THE LECTURE 10

DATA WAREHOUSES AND OLAP

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	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: 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
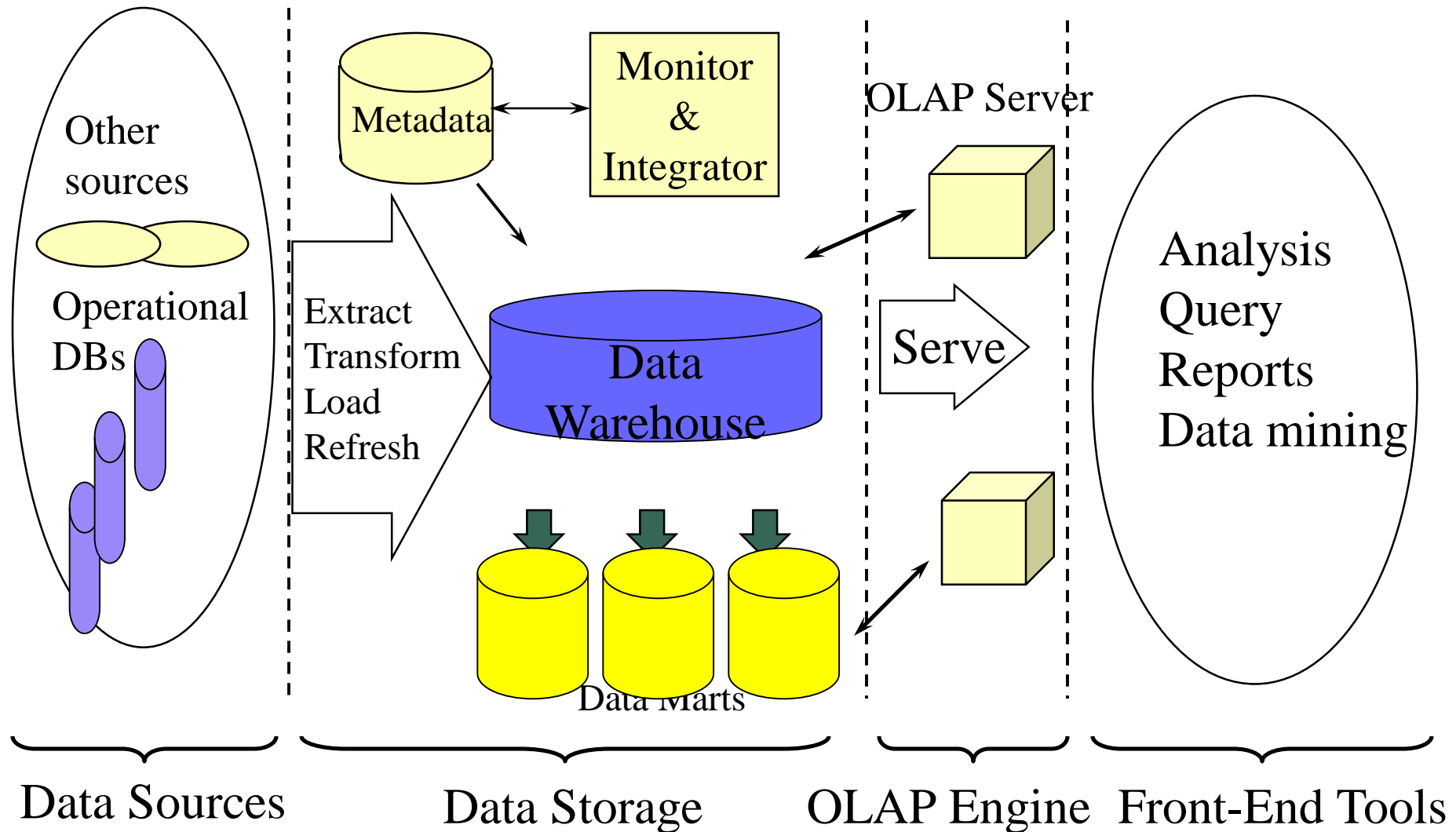
THREE DATA WAREHOUSE MODELS

- 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 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

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

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 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

