

Data Mining and Data Warehousing

Unit 2

Data Warehouse for Data mining

Data Warehouse

- A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured queries, and decision making
- Data warehousing is the process of constructing and using a data warehouse
- Nature of Data Warehouse
 - Subject Oriented
 - Integrated
 - Non Volatile
 - Time variant

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

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

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*

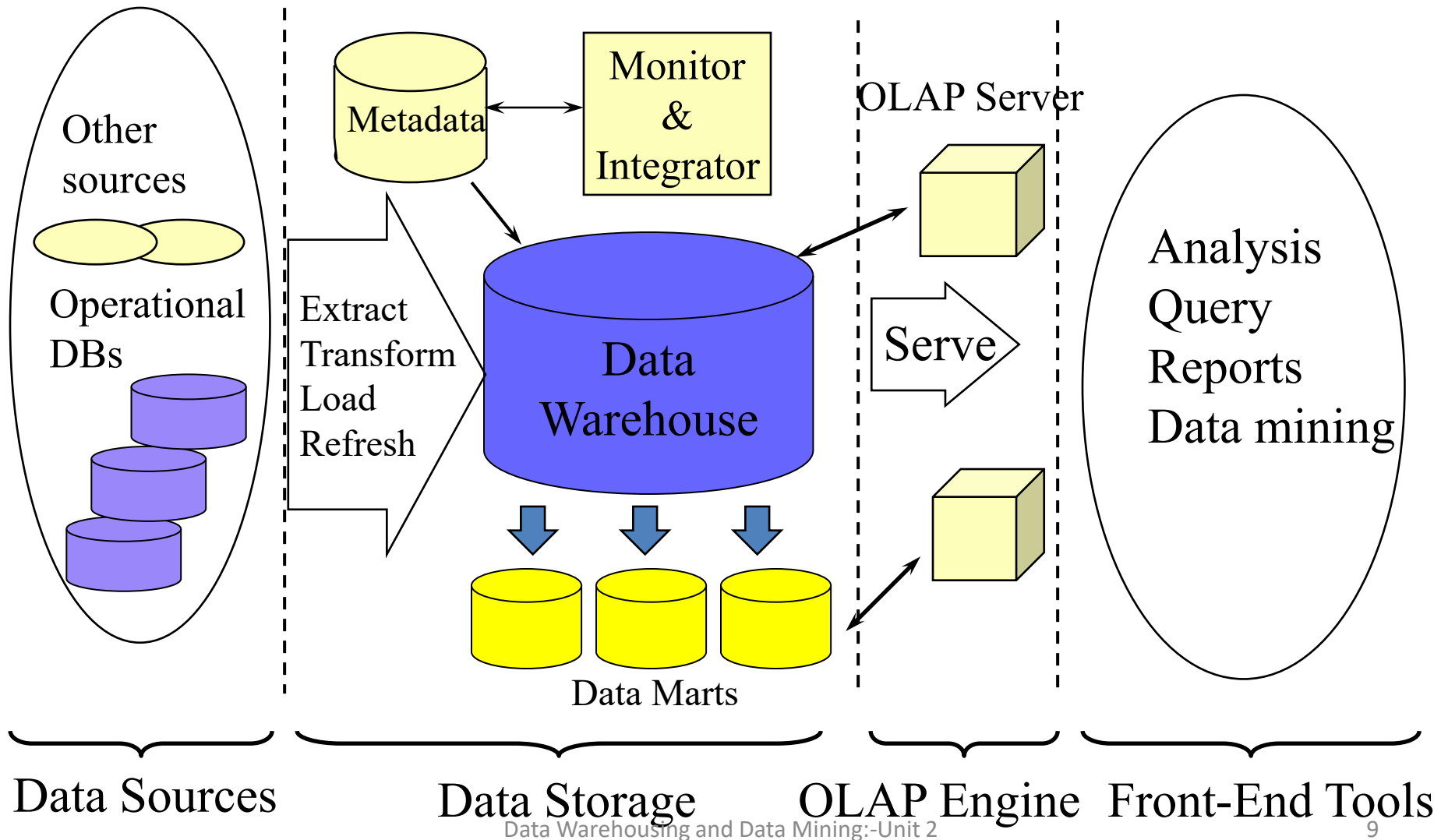
Database Vs Data Warehouse

- Data warehouse technology includes a set of concepts and methods that offer the users useful information for decision making
- The necessity to build a data warehouse arises from the necessity to improve the quality of information in the organization
- A database is an application-oriented collection of data
- A data warehouse is a subject-oriented collection of data

Data Warehouse views

- 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 Architecture.....



A Three Tier Data Warehouse Architecture

- **Data Sources:**

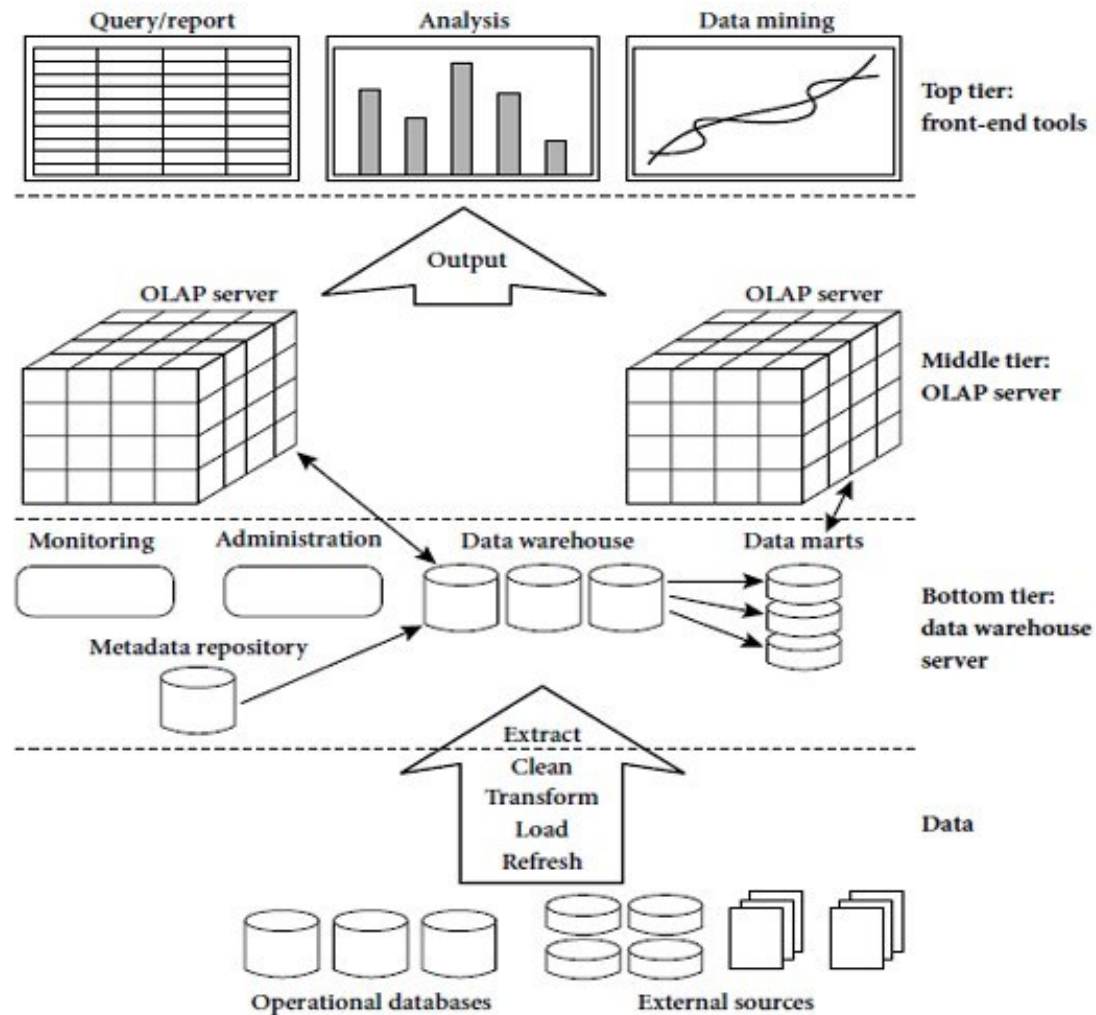
A data warehouse system uses heterogeneous sources of data either from operational databases or from some external sources.

