

---

Università degli Studi di Milano  
Master Degree in Computer Science

# Information Management course

Teacher: Alberto Ceselli

Lecture 14 : 18/11/2014

# Data Mining: Concepts and Techniques

---

(3<sup>rd</sup> ed.)


## — Chapter 4 —

Jiawei Han, Micheline Kamber, and Jian Pei  
University of Illinois at Urbana-Champaign &  
Simon Fraser University

©2011 Han, Kamber & Pei. All rights reserved.

# Chapter 4: Data Warehousing and On-line Analytical Processing

---

- Data Warehouse: Basic Concepts 
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- *Data Warehouse Implementation*
- *Data Generalization by Attribute-Oriented Induction*
- Summary

# 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

# 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—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

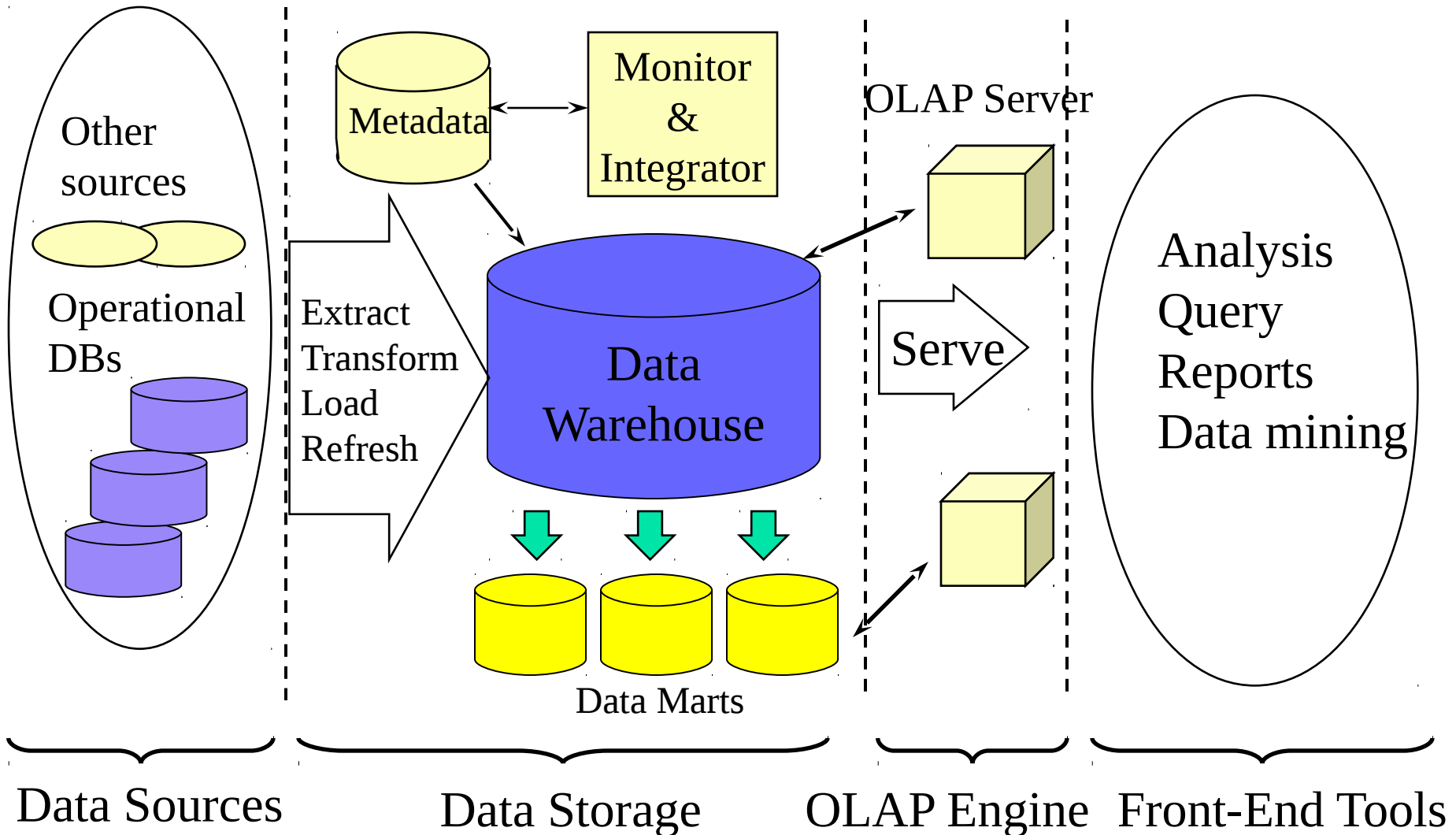
	<b>OLTP</b>	<b>OLAP</b>
<b>users</b>	clerk, IT professional	knowledge worker
<b>function</b>	day to day operations	decision support
<b>DB design</b>	application-oriented	subject-oriented
<b>data</b>	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated,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

# Data Warehouse: A Multi-Tiered Architecture



# Three Data Warehouse Models

---

- **Enterprise warehouse**
  - collects all of the information about subjects spanning the entire organization
- **Data Mart**
  - a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - Independent (captured from external sources)
    - Dependent (directly from company datawarehouse)
- **Virtual warehouse**
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized

# 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 indices and partitions

- **Refresh**

- propagate the updates from the data sources to the warehouse

# Metadata Repository

---

- **Meta data** is the data defining warehouse objects. It stores:
- Description of the **structure** of the data warehouse
  - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
- **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 DW
- Data related to **system performance**
  - warehouse schema, view and derived data definitions
- **Business data**
  - business terms and definitions, ownership of data ...

# Chapter 4: Data Warehousing and On-line Analytical Processing

---

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- *Data Warehouse Implementation*
- *Data Generalization by Attribute-Oriented Induction*
- Summary