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.
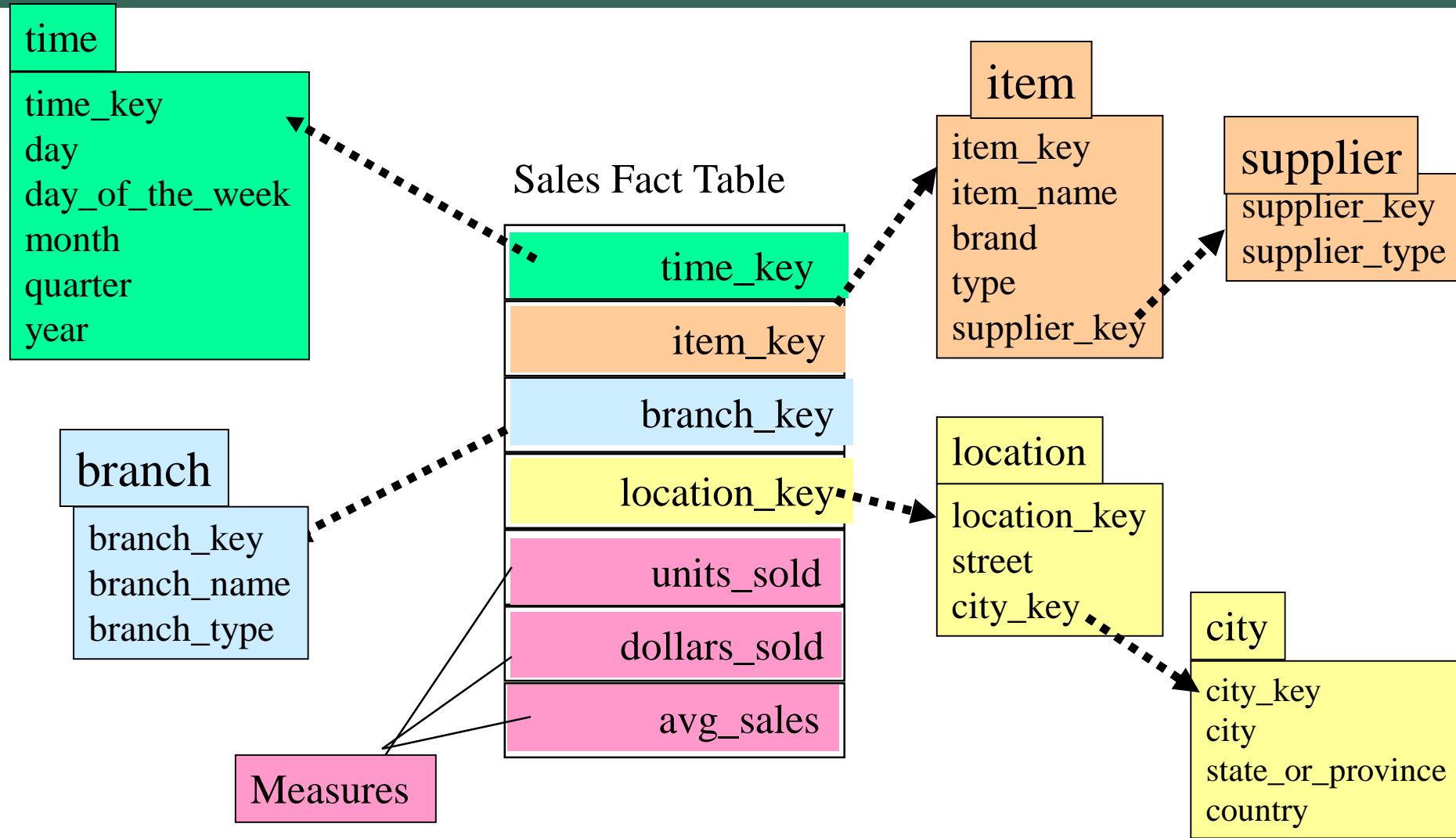
CUBE: A LATTICE OF CUBOIDS

