SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

Data Mining and Data Warehousing

SIT1301

UNIT 1: DATA WAREHOUSING

Data warehousing Components – Building a Data warehouse – Multi Dimensional Data Model – OLAP operations in Multi Dimensional Data model-Three Tier Data warehouse architecture- Schemas for multi dimensional data model-Online Analytical Processing (OLAP)- OLAP vs OLTP Integrated OLAM and OLAP Architecture

DATA WAREHOUSE COMPONENTS

What is a Data Warehouse?

- A Data warehouse is an information system that contains historical and commutative data from single or multiple sources. Data Warehouse Concepts simplify the reporting and analysis process of organizations.
- A process of transforming data into information and making it available to users in a timely enough manner to make a difference
- Data warehousing provides architecture and tools for business executives to systematically organize, understand and user their data to make strategic decisions.
- A Data Warehouse is a group of data specific to the entire organization, not only to a particular group of users.
- It is not used for daily operations and transaction processing but used for making decisions.

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*

A Data Warehouse can be viewed as a data system with the following attributes:

- It is a database designed for investigative tasks, using data from various applications.
- It supports a relatively small number of clients with relatively long interactions.
- It includes current and historical data to provide a historical perspective of information.
- Its usage is read-intensive.
- It contains a few large tables.

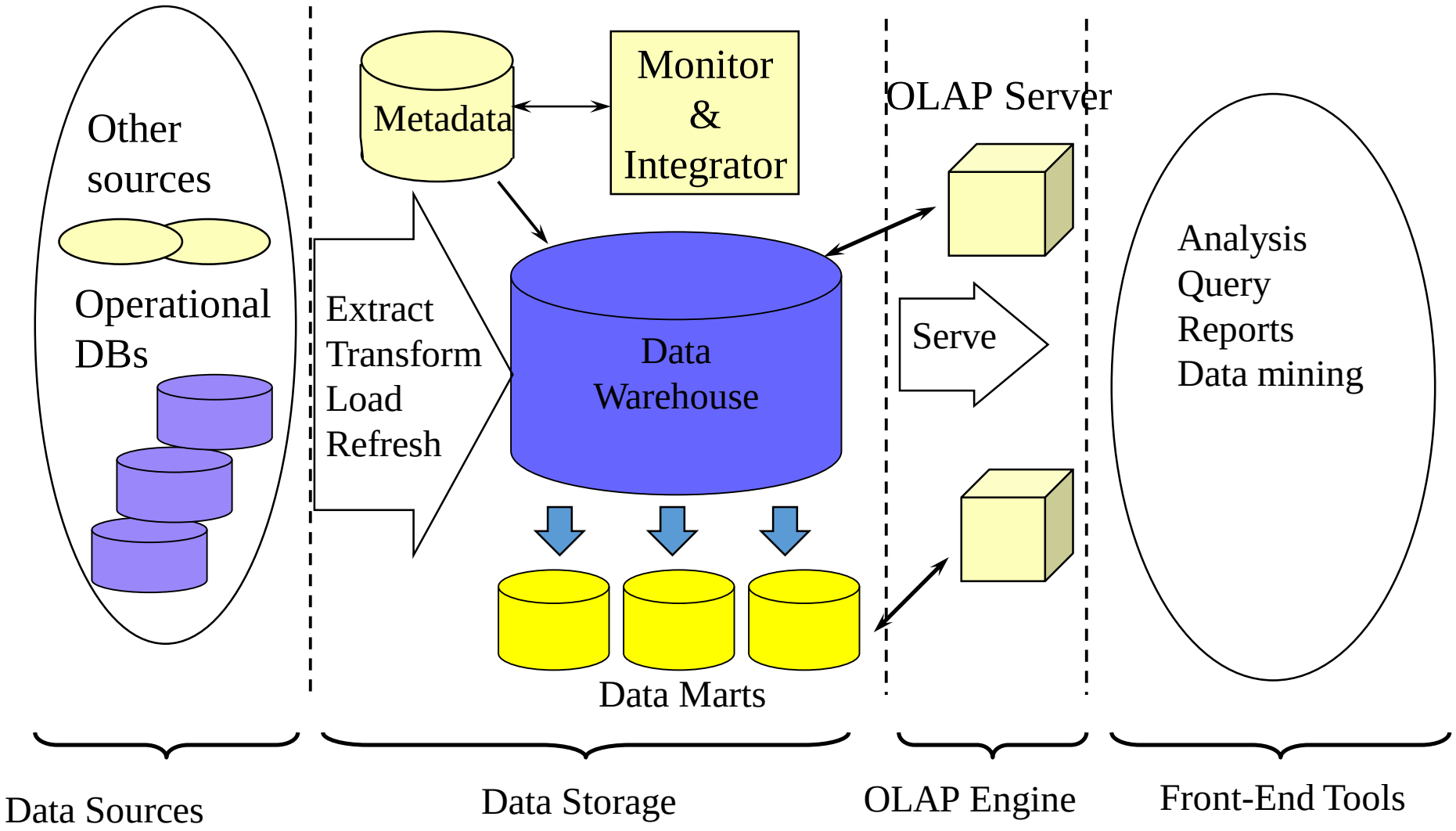
DATA WAREHOUSE COMPONENTS & ARCHITECTURE

Three different kinds of systems that are required for a data warehouse are:

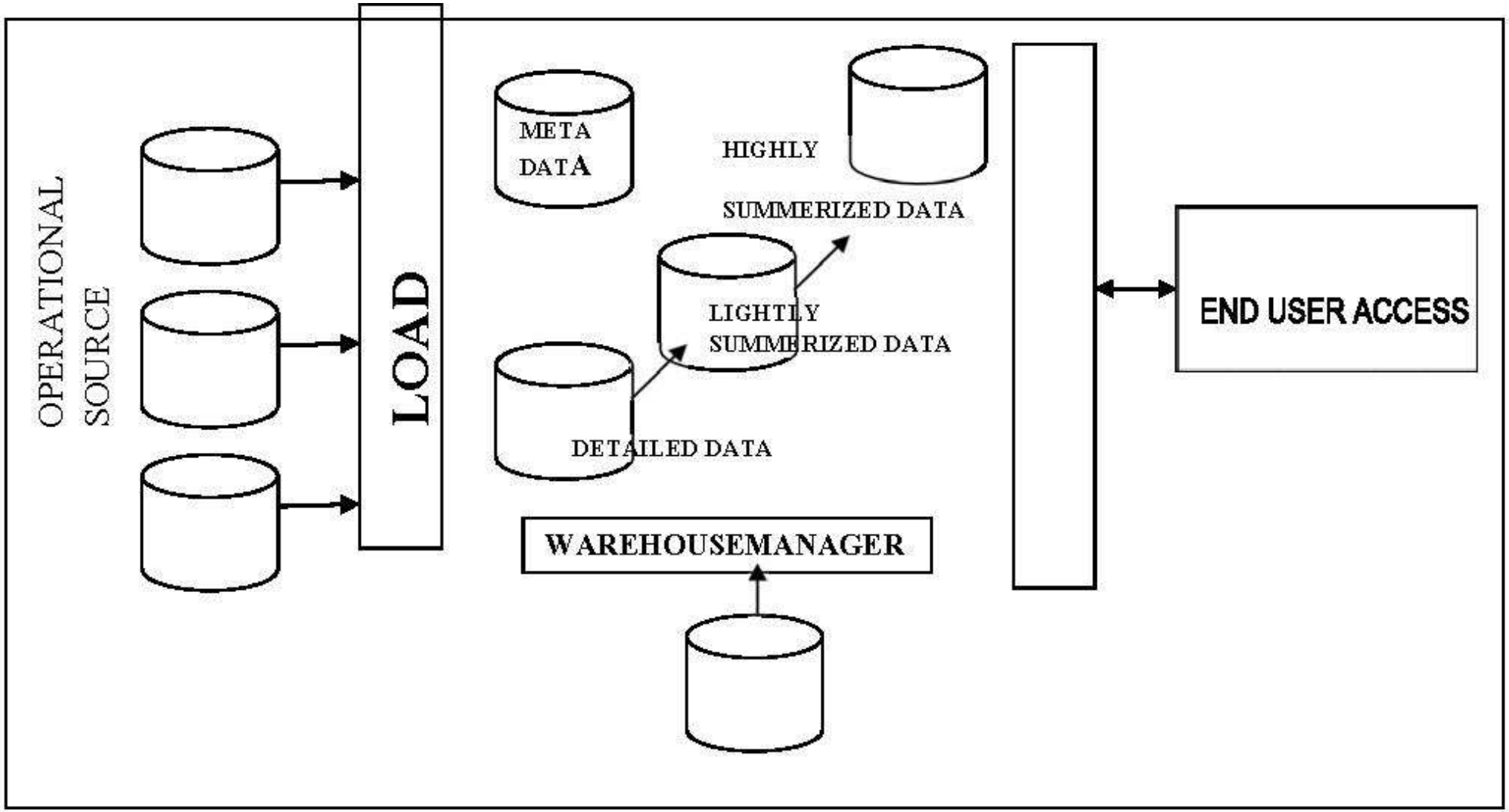
1. Source Systems
2. Data Staging Area
3. Presentation servers

- The data gets transmitted from source systems to presentation servers via the data staging area.
- This entire process is popularly known as ETL (Extract, Transform, And Load) or ETT (Extract, Transform, And Transfer).
- The ETL tool of Oracle is called Oracle Warehouse Builder (OWB) and MS SQL Server's ETL tool is called Data Transformation Services (DTS).

Data Warehouse: A Multi-Tiered Architecture



A typical architecture of a data warehouse is shown below:



Each component and the tasks performed by them are explained below:

1. OPERATIONALDATA

The sources of data for the data warehouse is supplied from:

(i) The data from the mainframe systems

(ii) Data can also come from the relational DBMS like Oracle, Informix.

(iii) In addition to these internal data, operational data also includes external data

2. LOADMANAGER

The load manager performs all the operations associated with extraction and loading data in to the data warehouse.

3. WAREHOUSE MANAGER

The warehouse manager performs all the operations associated with the management of data in the warehouse.. The operations performed by warehouse manager include:

(i) Analysis of data to ensure consistency

(ii) Transformation and merging the source data from temporary storage into data warehouse tables

(iii) Create indexes and views on the basetable.

(iv) Denormalization

(v) Generation ofaggregation

(vi) Backing up and archiving ofdata

4. QUERYMANAGER

The query manager performs all operations associated with management of user queries.

5. DETAILED DATA

This area of the warehouse stores all the detailed data in the database schema.

6. LIGHTLY AND HIGHLY SUMMERIZEDDATA

The area of the data warehouse stores all the predefined lightly and highly summarized (aggregated) data generated by the warehouse manager.

7. ARCHIVE AND BACK UPDATA

This area of the warehouse stores detailed and summarized data for the purpose of archiving and back up. The data is transferred to storage archives such as magnetic tapes or optical disks.

8. METADATA

The data warehouse also stores all the Meta data (data about data) definitions used by all processes in the warehouse. It is used for variety of purposed including:

(i) The extraction and loading process

(ii) The warehouse management process

(iii) As part of Query Management process

6. LIGHTLY AND HIGHLY SUMMERIZEDDATA

The area of the data warehouse stores all the predefined lightly and highly summarized (aggregated) data generated by the warehouse manager.

7. ARCHIVE AND BACK UPDATA

This area of the warehouse stores detailed and summarized data for the purpose of archiving and back up. The data is transferred to storage archives such as magnetic tapes or optical disks.

