UNIT-2

Data Warehouse

- Data Warehouse : Basic Concepts
  - What is Data Warehouse
  - Difference between operational database system and data warehouses
  - Multitier Architecture
- Data Warehouse Modeling: Data cube and OLAP
  - Data Cube: Multidimensional Data Model
  - Typical OLAP Operations
- Data Warehouse Design and Usage
  - A business Analysis Framework for Data Warehouse Design
  - Data warehouse Design Process
  - Data Warehouse USAGE FOR Information Processing
  - From OLAP to Multidimensional data Mining

Unit –2 Data Warehouse

2.1 Data Warehouse: Basic Concepts:

➢ Basic Concepts:

- If we talk about, a Data Base Management System then it stores data in the form of tables, uses ER model and its goal is to fulfill the ACID properties. For example, a database system of an institute has tables for courses, students, faculty members etc.
- While, a Data Warehouse is separate from DBMS. It stores huge amount of data, which is typically collected from multiple heterogeneous sources like files, DBMS etc. The goal is to produce statistical results that may help in decision making. For example, an institute might want to see overall results, course wise placement records that have been improved over last few years etc.
- An ordinary Database can store MBs to GBs of data for a specific purpose. For storing data of TB size, the storage shifted to Data Warehouse. A transactional database doesn't offer itself to analyze. To effectively perform analysis, an organization keeps a central Data Warehouse to closely study its business by organizing, understanding and using its historic data for taking strategic decisions and scrutinizing trends.
- The data warehouse enables the organization to make use of an enterprise-wide data store to link information from diverse sources and make the information accessible to the users for strategic analysis that includes trend analysis, forecasting, competitive analysis, and targeted market research etc.
- The basic concept of a Data Warehouse is to facilitate a single version of truth for a company for decision making and forecasting. A Data warehouse is an information system that contains historical and commutative data from single or multiple sources. Data Warehouse concept, simplifies reporting and analysis process of the organization.
- Data warehouse allows business users to quickly access critical data from some sources all in one place. It provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and query.
- It helps to integrate many sources of data to reduce stress on the production system. It helps to reduce total turnaround time for analysis and reporting.
- It allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.
- It stores a large amount of historical data. This helps users to analyze different time periods and trends to make future predictions.

Unit –2 Data Warehouse

2.1 Data Warehouse: Basic Concepts:

➢ Basic Concepts:

- If we talk about, a Data Base Management System then it stores data in the form of tables, uses ER model and its goal is to fulfill the ACID properties. For example, a database system of an institute has tables for courses, students, faculty members etc.
- While, a Data Warehouse is separate from DBMS. It stores huge amount of data, which is typically collected from multiple heterogeneous sources like files, DBMS etc. The goal is to produce statistical results that may help in decision making. For example, an institute might want to see overall results, course wise placement records that have been improved over last few years etc.
- An ordinary Database can store MBs to GBs of data for a specific purpose. For storing data of TB size, the storage shifted to Data Warehouse. A transactional database doesn't offer itself to analyze. To effectively perform analysis, an organization keeps a central Data Warehouse to closely study its business by organizing, understanding and using its historic data for taking strategic decisions and scrutinizing trends.
- The data warehouse enables the organization to make use of an enterprise-wide data store to link information from diverse sources and make the information accessible to the users for strategic analysis that includes trend analysis, forecasting, competitive analysis, and targeted market research etc.
- The basic concept of a Data Warehouse is to facilitate a single version of truth for a company for decision making and forecasting. A Data warehouse is an information system that contains historical and commutative data from single or multiple sources. Data Warehouse concept, simplifies reporting and analysis process of the organization.
- Data warehouse allows business users to quickly access critical data from some sources all in one place. It provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and query.
- It helps to integrate many sources of data to reduce stress on the production system. It helps to reduce total turnaround time for analysis and reporting.
- It allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.
- It stores a large amount of historical data. This helps users to analyze different time periods and trends to make future predictions.

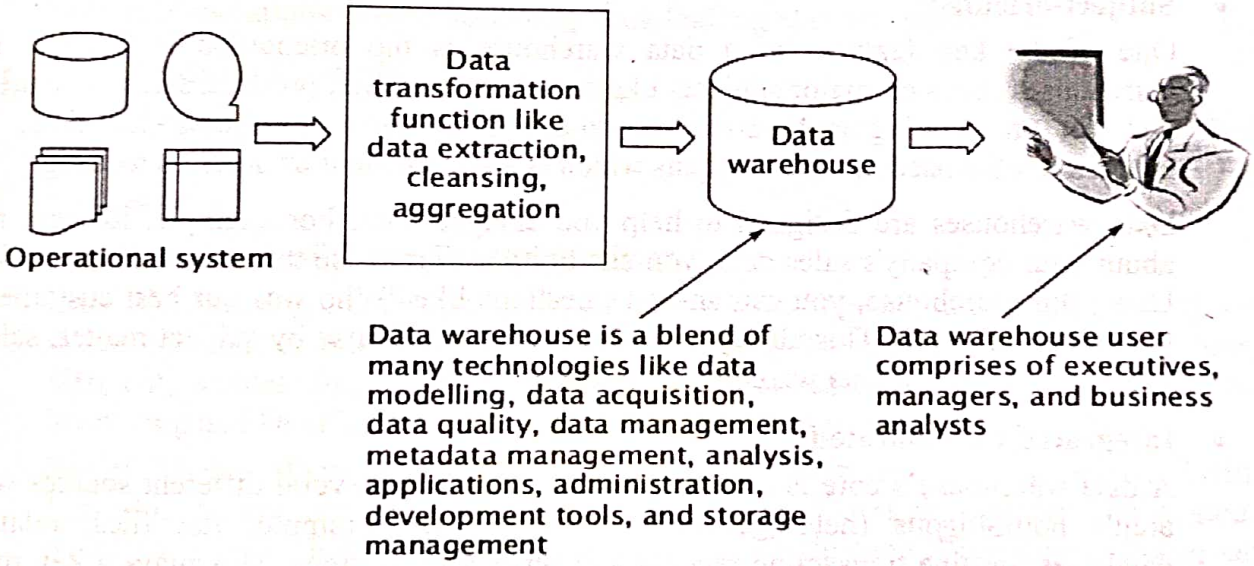
➢ **Database vs Data Warehouse:**

Database	DW
It is designed to record.	It is designed to analyze.
It is based on operational/transactional processing. It captures & maintains the data.	It is based on analytical processing. It explores the data.
Primitive and highly detailed data. Small in size – MB to GB.	Summarized and consolidated data. Larger in size – GB to TB.
It stores current data which is used for daily transactions.	It maintains historical data over time for trend analysis, making future predictions and to support in decision making.
Generally, it is application specific. For example, a database stores related data such as students detail in an institute.	Generally, it is integrated at organization level by combining data from different databases. For example, a data warehouse integrates data from one or more databases so analysis like sorting institutions in a city based on their performances can be done.
Construction of database is not so expensive.	Construction of a data warehouse can be expensive.
ER modeling techniques are used for designing.	Data modelling techniques are used for designing.
Simple transaction queries are used.	Complex queries are used for analysis purpose.

2.3.5 What is Data Warehouse?

- A Data Warehousing (DW) is the process for collecting and managing data from varied sources to provide meaningful business insights. It is typically used to connect and analyze business data from heterogeneous sources. It the core of the BI system which is built for data analysis and reporting.
- It is an electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users in a timely manner to make a difference.

- Data Warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. It is a valuable tool in today's competitive, fast evolving world. Many industry feel that data warehousing is a marketing weapon that help in retaining customer by learning more about their needs.
- Loosely speaking, a data warehouse refers to a database repository that is maintained separately from an organization's operational databases. Data warehouse systems allow for the integration of a variety of application systems. They support information processing by providing a solid platform of consolidated historical data analysis.
- The implementation of a data warehouse brings together information from disparate sources and puts the information into a format that is conducive for making business decisions.



- The basic operations of data warehousing are to extract data from the operational systems, to include relevant data from outside sources like magazines, journals, reports of other organizations in the same industry, to remove inconsistencies and transform and clean the data, to store the data in such a way so that it is for easy access for decision making etc.
- According to **William H. Inmon**, a leading architect in the construction of data warehouse systems, **"A Data Warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process."** These four characteristics of data warehouses distinguish them from other data repository systems, such as relational database systems, transaction processing systems, and file systems.

Characteristics of Data Warehouse:

- **Subject-oriented:**

  One of the key features of a data warehouse is the orientation it follows. Data warehouses focus on major subjects like customer, supplier, product, sales, revenue etc. and not on ongoing and current/day-to-day operations of organization data. This enables it to be used for data analysis which is a key element of decision making.

  Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a warehouse that concentrates on sales. Using this warehouse, you can answer questions like "Who was our best customer for this item last year?" This ability to define a data warehouse by subject matter, sales in this case, makes the data warehouse subject oriented.

- **Integrated/Collaborated:**

  A data warehouse's core is its integration of data from several different sources which aren't homologous (heterogeneous) in nature, for example, flat files, relational databases, on-line transaction records and other such sources. This plays a key role in enhancing the efficiency of data analysis.

  Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

- **Time-variant:**

  What's the importance of data without a time-stamp? Data uploaded into a warehouse can be identified with a certain timeline making it a multidimensional historical view (past few years) whenever you access data.

  In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that historical data be moved to an archive. A data warehouse's focus on change over time is what is meant by the term time-variant.

- **Nonvolatile:**

  The data in a warehouse is of the non-volatile type which ensures that your previous data is not lost as new data is updated which separates them for operational databases which are subject to frequent changes.

  Nonvolatile means that, once entered into the warehouse, data should not change. This is logical because the purpose of a warehouse is to enable you to analyze what has occurred.

  Due to separation of physically stored data and application data from operational environment, a data warehouse does not require transaction processing, recovery and concurrency control mechanism.

  Activities like delete, update, and insert which are performed in an operational application environment are omitted in Data warehouse environment. It usually requires only two operations in data accessing: data loading and data access.

- **No Additional Controls:**

  As the warehouse is maintained separate and has a separate storage from the operational databases, it doesn't require any concurrency controls, tweaks in processing, recovery mechanisms.

➢ **Applications of Data Warehousing:**

Data Warehousing can be applicable where we have huge amount of data and we want to see statistical results that help in decision making. There can be many applications in different sectors like E-Commerce, Telecommunication, Transportation Services, Marketing and Distribution, Healthcare and Retail. Let us discuss few of them.

