Unit 2

Data Warehousing for Data Mining

Operational Database Systems

- The Operational Database is the source of information for the data warehouse. It includes detailed information used to run the day to day operations of the business. They cover most of the day-to-day operations of an organization such as purchasing, inventory, manufacturing, banking, payroll, registration and accounting.

- The data frequently changes as updates are made and reflect the current value of the last transactions.

- Operational Database Management Systems also called as OLTP (Online Transactions Processing Databases), are used to manage dynamic data in real-time

Data Warehouse

- Defined in many different ways:

  - 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. Inmo

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 → Integreted

- 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. Example, 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.
  - (a) Operational database:
    Current value data.
  - (b) Data warehouse data:
    Provide information from a historical perspective (e.g, past 5-10 years)

Data Warehouse → Non Volatile

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
    a) Does not require transaction processing, recovery, and concurrency control mechanisms
    b) Requires only two operations in data accessing: initial loading of data and access of data.

Data Warehousing → Two Distinct Issues

- How to get information into warehouse "Data warehousing"
- What to do with data once it's in warehouse "Warehouse DBMS"

→ Both rich research areas

→ Industry has focused on 2

Data Warehouse vs Database

	Data Warehouse	Database
Purpose	Analysis, Decision making	Day to day use
Support For	OLAP(on-line analytical processing)	OLTP(on-line transaction processing)
Data model	Multi-dimentional	Rational
Age of data	Current & time series	Current & real time
Data modification	Read/access only	Insert, update, delete
Type of data	Static	Dynamic
Amount of data per transaction	Larger	Smaller

Differences between Operational Database Systems and Data Warehouses

Operational Database	Data Warehouse
Operational systems are designed to support high-volume transaction processing.	Data warehousing systems are typically designed to support high-volume analytical processing (i.e., OLAP).
Operational systems are usually concerned with current data.	Data warehousing systems are usually concerned with historical data.
Data within operational systems are mainly updated regularly according to need.	Non-volatile, new data may be added regularly. Once Added rarely changed.
It is designed for real-time business dealing and processes.	It is designed for analysis of business measures by subject area, categories, and attributes.

Differences between Operational Database Systems and Data Warehouses

Operational Database	Data Warehouse
It is optimized for a simple set of transactions, generally adding or retrieving a single row at a time per table.	It is optimized for extent loads and high, complex, unpredictable queries that access many rows per table.
Operational systems are widely process-oriented.	Data warehousing systems are widely subject-oriented
Less Number of data accessed.	Large Number of data accessed.
Relational databases are created for Online Transactional Processing (OLTP)	Data Warehouse designed for on-line Analytical Processing (OLAP)

Data Warehouse Systems

- Data Warehouse Systems serve users or knowledge workers in the purpose of data analysis and decision-making.
- Such systems can organize and present information in specific formats to accommodate the diverse needs of various users.
- These systems are called as Online-Analytical Processing (OLAP) Systems.

Differences between OLTP and OLAP

	OLTP	OLAP
Characteristic	It is a system which is used to manage operational Data.	It is a system which is used to manage informational Data.
Users	Clients, and information technology professionals.	Knowledge workers, including managers, executives, and analysts.
System orientation	OLTP system is a customer-oriented, transaction, and query processing are done by clerks, clients, and information technology professionals.	OLAP system is market-oriented, knowledge workers including managers, do data analysts executive and analysts.
Database Size	100 MB-GB	100 GB-TB

Differences between OLTP and OLAP

	OLTP	OLAP
Database design	OLTP system usually uses an entity-relationship (ER) data model and application-oriented database design.	OLAP system typically uses either a star or snowflake model and subject-oriented database design.
Normalization	Fully Normalized	Partially Normalized
Processing Speed	Very Fast	It depends on the amount of files contained, batch data refresh, and complex query may take many hours, and query speed can be upgraded by creating indexes.

Data-warehousing architecture

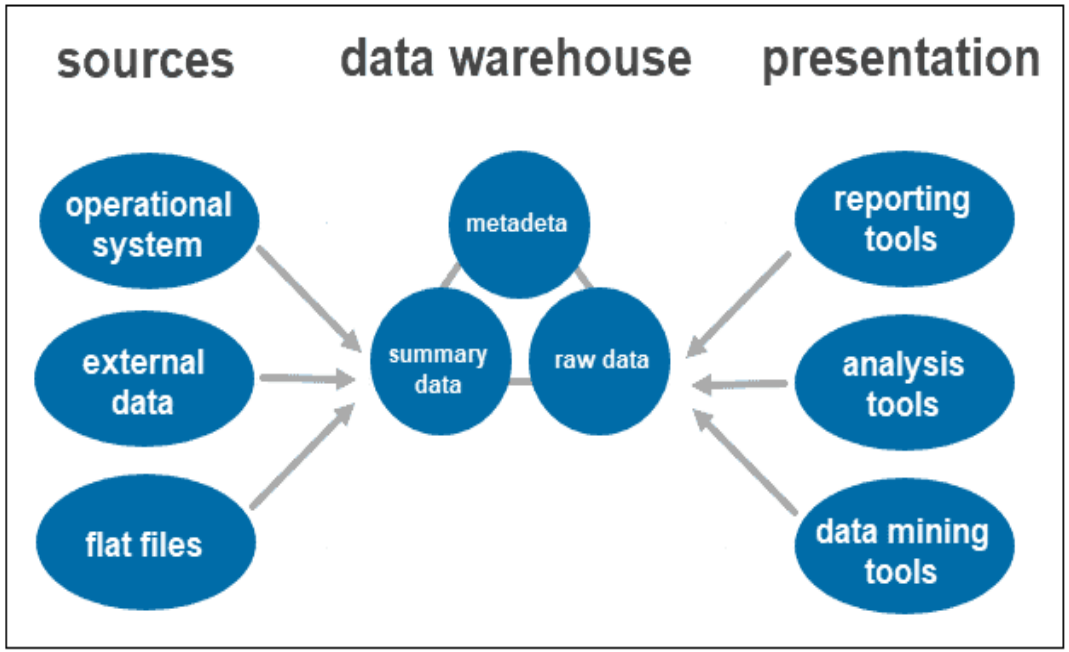
- In most of the organization, there occur large databases in operation for normal daily transactions called operational database. A data warehouse is a large database built from the operational database.
- "A **data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.**"