8. METADATA

The data warehouse also stores all the Meta data (data about data) definitions used by all processes in the warehouse. It is used for variety of purposed including:

(i) The extraction and loading process

(ii) The warehouse management process

(iii) As part of Query Management process

9. END-USER ACCESSTOOLS

The principal purpose of data warehouse is to provide information to the business managers for strategic decision-making. These users interact with the warehouse using end user access tools. The examples of some of the end user access tools can be:

(i) Reporting and QueryTools
(ii) Application DevelopmentTools
(iii) Executive Information Systems Tools
(iv) Online Analytical ProcessingTools
(v) Data Mining Tools

Building a Data warehouse

The ETL (Extract Transformation Load) process

Four major process of the data warehouses are

- Extract (data from the operational systems and bring it to the data warehouse),
- Transform (the data into internal format and structure of the data warehouse),
- Cleanse (to make sure it is of sufficient quality to be used for decision making)
- Load (cleanse data is put into the data warehouse).

9. END-USER ACCESSTOOLS

The principal purpose of data warehouse is to provide information to the business managers for strategic decision-making. These users interact with the warehouse using end user access tools. The examples of some of the end user access tools can be:

(i) Reporting and QueryTools
(ii) Application DevelopmentTools
(iii) Executive Information Systems Tools
(iv) Online Analytical ProcessingTools
(v) Data Mining Tools

Building a Data warehouse

The ETL (Extract Transformation Load) process

Four major process of the data warehouses are

- Extract (data from the operational systems and bring it to the data warehouse),
- Transform (the data into internal format and structure of the data warehouse),
- Cleanse (to make sure it is of sufficient quality to be used for decision making)
- Load (cleanse data is put into the data warehouse).

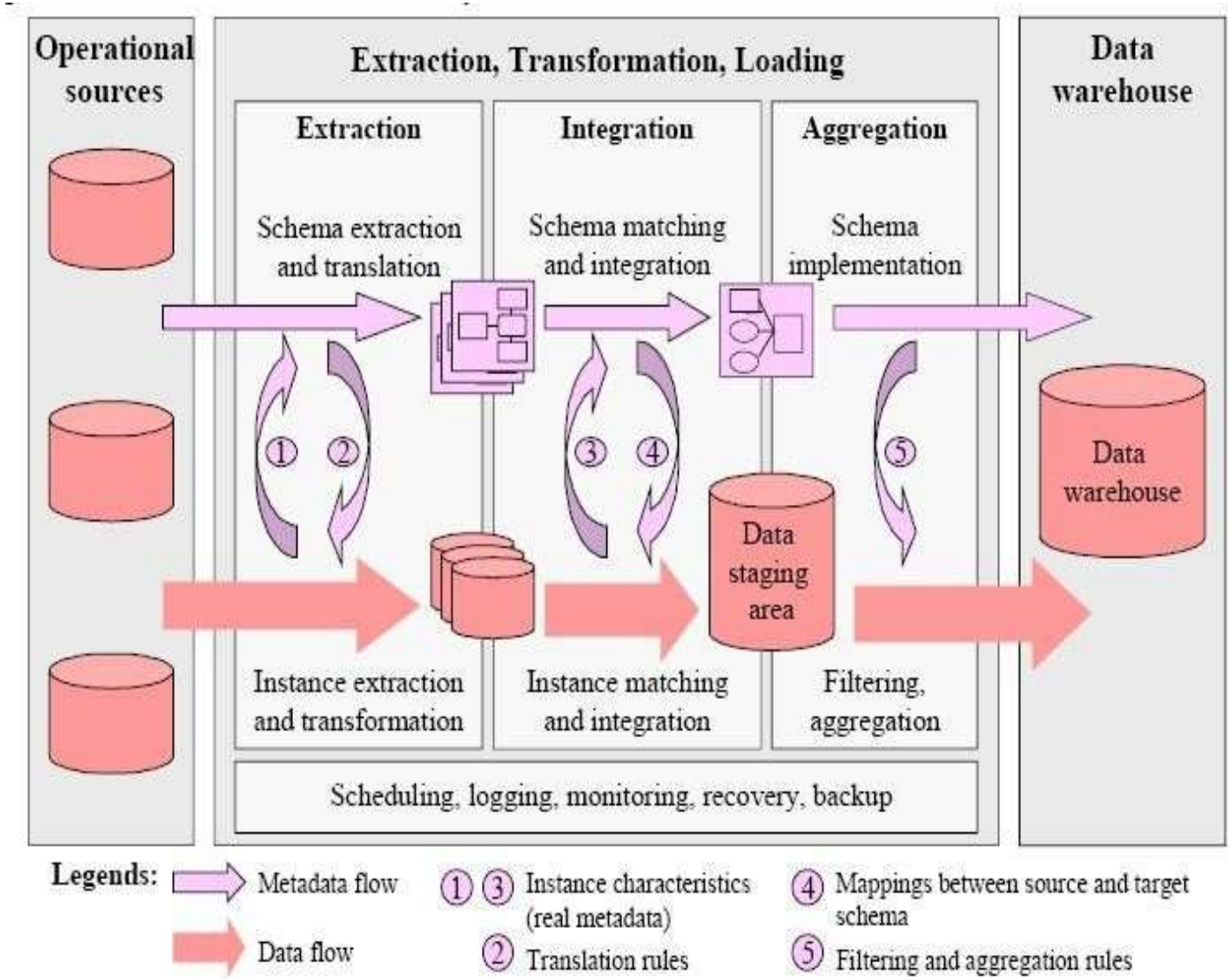
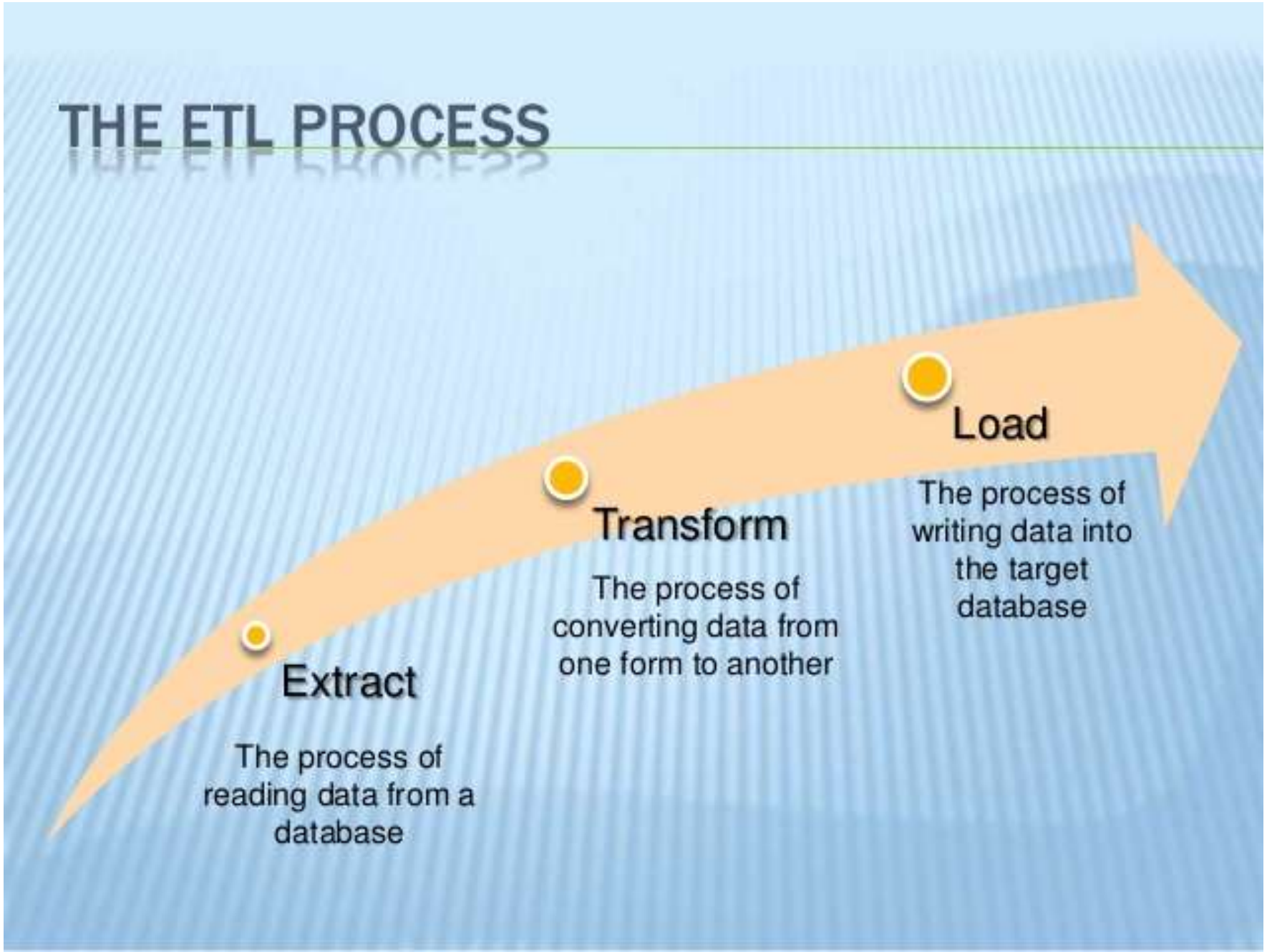


Figure 1. Steps of building a data warehouse: the ETL process

The four processes from extraction through loading often referred collectively as **Data Staging**.


THE ETL PROCESS
Extract
The process of reading data from a database
Transform
The process of converting data from one form to another
Load
The process of writing data into the target database

EXTRACT

- **Gathering the data**
  - Raw data that was written directly into the disk
  - Data written to flat files or relational tables from structured source systems
  - Data can be read multiple times, if needed
- **Cleansing the data**
  - Eliminate duplicates or fragmented data
  - Exclude unwanted / unneeded information

TRANSFORM

- Preparing the data to be housed in the data warehouse
- Converting the extracted data
- Standardization
- Verification/Validity checks
- Combining data
- Using rules and lookup tables

LOADING

- Storing the transformed data in the data warehouse
- Batch/Real-time processing
- Can follow star schema and snowflake schema.

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.