# From Tables and Spreadsheets to Data Cubes

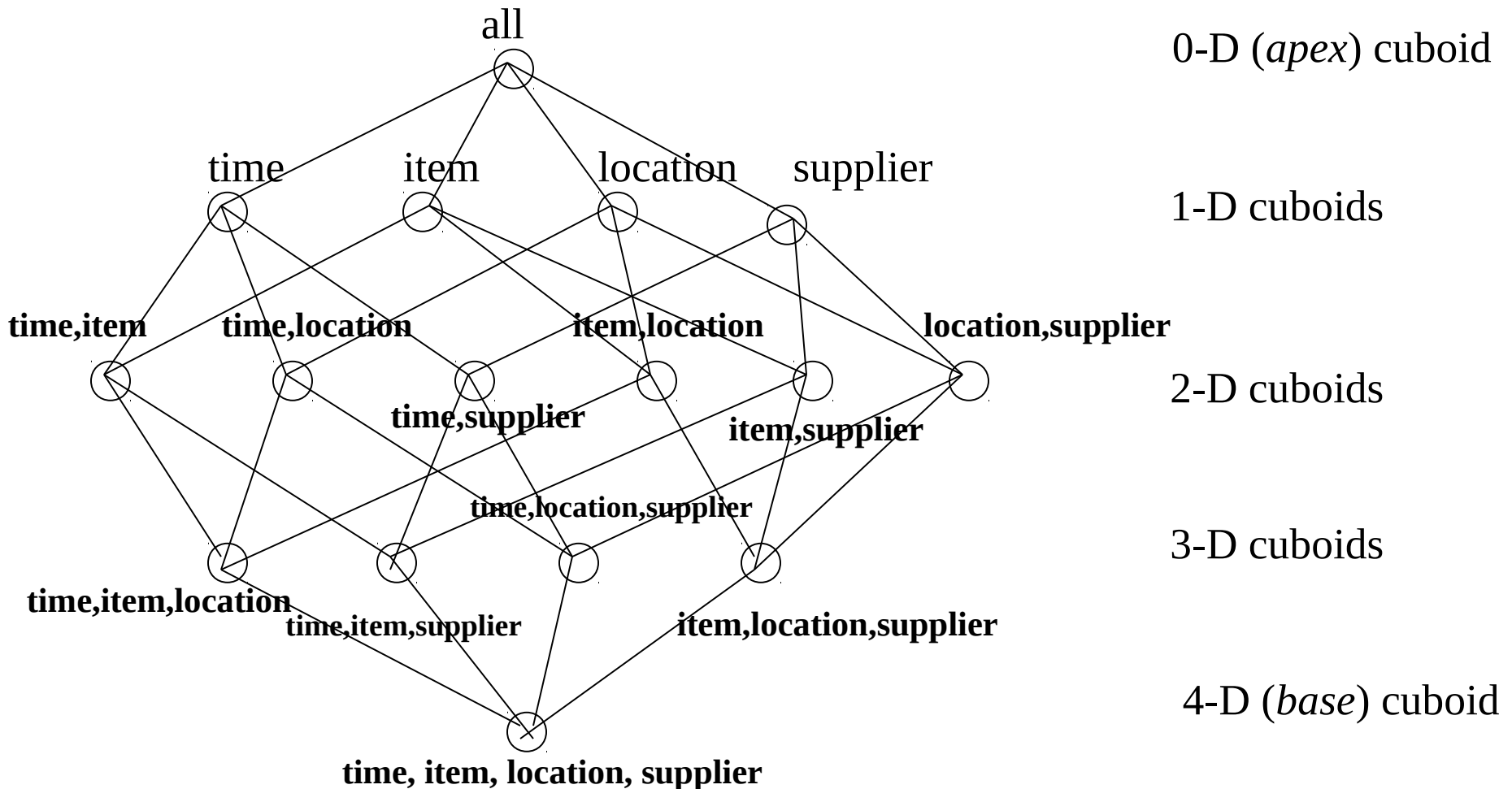
---

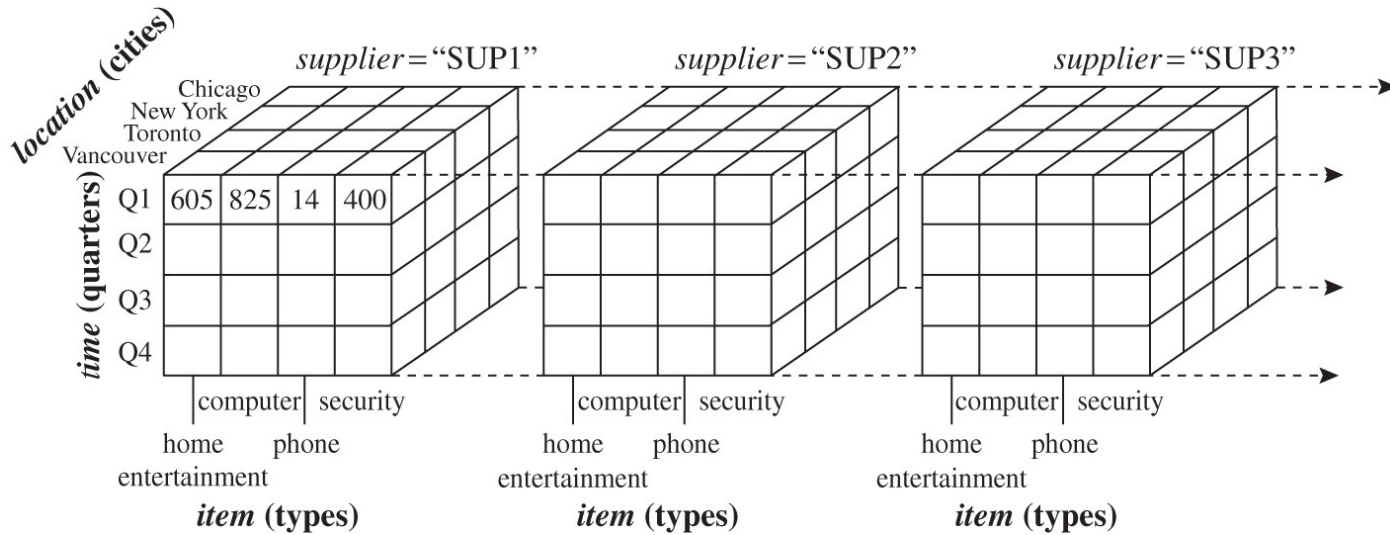
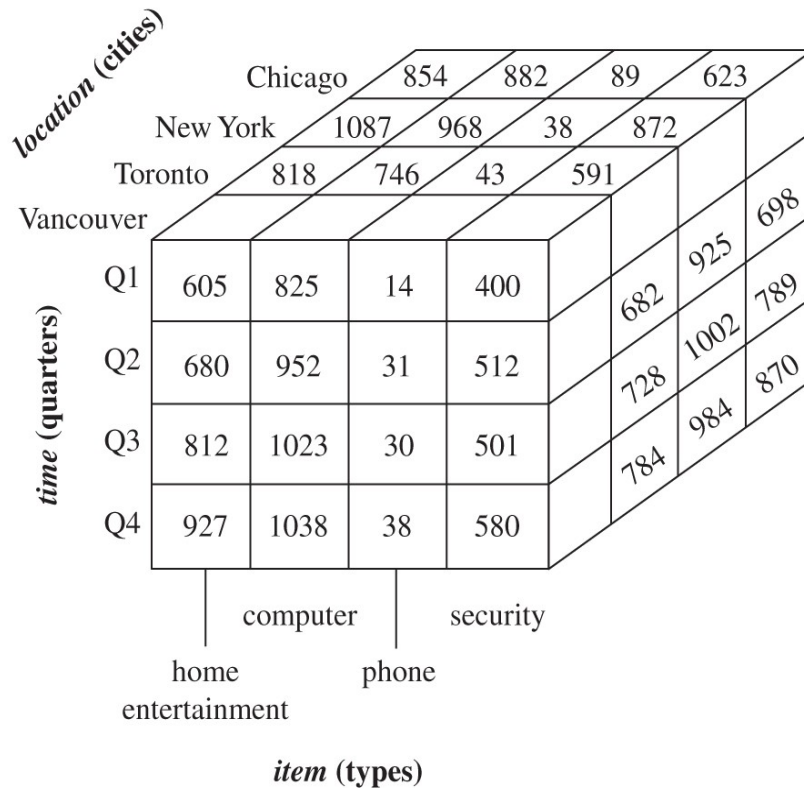
- A **data warehouse** is based on a **multidimensional data model** which views data in the form of a data cube
- A data cube, such as **sales**, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as **item** (**item\_name**, **brand**, **type**), or **time**(**day**, **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**.