Data-warehousing architecture

- A data warehouse is a complex system that stores historical and cumulative data used for forecasting, reporting, and data analysis.
- It involves collecting, cleansing, and transforming data from different data streams and loading it into fact/dimensional tables.
- There are three ways to construct a data warehouse system.
- These approaches are classified by the number of tiers in the architecture.
    a) Single-tier architecture
    b) Two-tier architecture
    c) Three-tier architecture

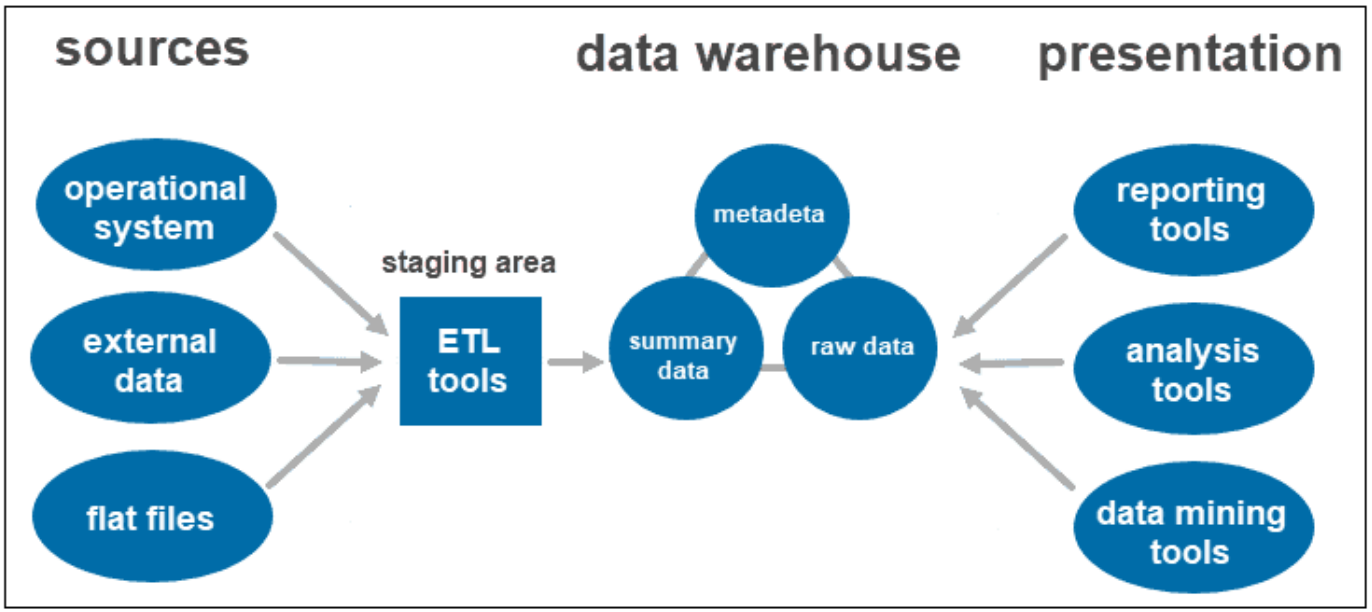
Data-warehousing → Single Tier Architecture

The single-tier architecture is not a frequently practiced approach



Data-warehousing → Two Tier Architecture

A two-tier architecture includes a staging area for all data sources, before the data warehouse layer.



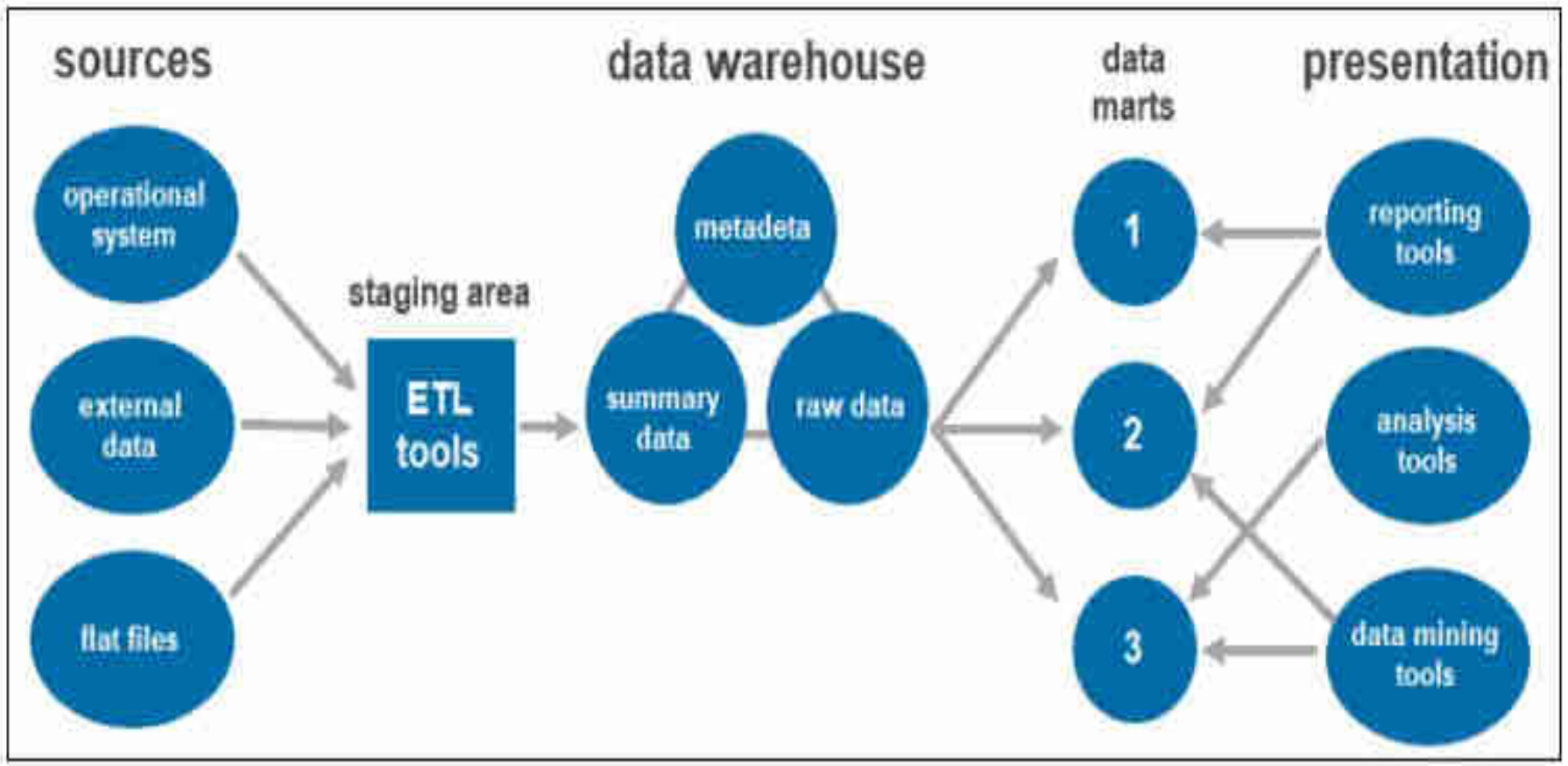
Data-warehousing → Two Tier Architecture