all — 0-D (*apex*) cuboid

time   item   location   supplier — 1-D cuboids

time,location   item,location   location,supplier

time,item   time,supplier   item,supplier — 2-D cuboids

time,location,supplier

time,item,location   time,item,supplier   item,location,supplier — 3-D cuboids

time, item, location, supplier — 4-D (*base*) cuboid
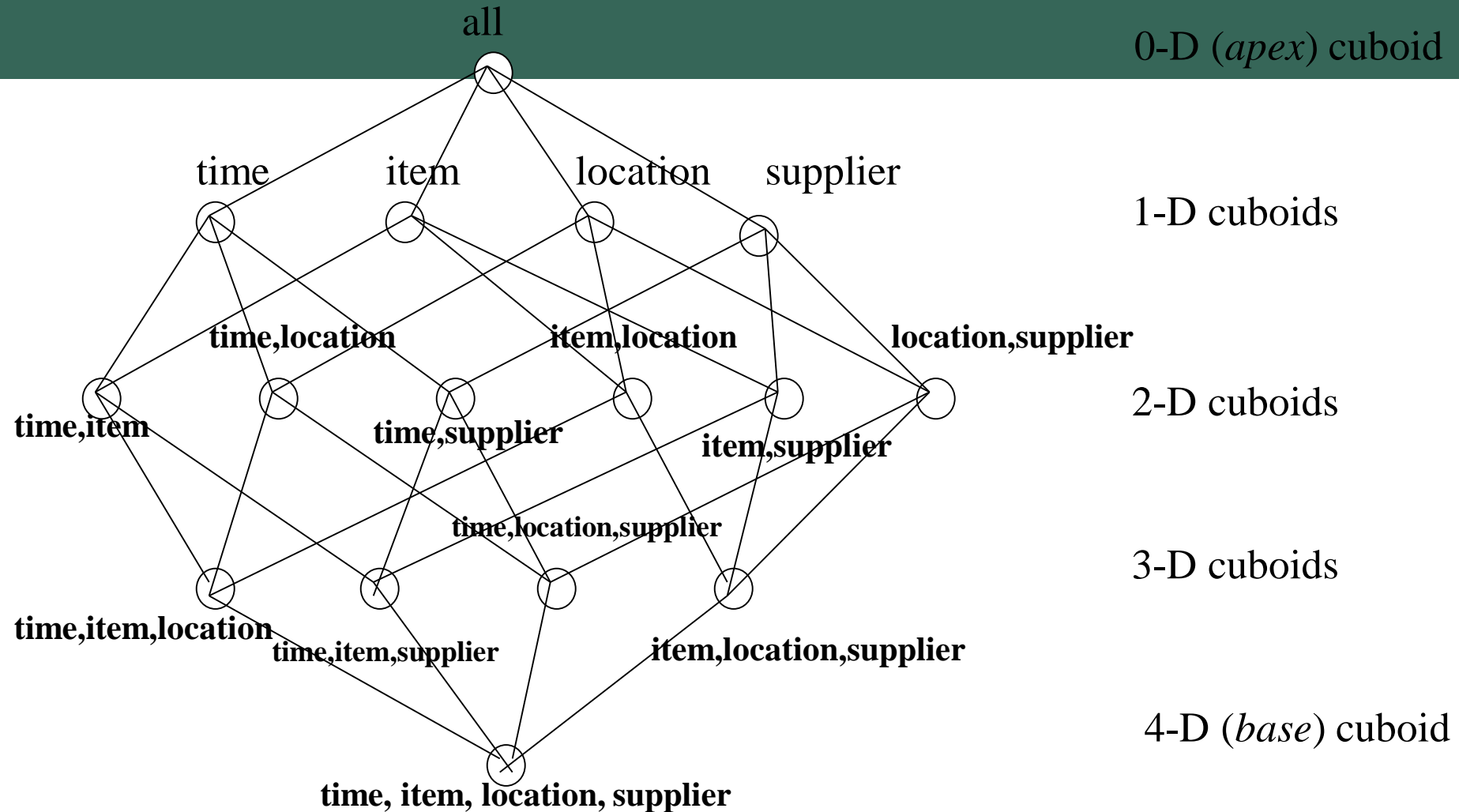
CONCEPTUAL MODELING OF DATA WAREHOUSES

- 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

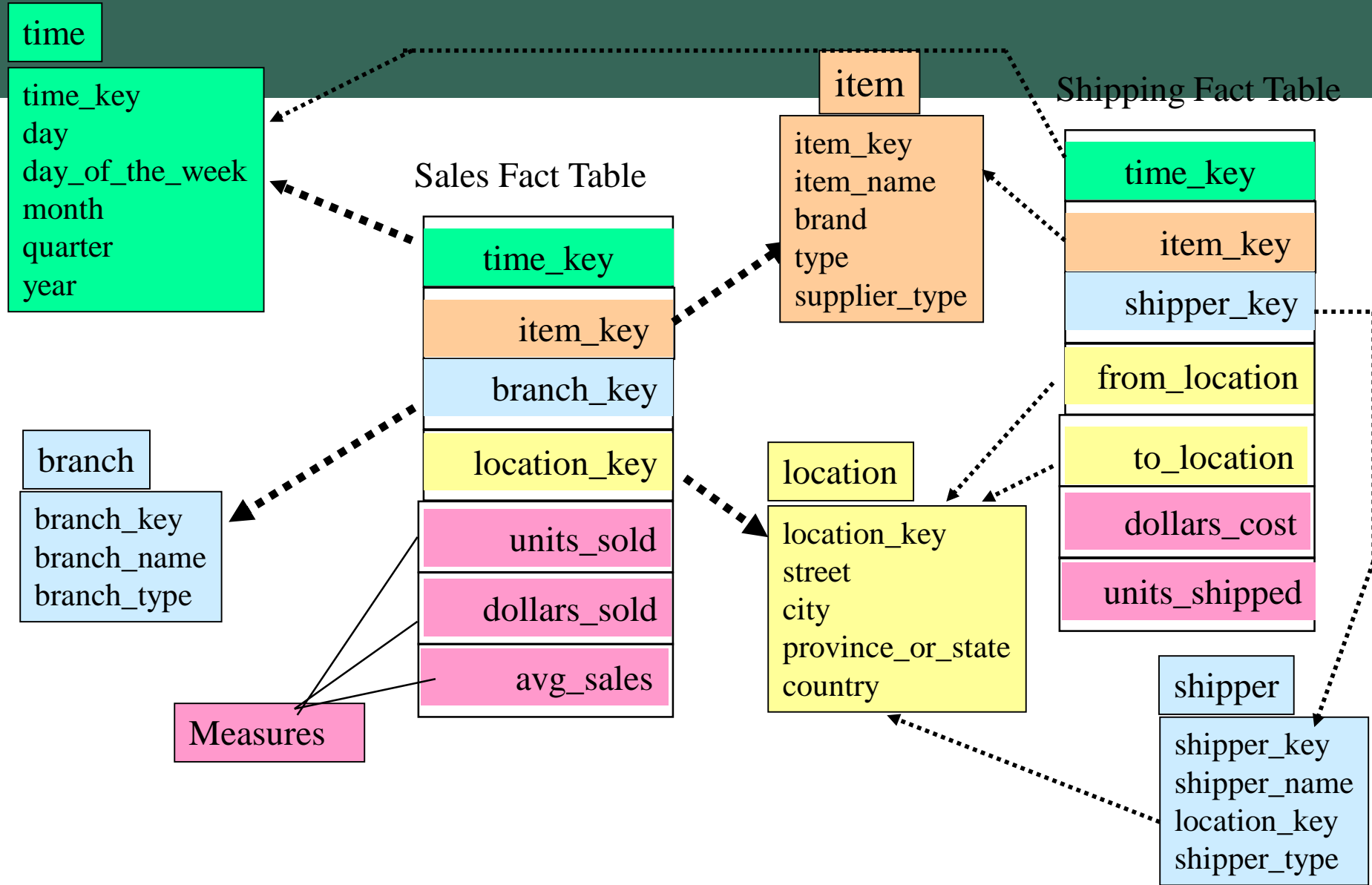 - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