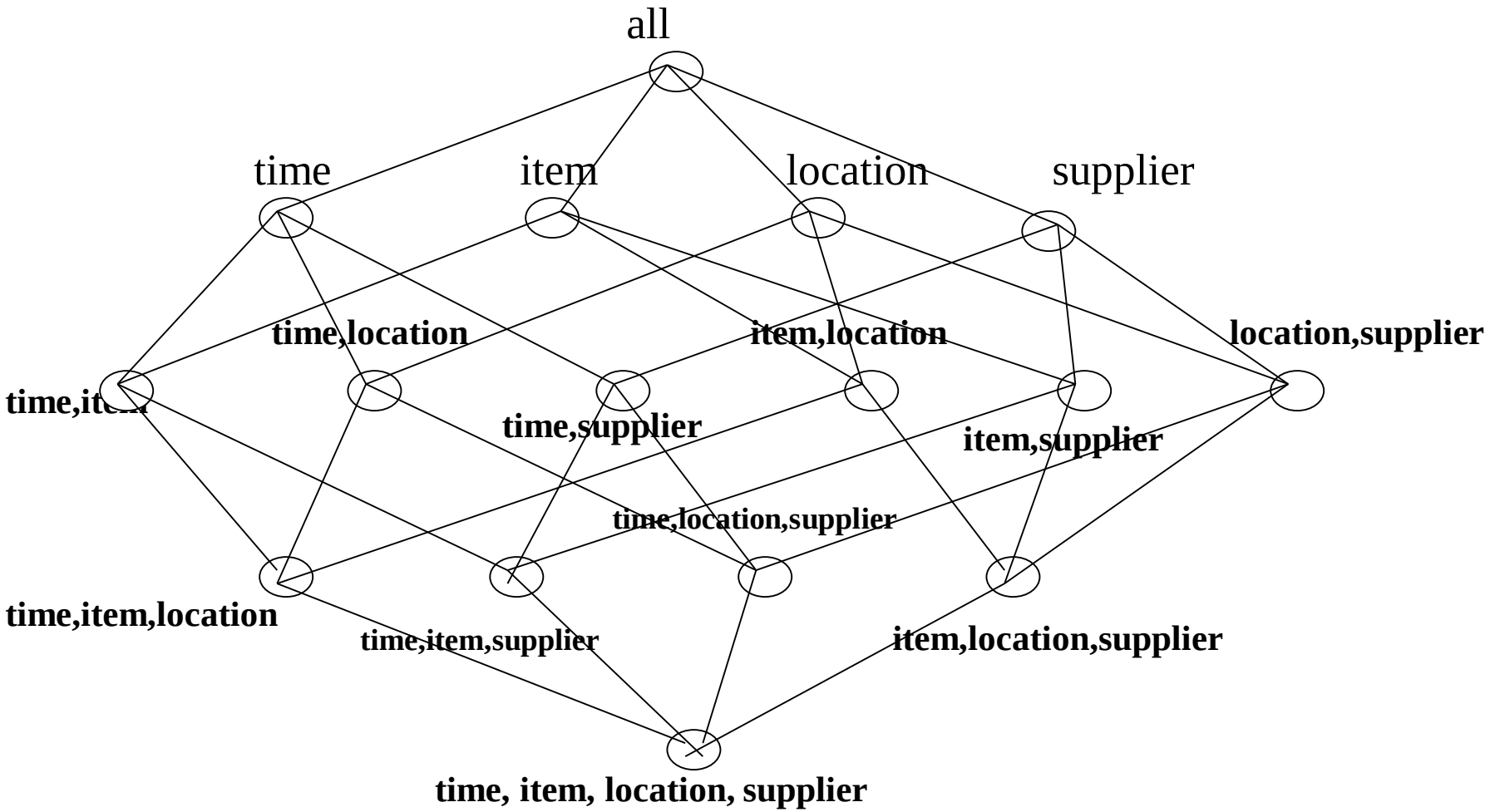
It is defined by Facts and Dimensions.

  - **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.

Multidimensional Data Model and its operation (OLAP operations)

The most popular data model for data warehouses is a multidimensional model. This model can exist in the form of a star schema, a snowflake schema, or a fact constellation schema. Let's have a look at each of these schema types.



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

4-D (*base*) cuboid

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

Multi-dimensional Schemas

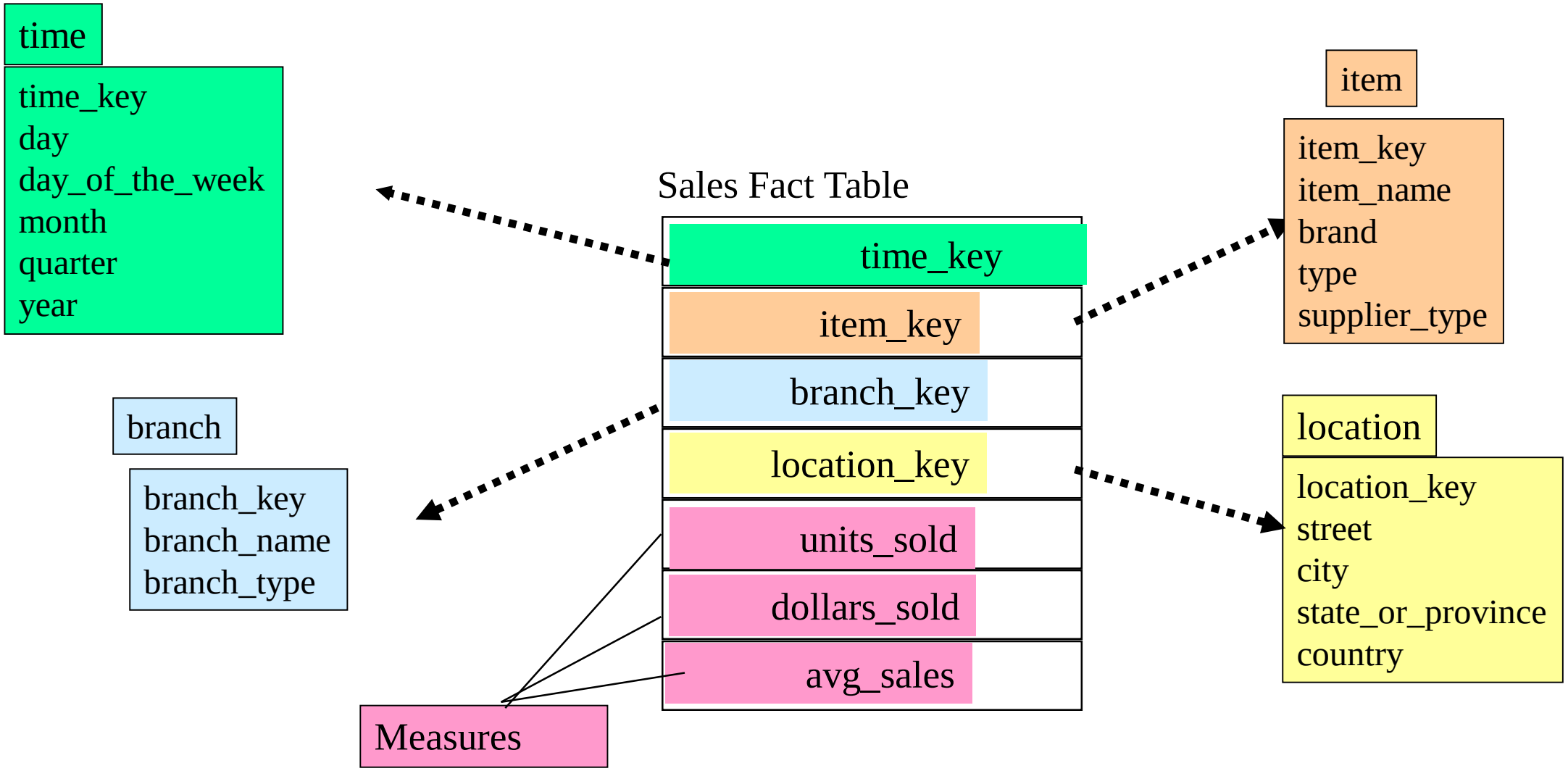
- Two common multi-dimensional schemas are

1. **Star schema**
   - Consists of a fact table with a single table for each dimension

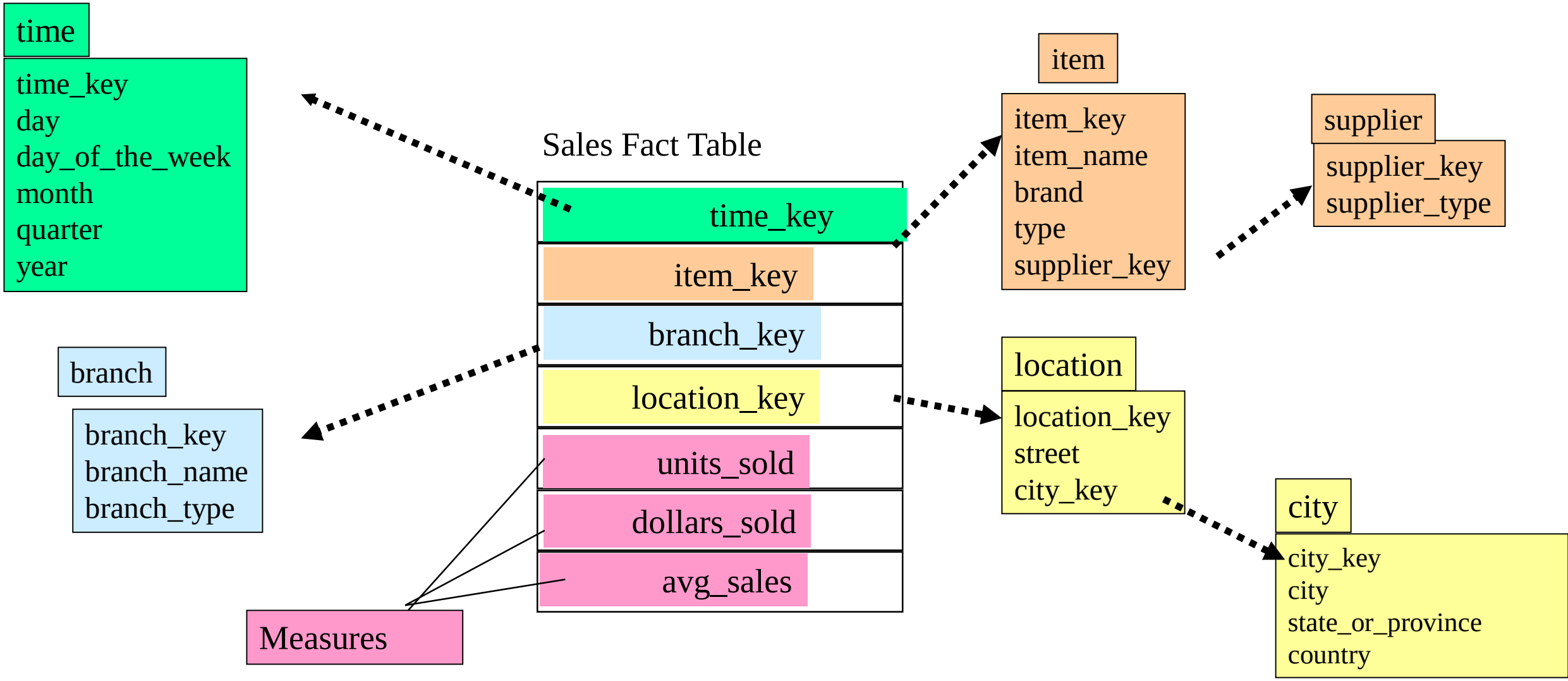
2. **Snowflake Schema**
   - It is a variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.