# Data Cube: A Lattice of Cuboids

---

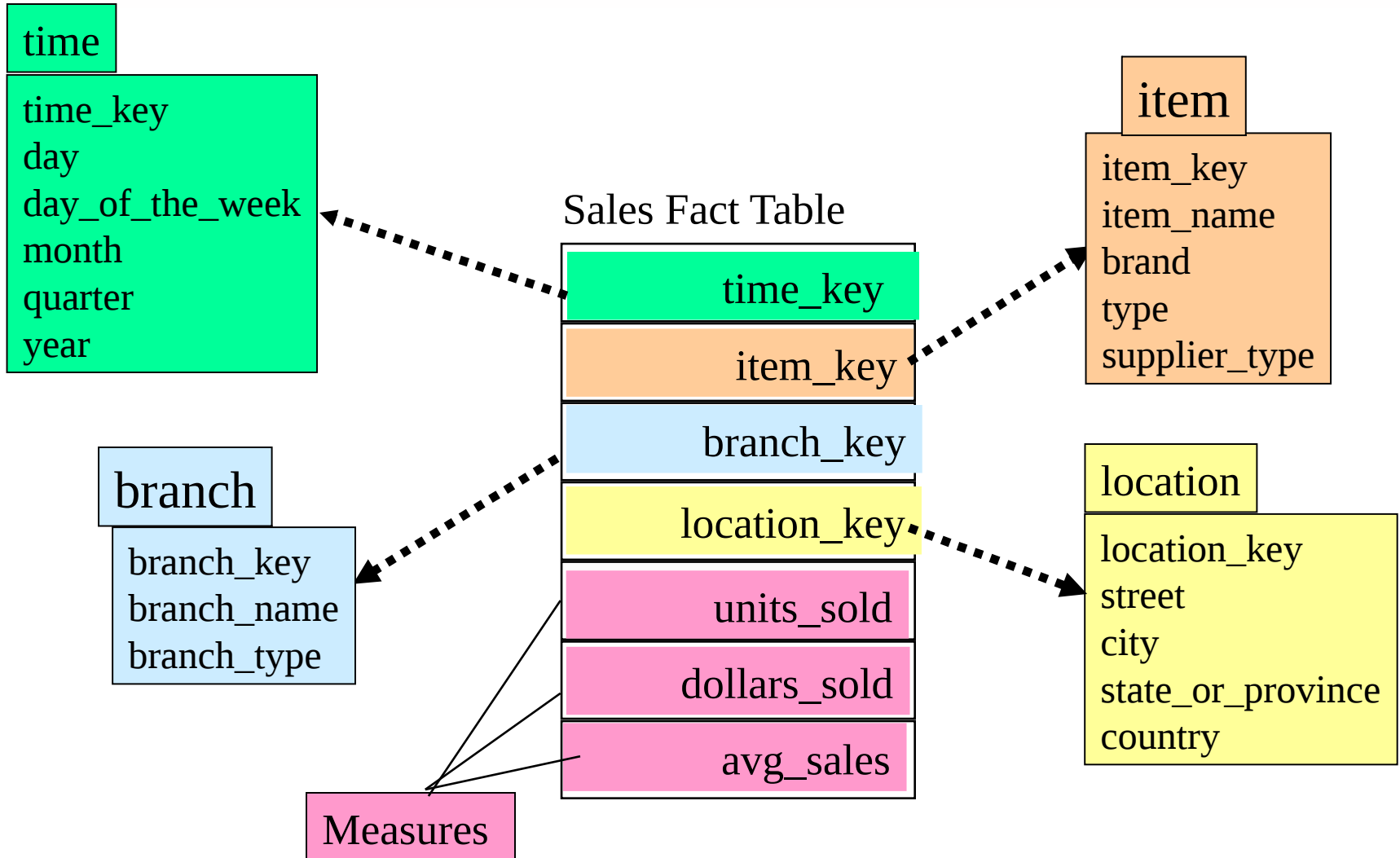




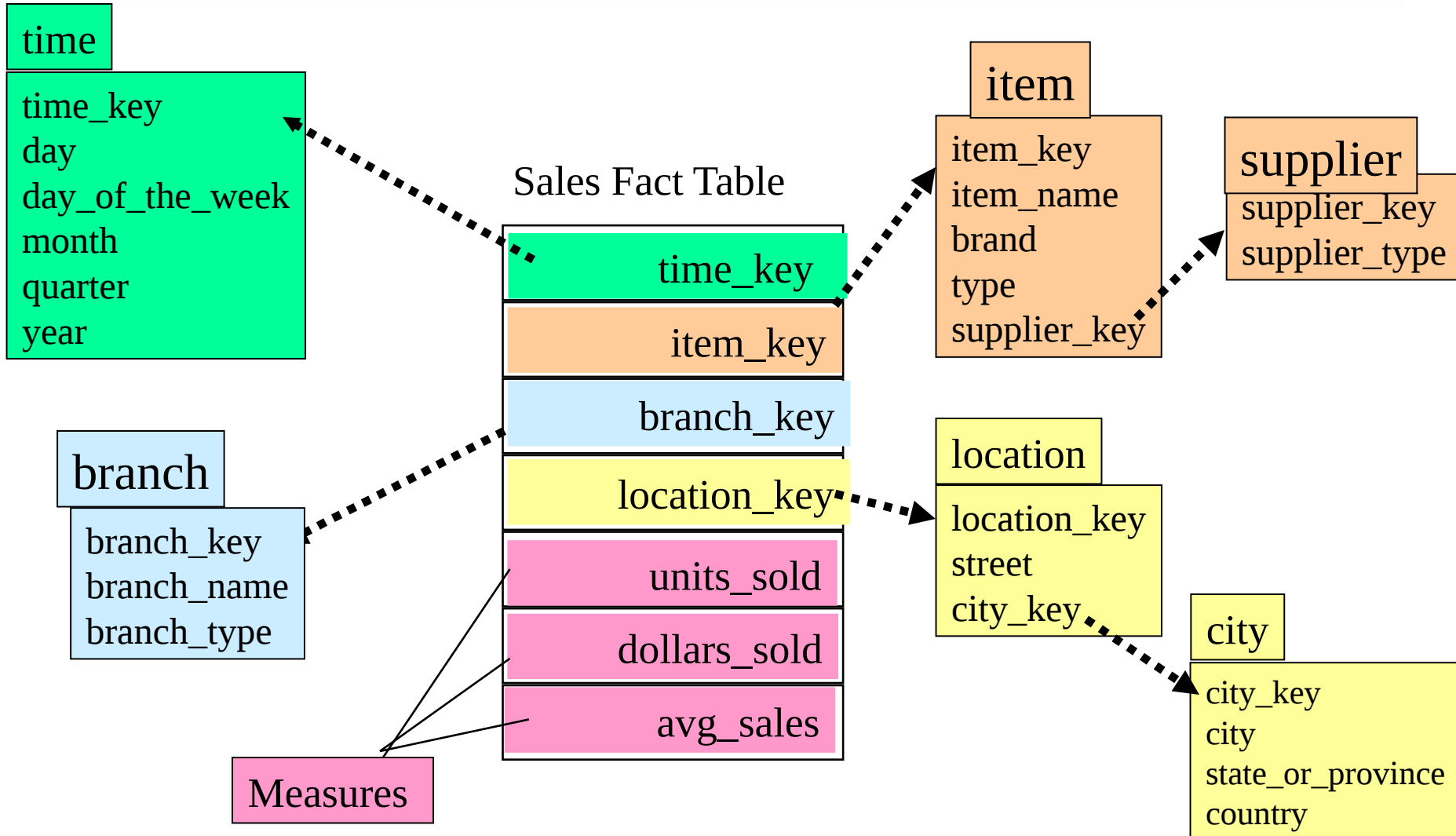
# Conceptual Modeling of Data Warehouses