EXAMPLE OF **STAR SCHEMA**

time

time_key
day
day_of_the_week
month
quarter
year

item

item_key
item_name
brand
type
supplier_type

branch

branch_key
branch_name
branch_type

location

location_key
street
city
state_or_province
country

Sales Fact Table

| time_key |
| item_key |
| branch_key |
| location_key |
| units_sold |
| dollars_sold |
| avg_sales |
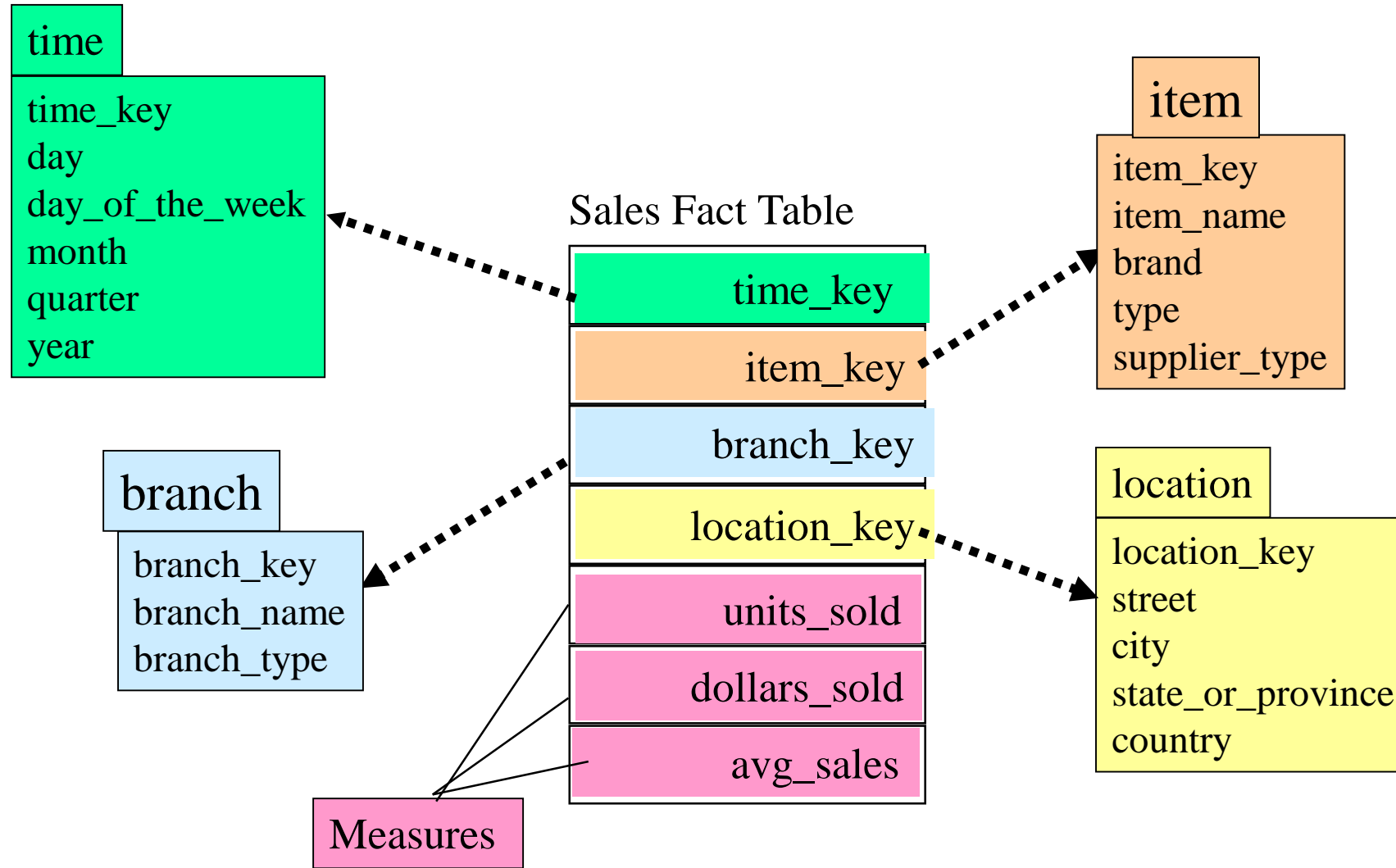
Measures