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

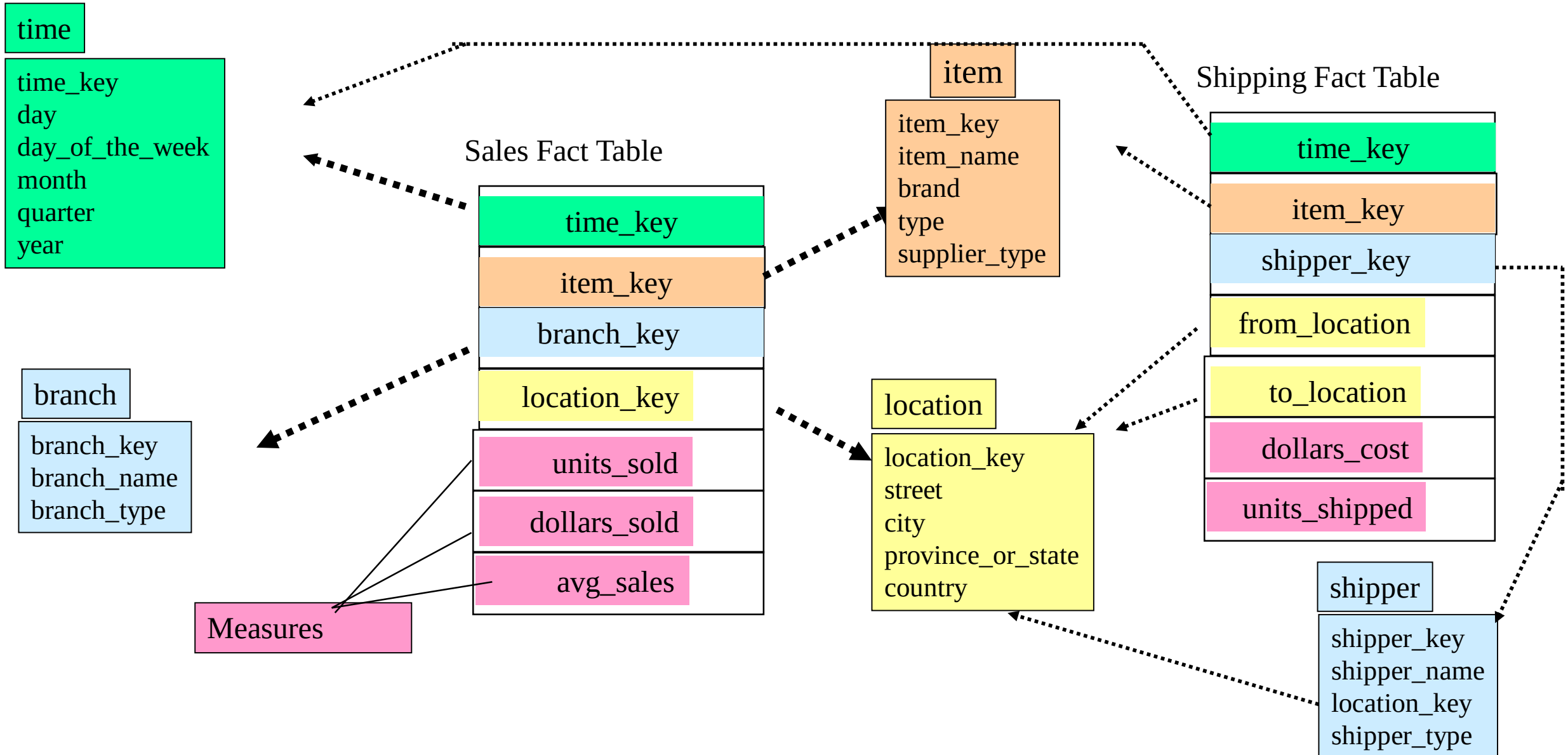
Example of **Star Schema**



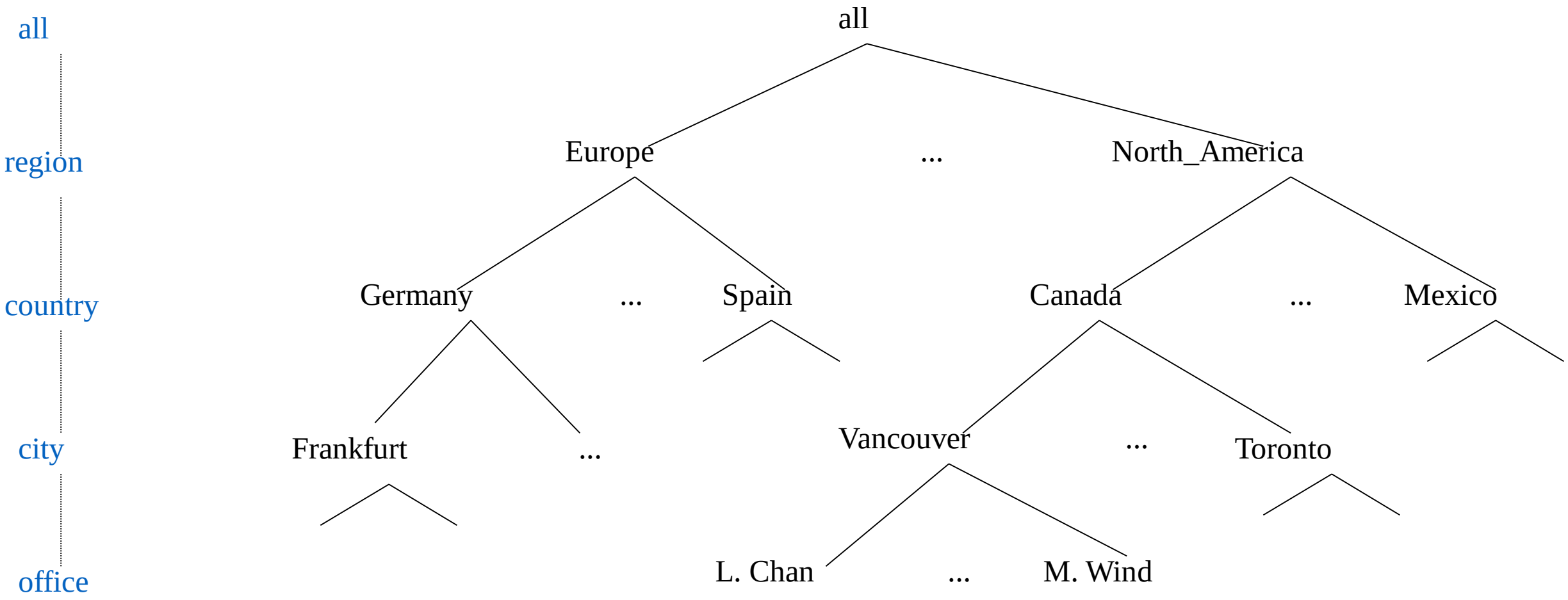
Example of **Snowflake Schema**



Example of **Fact Constellation**



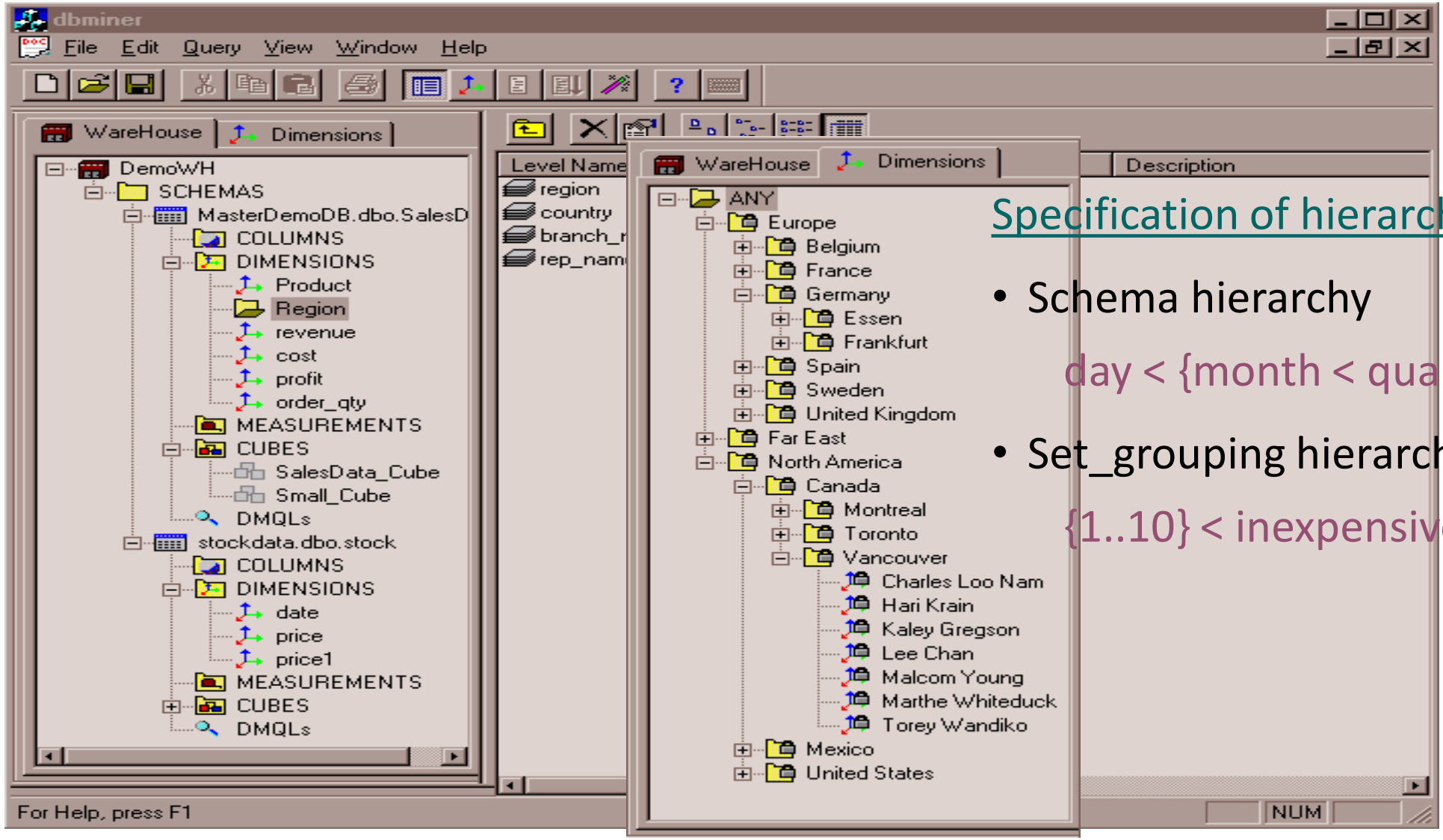
A Concept Hierarchy: **Dimension** (location)



Data Cube Measures: 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()