- 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

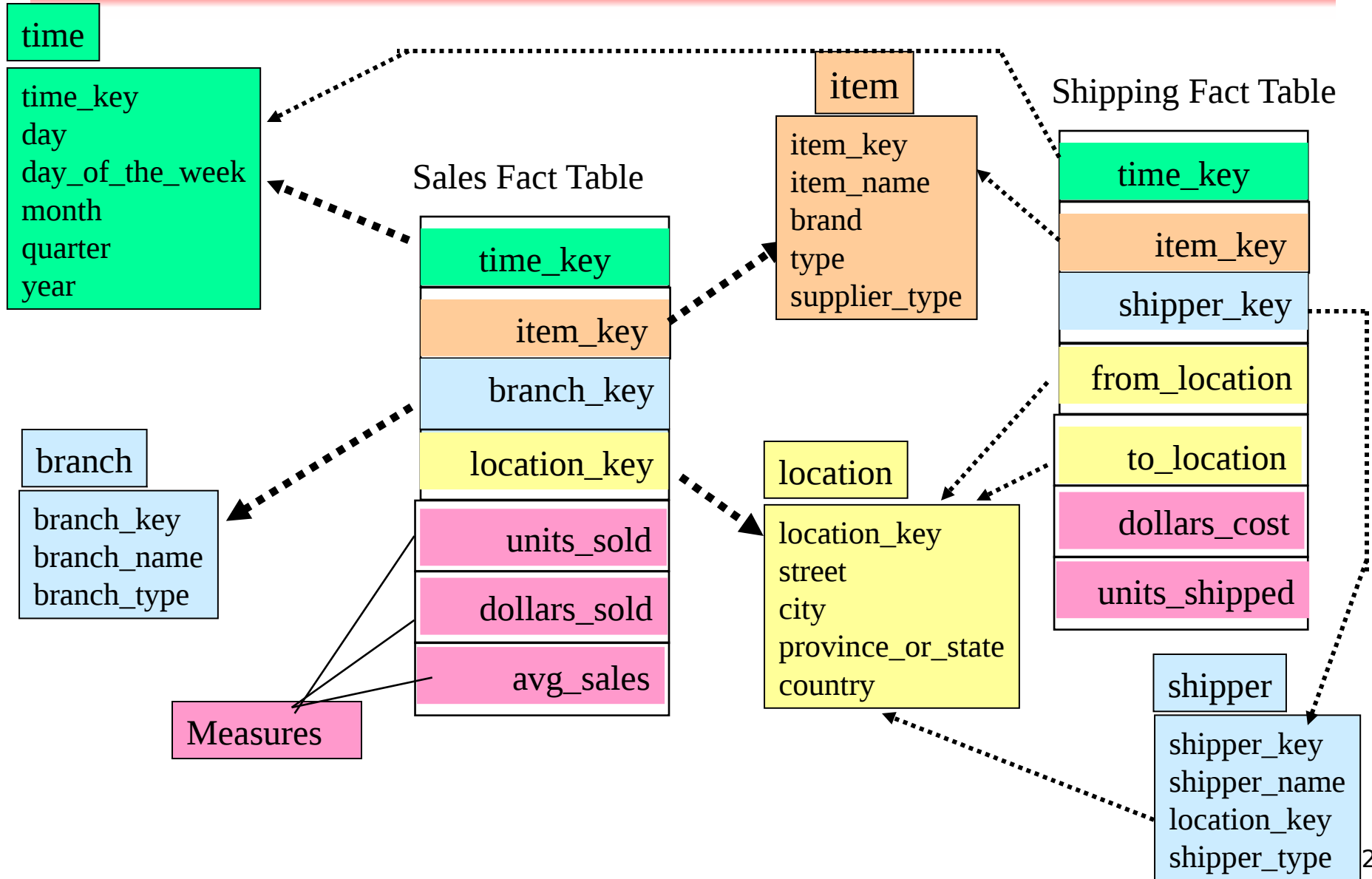
# Example of Star Schema



# Example of Snowflake Schema

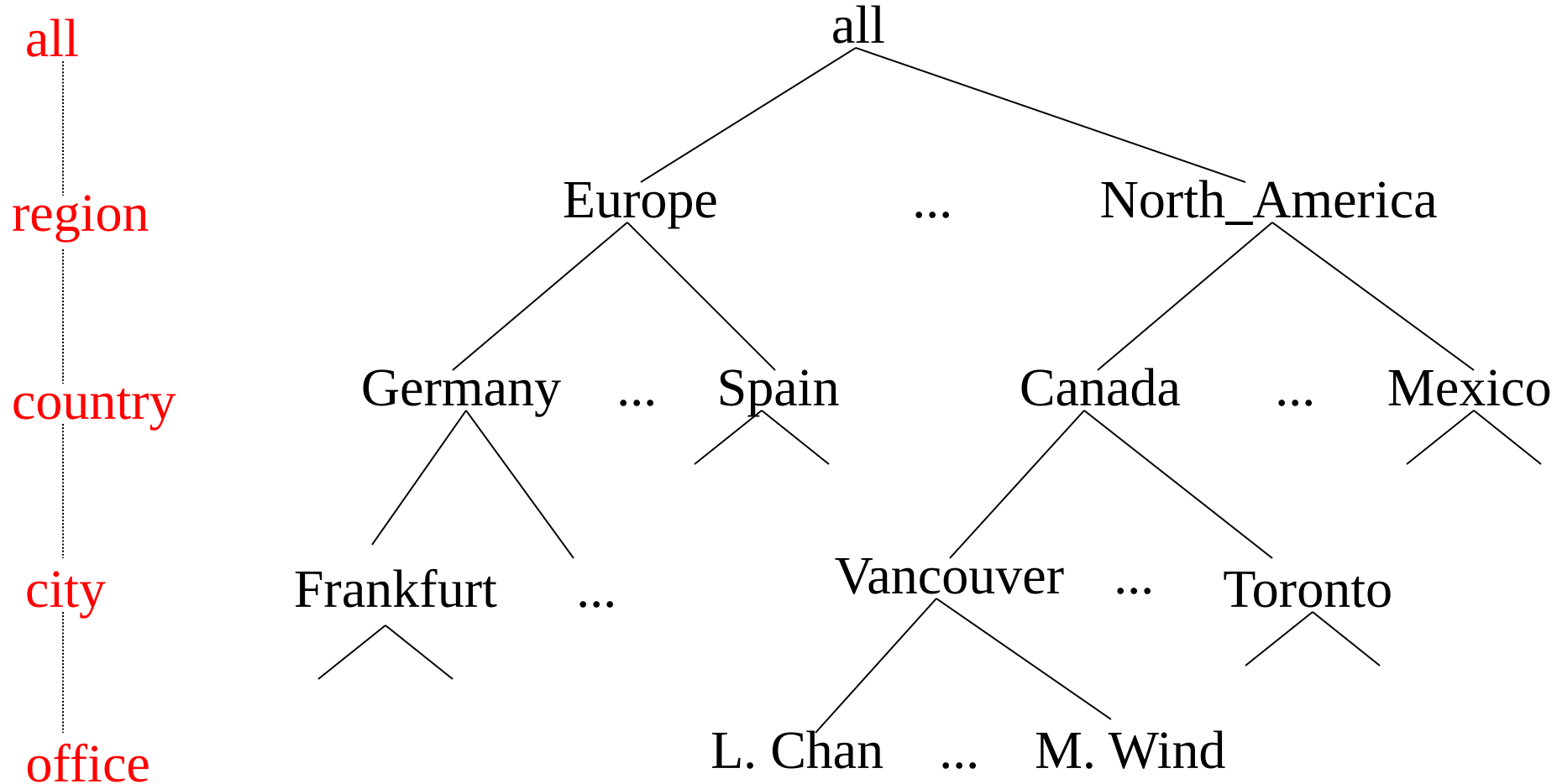


# Example of Fact Constellation



# A Concept Hierarchy: **Dimension** (location)

---



# Data cube measures

---

- Measure: a numeric function that can be evaluated at each point in the data cube space:
  - Fact
  - Aggregation of facts



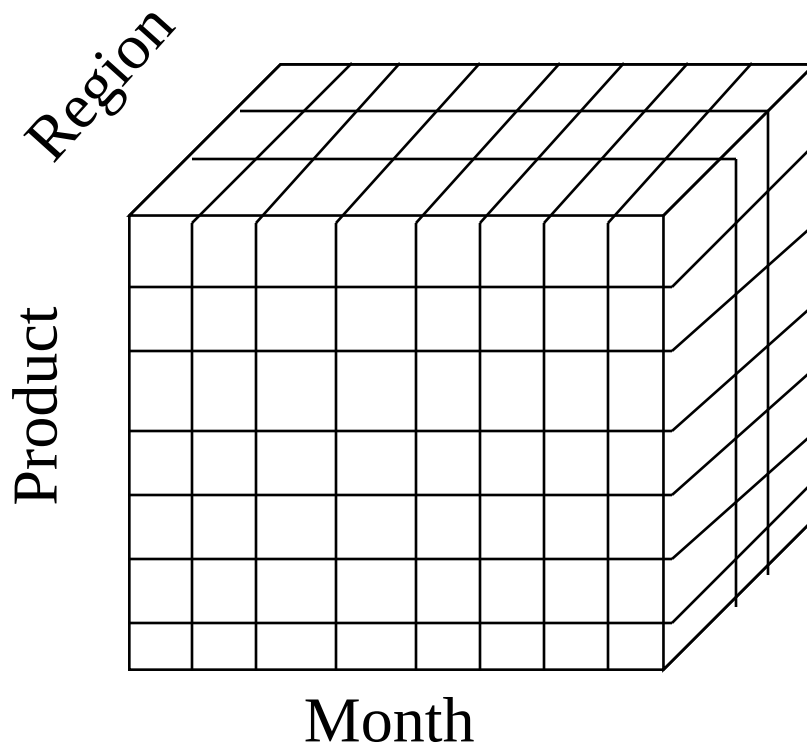
# Data Cube Measures: Three Categories

---

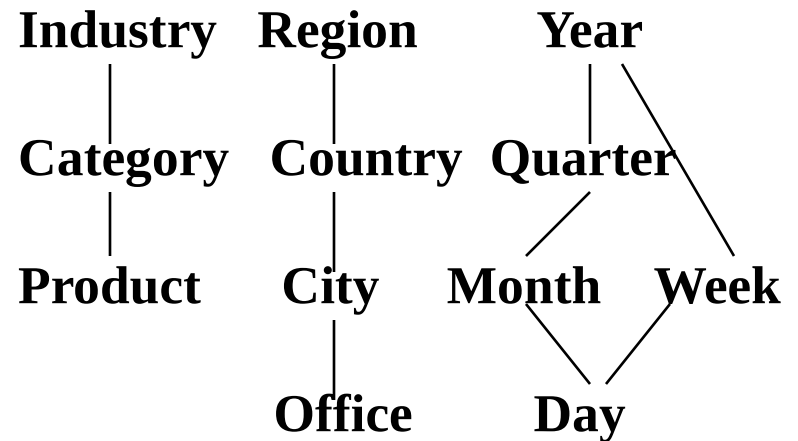