EXAMPLE OF **SNOWFLAKE SCHEMA**

EXAMPLE OF **FACT CONSTELLATION**

A CONCEPT HIERARCHY: **DIMENSION** (LOCATION)
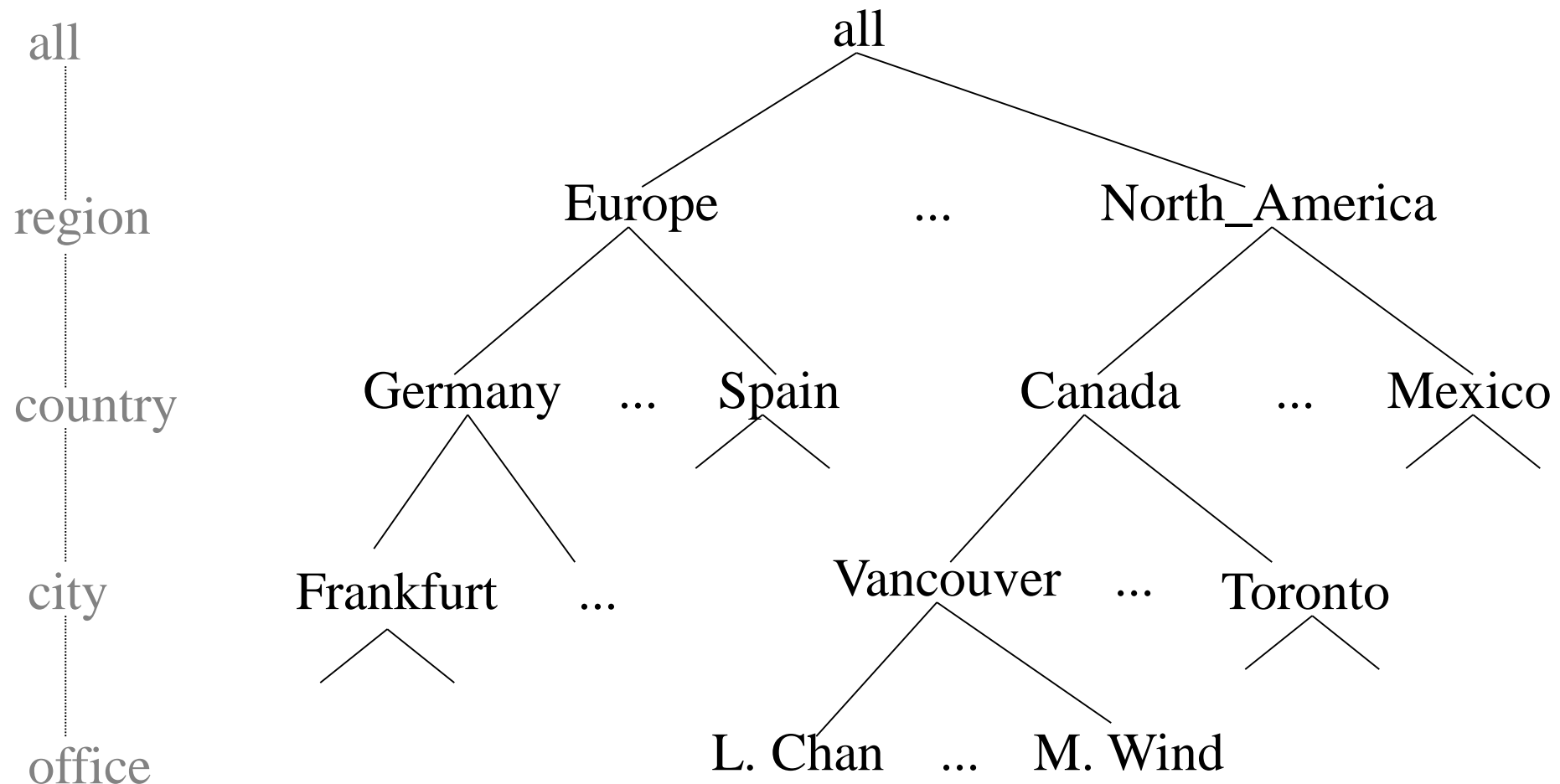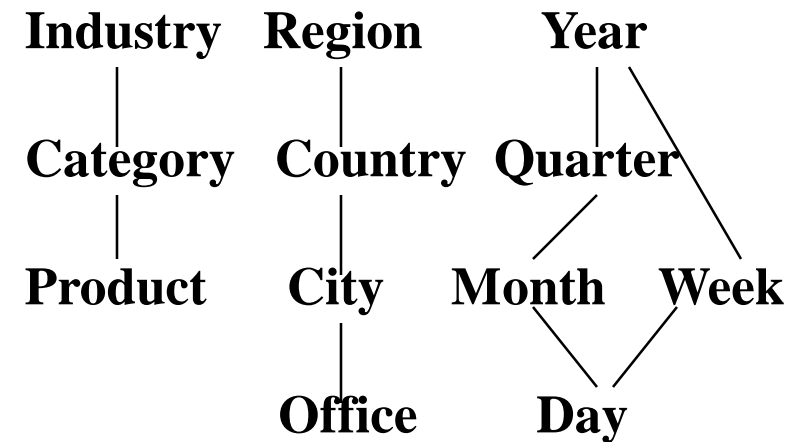
DATA CUBE MEASURES: THREE CATEGORIES