- By adding a staging area between the sources and the storage repository, you ensure all data loaded into the warehouse is cleansed and in the appropriate format.
- It includes ETL(Extract Transfer and Load).
- This approach has certain network limitations. Additionally, you cannot expand it to support a larger number of users.

Data-warehousing → Three Tier Architecture

- The three-tier approach is the most widely used architecture for data warehouse systems.
- It consists of three tiers:
  a) **The bottom tier** is the database of the warehouse, where the cleansed and transformed data is loaded.
  b) **The middle tier** is the application layer giving an abstracted view of the database. It arranges the data to make it more suitable for analysis. This is done with an OLAP server, implemented using the ROLAP or MOLAP model.
  c) **The top-tier** is where the user accesses and interacts with the data. It represents the front-end client layer. In this tier reporting tools, query, analysis or data mining tools is used to represent knowledge.

Data-warehousing → Three Tier Architecture

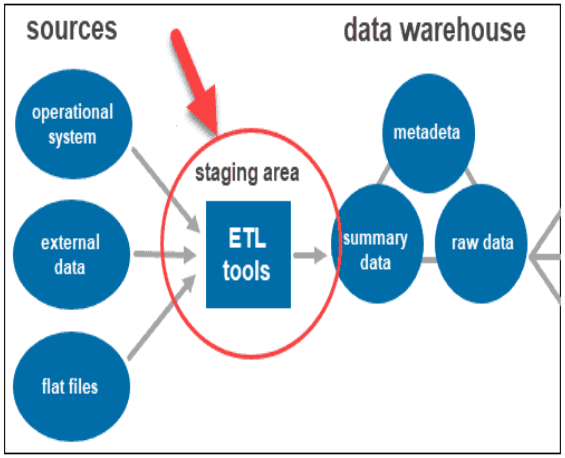


Data-warehouse → Components

- From the architectures outlined above, some components overlap, while others are unique to the number of tiers.
- Most important data warehouse components are:
  - **ETL Tool**
  - **Data**
  - **Access Tools**
  - **Data Marts**

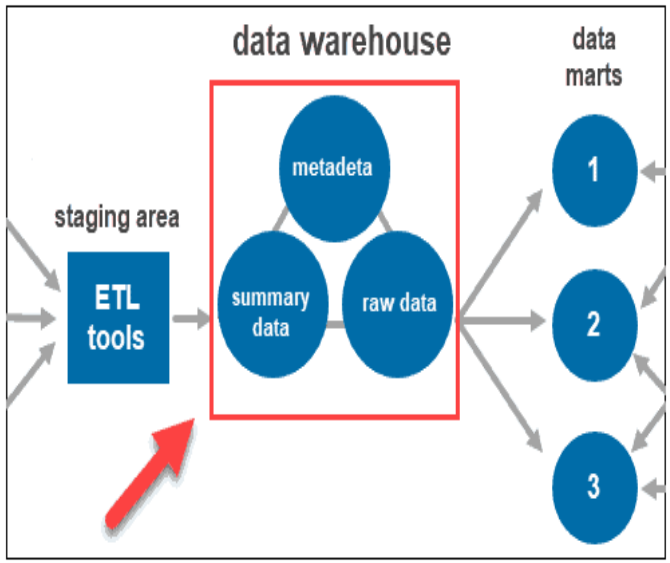
Data-warehouse → Components(ETL Tools)

- ETL stands for Extract, Transform, and Load.
- The staging layer uses ETL tools to extract the needed data from various formats and checks the quality before loading it into the data warehouse.
- The data coming from the data source layer can come in a variety of formats.
- Before merging all the data collected from multiple sources into a single database, the system must clean and organize the information.



Data-warehouse → Components(Data)

- Once the system cleans and organizes the data, it stores it in the data warehouse.
- The data warehouse represents the central repository that stores metadata, summary data, and raw data coming from each source.