Social Media Websites: The social networking websites like Facebook, Twitter, Linkedin etc. are based on analyzing large data sets. These sites gather data related to members, groups, locations etc. and store it in a single central repository. Being large amount of data, Data Warehouse is needed for implementing the same.

Banking: Most of the banks these days use warehouses to see spending patterns of account/card holders. They use this to provide them special offers, deals, etc. and to provide feedback to bankers regarding customer relationships and profitability.

Government: Government uses data warehouse to store and analyze tax payment which is used to detect tax thefts. Also to maintain and analyze health policy records, to share data with other entities, like insurance companies, NGOs, and medical aid services.

Finance: Used for evaluation of customer expenses trends, maintain transparency in transactions, predict defaulters and act accordingly, analyze and forecast different aspects of business, stock and bond performance etc.

Education: Used to store and analyze information about faculty and students, to maintain student portals to facilitate student activities, to extract information for

research grants and assess student demographics, to integrate information from differen sources into a single repository for analysis and strategic decision-making etc.

➢ **How is the organization using the information from data warehouses?**

Many organizations use this information to support business decision making activities, including. Increasing customer focus, which includes the analysis of customer buying patterns (such as buying preference, buying time, budget cycle).

Repositioning products and managing product portfolios by comparing the performance of sales by quarter, by year, and by geographic regions in order to fine tune production strategies

Analyzing operations and looking for sources of profit Managing customer – relationships.

Data Warehousing:

- It is the process of constructing and using data warehouses. The construction of a data warehouse requires data cleaning, data integration, and data consolidation.
- The utilization of a data warehouse often necessitates a collection of decision support technologies. This allows "knowledge workers" (e.g., managers, analysts, and executives) to use the warehouse to quickly and conveniently obtain an overview of the data, and to make sound decisions based on information in the warehouse.
- The term data warehousing is used to refer only to the process of data warehouse construction, while the term warehouse DBMS is used to refer to the management and utilization of data warehouses.
- Data warehousing is also very useful from the point of view of heterogeneous database integration. Organizations typically collect diverse kinds of data and maintain large databases from multiple, heterogeneous, autonomous, and distributed information sources.
- Data warehousing provides an alternative of traditional approach. Rather than using a query-driven approach, data warehousing employs an update-driven approach in which information from multiple, heterogeneous sources is integrated in advance and stored in a warehouse for direct querying and analysis.
- Unlike online transaction processing databases, data warehouses do not contain the most current information. However, a data warehouse brings high performance to the integrated heterogeneous database system because data are copied, preprocessed, integrated, annotated, summarized, and restructured into one semantic data store.
- Additionally, query processing in data warehouses does not interfere with the processing at local sources. Moreover, data warehouses can store and integrate historic information and support complex multidimensional queries.

- Data warehousing requires both business and technical expertise and involves the following activities:
- Accurate identification of business information that must be stored in the warehouse.
- Identification and prioritization of subject areas to be included in it.
- Defining the scope of each subject area.
- Development of a scalable architecture.
- Selection of the hardware/software/middleware components needed.
- Extracting, cleansing, aggregating, transforming, and validating the data to ensure accuracy and consistency.
- Providing user-friendly, powerful tools to the users with which they can gain access to the data warehouse.
- Giving adequate training to the users.
- Establishing a data warehouse helpdesk to support the users in their day-to-day tasks.
- Establishing procedures for maintenance and enhancement of the data warehouse.

2.3.6 Difference between operational database system and data warehouses:

- The major task of online operational database systems is to perform online transaction and query processing. These systems are called **online transaction processing (OLTP) systems**. OLTP covers most of the day-to-day operations of an organization, such as purchasing, inventory, manufacturing, banking, payroll, registration and accounting.
- Data warehouse systems at the other hand, serve users or knowledge workers in data analysis and decision making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of different users. These systems are known as **online analytical processing (OLAP) systems.**
- The major distinguishing features between OLTP and OLAP are:

- **Users and system orientation:**
  - → An OLTP system is customer oriented and is used for transaction and query processing by clerks, clients, IT 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 are too detailed. An OLAP system manages large amounts of historical data, and is in summarized form.
- **Database design:**
  - → An OLTP system usually adopts an 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.
- **Access patterns:**
  → The access patterns of an OLTP system consist mainly of short transactions. Such a system requires concurrency control and recovery mechanism. Whereas Accesses to OLAP systems are mostly read-only operations (because most data warehouses store historical rather than up-to-date information).

Comparison of OLTP and OLAP Systems:

Feature	OLTP (Operational Database)	OLAP (Data Warehouse)
Characteristic	used to manage operational Data	used to manage informational Data
Users	clerks, clients, and information technology professionals	knowledge workers, including managers, executives, and analysts
Orientation	Transaction	Analysis
Focus	data in (application oriented)	information out
Function	day-to-day operations	long-term informational, decision support
View	detailed, flat relational, complex data structure	summarized, multidimensional data structure
Unit of work	short, simple transaction	complex query
DB design	uses an ER data model and application-oriented DB design	uses either a star or snowflake model and subject-oriented DB design
Data contents	manages current, detailed, up-to-date data and are used for decision making	manages a large amount of historical data at different levels of granularity, makes the data more comfortable to use in informed decision making
Summarization	primitive and highly detailed	summarized and consolidated
Access	read/write	mostly read
Operations	index/hash on primary key	lots of scans

No. of records accessed	Tens	Millions
No. of users	Thousands	Hundreds
DB size	100 MB to 100 GB	100 GB to 100 TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time
Processing speed	very fast	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

2.3.7 Multitier Architecture:

Design and construction of Data Warehouse:

To design an effective data warehouse, we need to understand and analyze business needs and construct a business analysis framework. Four different views regarding the design of a data warehouse are:

The top-down view: allows selection of the relevant information necessary for the data warehouse

The data source view: exposes the information being captured, stored, and managed by operational systems

The data warehouse view: consists of fact tables and dimension tables

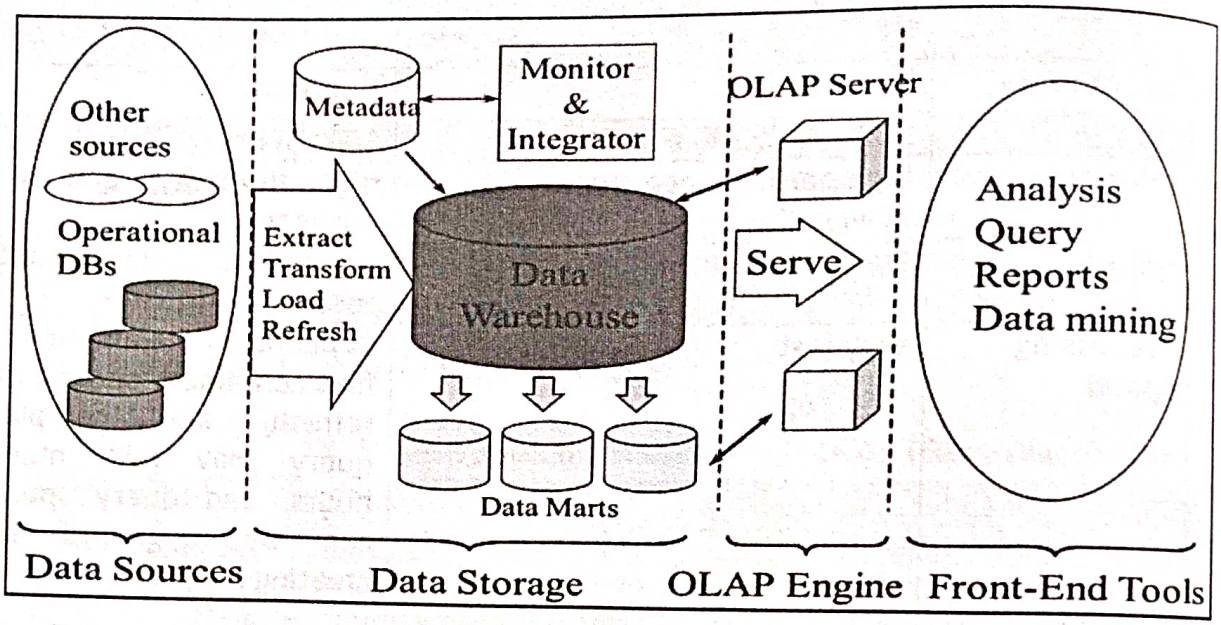
The business query view: sees the perspectives of data in the warehouse from the view of end-user

A Three-Tier Data Warehouse Architecture:

(1) Bottom tier is a warehouse database server

(2) Middle tier is a OLAP server

(3) Top tier is a front-end client layer



Bottom Tier:

→ A bottom-tier that consists of the Data Warehouse server, which is almost always an RDBMS.

→ It may include several specialized data marts and a metadata repository.