View of Warehouses and Hierarchies



Specification of hierarchies

- Schema hierarchy

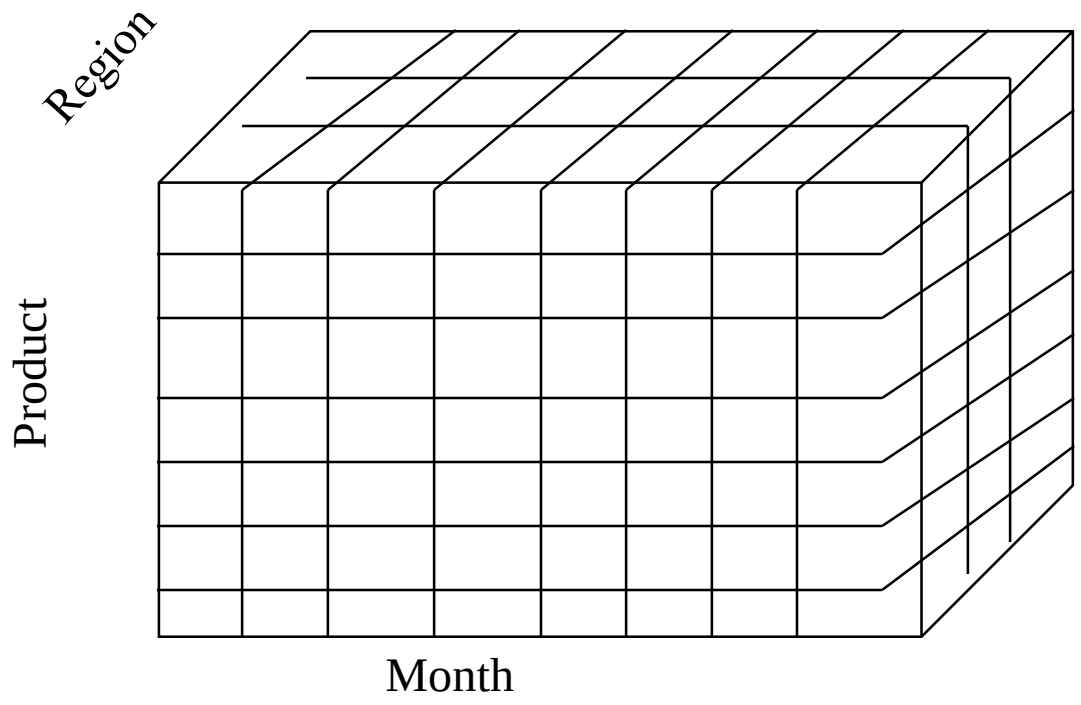
  day < {month < quarter; week} < year

- Set_grouping hierarchy

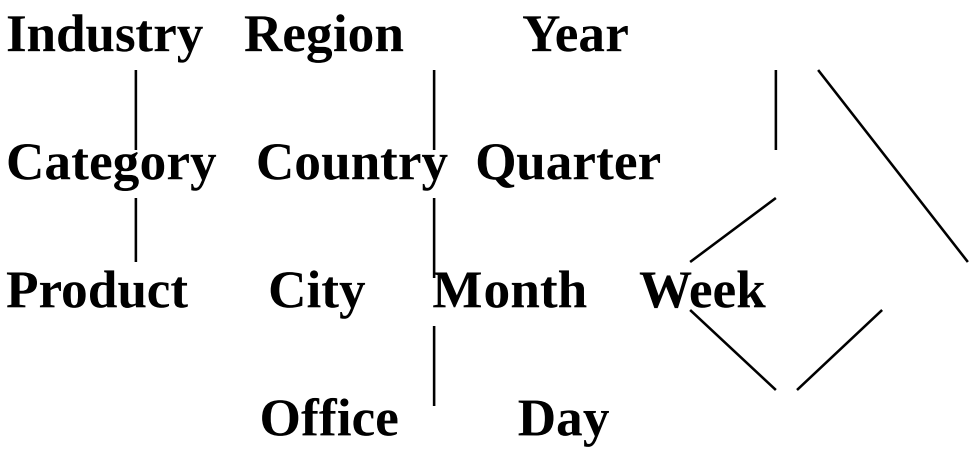
  {1..10} < inexpensive

Multidimensional Data

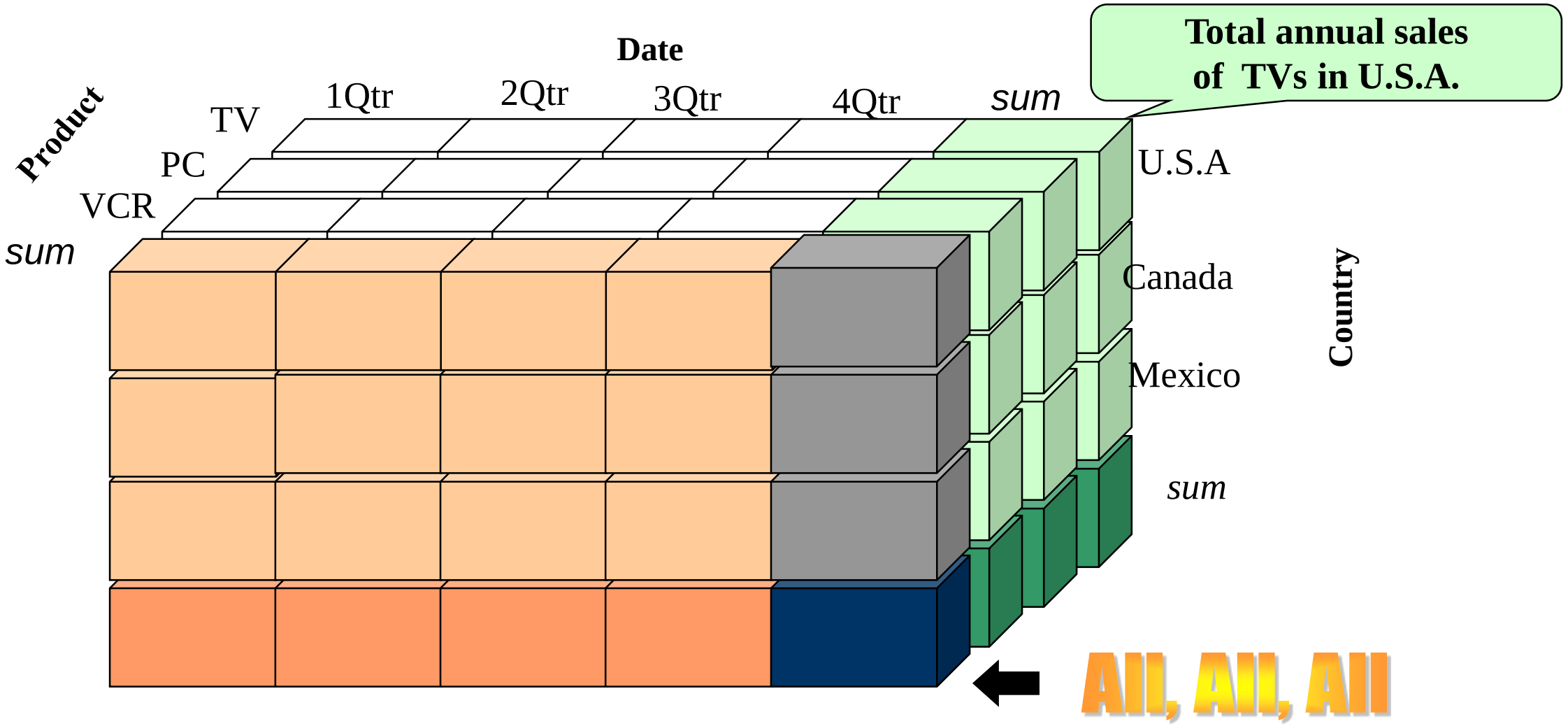
- Sales volume as a function of product, month, and region



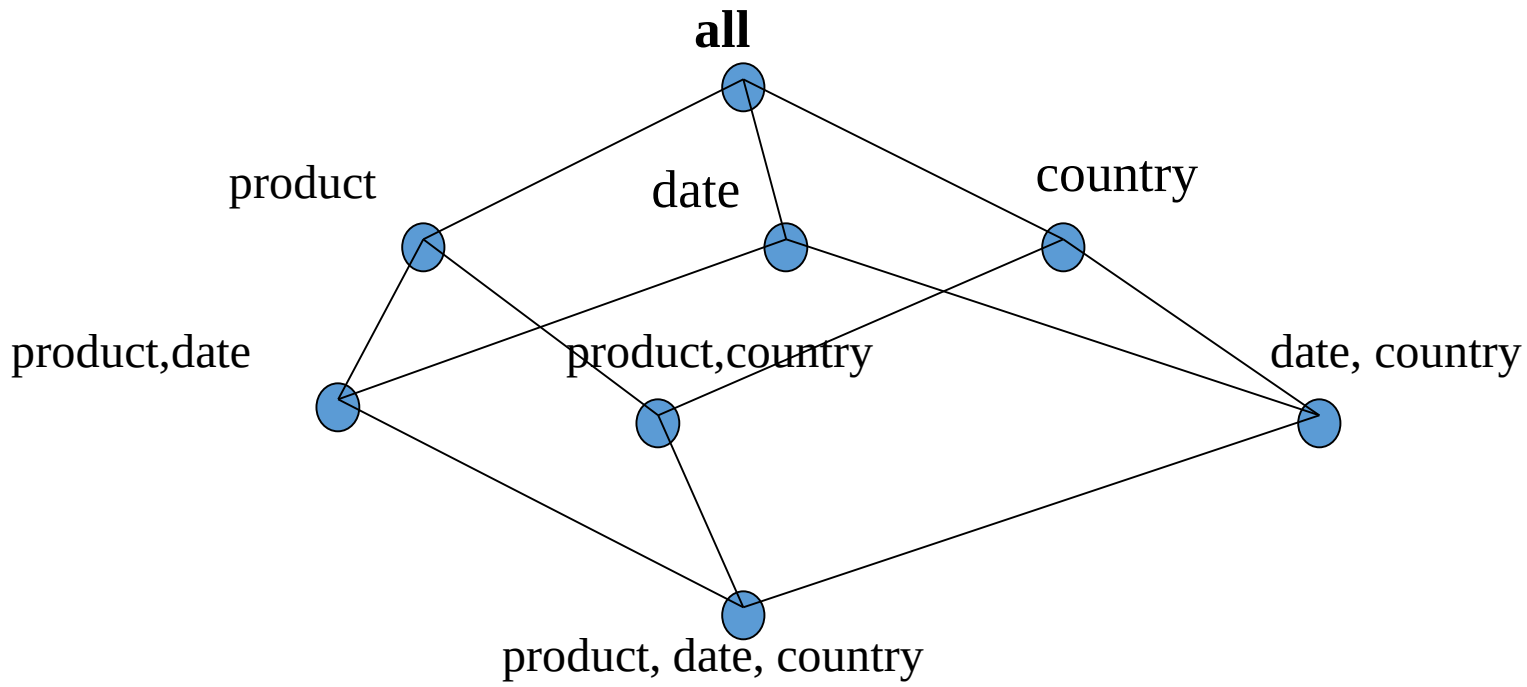
Dimensions: *Product, Location, Time*
Hierarchical summarization paths