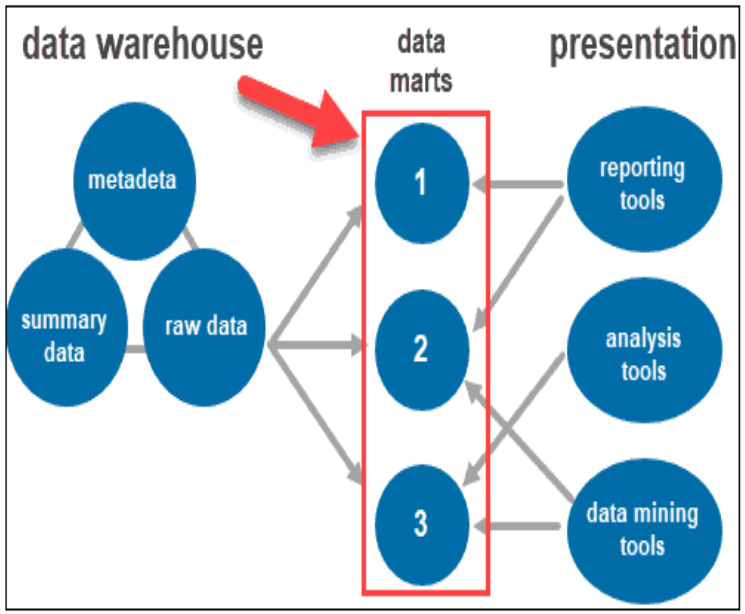


Data-warehouse → Components(Access Tools)

- Users interact with the gathered information through different tools and technologies.
- They can analyze the data, gather insight, and create reports.

Data-warehouse → Components(Data Marts)

- A **data mart contains a subset of corporate-wide data that is of value to a specific group of users.**
- The scope is confined to specific selected subjects. For example, a marketing data mart may confine its subjects to customer, item, and sales. The data contained in data marts tend to be summarized.
- Data marts are usually implemented on low-cost departmental servers that are Unix/Linux or Windows based.



Data-warehouse → Decision Support System

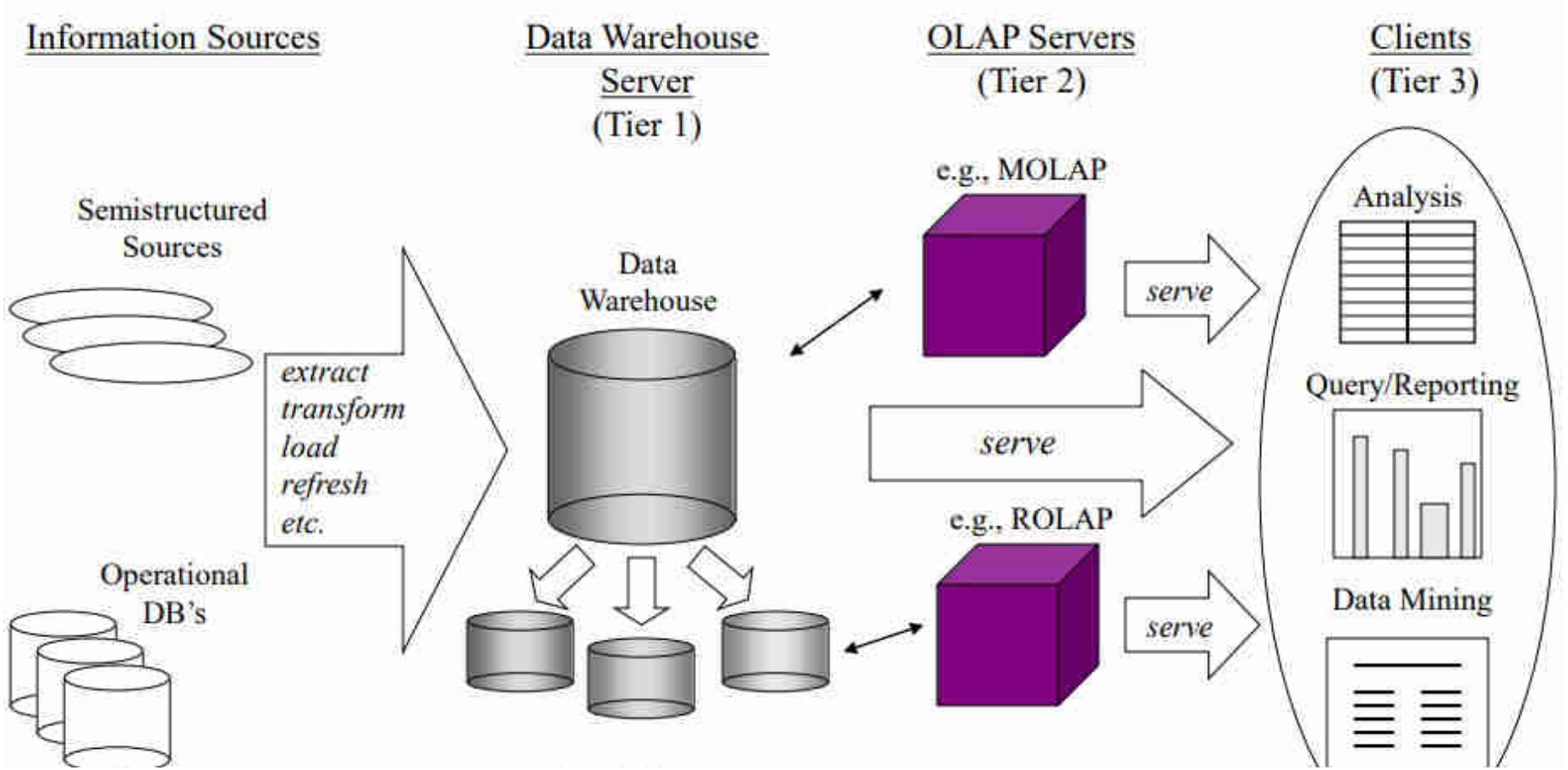
- Information technology to help the knowledge worker (executive, manager, analyst) make faster & better decisions
  - "What were the sales volumes by region and product category for the last year?"
  - "How did the share price of comp. manufacturers correlate with quarterly profits over the past 10 years?"