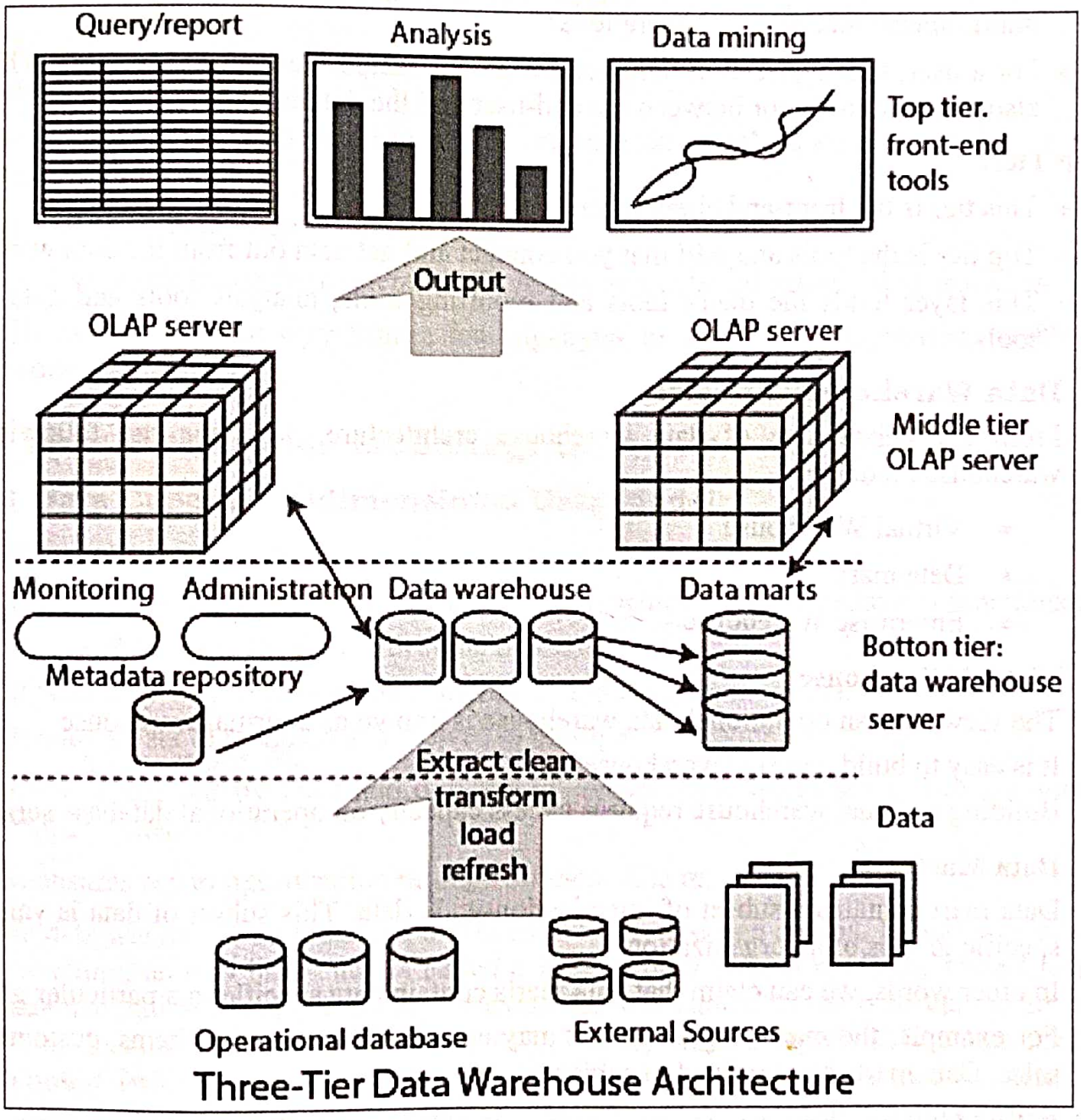
(A Data Mart is a subset of a data warehouse in which a summarized or highly focused portion of the organization's data is placed in separate database for a specific population of users. For example, A company might develop marketing & sales data marts to deal with customer information.

Metadata repository is an integral part of a data warehouse system. It is a database created to store metadata. Metadata are data about data which can be categorized into business, technical or operational. When used in a data warehouse, metadata are the data that define warehouse objects. It should contain data warehouse structure, Operational metadata, algorithms used for summarization, Mapping from the operational environment to the data warehouse, Data related to system performance, Business metadata)

→ We use the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the Extract (which typically gathers data from multiple, heterogeneous, and external sources), Clean (which detects errors in the data and rectifies them when possible), Transform (which converts data from legacy or host format to warehouse format), Load (which sorts, summarizes, consolidates, computes views, checks integrity, and builds indices and partitions), and Refresh (which propagates the updates from the data sources to the warehouse) functions.

→ Data from operational databases and external sources (such as user profile data provided by external consultants) are extracted using application program interfaces called a gateway. A gateway is provided by the underlying DBMS and allows customer programs to generate SQL code to be executed at a server.

→ Examples of gateways contain ODBC (Open Database Connection) and OLE-DB (Object Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection).



Three-Tier Data Warehouse Architecture

Middle Tier:

→ In the middle tier, we have the OLAP Server that can be implemented in either of the following ways.

→ By **Relational OLAP (ROLAP)**, which is an extended relational database management system? The ROLAP maps the operations on multidimensional data to standard relational operations.

→ By **Multidimensional OLAP (MOLAP)** model, which directly implements the multidimensional data and operations?

→ For a user, this application tier presents an abstracted view of the database. This layer also acts as a mediator between the end-user and the database.

Top Tier:

→ This tier is the front-end client layer.

→ Top tier is the tools and API that you connect and get data out from the data warehouse.

→ This layer holds the query tools and reporting tools, analysis tools and data mining tools.

➢ **Data Warehouse Models:**

From the perspective of data warehouse architecture, we have the following data warehouse models –

- Virtual Warehouse
- Data mart
- Enterprise Warehouse

- **Virtual Warehouse:**

The view over an operational data warehouse is known as a virtual warehouse.

It is easy to build a virtual warehouse.

Building a virtual warehouse requires excess capacity on operational database servers.

- **Data Mart:**

Data mart contains a subset of organization-wide data. This subset of data is valuable to specific groups of an organization.

In other words, we can claim that data marts contain data specific to a particular group.

For example, the marketing data mart may contain data related to items, customers, and sales. Data marts are confined to subjects.

Points to remember about data marts –

Window-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.

The implementation data mart cycles is measured in short periods of time, i.e., in weeks rather than months or years.

The life cycle of a data mart may be complex in long run, if its planning and design are not organization-wide.

Data marts are small in size.

Data marts are customized by department.

The source of a data mart is departmentally structured data warehouse.

Data mart is flexible.

- **Enterprise Warehouse:**

An enterprise warehouse collects all the information and the subjects spanning an entire organization.

It provides us enterprise-wide data integration.

The data is integrated from operational systems and external information providers.

This information can vary from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

2.4 Data Warehouse Modeling: Data cube and OLAP:

2.2.1 Data Cube: Multidimensional Data Model:

❖ **What is Data Cube?**

Data Cube is a structure that enables OLAP to achieve multidimensional functionality.

It is used to represent data with some measure.

It is an easy way to look at the complex data into a simple format.

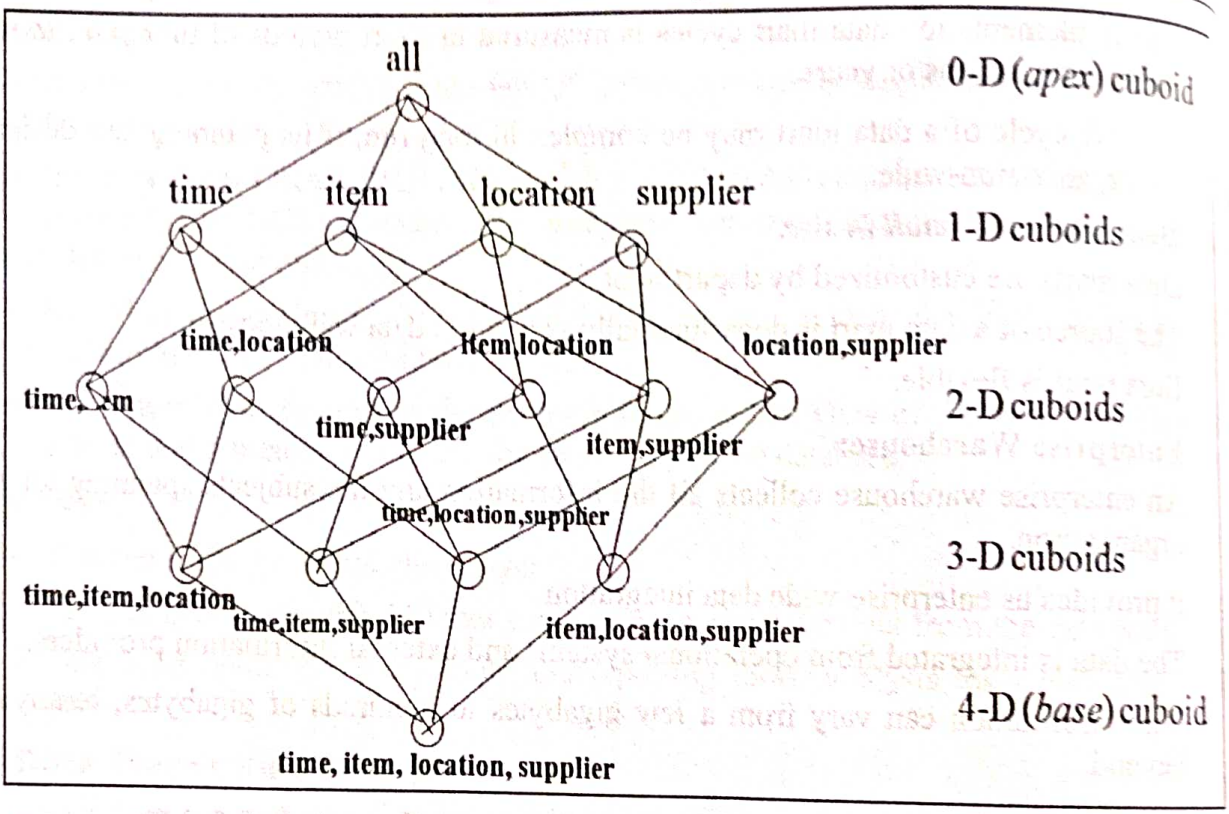
It allows data to modeled and viewed in multiple dimensions.

Cube is defined by facts and dimensions. It can be 2-dimensional, 3-dimensional or higher-dimensional.

It ensures report optimization through efficient data retrieval.

In data warehouse architecture, a data cube is referred as a cuboid. In data warehousing literature, an n-D base cube is called a **base cuboid** (which holds the lowest level of summarization means detailed. For example, 4-D cuboid is the base cuboid for time, item, location and supplier dimensions.). The top most 0-D cuboid is called the **apex cuboid** (which holds the highest level of summarization means not in detail. For example, total sales of AllElectronics). The lattice of cuboids forms a **data cube.**

A Lattice of Cuboids: Each cuboid represents a different degree of summarization.



Data cubes have categories of data called **dimension** (represents descriptive categories of data such as time or location) and **measure** (represents some fact or number such as cost or units of service).

❖ 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. For example, a data cube, such as sales, allows data to be modeled and viewed in multiple dimensions.

Dimensions are a fact that defines a data cube. Dimensions are the perspectives or entities with respect to which an organization wants to keep records. **Dimension tables**, such as item (item_name, brand, type, supplier_type), or time(day, week, month, quarter, year) or branch, or location. These dimensions enable the store to keep track of things like monthly sales of items, and the branches and locations at which the items were sold. Each dimension may have a table identify with it, known as a dimensional table, which describes the dimensions.