- **Bottom Tier:**

The bottom tier of the architecture is the data warehouse database server. It is the relational database system. Data is feed into bottom tier by some back end tools and utilities. The back end tools and utilities perform the following functions:

- **Data extraction:** gathers data from multiple, heterogeneous and external sources.
- **Data cleaning:** Detect errors in data and correct them when possible.
- **Data transformation:** converts data from legacy or host format to warehouse format.
- **Load:** which sorts, summarizes, checks integrity, and builds indices and partitions.
- **Refresh:** which involves updating from data sources to the warehouse.
- The data are extracted using application program interfaces known as gateways. Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection).
- This tier also contains a metadata repository, which stores information about the data warehouse and its contents.

A Three Tier Data Warehouse Architecture

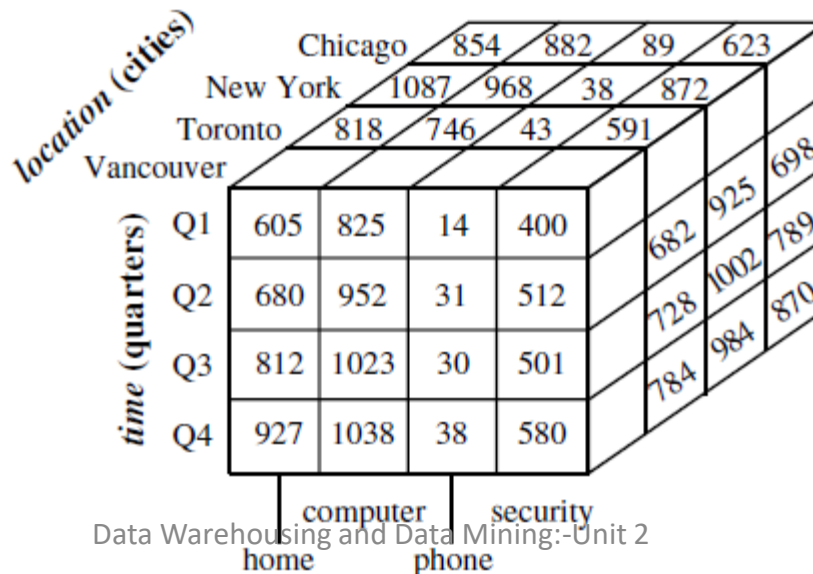


A Three Tier Data Warehouse Architecture

- **Middle Tier:**
 - Middle tier is an OLAP server that can be implemented using either relational OLAP (ROLAP) model or multidimensional OLAP (MOLAP) model.
 - ROLAP is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.
 - MOLAP directly implements multidimensional data and operations.
- **Top Tier:**
 - The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on). The top tier layer holds following tools:
 - Query and Reporting tools: Production reporting tool.
 - Analysis tools: Prepare charts based on analysis.
 - Data mining tools: Discover hidden knowledge, pattern.

Multidimensional Data Model

<i>location = "Chicago"</i>					<i>location = "New York"</i>					<i>location = "Toronto"</i>					<i>location = "Vancouver"</i>				
<i>item</i>					<i>item</i>					<i>item</i>					<i>item</i>				
<i>home</i>					<i>home</i>					<i>home</i>					<i>home</i>				
<i>time</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>		<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>		<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	
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			



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
- Virtual warehouse
 - A set of views over operational databases
 - Only some of the possible summary views may be materialized

Components of Data Warehouse

- **Load Manager**
 - Responsible for loading the extracted clean data into data warehouse
- **Query Manager**
 - Responsible for handling the user query
 - Scheduling the query
 - Directing the query to the relevant data cube
- **Warehouse Manager**
 - Responsible for maintaining the integrity and consistency of data warehouse
 - Disaster management
 - Regular back up

Meta Data Repository

- Metadata are data about data. When used in a data warehouse, Metadata describes and contextualizes other data. It provides information about the content, format, structure, and other characteristics of data, and can be used to improve the organization, discoverability, and accessibility of data
- Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.
- Metadata can be stored in various forms, such as text, XML, or RDF, and can be organized using metadata standards and schemas. write this in simple language so that layman can also understand
- A metadata repository should contain the following:
 - **Descriptive metadata:** A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents.
 - **Operational metadata,** which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails).

Meta Data Repository

- The algorithms used for summarization, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports.
- The mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control).
- Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles.
- Business metadata, which include business terms and definitions, data ownership information, and charging policies.

OLTP

- **OLTP (On-line Transaction Processing)** is characterized by a large number of short on-line transactions (INSERT, UPDATE, DELETE)
- The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second
- In OLTP database there is detailed and current data, and schema used to store transactional databases

OLAP

- **OLAP (On-line Analytical Processing)** is characterized by relatively low volume of transactions
- Queries are often very complex and involve aggregations
- For OLAP systems a response time is an effectiveness measure
- OLAP applications are widely used by Data Mining techniques
- In OLAP database there is aggregated, historical data, stored in multi-dimensional schemas

Operational Database System (OLTP System)	Data Warehouse (OLAP System)
Operational system are generally designed to support high-volume transaction processing.	Data warehousing systems are generally designed to support high volume analytical processing. (i.e. OLAP)
It is used for day-to-day operations.	It is used for long-term informational requirements and decision support.
Operational data are the original sources of the data.	Data comes from various OLTP Databases.
In operational system data is stored with a functional or process orientation.	In data warehousing systems data is stored with a subject orientation.
It provides detailed and flat relational view of data.	It provides summarized and multidimensional view of data.
It focuses on "Data In".	It focuses on Information out.
The tables and joins are complex since they are normalized (for RDMS). This is done to reduce redundant data and to save storage space.	The tables and joins are simple since they are de-normalized. This is done to reduce the response time for analytical queries.
Performance is low for analysis queries.	High performance for analytical queries.
Data within operational systems are generally updated regularly.	Data within a data warehouse is non- volatile, meaning when new data is added old data is not erased so rarely updates.
Data volumes are less and historical data is generally not maintained.	It involves large data volumes and historical data.
Simple queries are capable of fetching the data.	Complex queries are required to fetch data.
Processing speed is fast.	Processing speed is slow because of large size.
The common users are clerk, DBA, database professional.	The common users are knowledge worker (e.g. manager, executive, analyst)

OLAP Operations

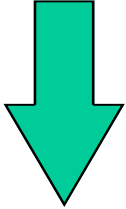
- OLAP provides a user-friendly environment for interactive data analysis.
- A number of OLAP data cube operations exist to materialize different views of data, allowing interactive querying and analysis of the data
- **Operations**
 - Roll Up (Drill Up)
 - Roll Down (Drill Down)
 - Slice
 - Dice
 - Pivot (*allows an analyst to rotate the cube in space to see its various faces. For example, cities could be arranged vertically and products horizontally while viewing data for a particular quarter.*)

OLAP Operations...

- **Roll Down (Drill Down)**
 - From higher level summary to lower level summary or detailed data, or introducing new dimensions
- **Roll Up (Drill Up)**
 - A roll-up involves summarizing the data along a dimension
 - Takes the current aggregation level of fact values and does a further aggregation on one or more of the dimensions

OLAP Operations...