- 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()

- 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()

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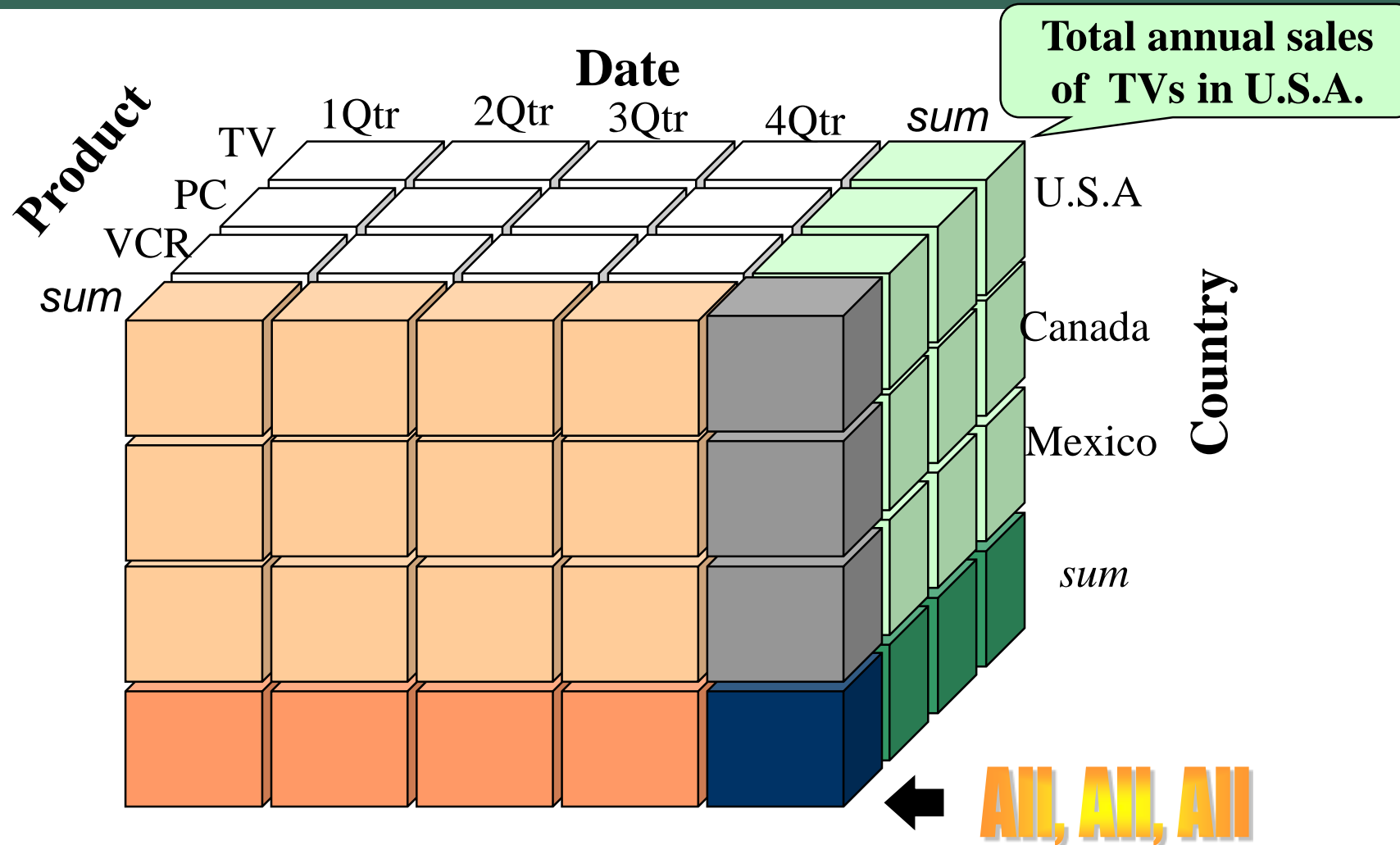 - E.g., median(), mode(), rank()