(Courtesy: Below examples – tables and figures are taken from the syllabus text-book e-copy Data Mining Concepts and Techniques by Morgan-Kaufmann, 3rd edition as one of the references.)

- **2-D view of sales data:**

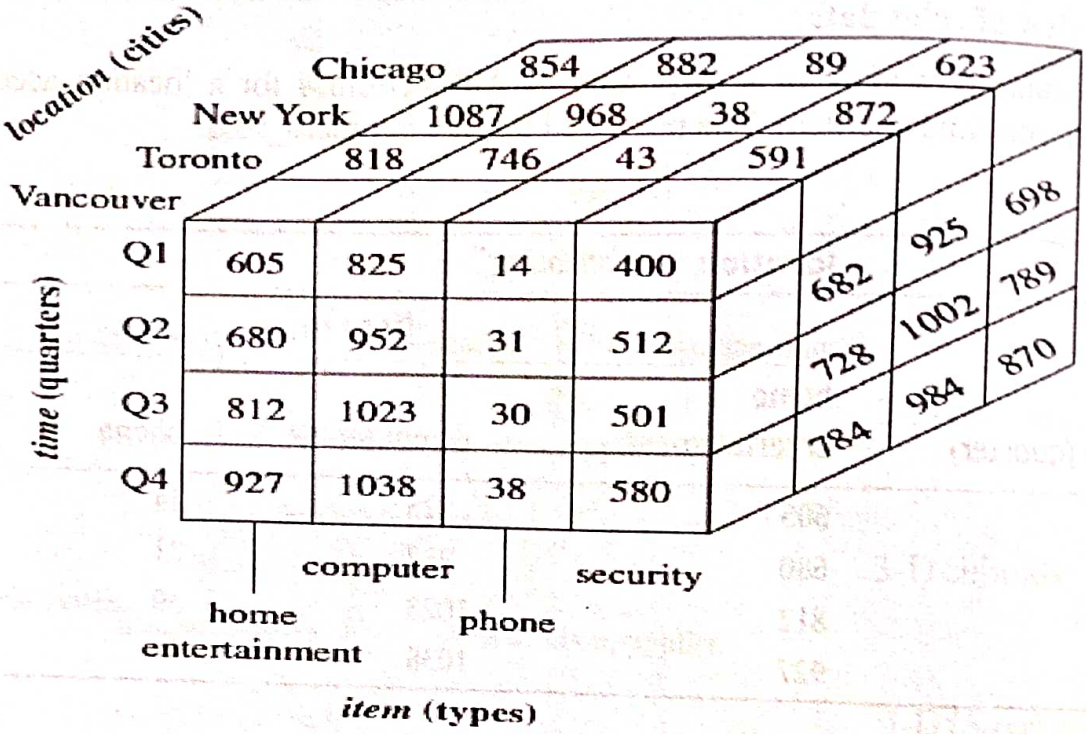
For example, a view of a sales data of AllElectronics for a location according to the dimension time and item. The measured display in rupees_sold.

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

- **3-D view of sales data:**

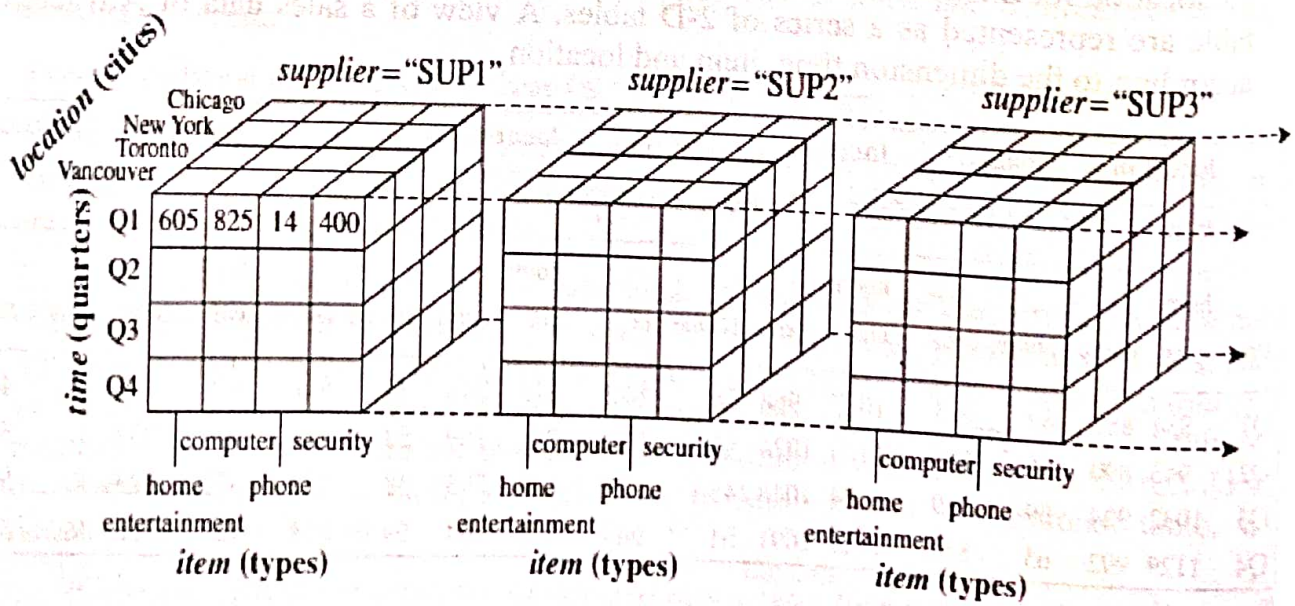
For example, suppose we would like to view the data according to time, item as well as the location for different cities. The measured display in rupees_sold. The 3-D data of the table are represented as a series of 2-D tables. A view of a sales data of AllElectronics according to the dimension time, item and location.

	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

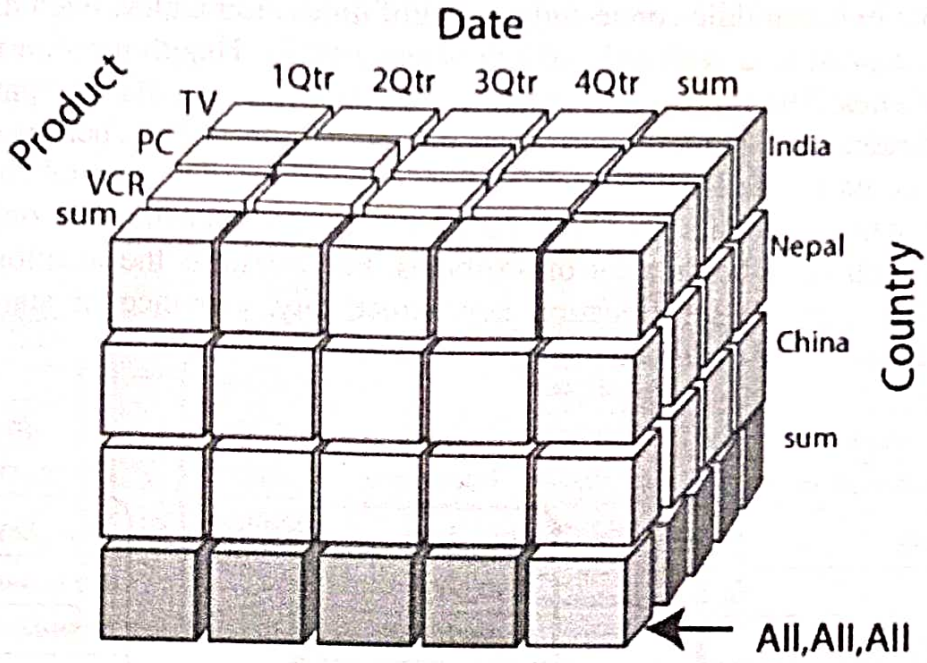


- **4-D view of sales data:**

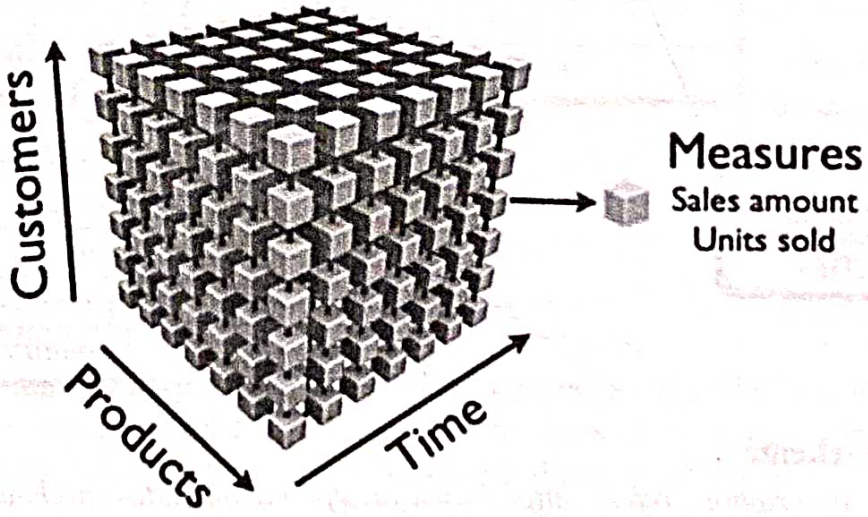
For example, a view of a sales data of AllElectronics according to the dimension time item, location and supplier.



A data cube enables data to be modeled and viewed in multiple dimensions. A multidimensional data model is organized around a central theme, like sales and transactions. A fact table represents this theme. **Fact table** contains measures (such as rupees_sold) and keys to each of the related dimension tables.