- On-line analytical processing (OLAP) is an element of decision support systems (DSS)

Data-warehouse → Three Tier Decision Support System

- Warehouse database server
  - Almost be always a relational DBMS, rarely flat files
- OLAP servers
  - Relational OLAP (ROLAP): extended relational DBMS that maps operations on multidimensional data to standard relational operators
  - Multidimensional OLAP (MOLAP): special-purpose server that directly implements multidimensional data and operations
- Clients
  - Query and reporting tools
  - Analysis tools
  - Data mining tools

The Complete Decision Support System



Why Separate is Data Warehouse Needed?

Why not perform online analytical processing directly on operational databases instead of spending additional time and resources to construct a separate data warehouse?

For High performance for both system

a) DBMS → tuned for OLTP : searching for particular records, indexing, hashing, recovery

b) Warehouse → tuned for OLAP : Complex OLAP queries, multidimensional view, summerization and aggregation

- **Requirements analysis and capacity planning:**

  In the first step in data warehousing involves defining enterprise needs, defining architecture, carrying out capacity planning and selecting the hardware and software tools. This step will involve consulting senior management as well as the various stakeholders.

- **Hardware integration:**

  Once the hardware and software have been selected, they need to be put together by integrating the servers, the storage devices and the client software tools.

- **Modeling:**

  Modeling is a major step that involves designing the warehouse schema and views. This may involve using a modeling tool if the data warehouse is complex.

- **Physical modeling:**

  For the data warehouse to perform efficiently, physical modeling is required. This involves designing the physical data warehouse organization, data placement, data partitioning, deciding on access methods and indexing.

- **Sources:**

  The data for the data warehouse is likely to come from a number of data sources. This step involves identifying and connecting the sources using gateways, ODBC drives or other wrappers.

- **ETL:**

  The data from the source systems will need to go through an ETL process. The step of designing and implementing the ETL process may involve identifying a suitable ETL tool vendor and purchasing and implementing the tool. This may include customizing the tool to suit the needs of the enterprise.

- **Populate the data warehouse:**

  Once the ETL tools have been agreed upon, testing the tools will be required. Once everything is working satisfactorily, the ETL tools may be used in populating the warehouse given the schema and view definitions.

- **User applications:**

  For the data warehouse to be useful there must be end-user applications. This step involves designing and implementing applications required by the end users.

- **Roll-out the warehouse and applications:**

  Once the data warehouse has been populated and the end-user applications tested, the warehouse system and the applications may be rolled out for the user community to use.

- Build incrementally
- Senior management support
- Ensure quality
- Corporate strategy
- Training
- Adaptability
- Joint management

Data Warehouse Models

From the architecture point of view, there are three data warehouse models:

a) The enterprise warehouse
b) The data mart
c) The virtual warehouse

Data Warehouse Models → Enterprise Warehouse

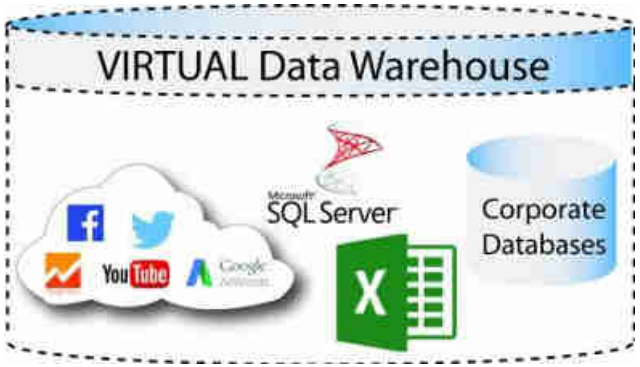
- An enterprise warehouse collects all of the information about subjects spanning the entire organization.
- It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope.
- It typically contains detailed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

Data Warehouse Models → Data Mart

???

Data Warehouse Models → Virtual Warehouse

- Data Virtualization makes all data, regardless of where it's located, look as if it is one place and in a consistent format. It provides access to data directly from one or more disparate data sources, without physically moving the data. The technical aspects of location, structure, and access language are transparent to the analyst.
- A virtual data warehouse is a set of separate databases, which can be queried together, so a user can effectively access all the data as if it was stored in one data warehouse



Multidimensional Data Model

- Multidimensional Data Model can be defined as a method for arranging the data in the database, with better structuring and organization of the contents in the database. Unlike a system with one dimension such as a list, the Multidimensional Data Model can have two or three dimensions of items from the database system
- Data warehouses and OLAP tools are based on a multidimensional data model. This model views data in the form of a *data cube*.

- ***What is a data cube?***

  A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

Multidimensional Model → Data Cube

- A data cube is defined by facts and dimensions.
  - Facts are data which data warehouse focus on.
    → Fact tables contain numeric measures and keys to each of the related dimension tables
  - Dimensions are perspective with respect to facts.
    → Dimension tables describe the dimension with attributes . For example, item(item_name, brand, type)

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
. . .	. . .	. . .	. . .	. . .	. . .

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
. . .	. . .	. . .

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .

items_sold

trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
. . .	. . .	. . .

works_at

empl_ID	*branch_ID*
E55	B1
. . .	. . .