MULTIDIMENSIONAL DATA
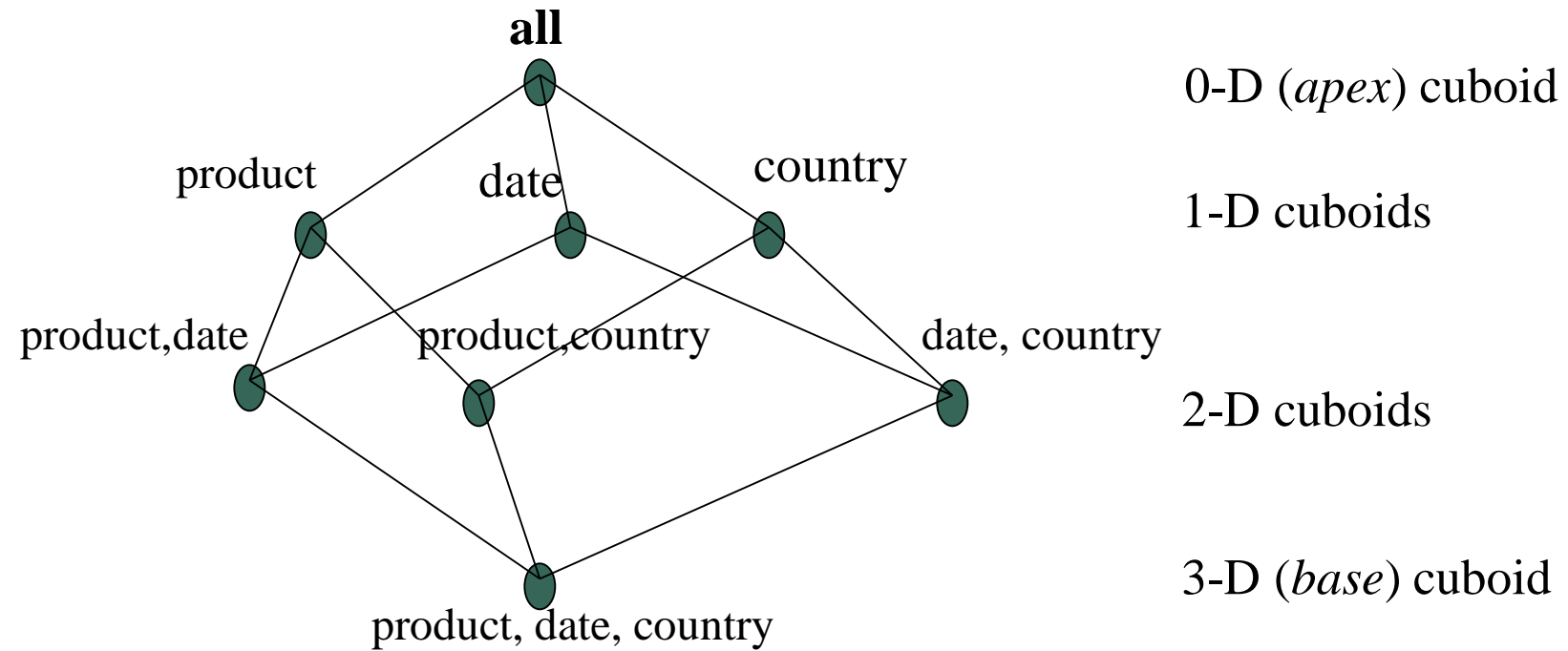
- Sales volume as a function of product, month, and region
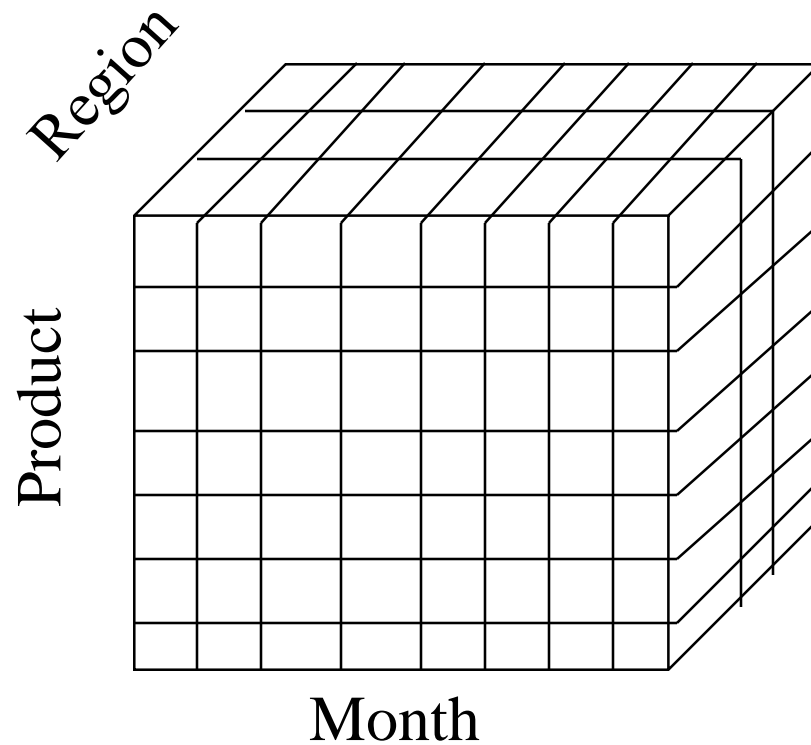
**Dimensions:** *Product, Location, Time*
Hierarchical summarization paths