Let us take an another example: a Sales cube that has Customers, Products and the Time dimensions and has two measures - Sales amount and Units sold. Each detailed cube "cell" contains sales amount and units sold for particular product, customer and time period:

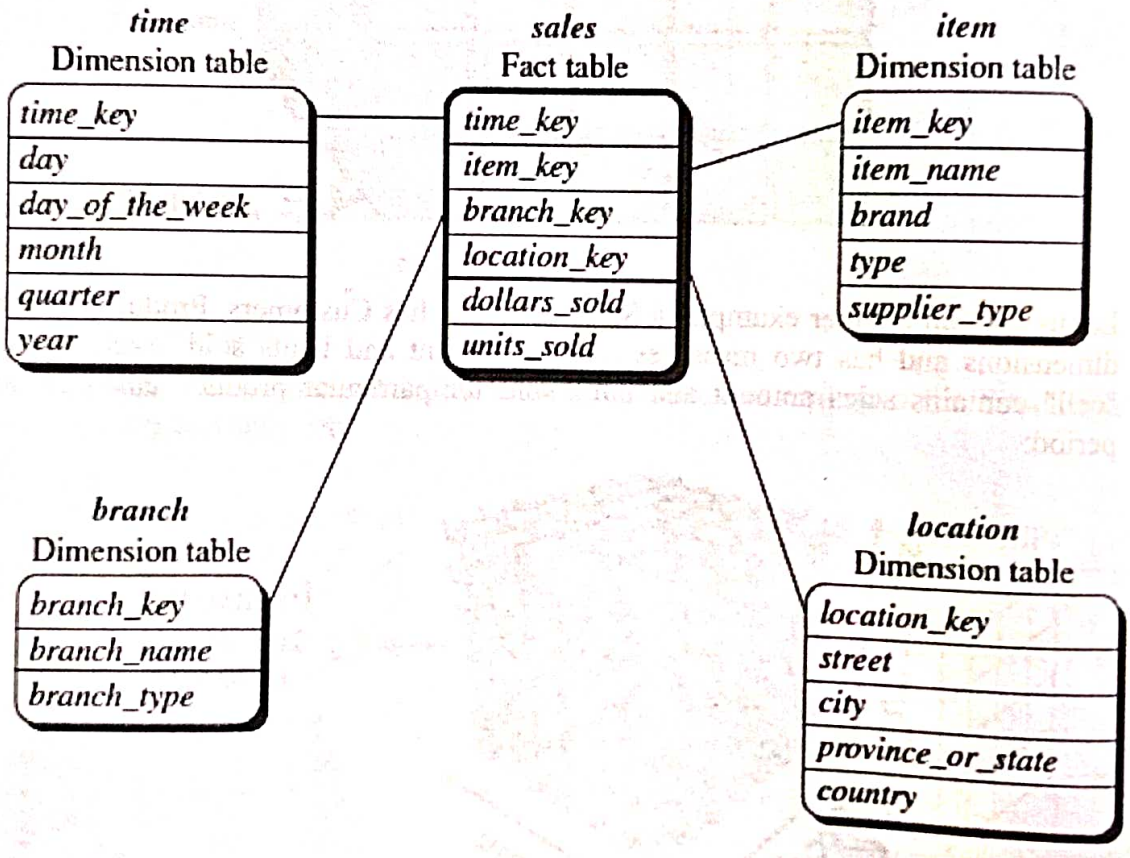


It is easy to illustrate cube with three dimensions but you can have as many dimensions and as many measures as you need in one data cube. All measures in the same cube share the same dimensions.

- **Conceptual Modeling of Data Warehouses:**

Star schema: (For more description refer: https://www.javatpoint.com/data-warehouse-what-is-star-schema)

A fact table in the middle connected to a set of dimension tables. Each dimension in a sc schema is represented with only one-dimension table. This dimension table contains th set of attributes. The following diagram shows the sales data of a company with respect t the four dimensions, namely time, item, branch, and location. There is a fact table at th center. It contains the keys to each of four dimensions. The fact table also contains th attributes, namely dollars sold and units sold. Each dimension has only one dimensio table and each table holds a set of attributes. For example, the location dimension tabl contains the attribute set {location_key, street, city, province_or_state, country}. Thi constraint may cause data redundancy.

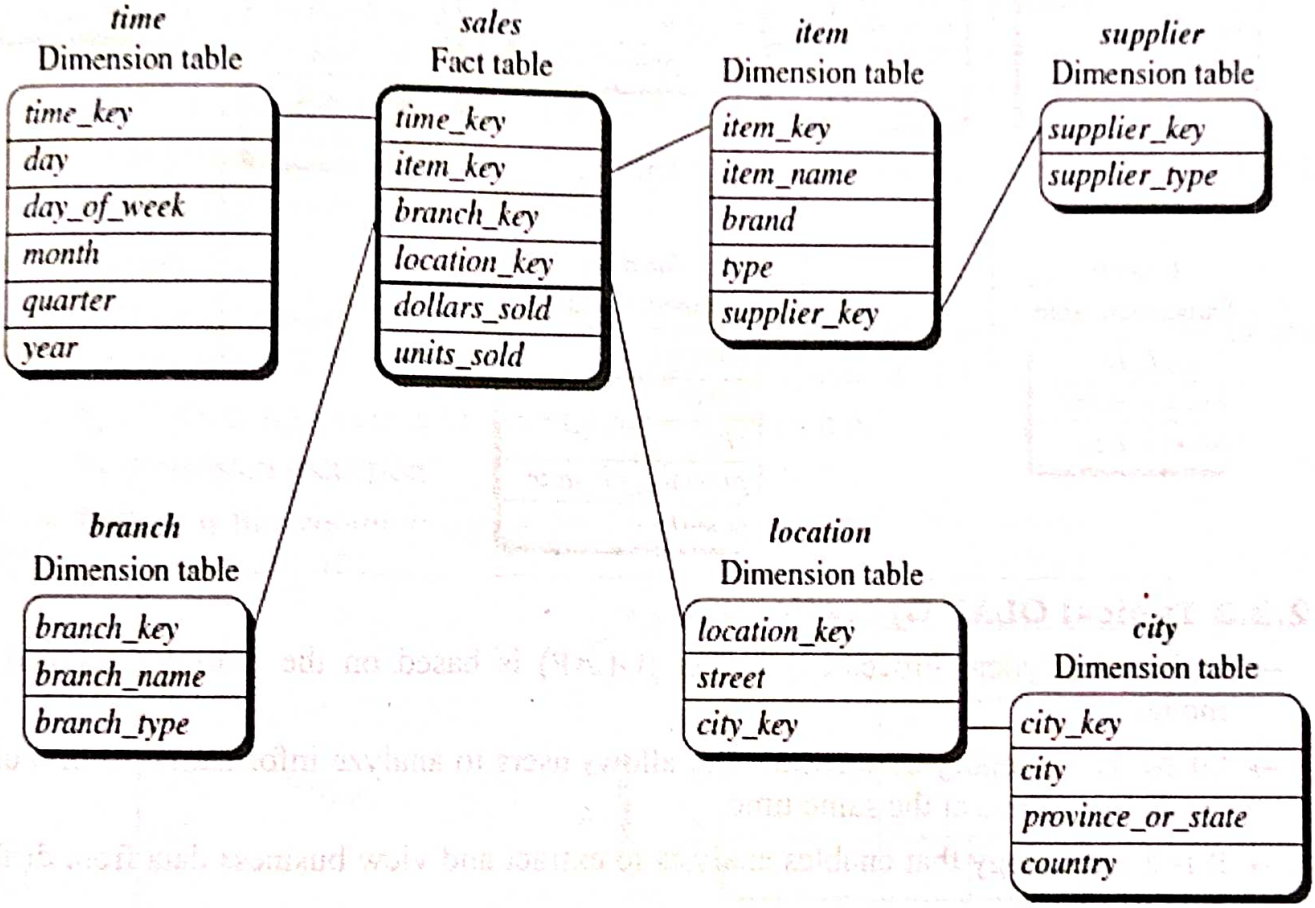


- **Snowflake schema:**

(For more description refer: https://www.javatpoint.com/data-warehouse-what-is-snowflake-schema)

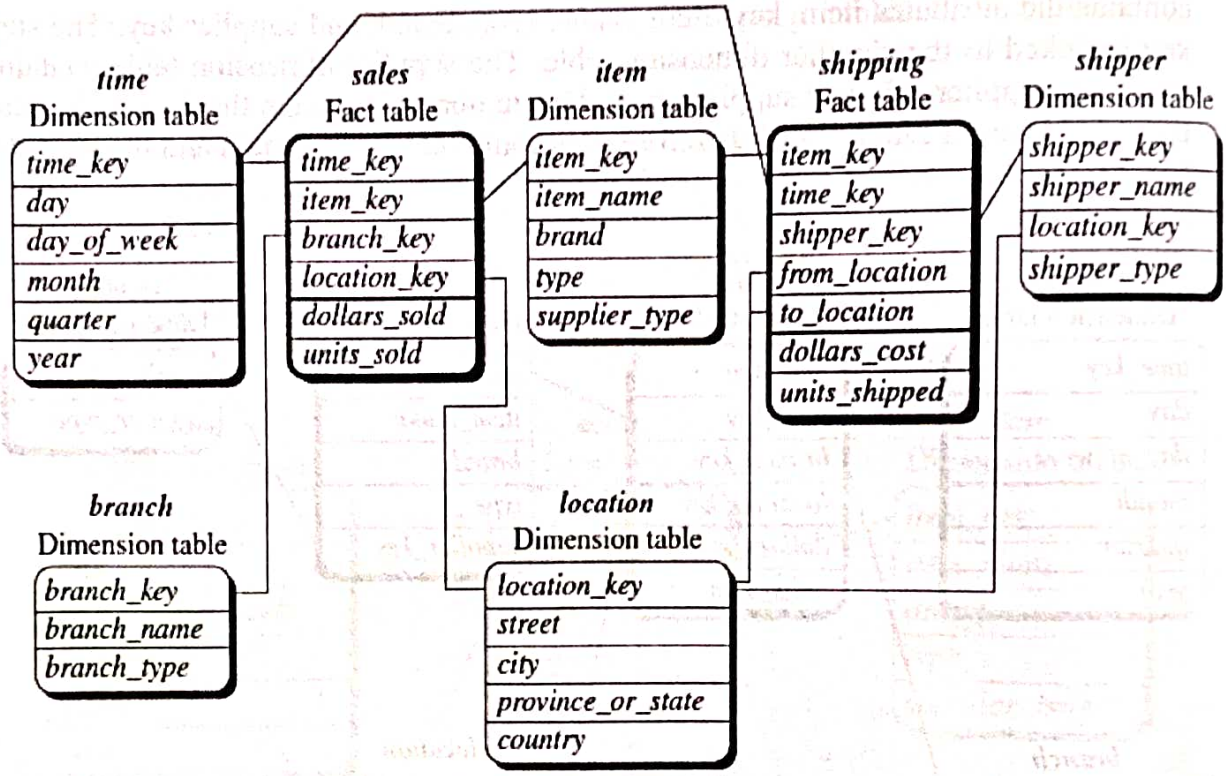
A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake. Some dimension tables in the Snowflake schema are normalized. The normalization splits up the data into additional tables. Unlike Star schema, the dimensions table in a snowflake schema is normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table. Now the item dimension table

contains the attributes item_key, item_name, type, brand, and supplier-key. The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type. Due to normalization in the Snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.