- **Distributive**: if the result derived by applying the function to  $n$  aggregate values is the same as that derived by applying the function on all the data without partitioning
  - E.g., `count()`, `sum()`, `min()`, `max()`
- **Algebraic**: if it can be computed by an algebraic function with  $M$  arguments (where  $M$  is a bounded integer), each of which is obtained by applying a distributive aggregate function
  - E.g., `avg() = sum() / count()`, `min_N()` ...
- **Holistic**: if there is no constant bound on the storage size needed to describe a subaggregate.
  - E.g., `median()`, `mode()`, `rank()`

# Multidimensional Data

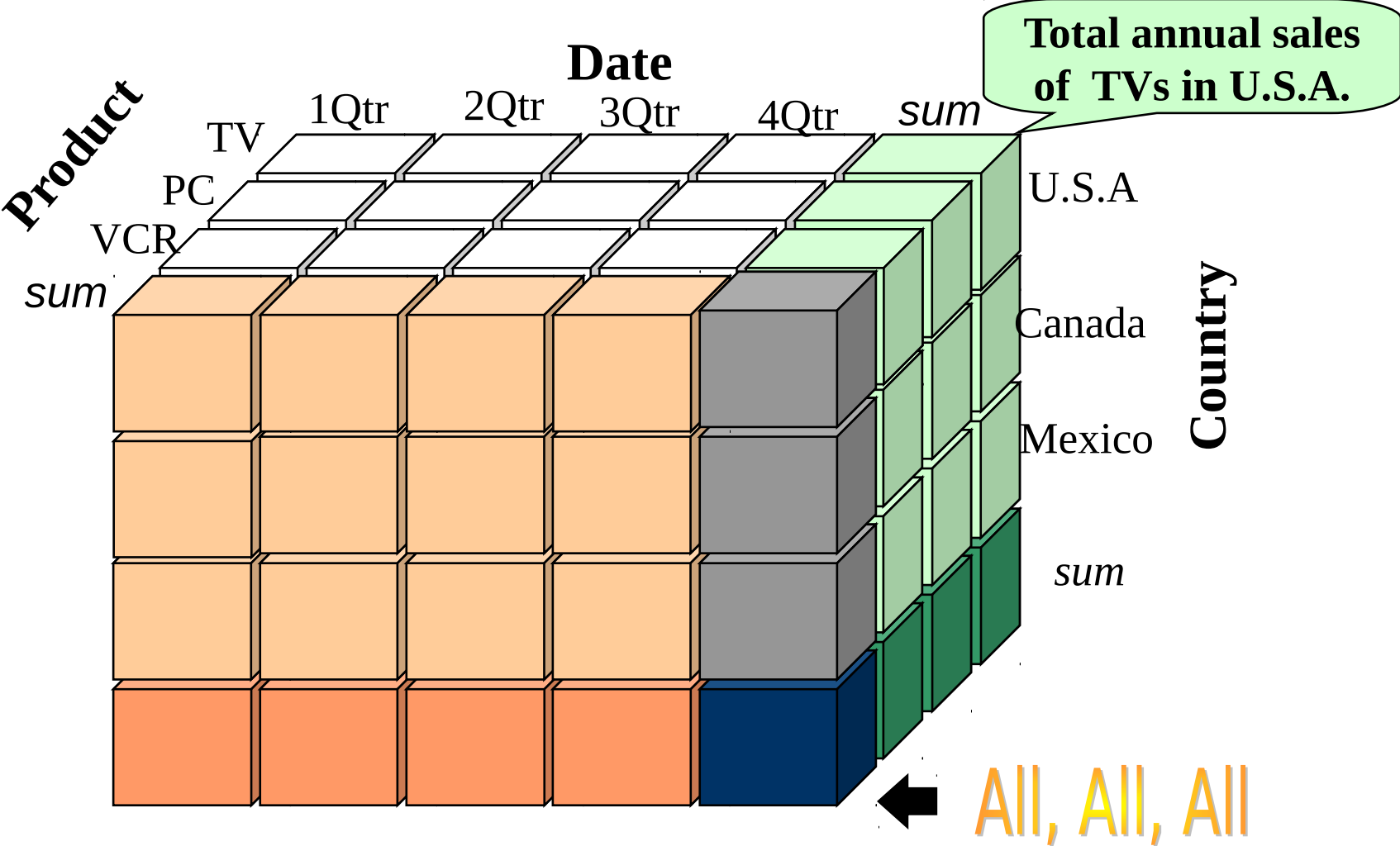
- Sales volume as a function of product, month, and region