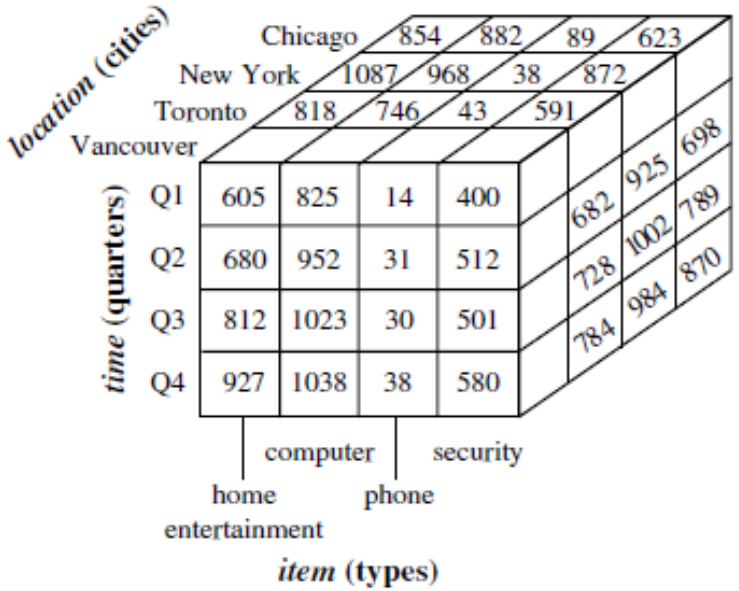
From table and Spreadsheet to Data cube

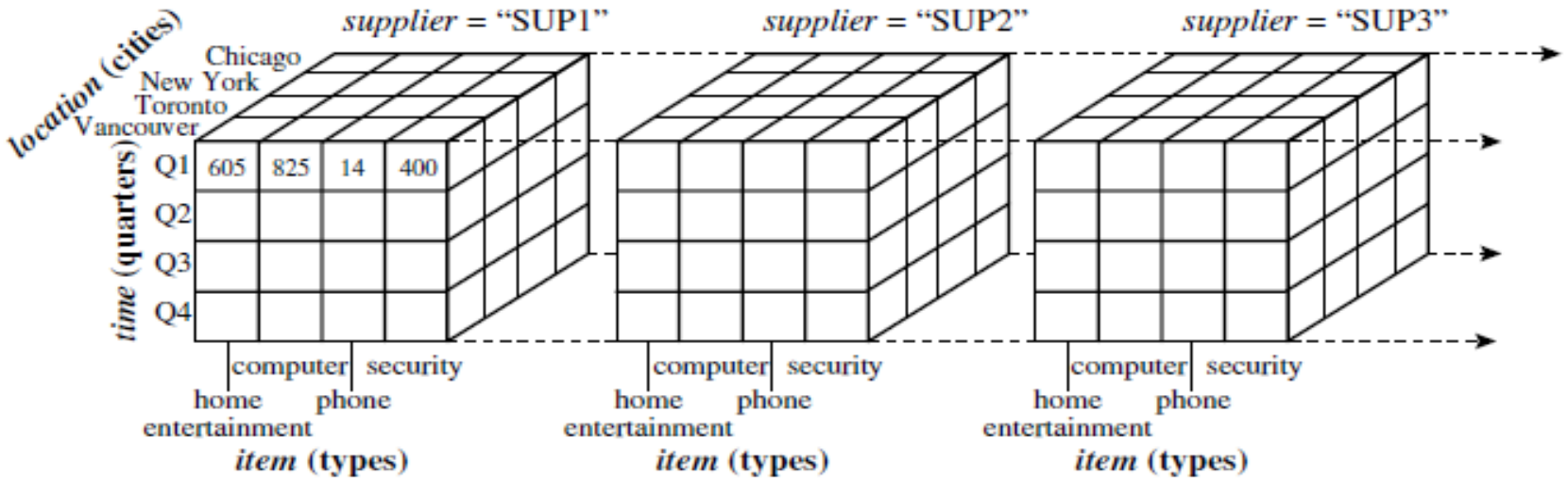
	location = "Vancouver"			
	item (type)			
time (quarter)	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

A 2-D view of sales data for *AllElectronics* according to the dimensions *time* and *item*, where the sales are from branches located in the city of Vancouver

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

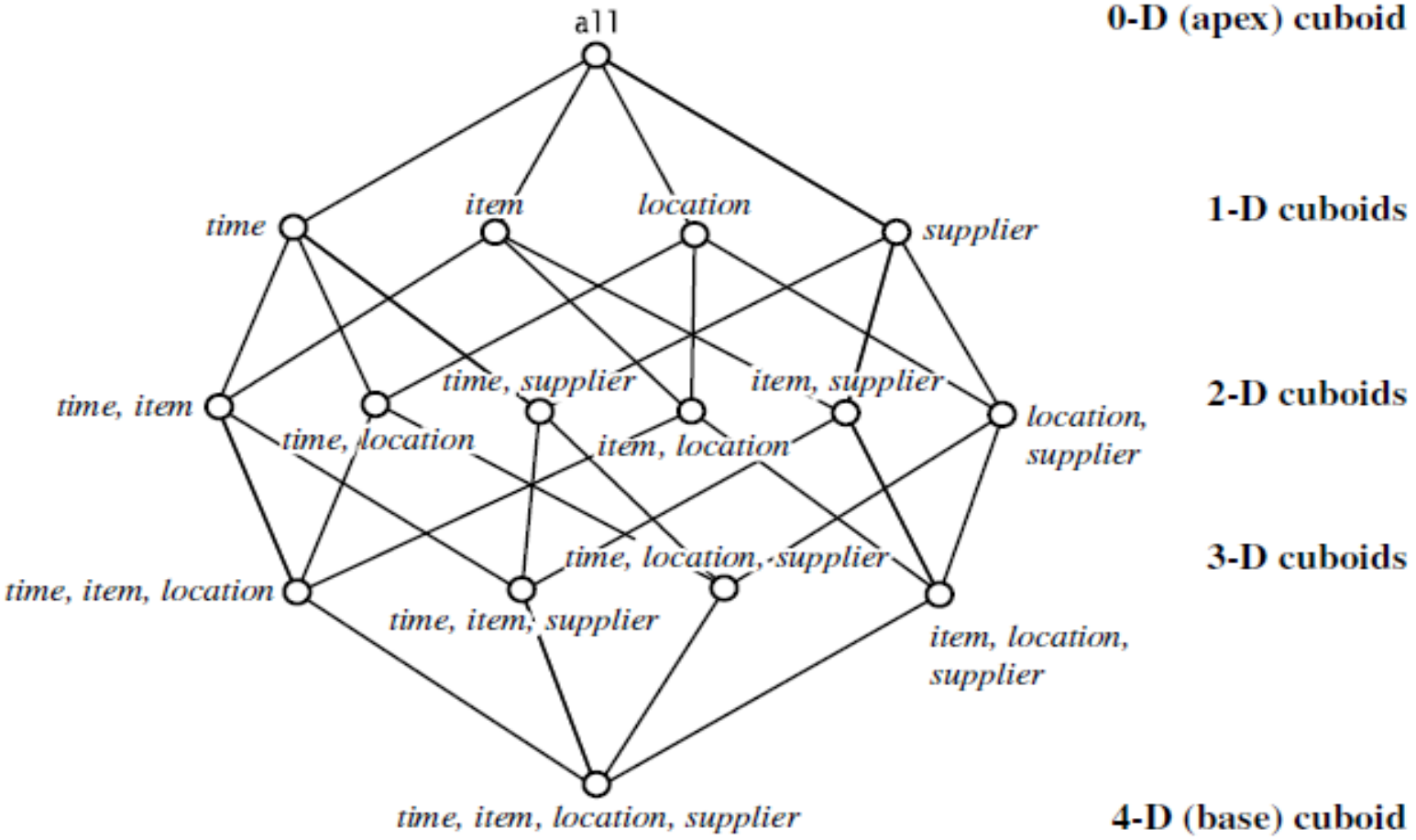
A 3-D data cube representation of the data according to the dimensions *time*, *item*, and *location*.