- **Fact constellations:**

  Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation. A fact constellation has multiple fact tables. It is also known as galaxy schema. The following diagram shows two fact tables, namely sales and shipping. The sales fact table is same as that in the star schema. The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location. The shipping fact table also contains two measures, namely dollars sold and units sold. It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.



2.2.2 Typical OLAP Operations:

→ Online Analytical Processing Server (OLAP) is based on the multidimensional data model.

→ OLAP is a category of software that allows users to analyze information from multiple database systems at the same time.

→ It is a technology that enables analysts to extract and view business data from different points of view.

→ It allows managers and analysts to get an insight of the information through fast, consistent and interactive access.

→ A number of OLAP data cube operations exist to demonstrate different views, allowing interactive queries and search of the record at hand.

→ OLAP offers analytical modeling capabilities, including a calculation engine for deriving ratios, variance, and so on, and for computing measures across multiple dimensions.

→ It can generate summarizations, aggregations, and hierarchies at each granularity level and at every dimension intersection.

→ OLAP also supports functional models for forecasting, trend analysis, and statistical analysis.

→ Following types of analytical operations in OLAP:

- ✓ **Roll-up (drill-up/consolidation/aggregation):** 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:** uses relational SQL facilities to drill through the bottom level of the cube to its backend relational tables

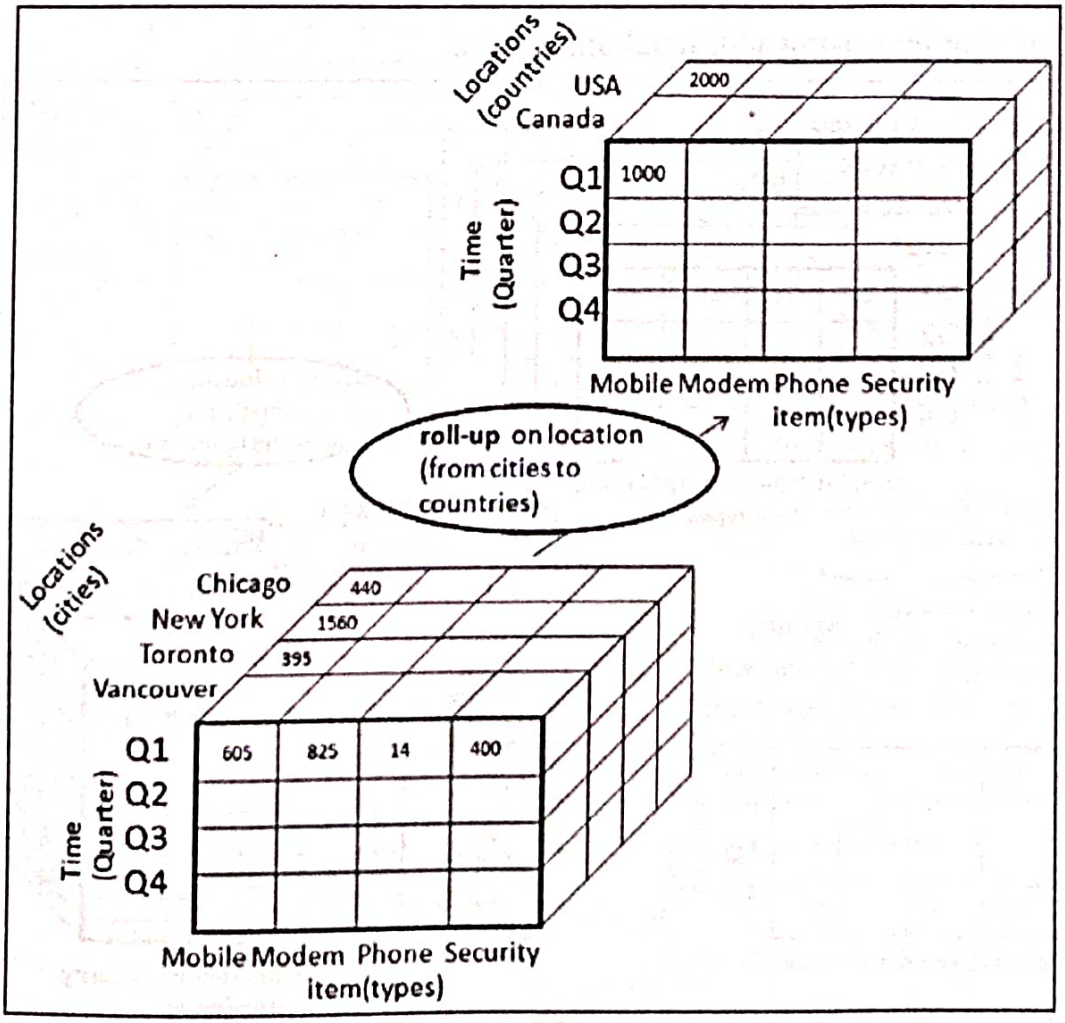
- **Roll-up:**

→ Roll-up or consolidation refers to data aggregation and computation in one or more dimensions. It is actually performed on an OLAP cube either_

by climbing up a concept hierarchy for a dimension or

by dimension reduction

→ Roll-up is like zooming-out on the data cubes.



→ Figure shows the result of roll-up operations performed on the dimension location.

→ The hierarchy for the location is defined as the Order street, city, province, or state country i.e. "street < city < province/state < country" .

→ The roll-up operation aggregates the data by ascending the location hierarchy from the level of the city to the level of the country. The data is grouped into cities rather than countries.

→ When roll-up is performed, one or more dimensions from the data cube are removed.

→ For example, consider a sales data cube containing only the location and time dimensions. Roll-up may be performed by removing the time dimension, resulting in an aggregation of the total sales by location, rather than by location and by time.

- **Drill-down:**

→ The drill-down operation is the reverse operation of roll-up.

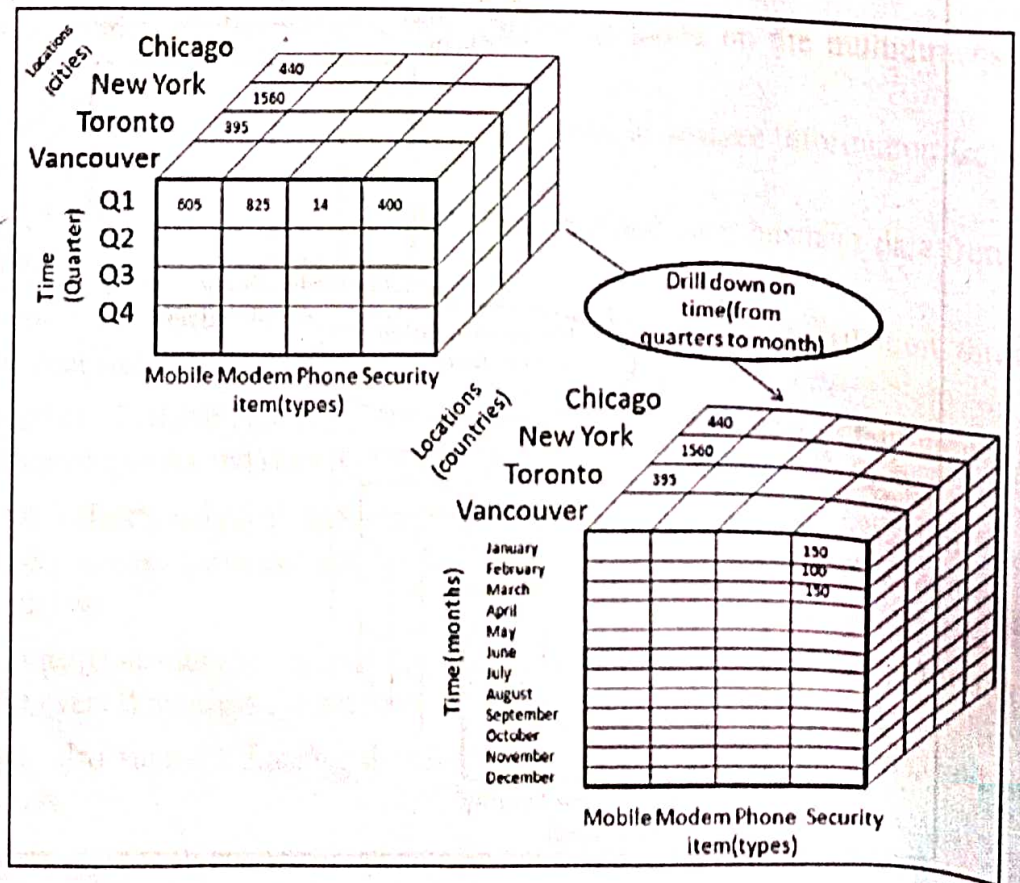
→ Drill-down is like zooming-in on the data cube. Here, data is fragmented into smaller parts.

→ It navigates from less detailed record to more detailed data.

→ It can be performed either_

by moving/stepping down a concept hierarchy for a dimension or

by adding/increasing additional dimensions



→ Figure shows a drill-down operation performed on the dimension time by stepping down a concept hierarchy which is defined as day, month, quarter, and year i.e. day < month < quarter < year. Here, Quarter Q1 is drilled down to months January, February, and March. Corresponding sales are also registers.

→ Drill-down appears by descending the time hierarchy from the level of the quarter to a more detailed level of the month.

→ When drill-down is performed, one or more dimensions from the data cube are added. For example, adding one more dimension that is customer group. It navigates the data from less detailed data to highly detailed data.