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

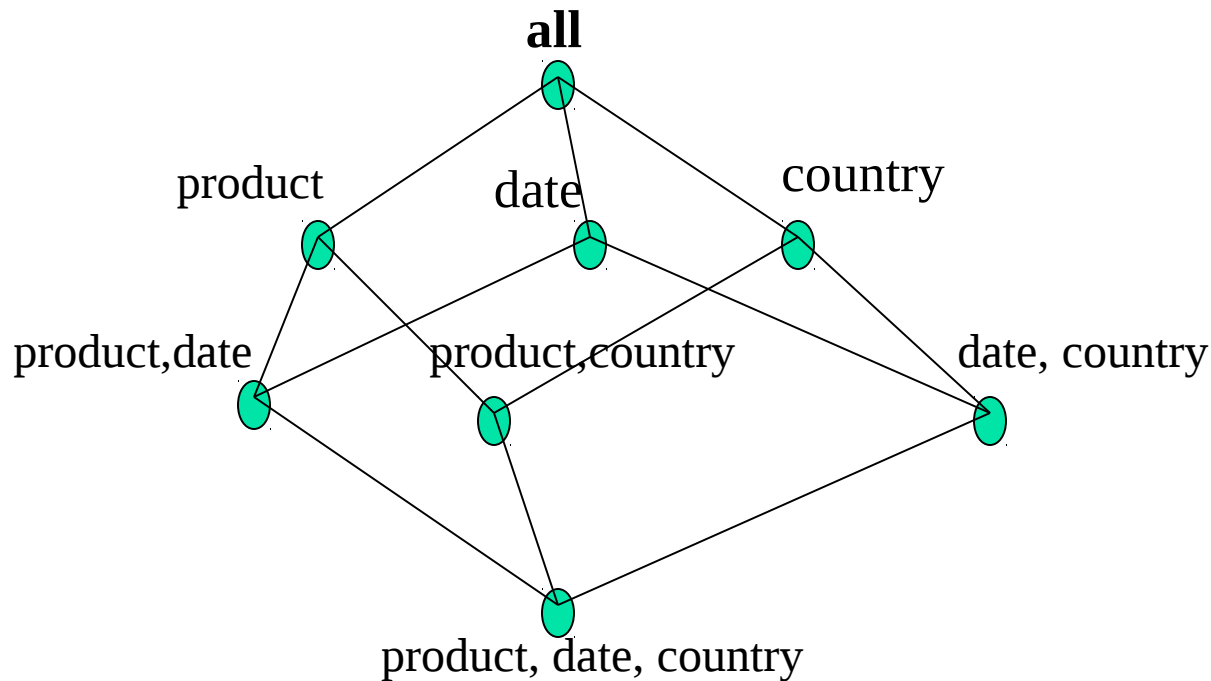


# A Sample Data Cube



# Cuboids Corresponding to the Cube

---



0-D (*apex*) cuboid

1-D cuboids

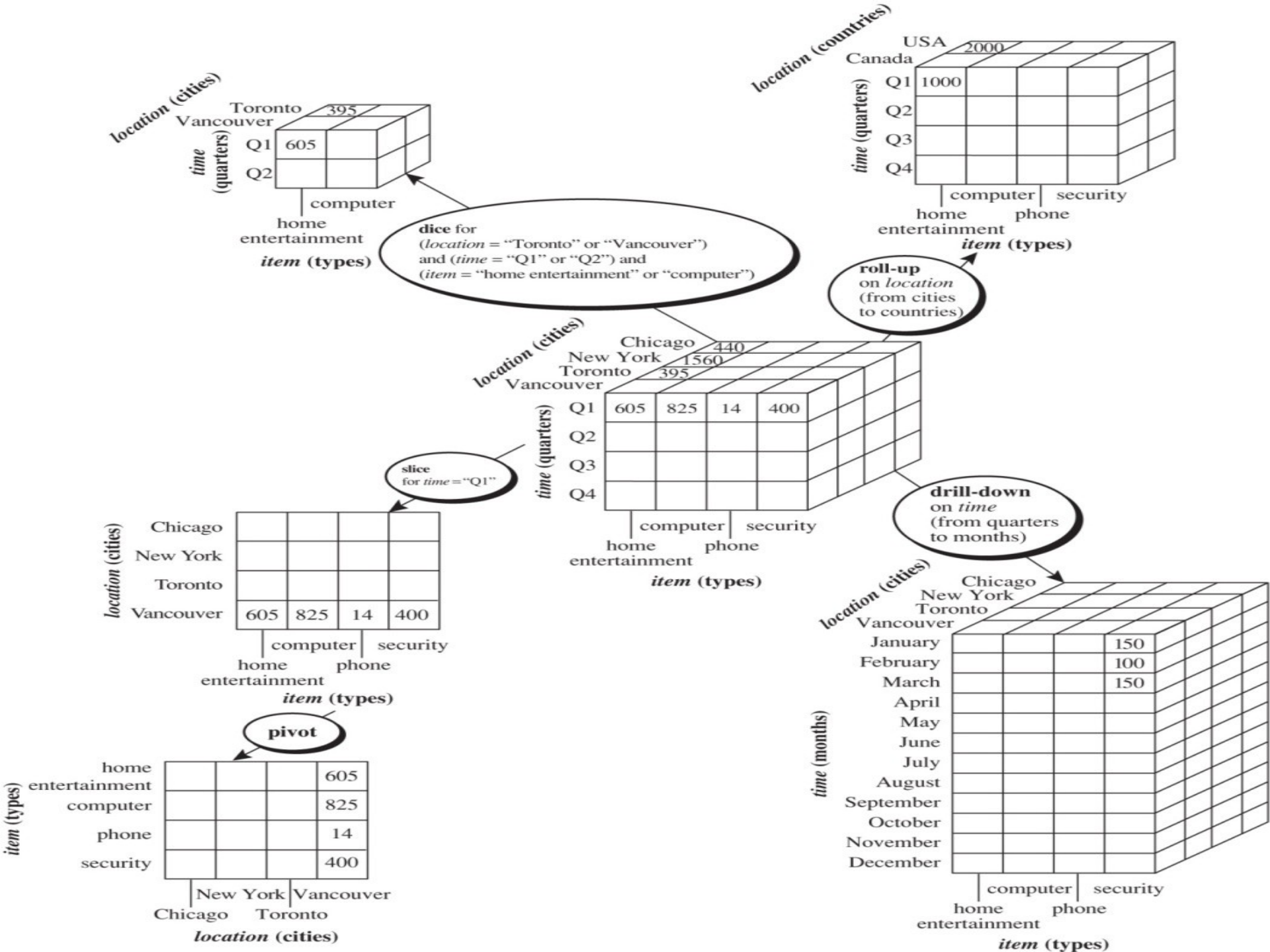
2-D cuboids

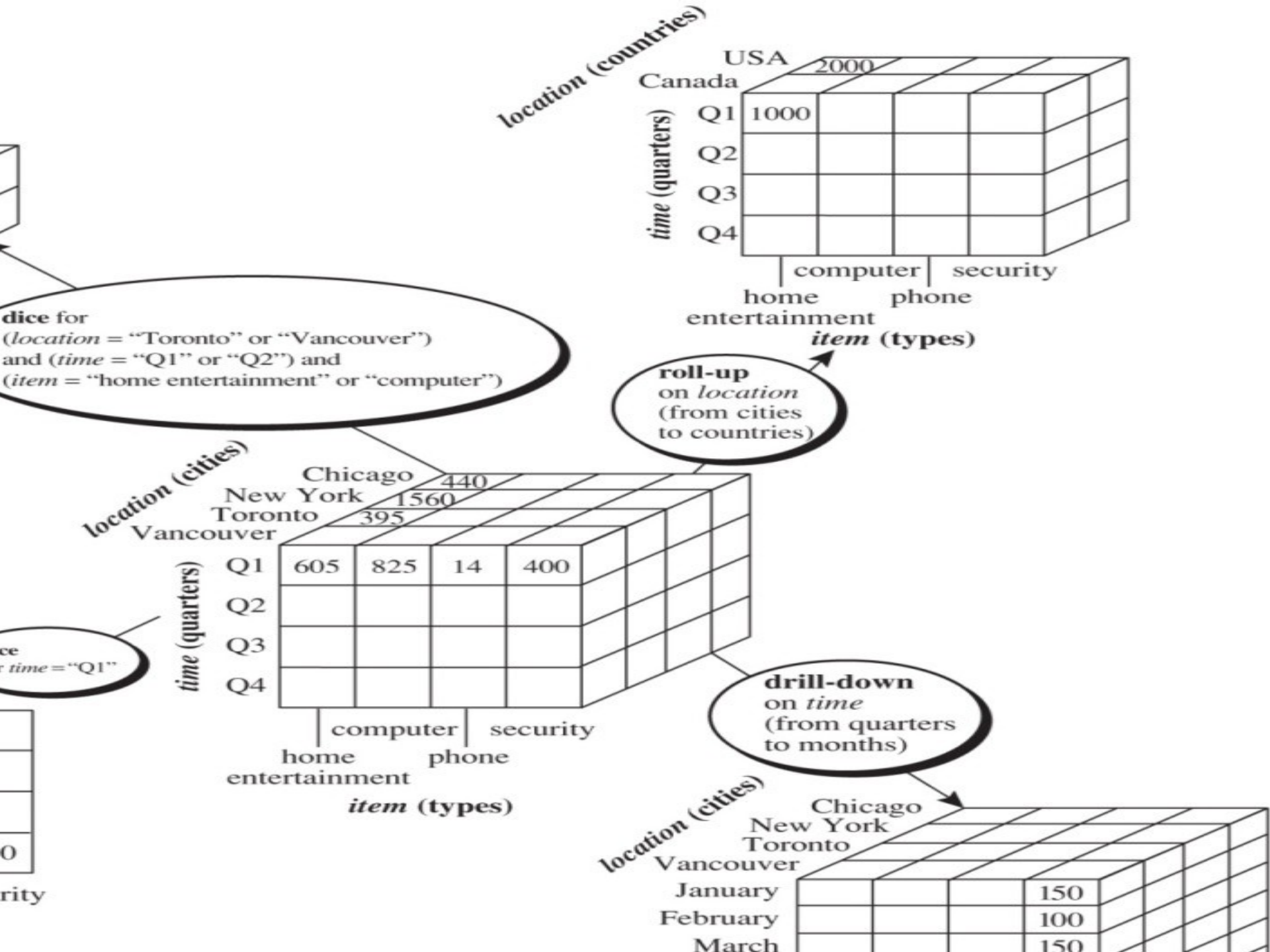
3-D (*base*) cuboid

# 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*



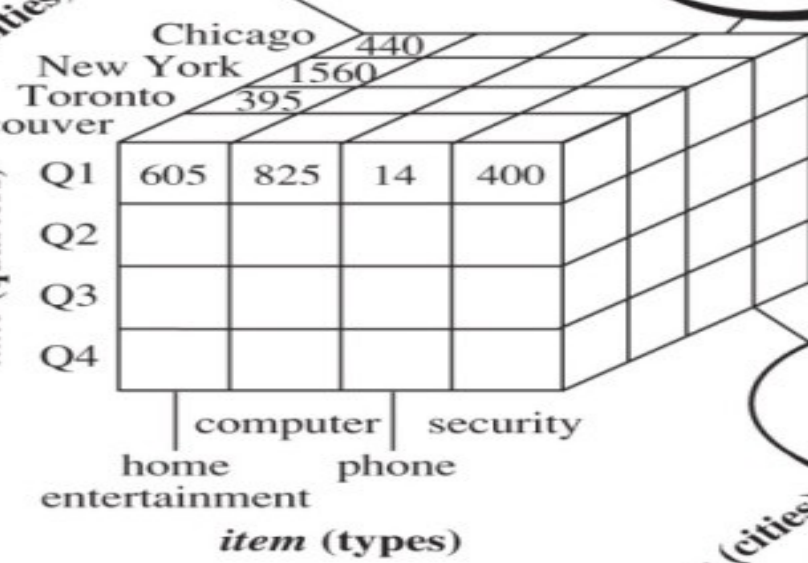


(item = "home entertainment" or "computer")

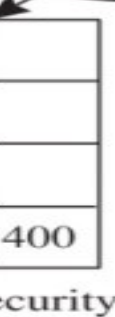
drill-up  
on location  
(from cities  
to countries)

location (cities)

time (quarters)



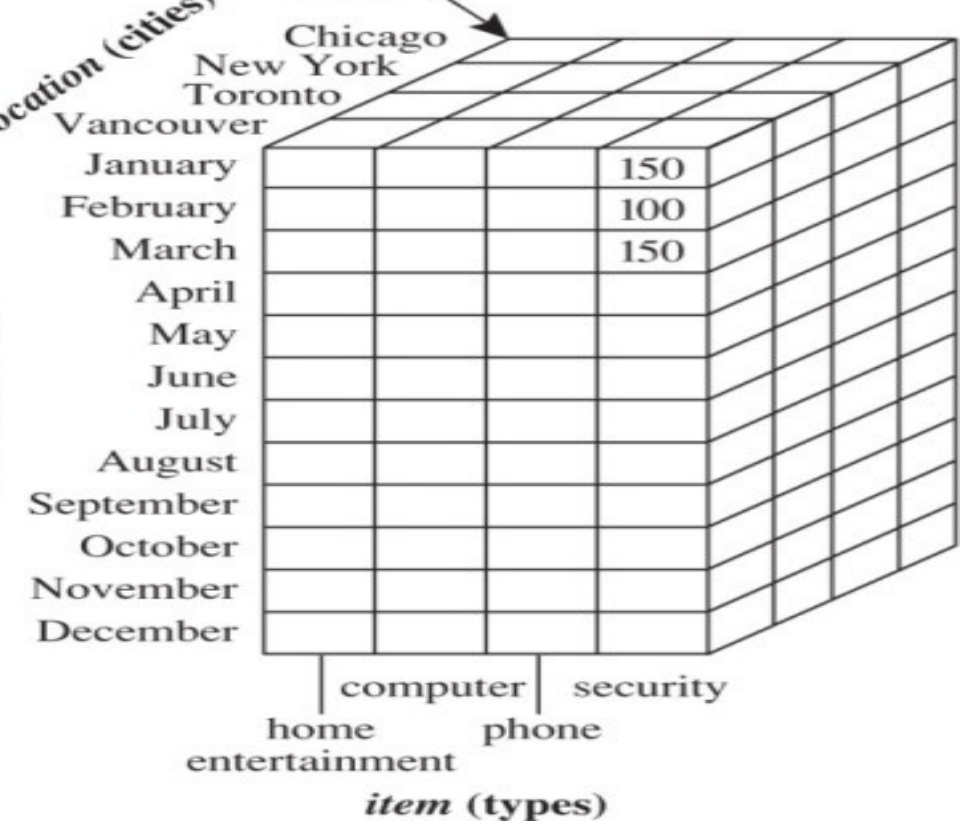
slice  
for time = "Q1"