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

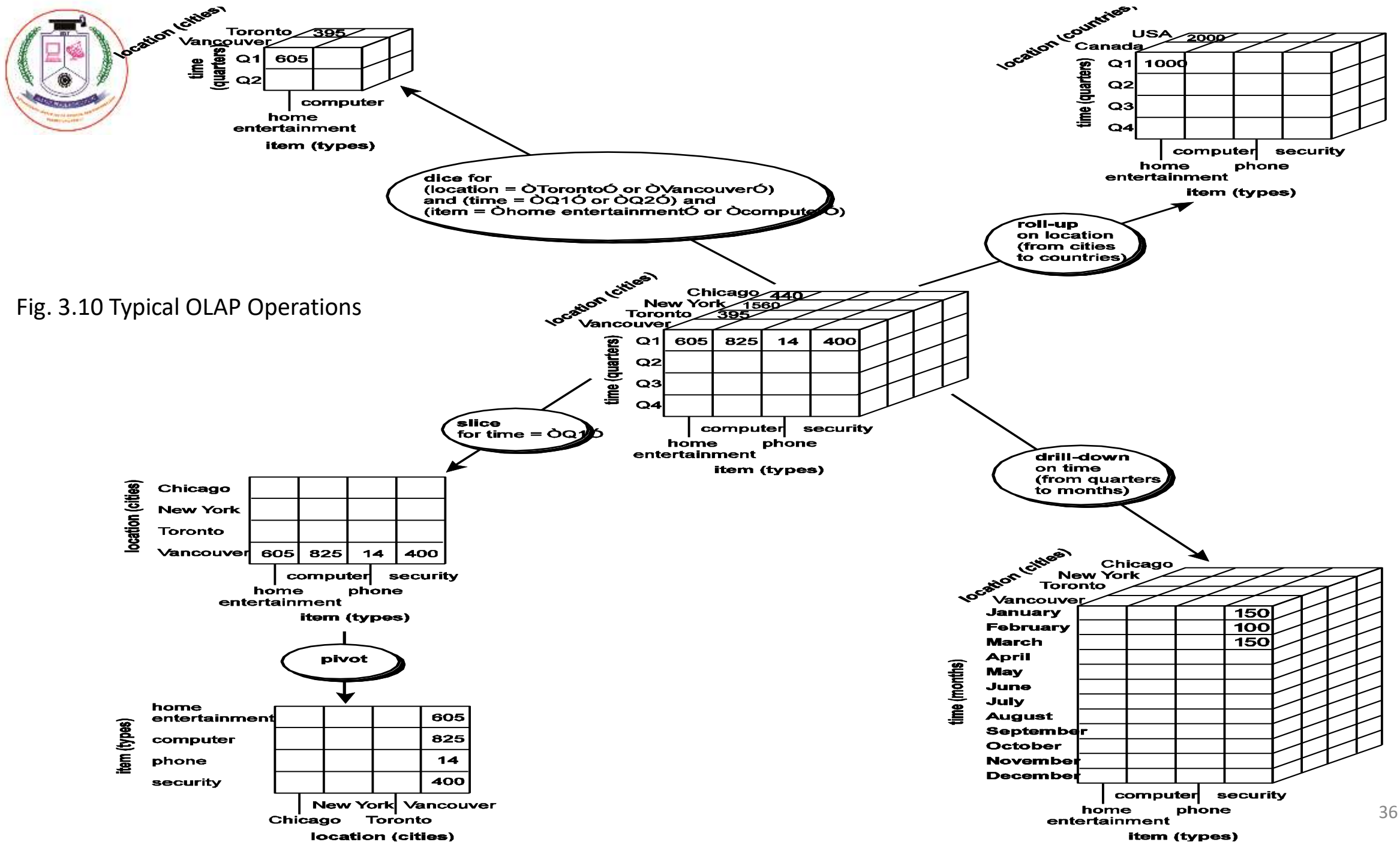
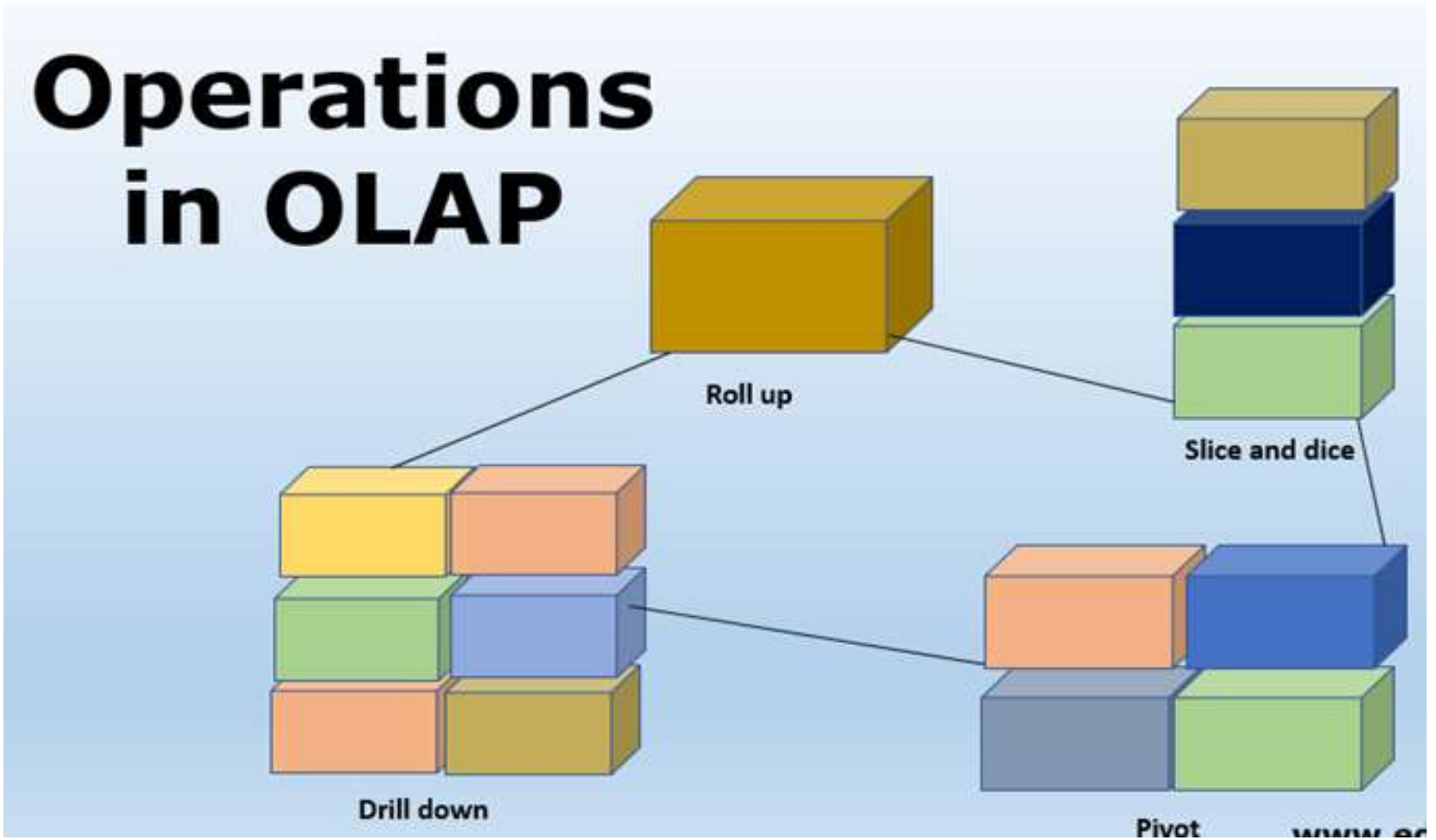
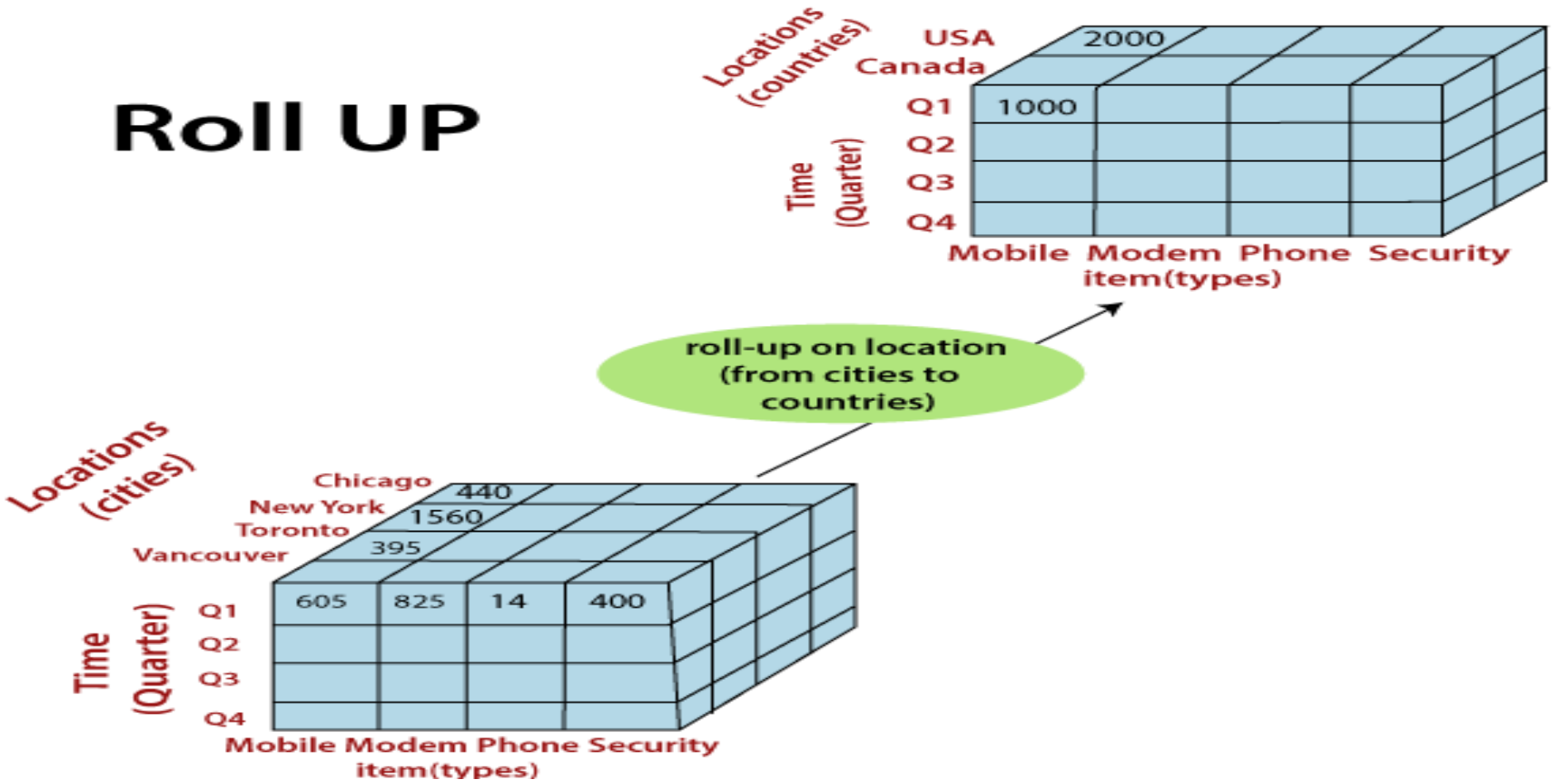