Industry   Region   Year

Category   Country   Quarter

Product   City   Month   Week

Office   Day

A SAMPLE DATA CUBE

CUBOIDS CORRESPONDING TO THE CUBE

all

product   date   country

product,date   product,country   date, country

product, date, country

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 to its back-end relational tables (using SQL)*

BROWSING A DATA CUBE
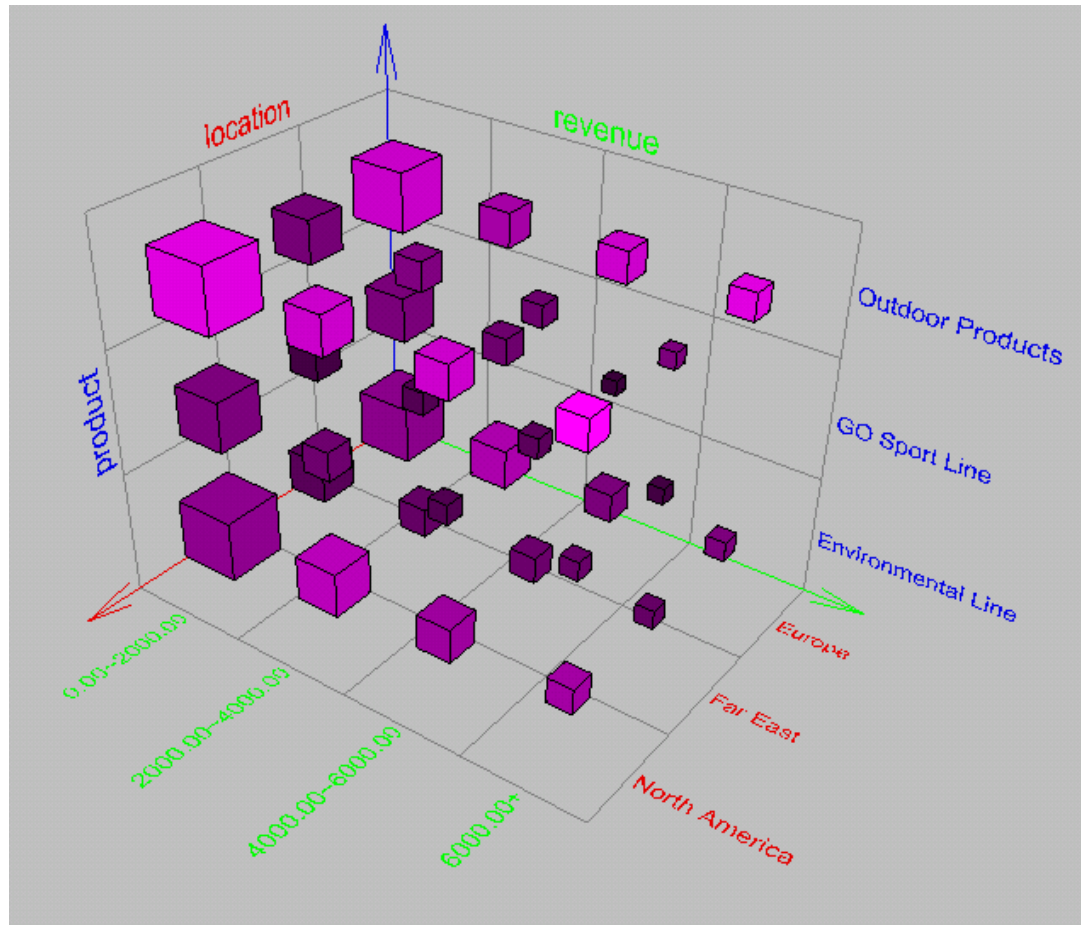
- Visualization
- OLAP capabilities
- Interactive manipulation