*Drill
Down*

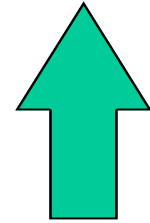


Total Sales

Total Sales per city

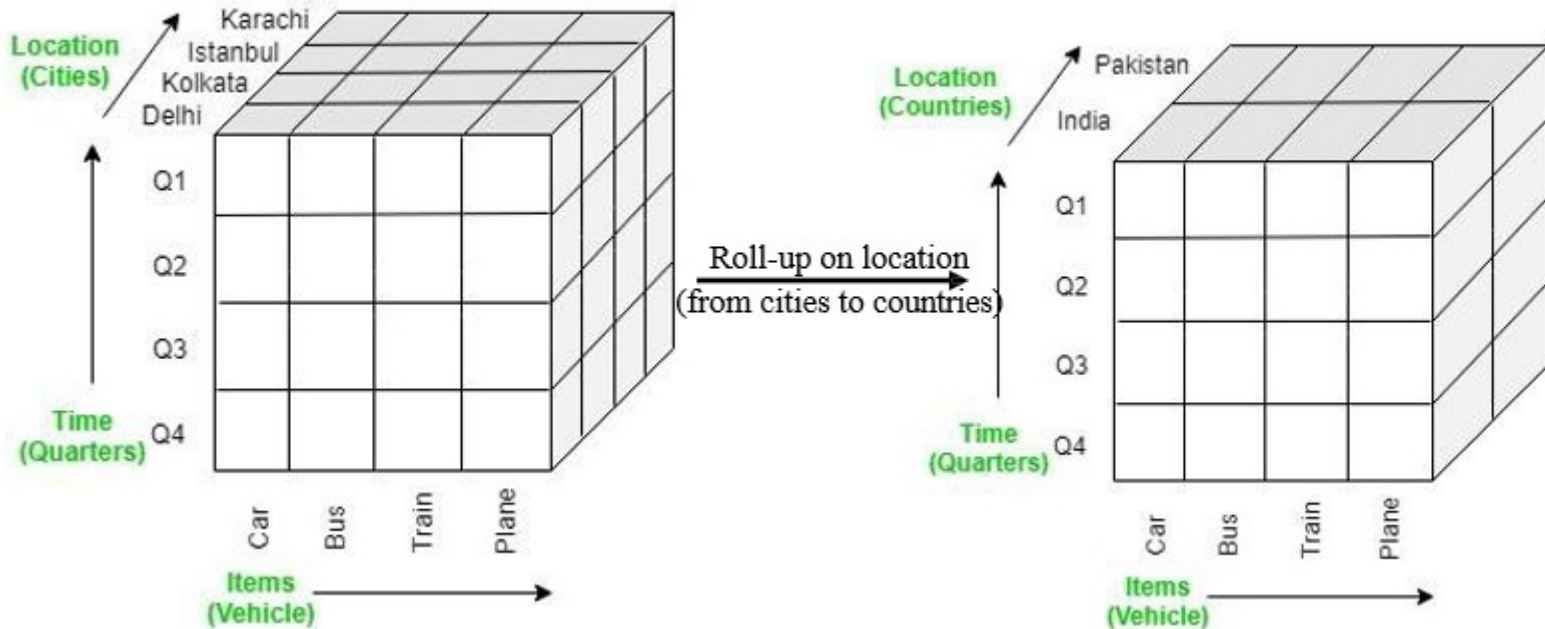
Total Sales per city per store

Total Sales per city per store per month

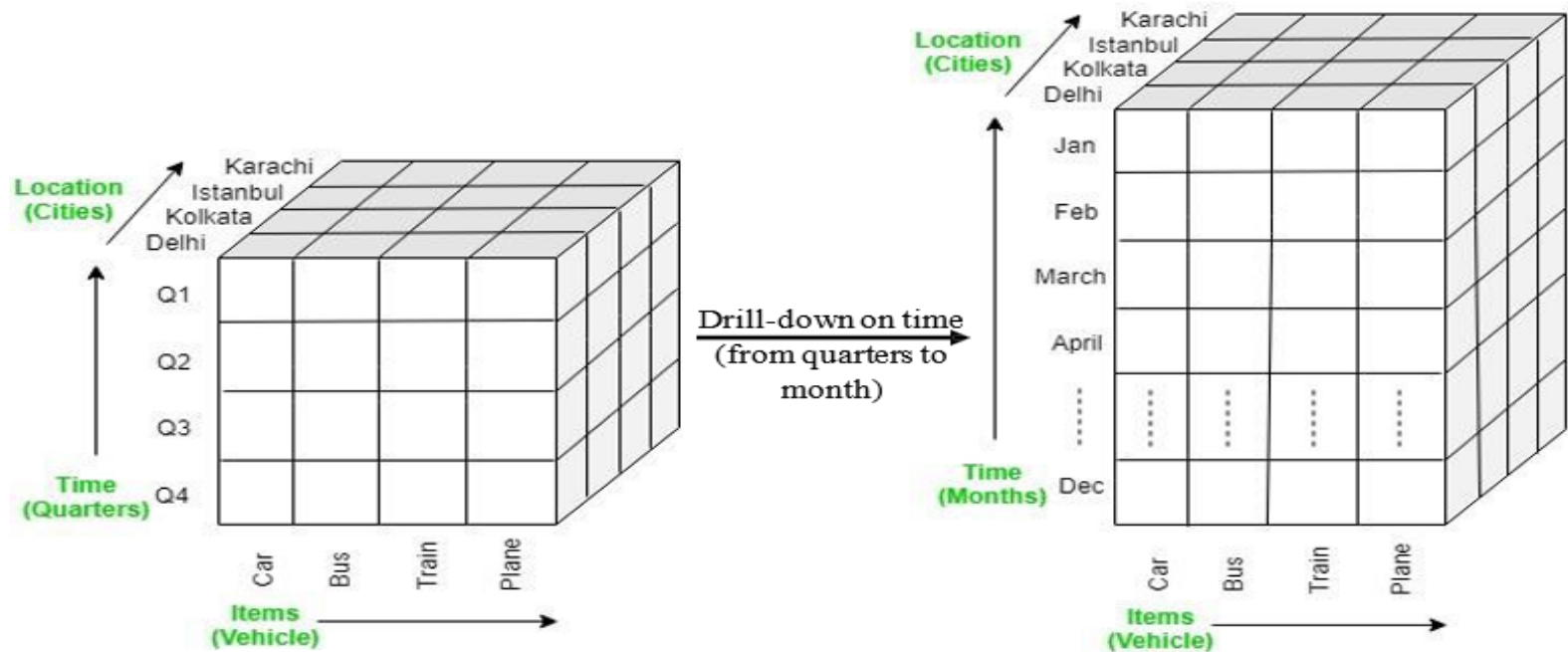


*Drill
Up*

Example:



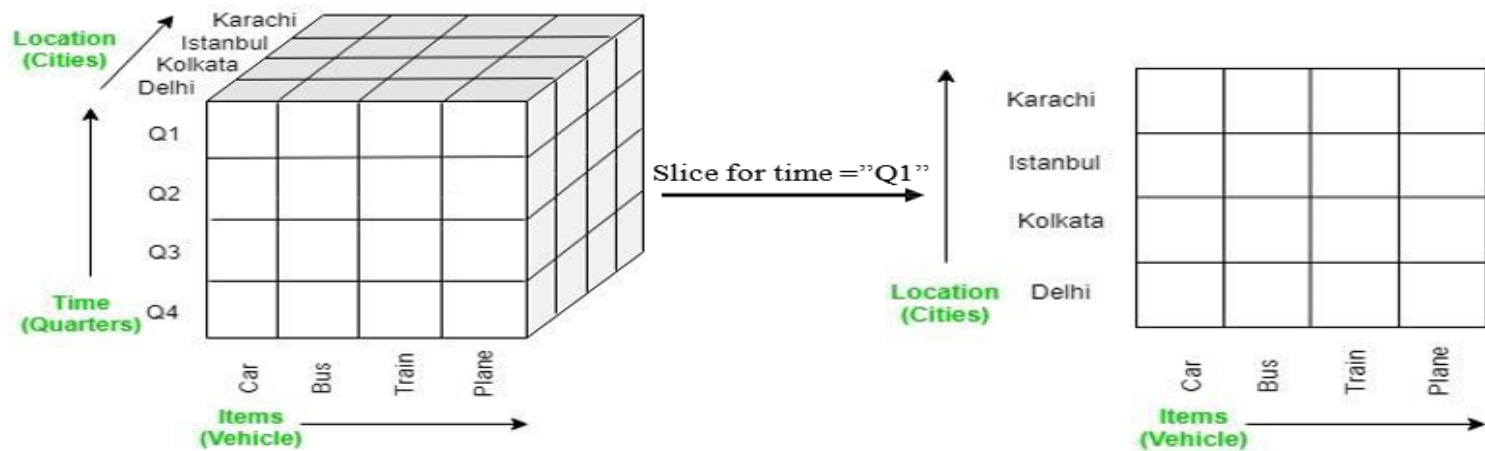
- In this example, the roll-up operation is performed by climbing up in the concept hierarchy of *Location* dimension (City -> Country).