A 4-D data cube representation of sales data, according to the dimensions *time*, *item*, *location*, and *supplier.*

Lattice of cuboids



The cuboid that holds the lowest level of summarization is called the **base cuboid**. For example, the 4-D cuboid in Figure is the base cuboid for the given *time*, *item*, *location*, and *supplier* dimensions. The 0-D cuboid, which holds the highest level of summarization, is called the **apex cuboid.**

Curse of Dimensionality

a) Even in the simplest case of d binary variables, the number of possible combinations already is 2^d, exponential in the dimensionality.
b) There is an exponential increase in volume associated with adding extra dimensions to a mathematical space.
c) When solving dynamic optimization problems by numerical backward induction, the objective function must be computed for each combination of values

Metadata Repository → Metadata

- Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects.
- Metadata are created for the data names and definitions of the given warehouse.
- Additional metadata are created and captured for timestamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

Metadata Repository

A metadata repository should contain the following:

- A description of the data warehouse structure, 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)
- The algorithms used for summarization, which include measure and dimension definition algorithms

Data Warehouse Manager

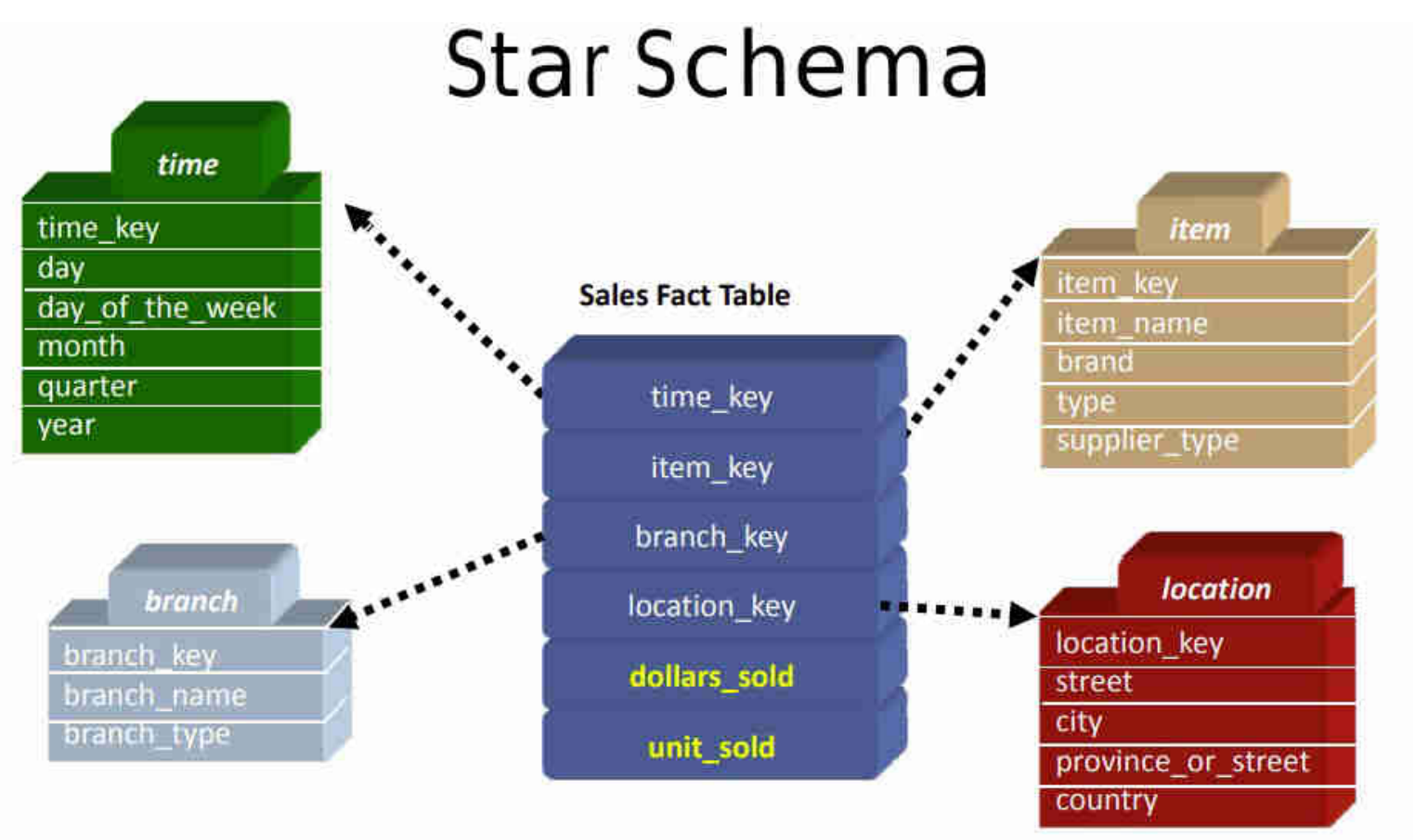
- Data Warehousing Manager manages the daily activities of the team responsible for the design, implementation, maintenance, and support of data warehouse systems and related data marts.