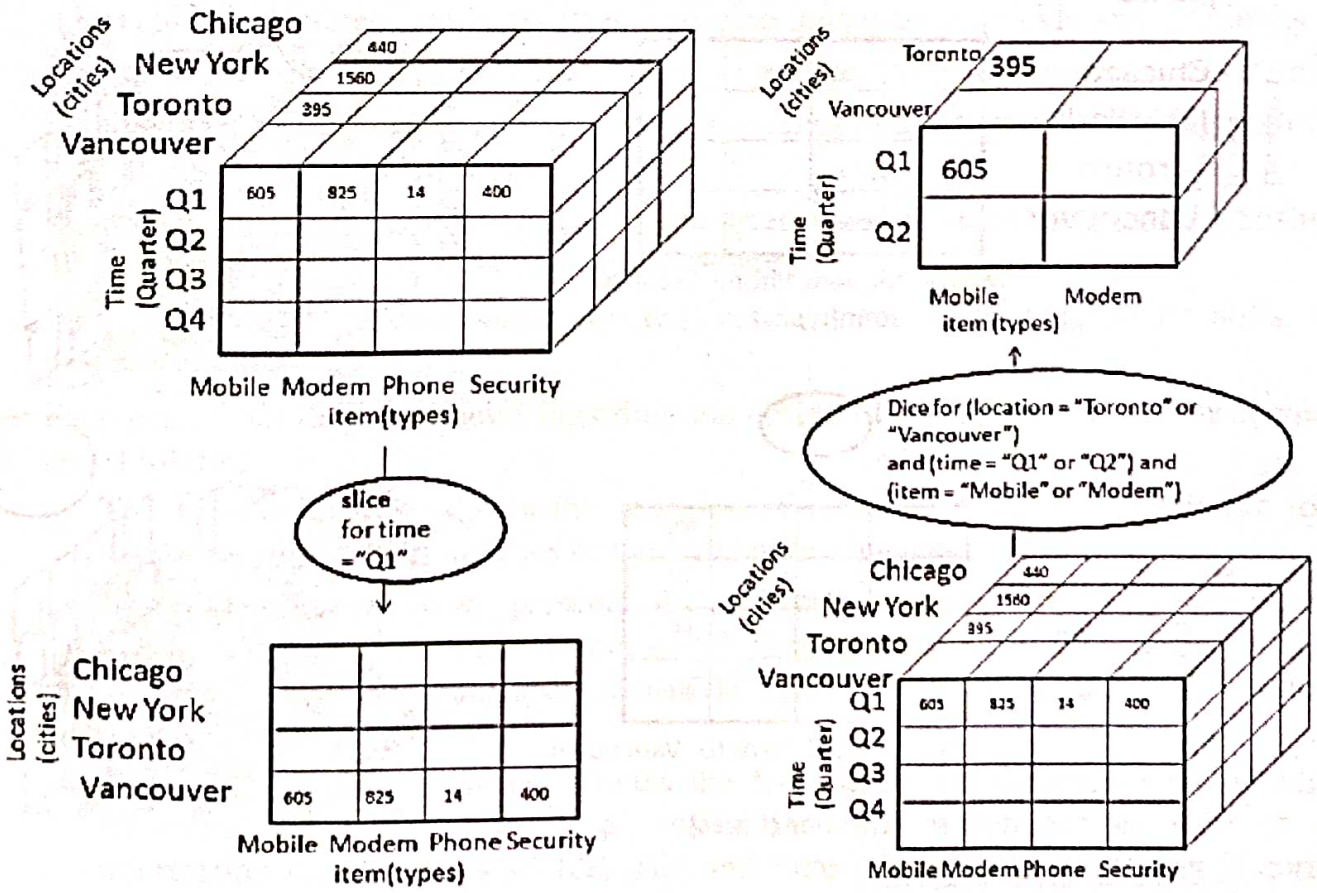
- **Slice and dice:**

→ A **slice** is a subset of the cubes corresponding to a single value for one or more members of the dimension.

→ For example, a slice operation is executed when the customer wants a selection on one dimension of a three-dimensional cube resulting in a two-dimensional site.

→ Here, Slice is performed for the dimension "time" using the criterion time = "Q1".

→ So, the Slice operations perform a selection on one dimension of the given cube, thus resulting in a subcube.



→ **Dice** operation is similar to a slice.

→ The difference in dice is you select 2 or more dimensions that result in the creation o a subcube.

→ The dice operation on the cube based on the following selection criteria involves thre dimensions.

(location = "Toronto" or "Vancouver")

(time = "Q1" or "Q2")

(item =" Mobile" or "Modem")

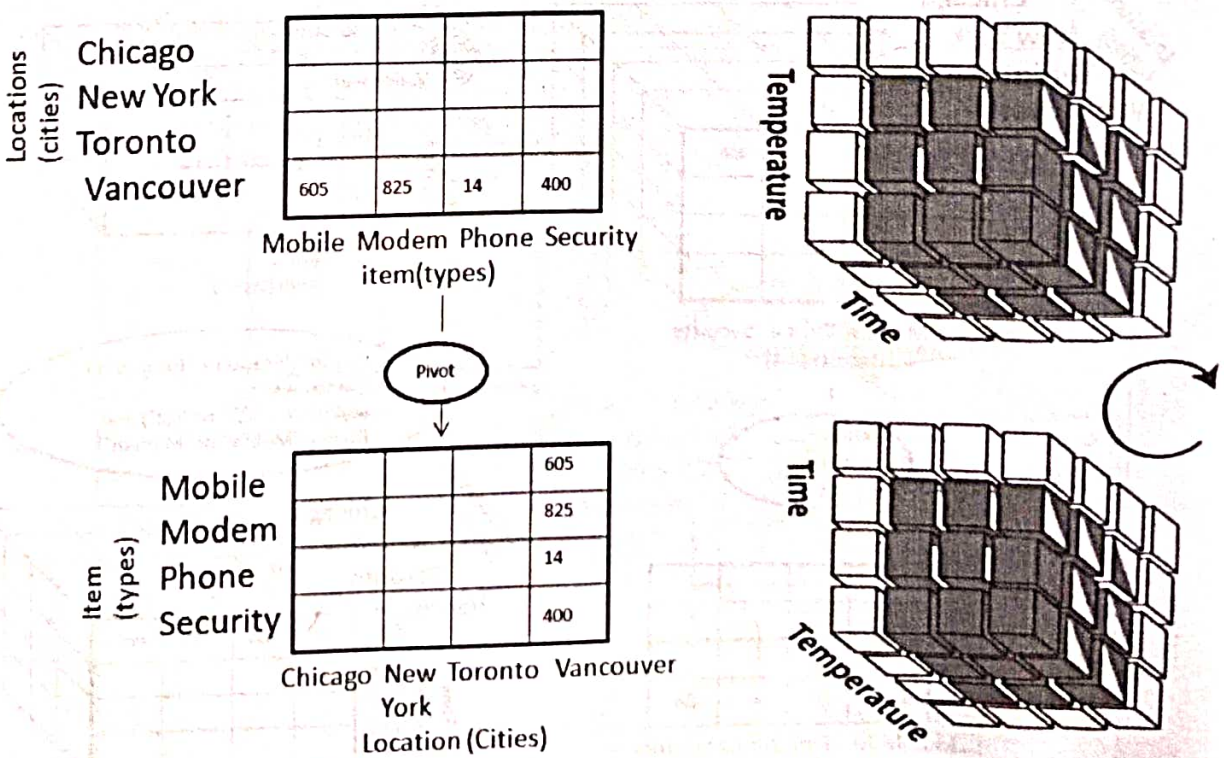
- **Pivot:**

→ Pivot is a visualization operations which rotates the data axes in view to provide a alternative presentation of the data.

→ It may contain swapping the rows and columns or moving one of the row-dimension into the column dimensions.

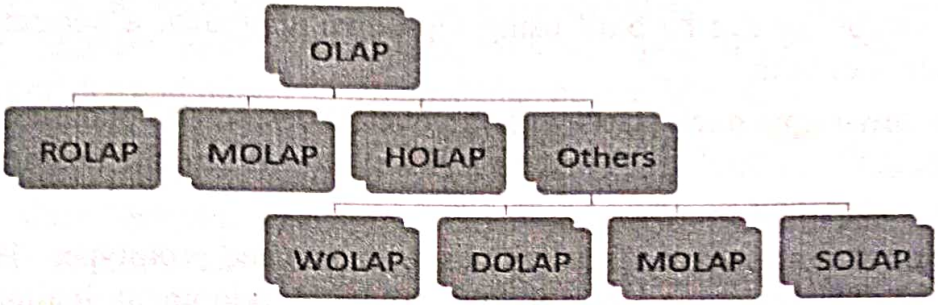
→ For example, here, pivot operation where the item and location axes in a 2-D slice ar rotated.

→ You can rotate the axes in a 3-D cube, or transforming a 3-D cube into a series of 2-D planes.



Types of OLAP systems:

OLAP Hierarchical Structure



Relational OLAP(ROLAP), Multidimensional OLAP (MOLAP), Hybrid Online Analytical Processing (HOLAP), Desktop OLAP (DOLAP), Web OLAP (WOLAP), Mobile OLAP, Spatial OLAP

2.5 Data Warehouse Design and Usage:

2.3.1 A business Analysis Framework for Data Warehouse Design:

→ The business analyst get the information from the data warehouses to measure the performance and make critical adjustments in order to win over other business holders in the market. Data warehouse offers the following advantages:

it can enhance business productivity as gathering information quickly and efficiently

it helps us manage customer relationship by a consistent view of customers and items

it helps in bringing down the costs by tracking trends and patterns over a long period in a consistent and reliable manner

→ To design an effective data warehouse we need to understand and analyze business needs and construct a business analysis framework.

→ Building and using a data warehouse requires business skills, technology skills, and program management skills.

→ Each person has different views regarding the design of a data warehouse. These views are as follows:

- **The top-down view:** allows the selection of relevant information desirable for a data warehouse which matches current and future business needs
- **The data source view:** presents the information being captured, stored, and managed by the operational system which may be documented at various levels of detail and accuracy, from individual data source tables to integrated data source tables
- **The data warehouse view:** includes the fact tables and dimension tables which represents the information stored inside the data warehouse as well as the information regarding the source, date and time of origin for providing historical context

- **The business query view** – presents data from the viewpoint of the end-user

2.3.2 Data warehouse Design Process:

→ A data warehouse can be built using top-down approach, a bottom-up approach or combination of both.

The **top-down approach** starts with the overall design and planning. It is useful in the cases where the technology is mature and well known and the business problem is clear and well understood.