- In this example, the drill down operation is performed by moving down in the concept hierarchy of *Time* dimension (Quarter -> Month).

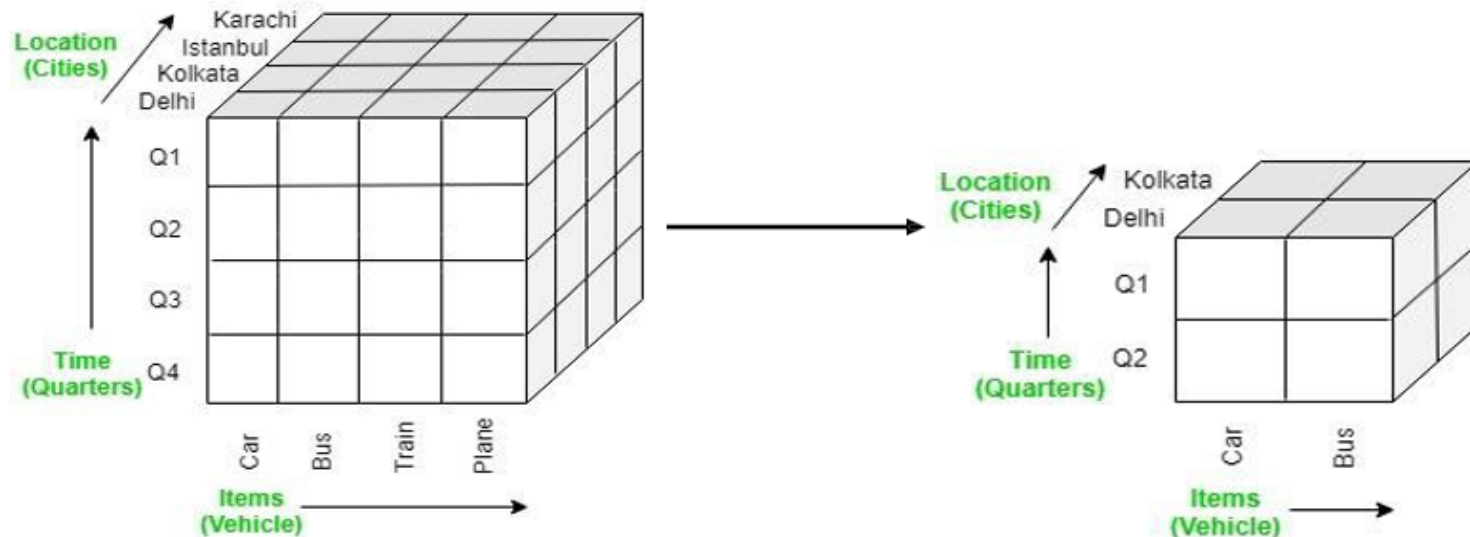
Slice and Dice

- **Slice**: a selection on one dimension of the cube resulting in subcube
- *Example:*



Slice and Dice

- **dice**: defines a subcube by performing a selection on two or more dimensions



In this example, a sub-cube is selected by selecting following dimensions with criteria:

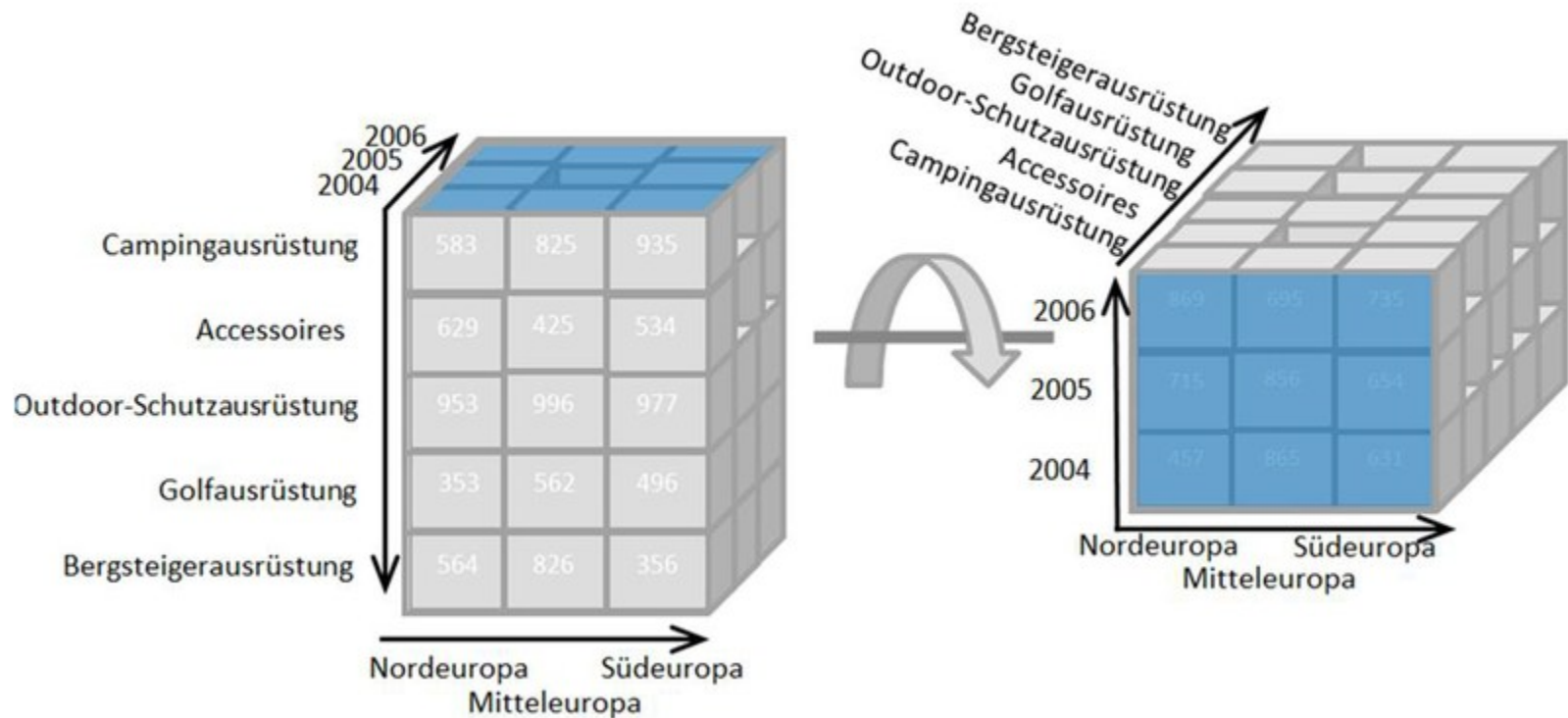
Location = "Delhi" or "Kolkata"

Time = "Q1" or "Q2"

Item = "Car" or "Bus"

Pivoting

- The pivot operation is also known as rotation. It rotates the data axis to view the data from different perspectives.