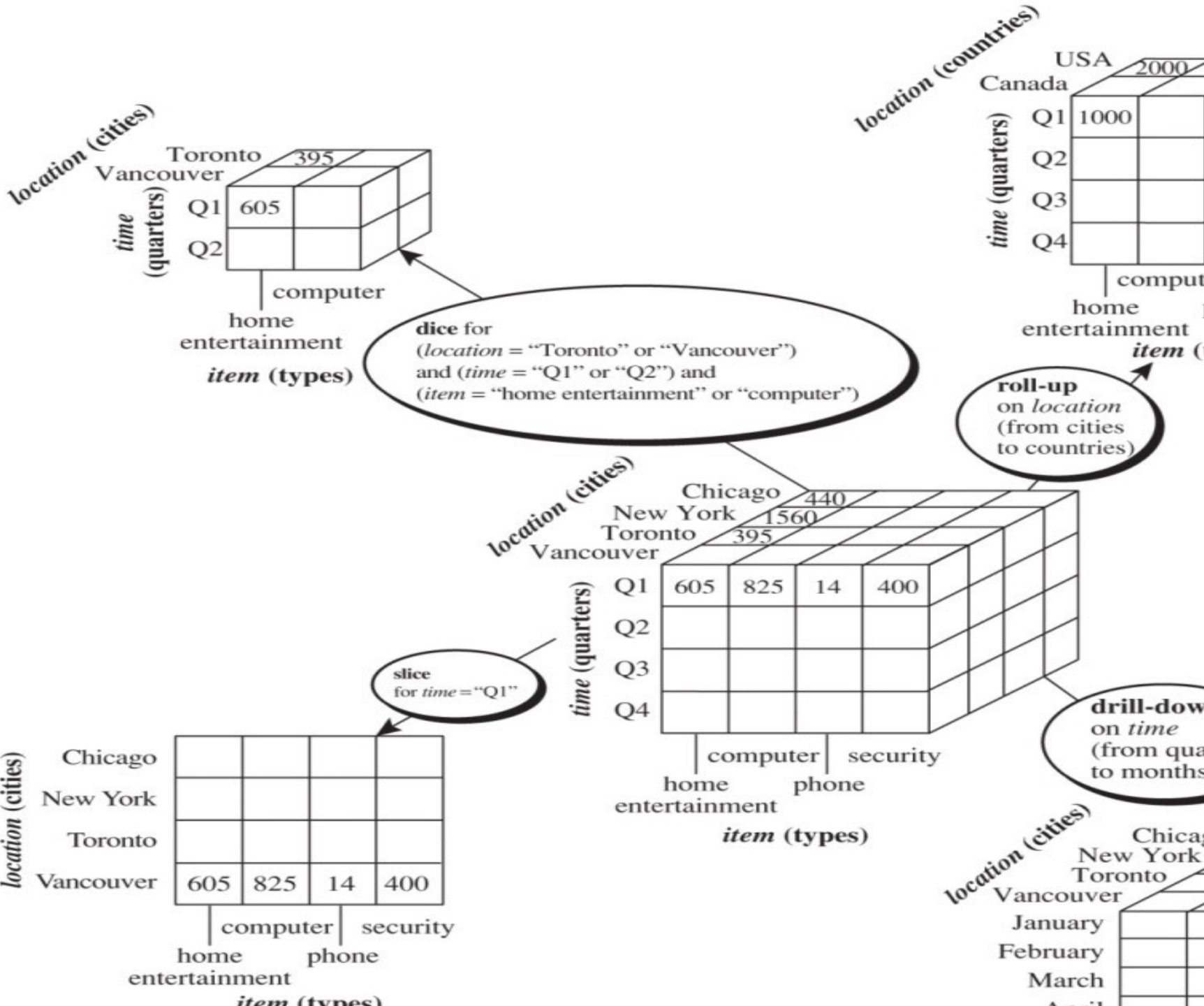
drill-down  
on time  
(from quarters  
to months)

location (cities)

time (months)





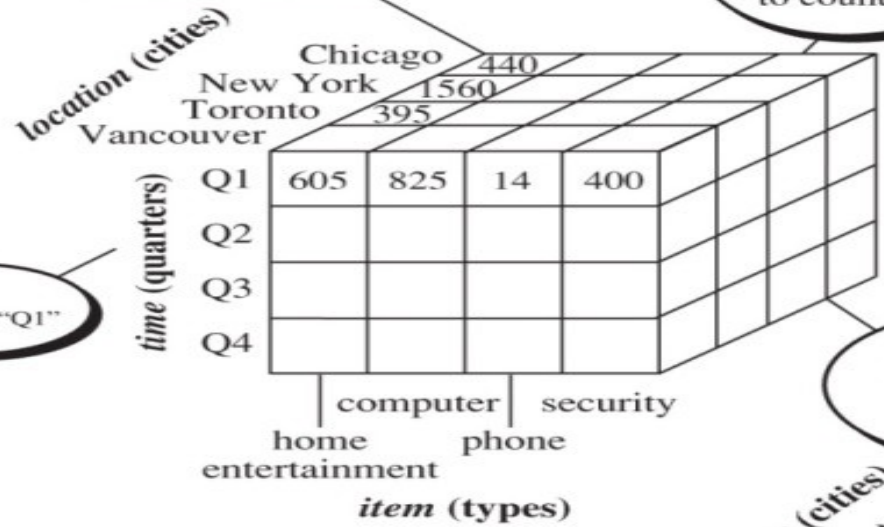


computer  
home entertainment  
**item (types)**

**dice** for  
(location = "Toronto" or "Vancouver")  
and (time = "Q1" or "Q2") and  
(item = "home entertainment" or "computer")

home entertainment  
**item (types)**

**roll-up**  
on location  
(from cities to countries)



**drill-down**  
on time  
(from quarters to months)

location (cities)

Chicago				
New York				
Toronto				
Vancouver	605	825	14	400
	home entertainment	computer	phone	security

item (types)

**slice**  
for time = "Q1"

**pivot**

item (types)

home entertainment				605
computer				825
phone				14
security				400
	New York	Vancouver		
	Chicago	Toronto		

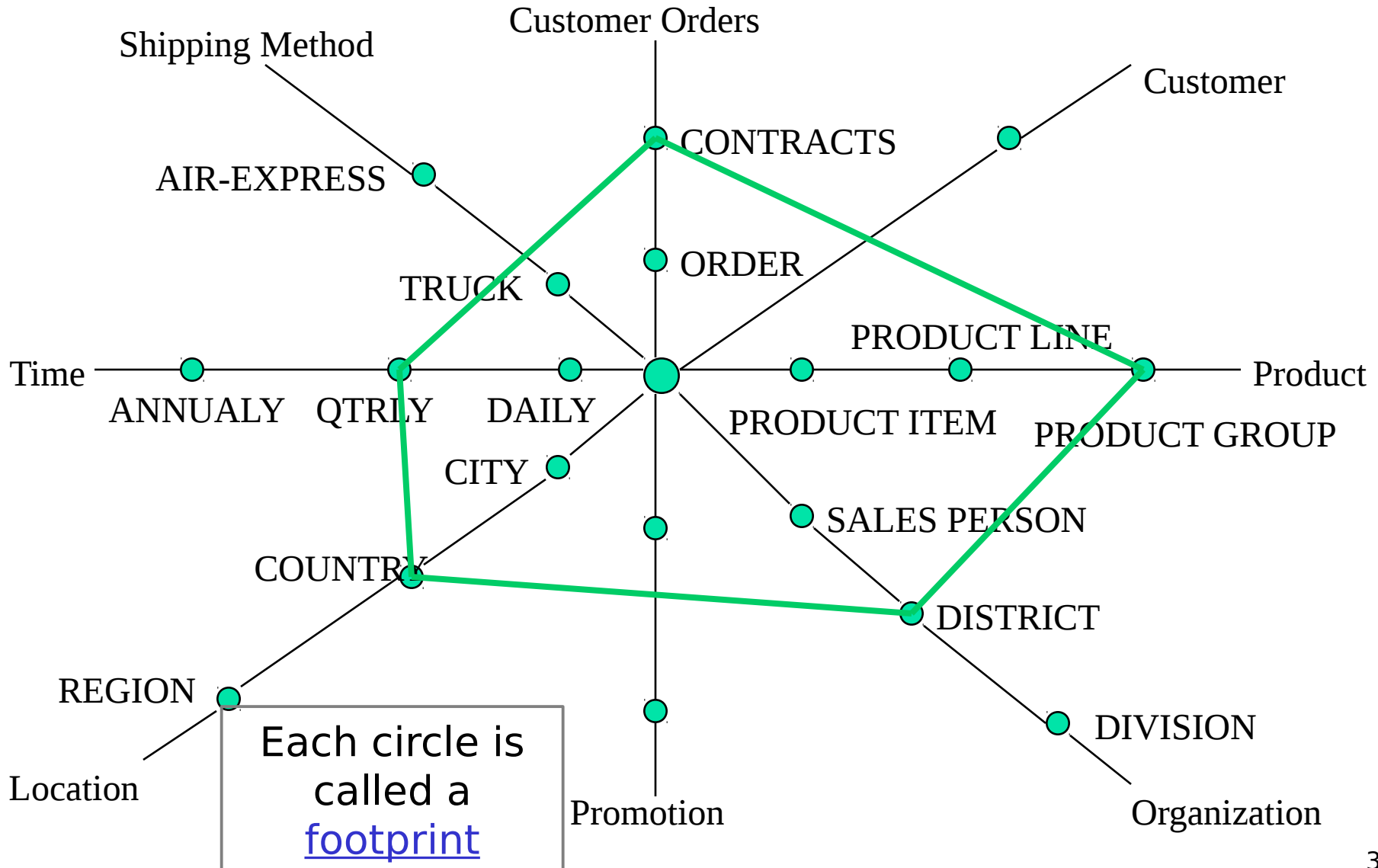
location (cities)

location (cities)