The **bottom-up approach** starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. It allows an organization to move forward at considerably less expense and to evaluate the technological benefits before making significant commitments.

(12)

In the **combined approach**, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.

From the **software engineering point of view**, the design and construction of a data warehouse may consist of the following steps: planning, requirements study, problem analysis, warehouse design, data integration and testing, and finally deployment of the data warehouse.

→ Large software systems can be developed using one of two methodologies: the waterfall method or the spiral method.

The **waterfall method** performs a structured and systematic analysis at each step before proceeding to the next, which is like a waterfall, falling from one step to the next.

The **spiral method** involves the rapid generation of increasingly functional systems with short intervals between successive releases.

→ This is considered a good choice for data warehouse development, especially for data marts, because the turnaround time is short, modifications can be done quickly, and new designs and technologies can be adapted in a timely manner.

→ In general, the warehouse **design process** consists of the following steps:

Choose a **business process** to model orders, invoices, shipments, inventory, administration, sales etc. Follow a **data warehouse model** for organizational business process which encompasses multiple complex object collections. Choose a **data mart model** for departmental process which focuses on the analysis of a kind of business process.

Choose the **business process grain**, which is the fundamental, atomic level of data to be represented in the fact table for the data warehouse design process. For example, individual transactions, individual daily snapshots and so on.

Choose the **dimensions** that will apply to each fact table record. For example, time, item, customer, supplier, transaction type, status etc.

Choose the **measures** that will populate each fact table record. For example, numeric additive quantities like rupees_sold and units_sold.

→ Once a data warehouse is **designed and constructed**, the initial **deployment** of the warehouse includes initial installation, roll-out planning, training, and orientation. Platform upgrades and maintenance must also be considered.

Data warehouse **administration** includes data refreshment, data source synchronization, planning for disaster recovery, managing access control and security, managing data growth, managing database performance, and data warehouse enhancement and extension.

Scope management includes controlling the number and range of queries, dimensions, and reports; limiting the data warehouse's size; or limiting the schedule, budget, or resources.

Various kinds of data warehouse **design tools** are available.

Data warehouse **development tools** provide functions to define and edit metadata repository contents (e.g., schemas, scripts, or rules), answer queries, output reports, and ship metadata to and from relational database system catalogs.

Planning and analysis tools study the impact of schema changes and of refresh performance when changing refresh rates or time windows.

2.3.3 Data Warehouse usage for Information Processing:

→ The data warehouse may be employed for knowledge discovery and strategic decision making using data mining tools.

→ In this context, the tools for data warehousing can be categorized into access and retrieval tools, database reporting tools, data analysis tools, and data mining tools.

→ Business users need know what exists in the data warehouse (through metadata), how to access the contents of the data warehouse, how to examine the contents using analysis tools, and how to present the results of such analysis.

→ There are three kinds of **data warehouse applications**: information processing, analytical processing, and data mining.

Information processing supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts, or graphs. A current trend in data warehouse information processing is to construct low-cost web-based accessing tools that are then integrated with web browsers.

Analytical processing supports basic OLAP operations, including slice-and-dice, drill-down, roll-up, and pivoting. It generally operates on historic data in both summarized and detailed forms. The major strength of OLAP over information processing is the multidimensional data analysis of data warehouse data.

Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

2.3.4 From OLAP to Multidimensional data Mining:

→ Multidimensional data mining/exploratory multidimensional data mining/On Line Analytical Mining-**OLAM** integrates OLAP with data mining to uncover knowledge in multidimensional databases.

→ Among many different paradigms and architectures of data mining systems, multidimensional data mining is particularly important for the following reasons:

- **High quality of data in data warehouses:**

  Most data mining tools need to work on integrated, consistent, and cleaned data, which requires costly data cleaning, data integration, and data transformation as preprocessing steps.

  A data warehouse constructed by such preprocessing serves as a valuable source of high-quality data for OLAP as well as for data mining.

  Data mining may serve as a valuable tool for data cleaning and data integration as well.

- **Available information processing infrastructure surrounding data warehouses:**

  Comprehensive information processing and data analysis infrastructures have been systematically constructed surrounding data warehouses, which include accessing, integration, consolidation, and transformation of multiple heterogeneous databases, ODBC/OLEDB connections, Web accessing and service facilities, and reporting and OLAP analysis tools.

  It is practical to make the best use of the available infrastructures rather than constructing everything from scratch.

- **OLAP-based exploration of multidimensional data:**

  Effective data mining needs exploratory data analysis.

  A user will often want to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge/results in different forms.

  Multidimensional data mining provides facilities for mining on different subsets of data and at varying levels of abstraction—by drilling, pivoting, filtering, dicing, and slicing on a data cube and/or intermediate data mining results. This, together with data/knowledge visualization tools, greatly enhances the power and flexibility of data mining.

- **Online selection of data mining functions:**

  Users may not always know the specific kinds of knowledge they want to mine.

  By integrating OLAP with various data mining functions, multidimensional data mining provides users with the flexibility to select desired data mining functions and swap data mining tasks dynamically.

☺ ☺ ☺ ☺ ☺

❖ **Answer the**

Q 1. Describe t
Q 2. Discuss fe
Q 3. Discuss m
Q 4. Differenti
Q 5. Write a n
Q 6. Write a n fact table
Q 7. Explain va
Q 8. Write a n
Q 9. Write a n
Q 10. Write a n
Q 11. Write a n
Q 12. Discuss d
Q 13. Explain different
Q 14. Discuss d
Q 15. Explain h
Q 16. What are informat motivati

❖ **Filling th**

Q 1. A ____ separate
Q 2. A Data ____ support
Q 3. ____ as purc account

Exercises

❖ Answer the following Questions in brief.

Q 1. Describe the term Data Warehouse. Also explain its characteristics.

Q 2. Discuss few applications of Data Warehousing.

Q 3. Discuss major distinguishing features between OLTP and OLAP.

Q 4. Differentiate between OLTP and OLAP.

Q 5. Write a note on three-tier data warehouse architecture.

Q 6. Write a note on data cube and its different views. Also discuss dimension table and fact table with a suitable example.

Q 7. Explain various types of analytical operations in OLAP.

Q 8. Write a note on roll-up operation.

Q 9. Write a note on drill-down operation.

Q 10. Write a note on slice and dice operation.

Q 11. Write a note on pivot operation.

Q 12. Discuss data warehouse design process.

Q 13. Explain a business analysis framework for data warehouse design. / Explain different views regarding the design of a data warehouse.

Q 14. Discuss data warehouse usage for information processing.

Q 15. Explain how multidimensional data mining/OLAM integrates OLAP.

Q 16. What are the differences between the three main types of data warehouse usage: information processing, analytical processing, and data mining? Discuss the motivation behind OLAP mining (OLAM).

❖ Filling the Blanks.

Q 1. A _____ refers to a database repository that is maintained separately from an organization's operational databases.

Q 2. A Data Warehouse is a _____, _____, _____, and _____ collection of data in support of management's decision making process.

Q 3. _____ covers most of the day-to-day operations of an organization, such as purchasing, inventory, manufacturing, banking, payroll, registration and accounting.

Q 4. _____ uses an entity-relationship (ER) data model and applicat oriented database design.

Q 5. _____ having summarized, multidimensional data structure.

Q 6. _____ manages a large amount of historical data at different level granularity.

Q 7. The _____ view exposes the information being captured, stor and managed by operational systems.

Q 8. The _____ view consists of fact tables and dimension tables.

Q 9. A _____ tier that consists of the Data Warehouse server, which almost always an RDBMS.

Q 10. A _____ is a subset of a data warehouse in which a summarized highly focused portion of the organization's data is placed in separate database f a specific population of users.

Q 11. In the _____ tier, we have the OLAP Server that can be implemented either ROLAP or MOLAP.

Q 12. _____ tier is the front-end client layer.

Q 13. _____ is a structure that enables OLAP to achie multidimensional functionality.

Q 14. An n-D base cube is called a _____ cuboid which holds the lowest lev of summarization means detailed.

Q 15. The top most 0-D cuboid is called the _____ cuboid which holds th highest level of summarization means not in detail.

Q 16. Data cubes have categories of data called _____ represent descriptive categories of data such as time or location and _____ represents some fact or number such as cost or units of service.

Q 17. _____ are the perspectives or entities with respect to which a organization wants to keep records.

Q 18. _____ table contains measures and keys to each of the relate dimension tables.

Q 19. _____ summarizes data by climbing up hierarchy or b dimension reduction.

Q 20. _____ leve

Q 21. The

Q 22. _____ frag

Q 23. A _____ or r

Q 24. _____ to p

Q 25. The

Q 26. The eac ste

Q 27. The fur

Q 28. The dat

Q 29. _____ rep

Q 30. _____ pa an

Answ

1.
3.
8.
13
17
20
24
28

Q 20. _____ uses relational SQL facilities to drill through the bottom level of the cube to its backend relational tables.

Q 21. The _____ operation is the reverse operation of roll-up.

Q 22. _____ is like zooming-in on the data cube. Here, data is fragmented into smaller parts.

Q 23. A _____ is a subset of the cubes corresponding to a single value for one or more members of the dimension.

Q 24. _____ is a visualization operations which rotates the data axes in view to provide an alternative presentation of the data.

Q 25. The _____ approach starts with experiments and prototypes.

Q 26. The _____ method performs a structured and systematic analysis at each step before proceeding to the next, which is like a waterfall, falling from one step to the next.

Q 27. The _____ method involves the rapid generation of increasingly functional systems, with short intervals between successive releases.

Q 28. The _____, which is the fundamental, atomic level of data to be represented in the fact table for data warehouse design process.

Q 29. _____ processing supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts, or graphs.

Q 30. _____ supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

Answer:

1. data warehouse **2.** subject-oriented, integrated, time-variant, nonvolatile
3. OLTP **4.** OLTP **5.** OLAP **6.** OLAP **7.** data source
8. data warehouse **9.** bottom **10.** Data Mart **11.** middle **12.** Top
13. Data Cube **14.** base **15.** apex **16.** dimension, measure
17. Dimensions **18.** Fact **19.** Roll-up/drill-up/consolidation/aggregation
20. Drill-through **21.** Drill-down **22.** Drill-down **23.** Slice
24. Pivot **25.** bottom-up **26.** Waterfall **27.** Spiral
28. business process grain **29.** Information **30.** Data mining