OLAP Server Architectures

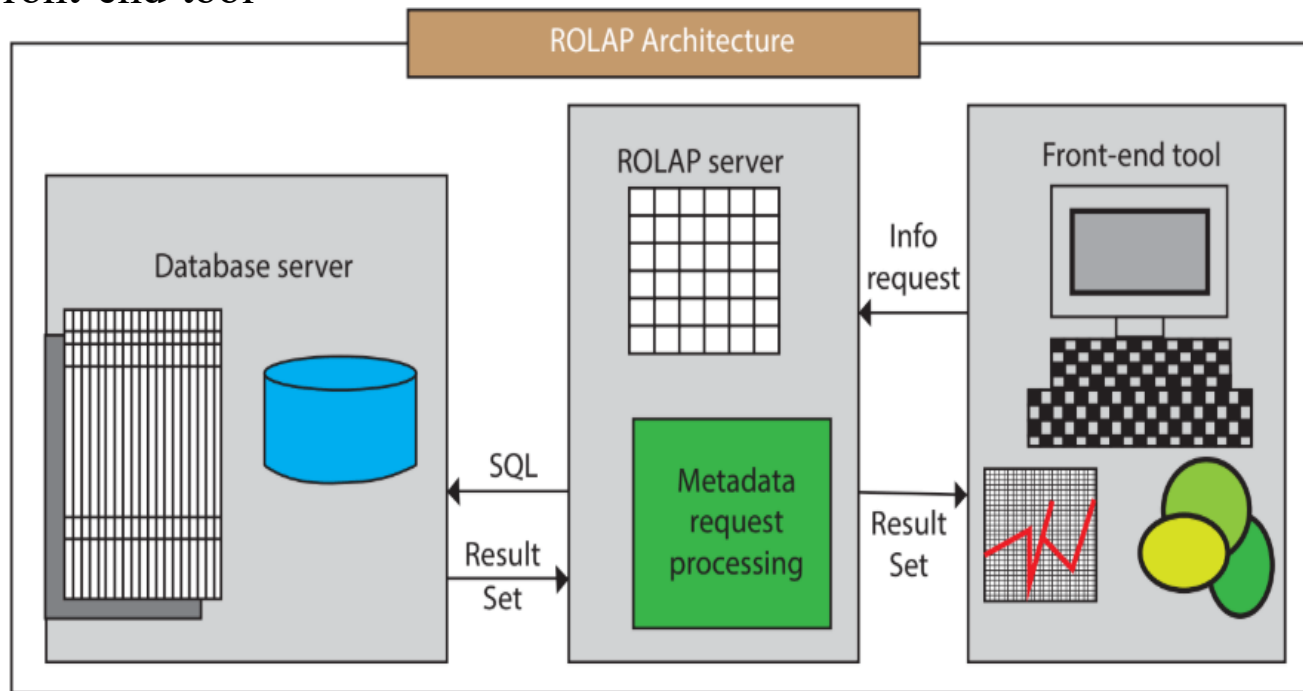
- [Relational OLAP \(ROLAP\)](#)
 - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware
 - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
 - Greater scalability
- [Multidimensional OLAP \(MOLAP\)](#)
 - Sparse array-based multidimensional storage engine
 - Fast indexing to pre-computed summarized data
- [Hybrid OLAP \(HOLAP\)](#) (e.g., Microsoft SQLServer)
 - Flexibility, e.g., low level: relational, high-level: array

ROLAP

- Relational OLAP (ROLAP) servers: These are the intermediate servers that stand in between a relational back-end server and client front-end tools.
- They use a relational or extended-relational DBMS to store and manage warehouse data, and OLAP middleware to support missing pieces.
- ROLAP servers include optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services.
- ROLAP technology tends to have greater scalability than MOLAP technology.

ROLAP

- ROLAP includes the following components:
 - Database server
 - ROLAP server
 - Front-end tool



ROLAP

- ***Advantages:***

- ROLAP servers can be easily used with existing RDBMS.
- ROLAP tools do not use pre-calculated data cubes.
- ROLAP server offers highly scalability.
- Can handle large amounts of information.

- ***Disadvantages:***

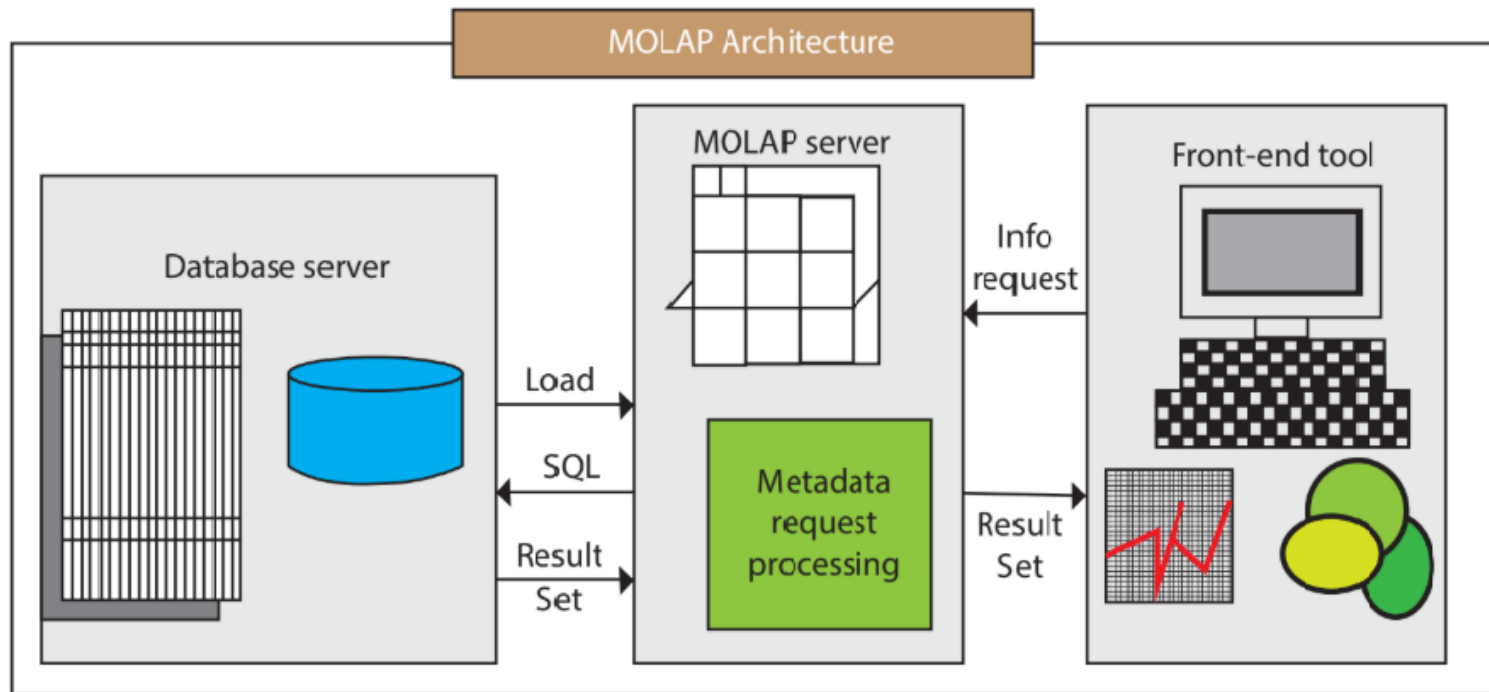
- ROLAP needs high utilization of manpower, software, and hardware resources.
- Query performance in this model is slow.
- SQL functionality is constrained.

MOLAP

- These servers support multidimensional views of data through array-based multidimensional storage engines. They map multidimensional views directly to data cube array structures. The advantage of using a data cube is that it allows fast indexing to pre computed summarized data. Notice that with multidimensional data stores, the storage utilization may be low if the data set is sparse. In such cases, sparse matrix compression techniques should be explored (Chapter 4).
- Many MOLAP servers adopt a two-level storage representation to handle dense and sparse data sets: denser subcubes are identified and stored as array structures, while sparse subcubes employ compression technology for efficient storage utilization

MOLAP

- ROLAP includes the following components:
 - Database server
 - MOLAP server
 - Front-end tool



MOLAP

- ***Advantages:***

- Fast information retrieval.
- Easier to use, therefore MOLAP is suitable for in experienced users.
- Suitable for slicing and dicing operations.
- Capable of performing complex calculations.