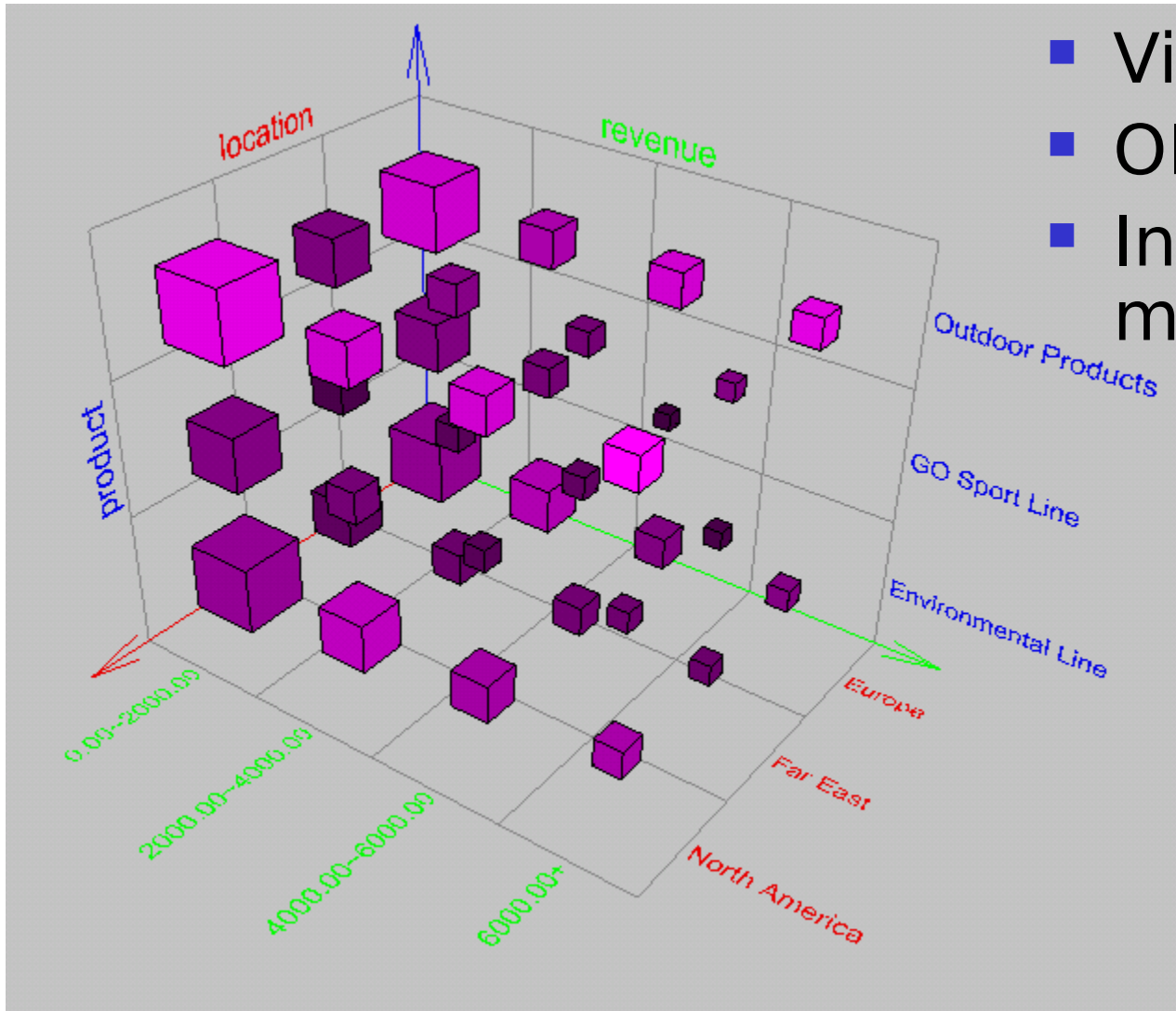
time (months)

Chicago				
New York				
Toronto				
Vancouver				
	January			
	February			
	March			
	April			
	May			
	June			
	July			
	August			
	September			
	October			
	November			
	December			
		home entertainment	computer	security

# A Star-Net Query Model




# Browsing a Data Cube



- Visualization
- OLAP capabilities
- Interactive manipulation

# Chapter 4: Data Warehousing and On-line Analytical Processing

---

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage 
- *Data Warehouse Implementation*
- *Data Generalization by Attribute-Oriented Induction*
- Summary

# Design of Data Warehouse: A Business Analysis Framework

---

- Four views regarding the design of a data warehouse
  - **Top-down view**
    - allows selection of the relevant information necessary for the data warehouse
  - **Data source view**
    - exposes the information being captured, stored, and managed by operational systems
  - **Data warehouse view**
    - consists of fact tables and dimension tables
  - **Business query view**
    - sees the perspectives of data in the warehouse from the view of end-user

# Data Warehouse Design Process

- **Top-down, bottom-up approaches or a combination** of both
  - Top-down: Starts with overall design and planning (mature)
  - Bottom-up: Starts with experiments and prototypes (rapid)
- **From software engineering point of view**
  - (1) planning (2) requirements study (3) problem analysis (4) warehouse design (5) data integration and testing (6) deployment
  - Waterfall: structured and systematic analysis at each step before proceeding to the next (better for data warehouse)
  - Spiral: rapid generation of increasingly functional systems, short turn around time, quick turn around (better for data marts)

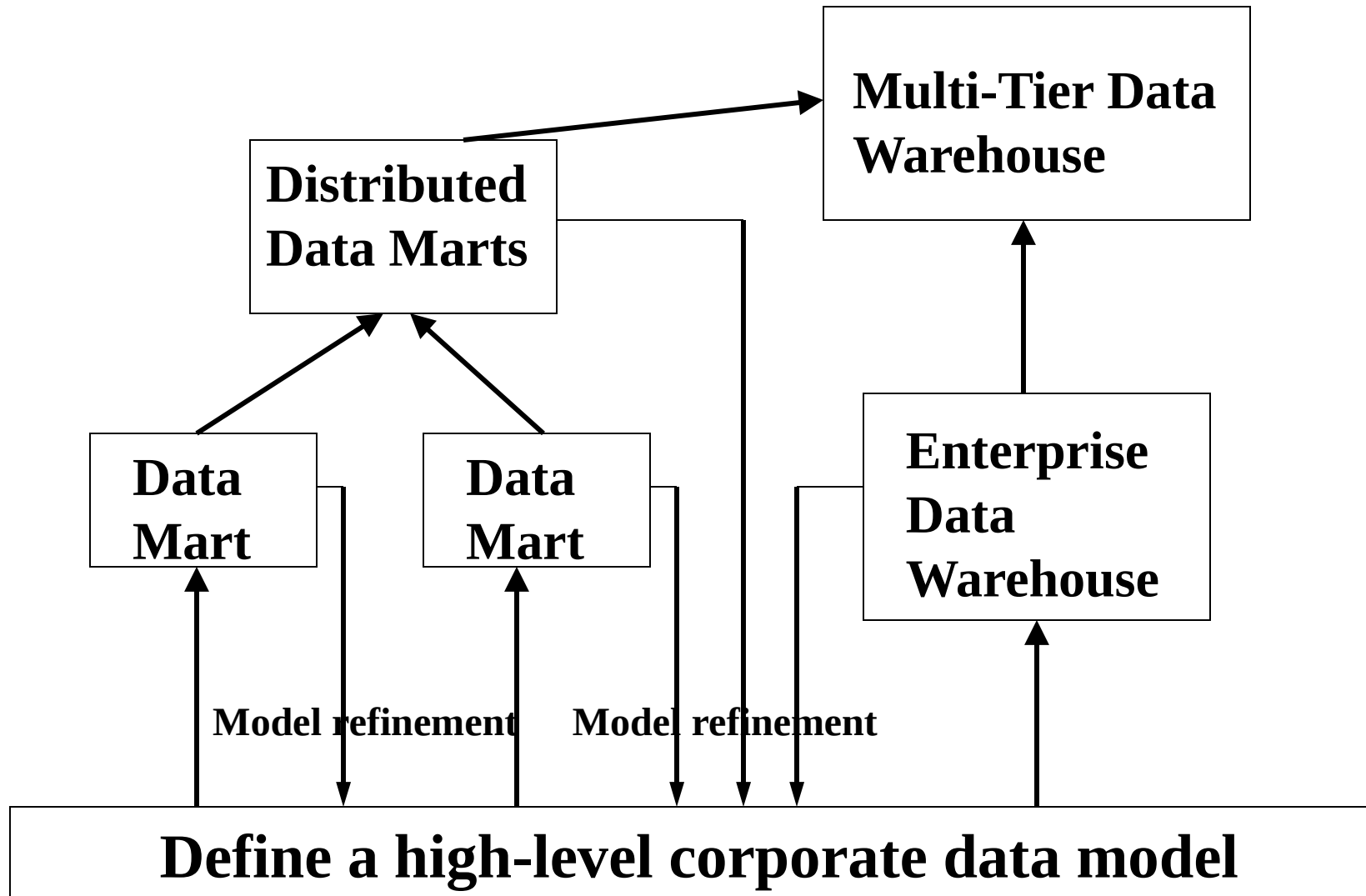
# Data Warehouse Design Process

---

- **Typical data warehouse design process**
  - Choose a **business process** to model, e.g., orders, invoices, etc.
  - Choose the ***grain (atomic level of data)*** of the business process
  - Choose the **dimensions** that will apply to each fact table record
  - Choose the **measure** that will populate each fact table record



# Data Warehouse Development: A Recommended Approach



# Data Warehouse Usage

---

- Three kinds of data warehouse applications
  - **Information processing**
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - **Analytical processing**
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting
  - **Data mining**
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

# From On-Line Analytical Processing (OLAP) to On Line Analytical Mining (OLAM)

- Why **online analytical mining**?
  - High quality of data in data warehouses
    - DW contains integrated, consistent, cleaned data
  - Available information processing structure surrounding data warehouses
    - ODBC, OLEDB, Web accessing, service facilities, reporting and OLAP tools
  - OLAP-based exploratory data analysis
    - Mining with drilling, dicing, pivoting, etc.
  - On-line selection of data mining functions
    - Integration and swapping of multiple mining functions, algorithms, and tasks