- Responsibilities:
    a) Monitor all long term objectives
    b) Train data warehouse staff
    c) Ensure appropriate maintenance and development of all data.
    d) Evaluate staff performance
    e) Administer database consolidation
    f) Administer all Service Level Agreements
    g) Maintain various big enterprises

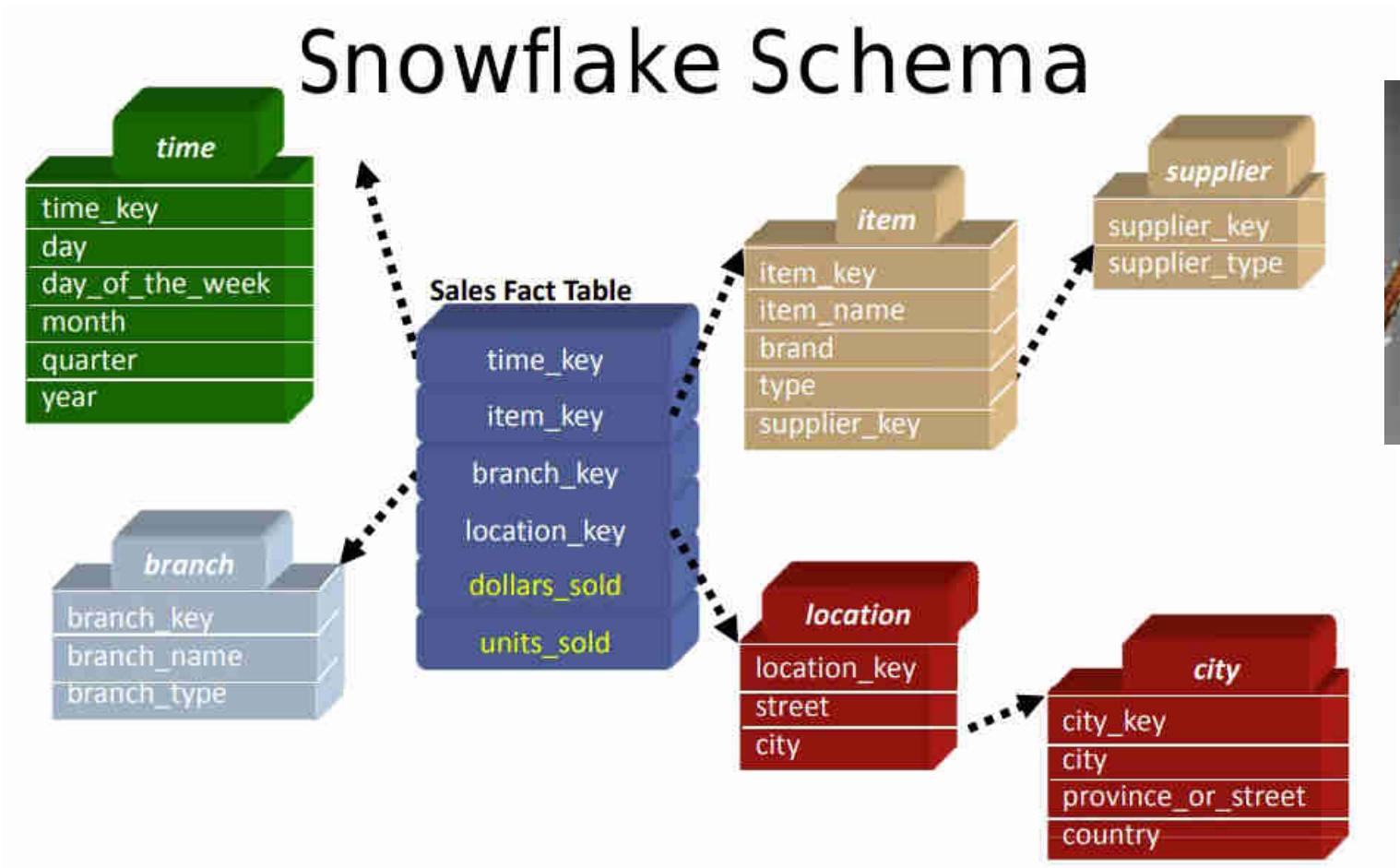
Schema → Conceptual Modeling of Data Warehouses

- Modeling data warehouse : dimension and measures
  - a) **Star Schema**
    A fact table in the middle connected to a set of dimension tables
  - b) **Snowflake schema**
    A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension table forming a shape similar to snowflakes
  - c) **Fact Constellation**
    Multiple fact tables share dimension tables, viewed as collection of stars, therefore called galaxy schema or fact constellation

Star Schema


Star Schema
time
time_key
day
day_of_the_week
month
quarter
year
Sales Fact Table
time_key
item_key
branch_key
location_key
dollars_sold
unit_sold
item
item_key
item_name
brand
type
supplier_type
branch
branch_key
branch_name
branch_type
location
location_key
street
city
province_or_street
country

Snowflake Schema



Fact Constellation Schema/Galaxy Schema