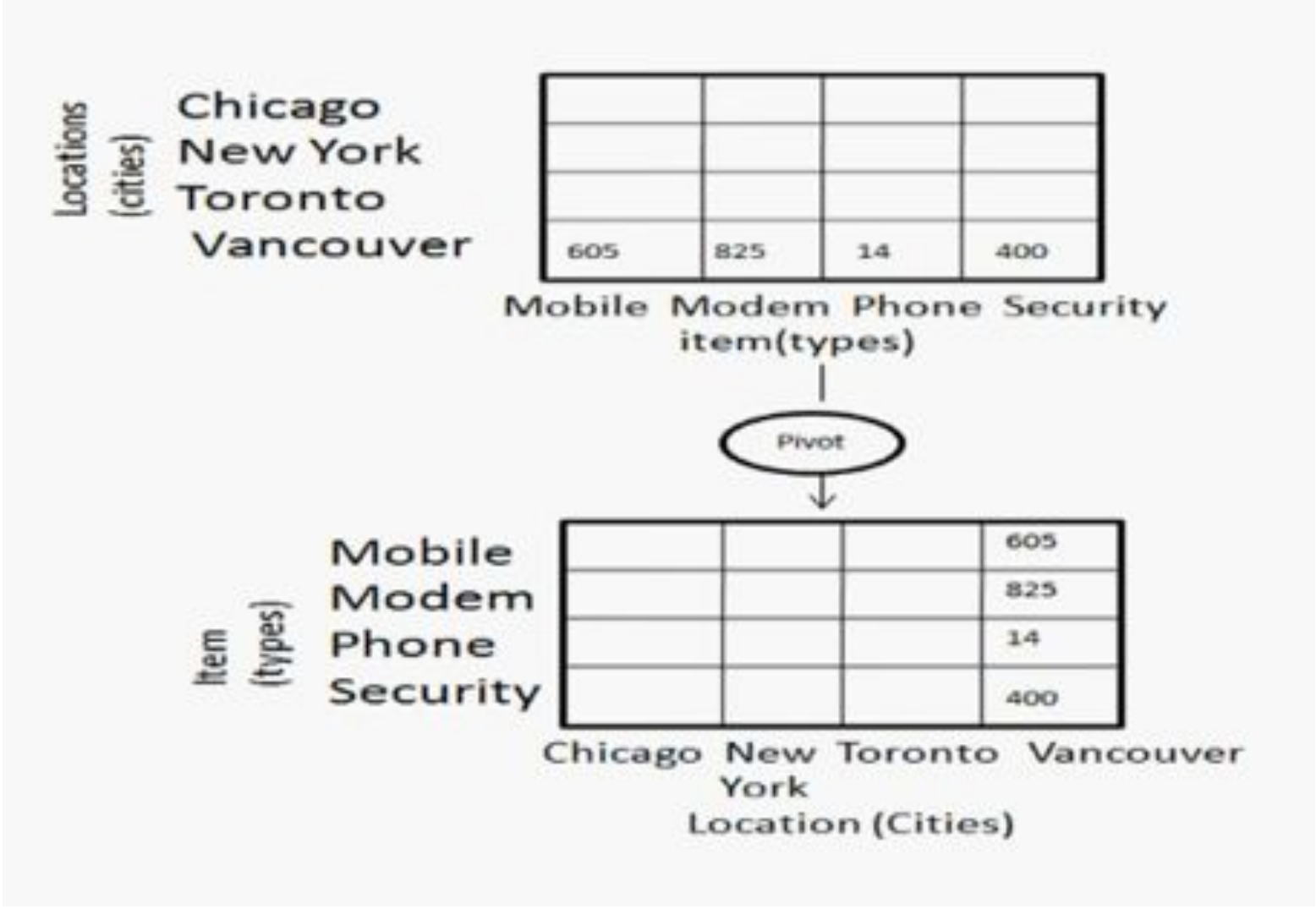
Slice


Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
Mobile Modem Phone Security
item(types)
slice
for time
="Q1"
Locations (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
Mobile Modem Phone Security
item(types)

Dice


Locations (cities)
Toronto
395
Vancouver
Q1
605
Q2
Time (Quarter)
Mobile
Modem
item (types)
Dice for (location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "Mobile" or "Modem")
Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
605
825
14
400
Q2
Q3
Q4
Time (Quarter)
Mobile
Modem
Phone
Security

Pivot


Locations (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
Mobile Modem Phone Security
item(types)
Pivot
Item (types)
Mobile
Modem
Phone
Security
605
825
14
400
Chicago New York Toronto Vancouver
Location (Cities)

Thank you