- ***Disadvantages:***

- MOLAP are not capable of containing detailed data.
- The storage utilization may be low if the data set is sparse.
- It is difficult to change the dimensions without re-aggregating.

HOLAP

- The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP.
- For example, a HOLAP server may allow large volumes of detail data to be stored in a relational database, while aggregations are kept in a separate MOLAP store. The Microsoft SQL Server 2000 supports a hybrid OLAP server

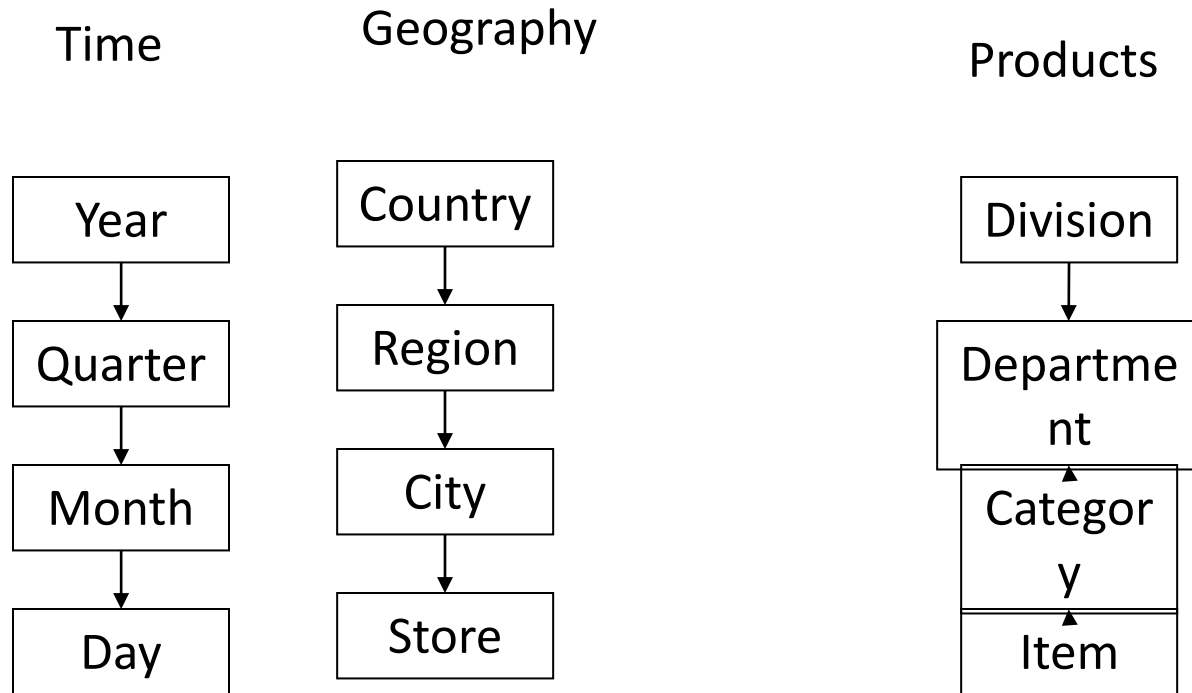
Multidimensional Data Model

- The **multidimensional data model** is an integral part of On-Line Analytical Processing, or OLAP.
- Because OLAP is on-line, it must provide answers quickly; analysts pose iterative queries during interactive sessions, not in batch jobs that run overnight.
- And because OLAP is also analytic, the queries are complex
- The multidimensional data model is designed to solve complex queries in real time.
- Logical view of the enterprise

Model Components

- Dimensions
- Attributes
- Facts
- Relationships

Multidimensional Data Model Example



Attributes

- Attributes are abstract items with business relevance that are created for convenient qualification or summarization of data on a report.
- Attribute can also be defined as column headings on a report that are not a calculation

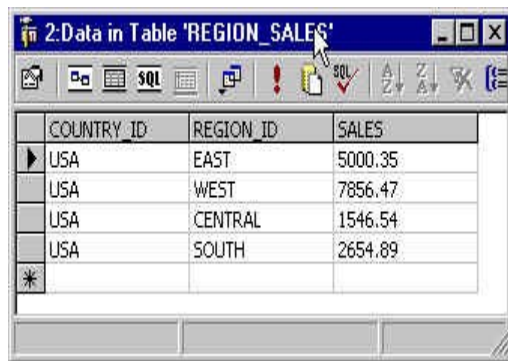
Country	Region	Metrics	Dollar Sales
		Year	
USA	North-East	1997	\$ 12,285.00
USA	North-East	1998	\$ 18,261.00
USA	Mid-Atlantic	1997	\$ 14,367.00
USA	Mid-Atlantic	1998	\$ 26,647.00
USA	South-East	1997	\$ 13,287.00
USA	South-East	1998	\$ 15,479.00
USA	Central	1997	\$ 7,141.00
USA	Central	1998	\$ 9,354.00
USA	South	1997	\$ 5,409.00
USA	South	1998	\$ 10,160.00
USA	North-West	1997	\$ 8,032.00
USA	North-West	1998	\$ 6,320.00
USA	South-West	1997	\$ 9,551.00
USA	South-West	1998	\$ 15,986.00
England	England	1997	\$ 5,595.00
England	England	1998	\$ 6,873.00
France	France	1997	\$ 4,428.00
France	France	1998	\$ 5,292.00
Germany	Germany	1997	\$ 3,617.00
Germany	Germany	1998	\$ 4,403.00

Attribute relationships

- One to One
 - Each customer has only one SSN.
- One to Many
 - Each customer can have several addresses.
- Many to Many
 - Each customer can buy many items, an item can be purchased by many customers (item means SKU, not the same physical object).
- Many to One
 - Several phone numbers can belong to one store, and one store only.

Facts

- Data columns (usually numeric) that can be used to perform calculations needed to answer business questions.
- Facts can be aggregated on different levels:



COUNTRY_ID	REGION_ID	SALES
USA	EAST	5000.35
USA	WEST	7856.47
USA	CENTRAL	1546.54
USA	SOUTH	2654.89

Aggregated
on Region
level

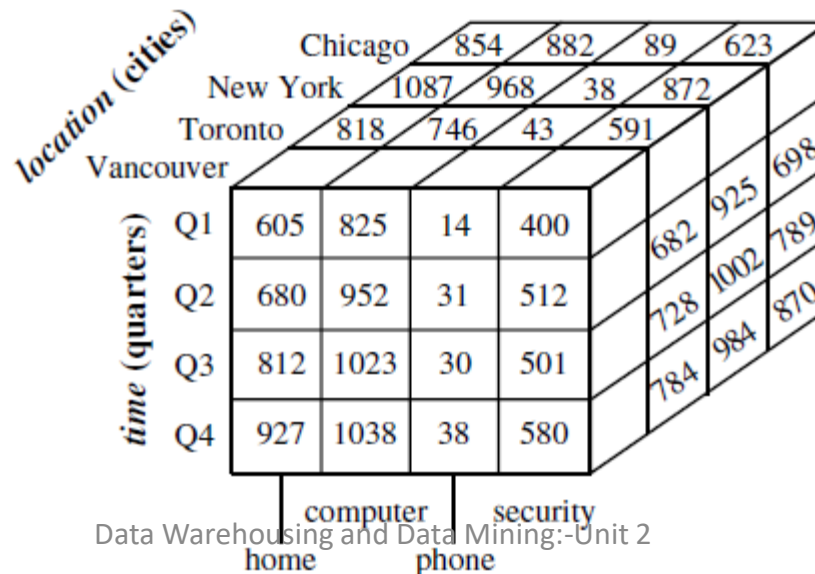


COUNTRY_ID	SALES
USA	35678.55
FRANCE	14200
UK	18550

Aggregated
on Country
level

Multidimensional Data Model

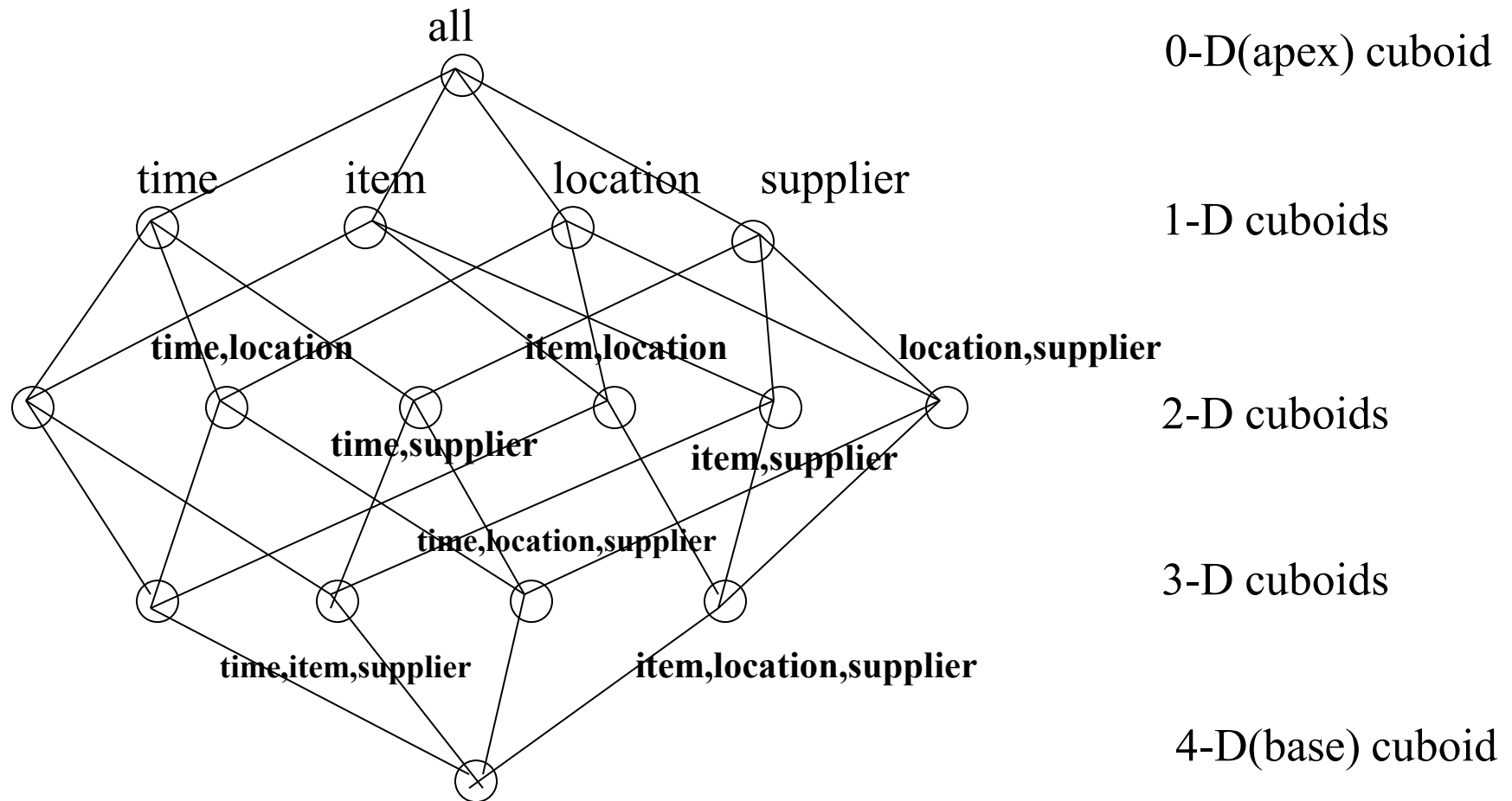
<i>location = "Chicago"</i>					<i>location = "New York"</i>					<i>location = "Toronto"</i>					<i>location = "Vancouver"</i>				
<i>item</i>					<i>item</i>					<i>item</i>					<i>item</i>				
<i>home</i>					<i>home</i>					<i>home</i>					<i>home</i>				
<i>time</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>		<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>		<i>ent.</i>	<i>comp.</i>	<i>phone</i>	<i>sec.</i>	
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			



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*.