Fig. 3.10 Typical OLAP Operations

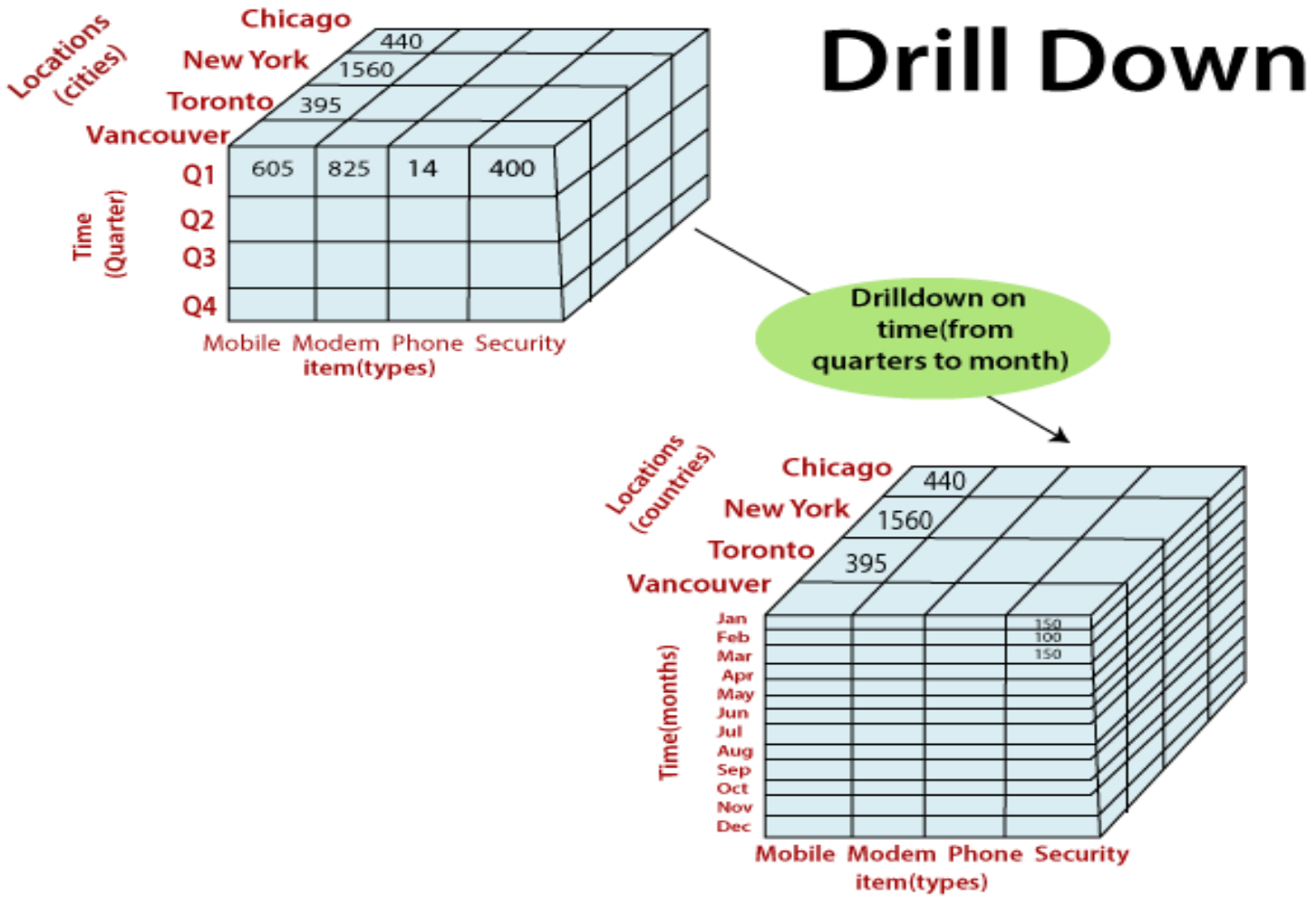
OLAP Operations



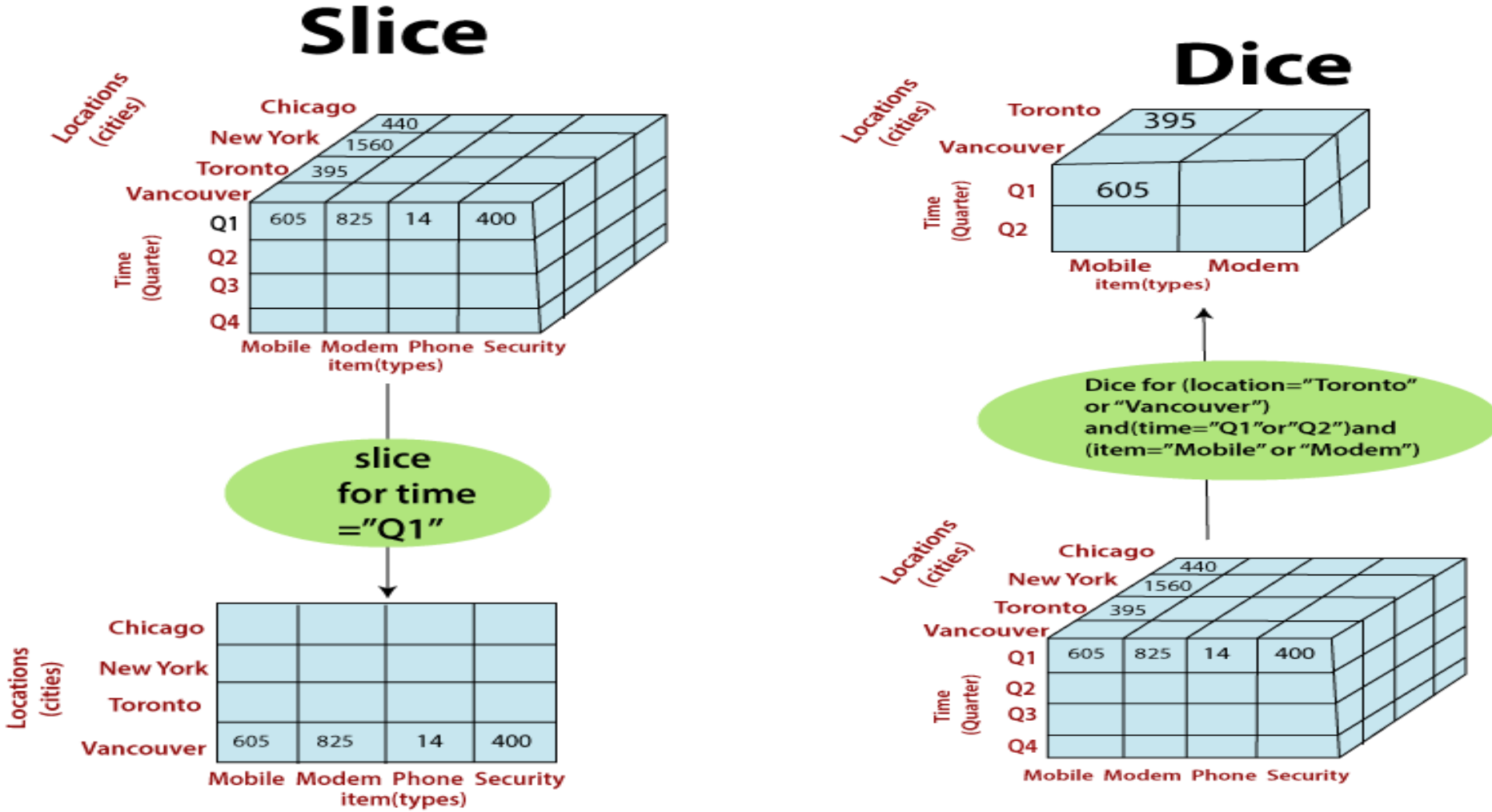
OLAP Operations



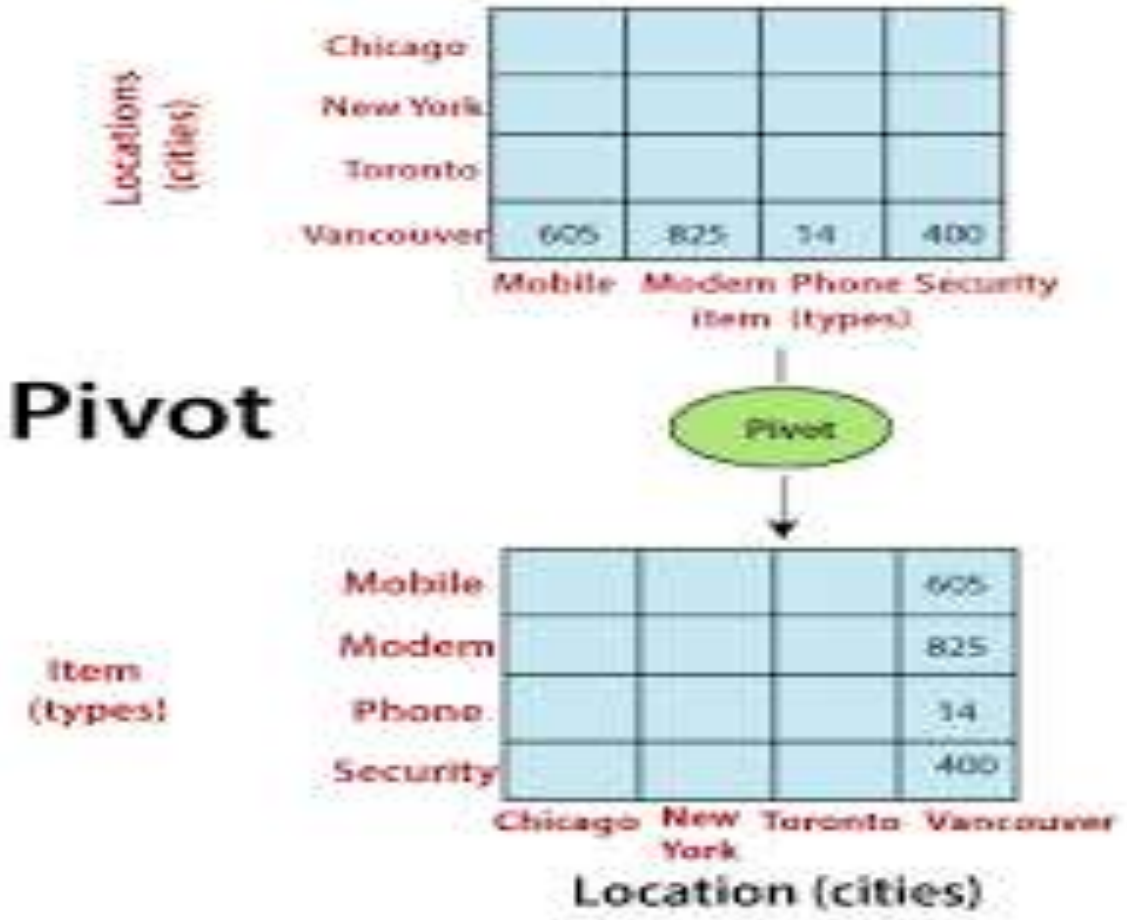
OLAP Operations



OLAP Operations



OLAP Operations



OLAP operations on multidimensional data.

Roll-up

The roll-up operation performs aggregation on a data cube, either by climbing-up a concept hierarchy for a dimension or by dimension reduction. Figure shows the result of a roll-up operation performed on the central cube by climbing up the concept hierarchy for location. This hierarchy was defined as the total order street < city < province or state<country.

Drill-down

Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping-down a concept hierarchy for a dimension or introducing additional dimensions. Figure shows the result of a drill-down operation performed on the central cube by stepping down a concept hierarchy for time defined as day < month < quarter < year. Drill-down occurs by descending the time hierarchy from the level of quarter to the more detailed level of month.

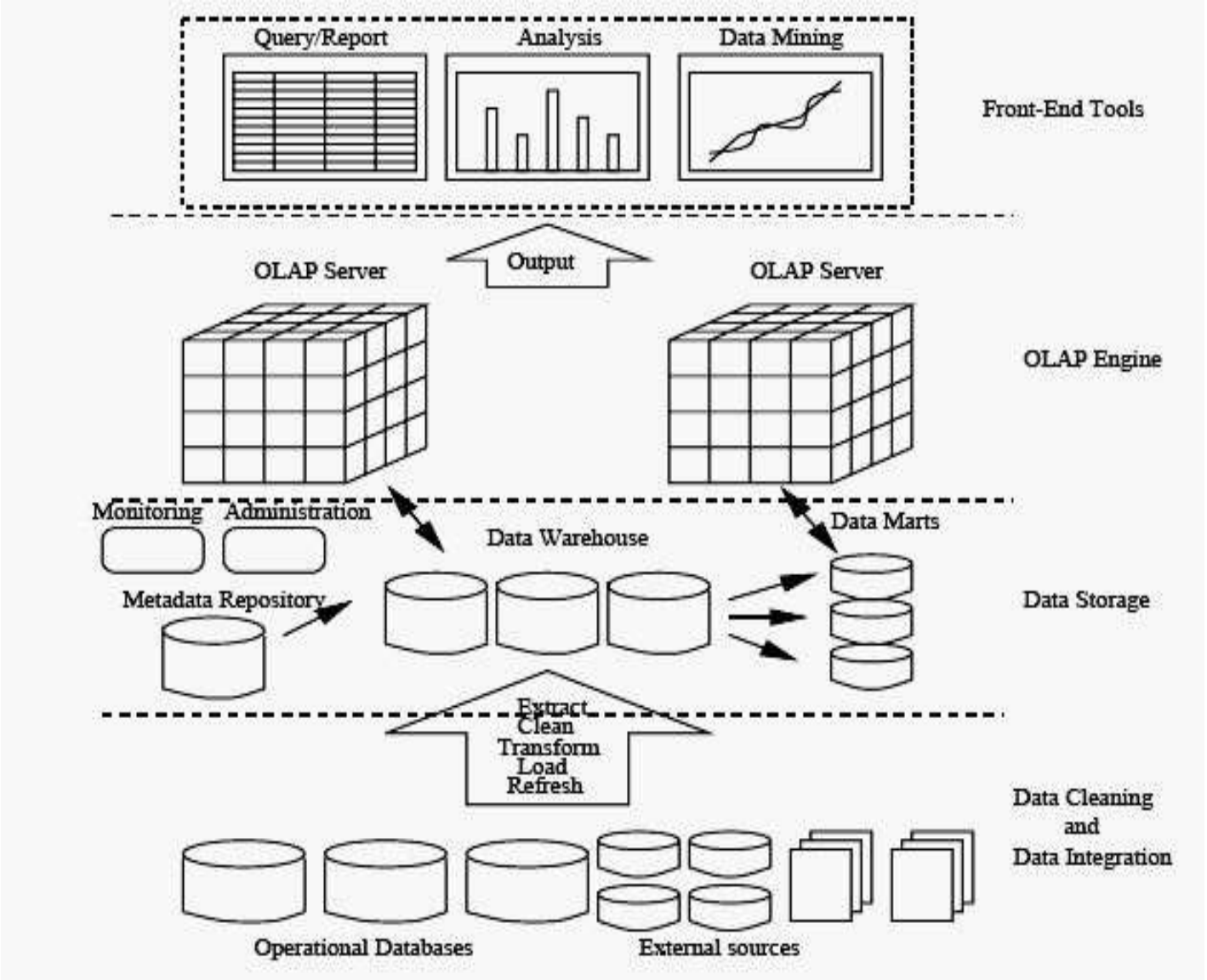
Slice and dice:

he slice operation performs a selection on one dimension of the given cube, resulting in a subcube. Figure shows a slice operation where the sales data are selected from the central cube for the dimension time using the criteria time=”Q2". The dice operation defines a subcube by performing a selection on two or more dimensions.

Pivot (rotate)

Pivot is a visualization operation which rotates the data axes in view in order to provide an alternative presentation of the data. Figure shows a pivot operation where the item and location axes in a 2-D slice are rotated.

Three-tier Data warehouse architecture



The bottom tier is a ware-house database server which is almost always a relational database system. The middle tier is an OLAP server which is typically implemented using either a Relational OLAP (ROLAP) model, (2) a Multidimensional OLAP (MOLAP) model. The top tier is a client, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

From the architecture point of view, there are three data warehouse models: the enterprise warehouse, the data mart, and the virtual warehouse.

Enterprise warehouse

An enterprise warehouse collects all of the information about subjects spanning the entire organization. It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope. It typically contains detailed data as well as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.

Data mart

A data mart contains a subset of corporate-wide data that is of value to a specific group of users. The scope is connected to specific, selected subjects. For example, a marketing data mart may connect its subjects to customer, item, and sales. The data contained in data marts tend to be summarized. Depending on the source of data, data marts can be categorized into the following two classes:

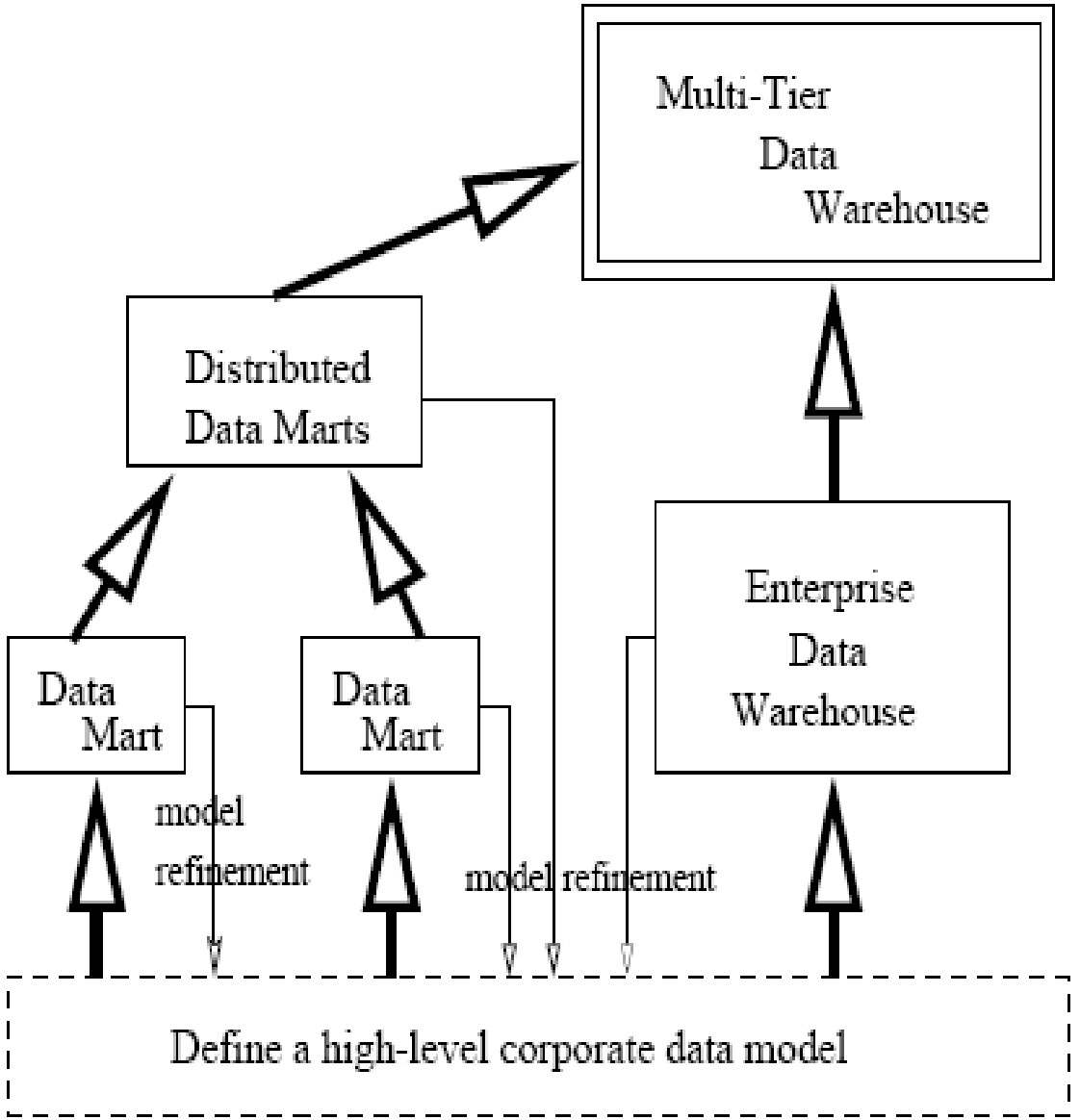
(i) Independent data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area.

(ii) Dependent data marts are sourced directly from enterprise data warehouses.

Virtual warehouse

A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized. A virtual warehouse is easy to build but requires excess capacity on operational database servers.

Figure: A recommended approach for data warehouse development.



Features of OLTP and OLAP

The major distinguishing features between OLTP and OLAP are summarized as follows.

Users and system orientation: An OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.

Data contents: An OLTP system manages current data that, typically, are too detailed to be easily used for decision making. An OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity. These features make the data easier for use in informed decision making.

Database design: An OLTP system usually adopts an entity-relationship (ER) data model and an application oriented database design. An OLAP system typically adopts either a star or snowflake model and a subject-oriented database design.

View: An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations. In contrast, an OLAP system often spans multiple versions of a database schema. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.

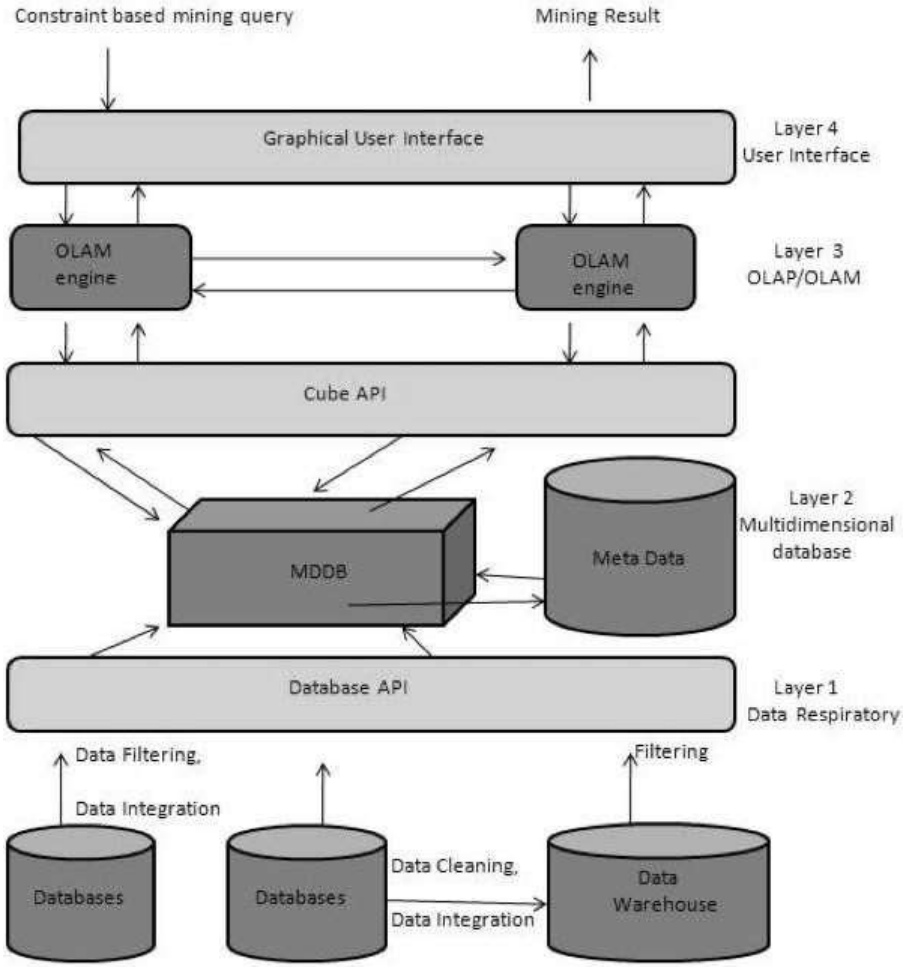
Access patterns: The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read-only operations although many could be complex queries.

Comparison between OLTP and OLAP systems.

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	E-R 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
# of records accessed	tens	millions
# 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

Integrated OLAP and OLAM Architecture

Online Analytical Mining integrates with Online Analytical Processing with data mining and mining knowledge in multidimensional databases. Here is the diagram that shows the integration of both OLAP and OLAM –



Importance of OLAM

OLAM is important for the following reasons −

- **High quality of data in data warehouses** − The data mining tools are required to work on integrated, consistent, and cleaned data. These steps are very costly in the preprocessing of data. The data warehouses constructed by such preprocessing are valuable sources of high quality data for OLAP and data mining as well.
- **Available information processing infrastructure surrounding data warehouses** − Information processing infrastructure refers to accessing, integration, consolidation, and transformation of multiple heterogeneous databases, web-accessing and service facilities, reporting and OLAP analysis tools.
- **OLAP−based exploratory data analysis** − Exploratory data analysis is required for effective data mining. OLAM provides facility for data mining on various subset of data and at different levels of abstraction.
- **Online selection of data mining functions** − Integrating OLAP with multiple data mining functions and online analytical mining provide users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

THANK YOU