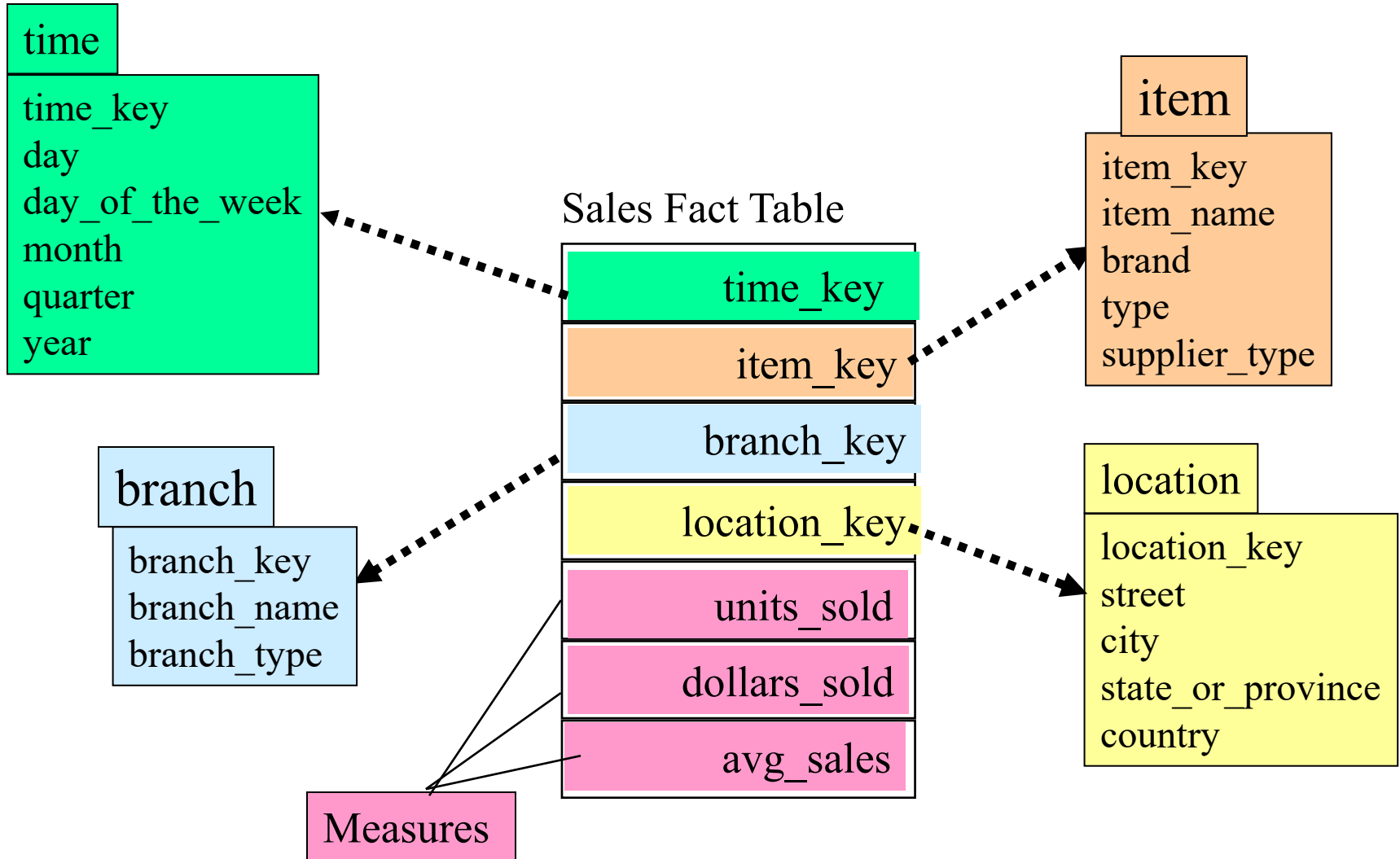
Cube: A Lattice of Cuboids



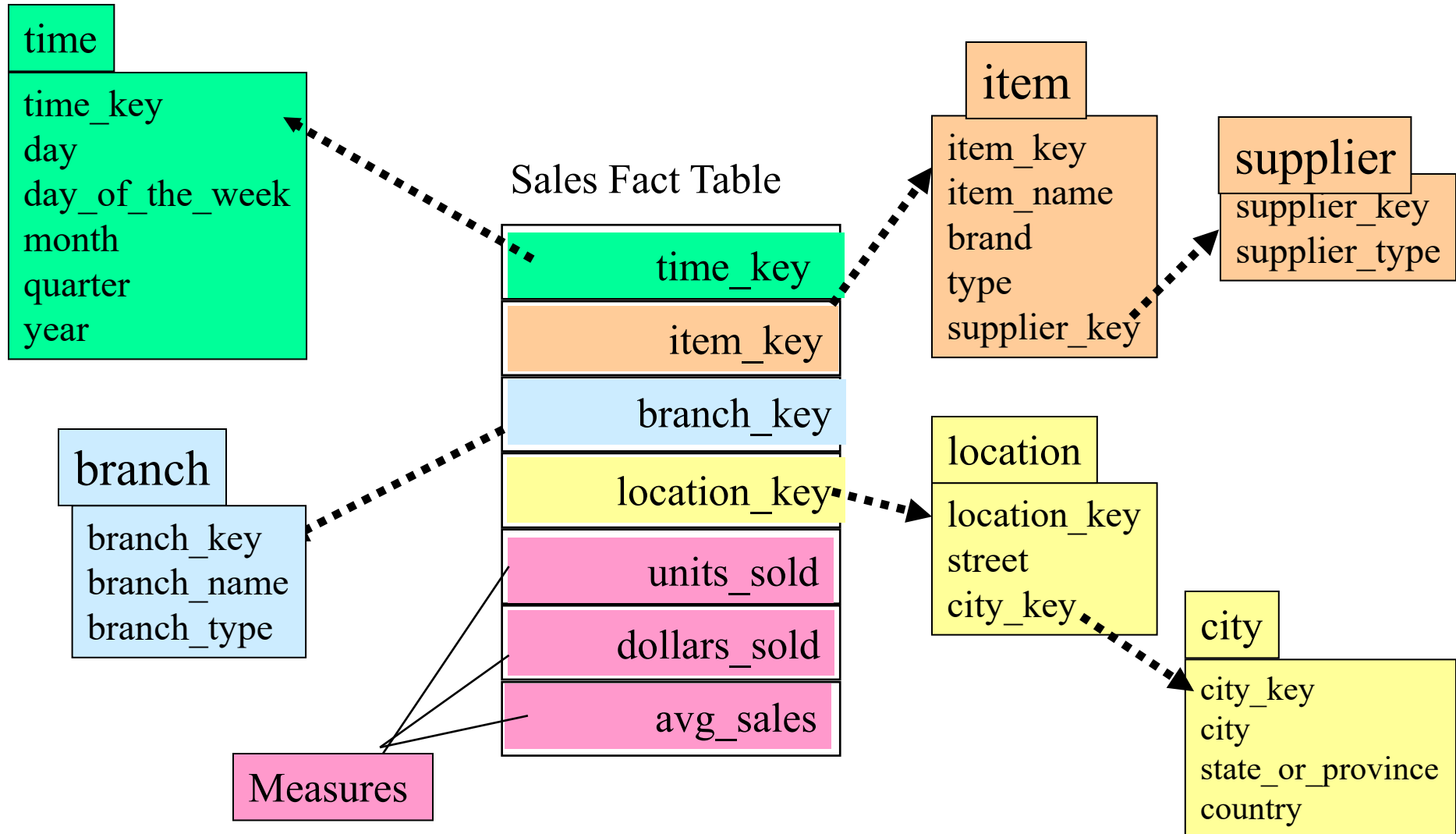
Conceptual Modeling of Data Warehouses

- 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

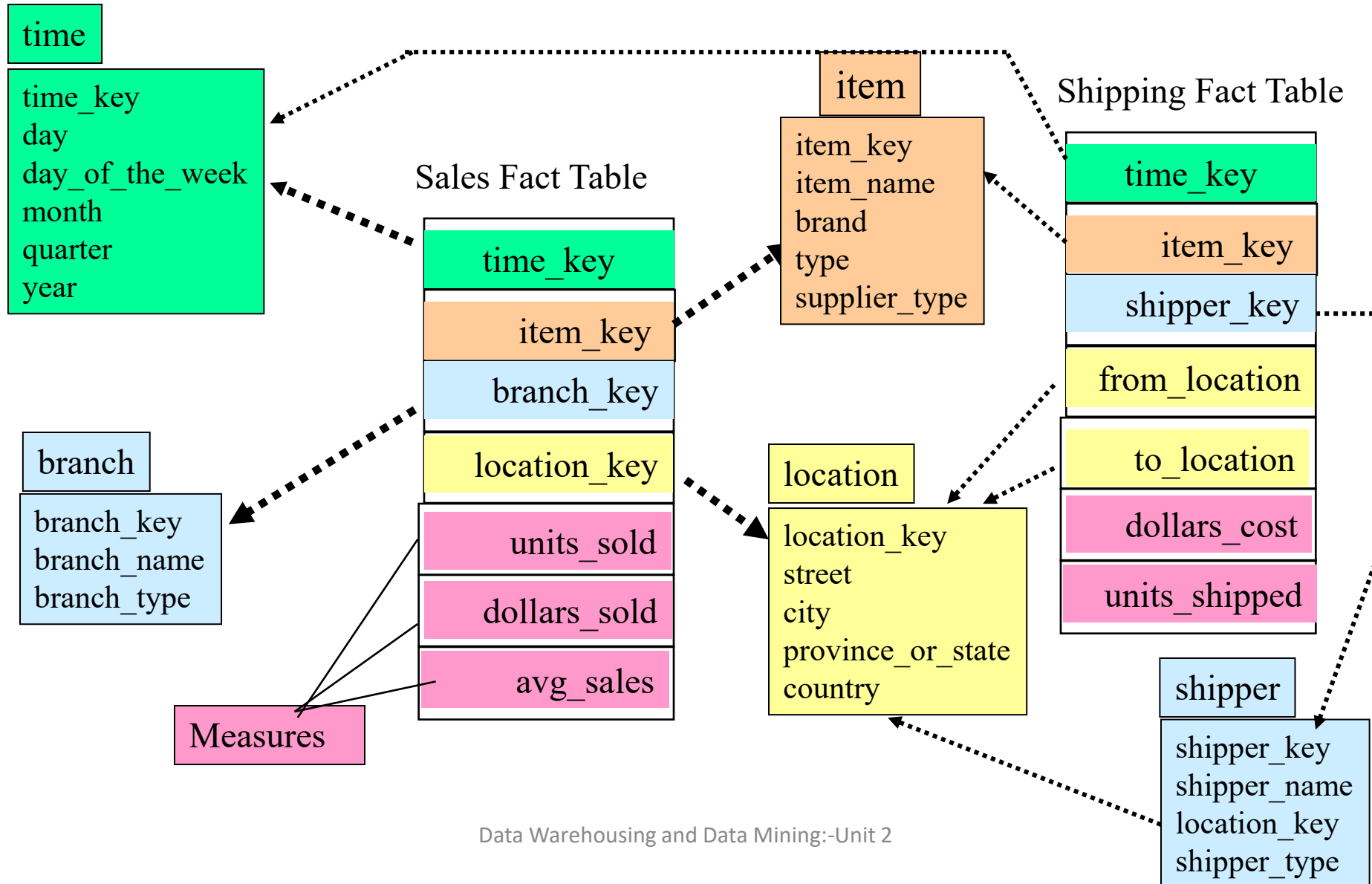
Example of Star Schema



Example of Snowflake Schema



Example of Fact Constellation



End of Session