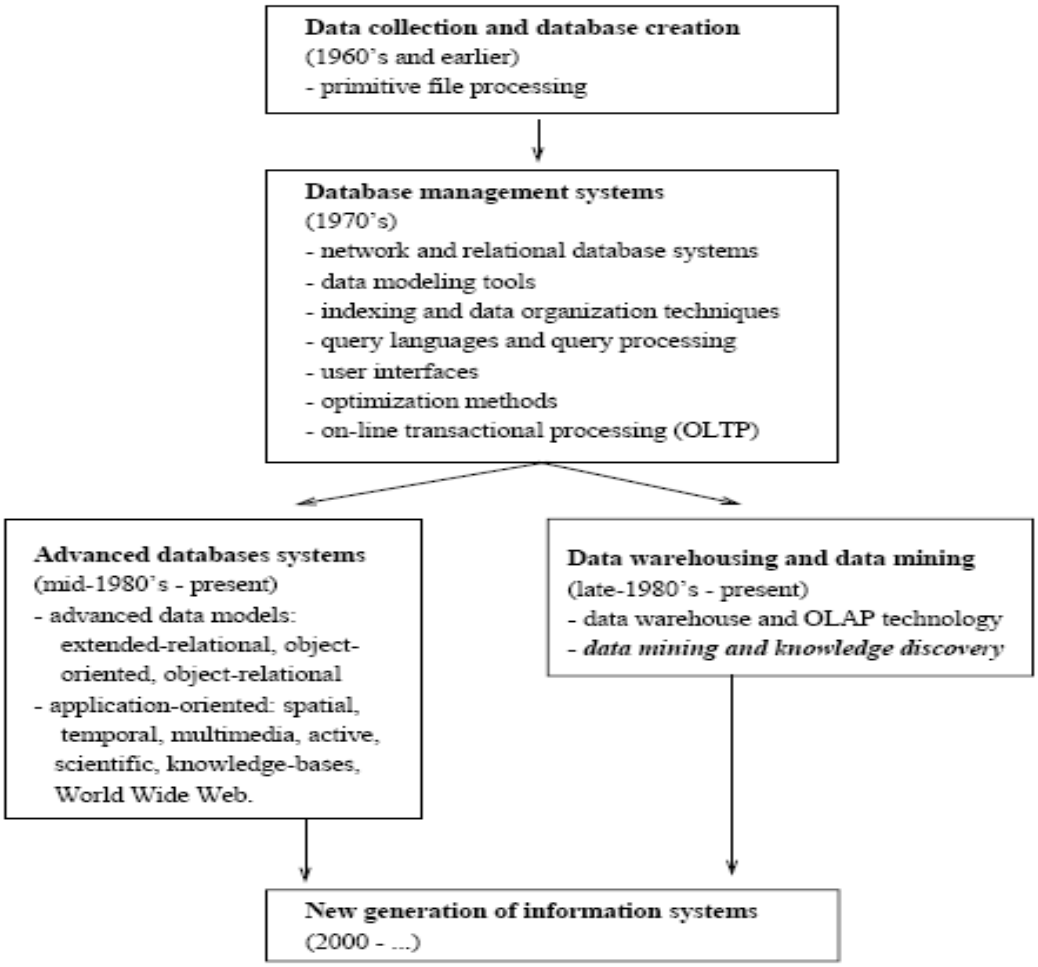
UNIT-I

## 1.1 What motivated data mining? Why is it important?

The major reason that data mining has attracted a great deal of attention in information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.

- The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration.

*The evolution of database system technology*




## 1.2 What is data mining

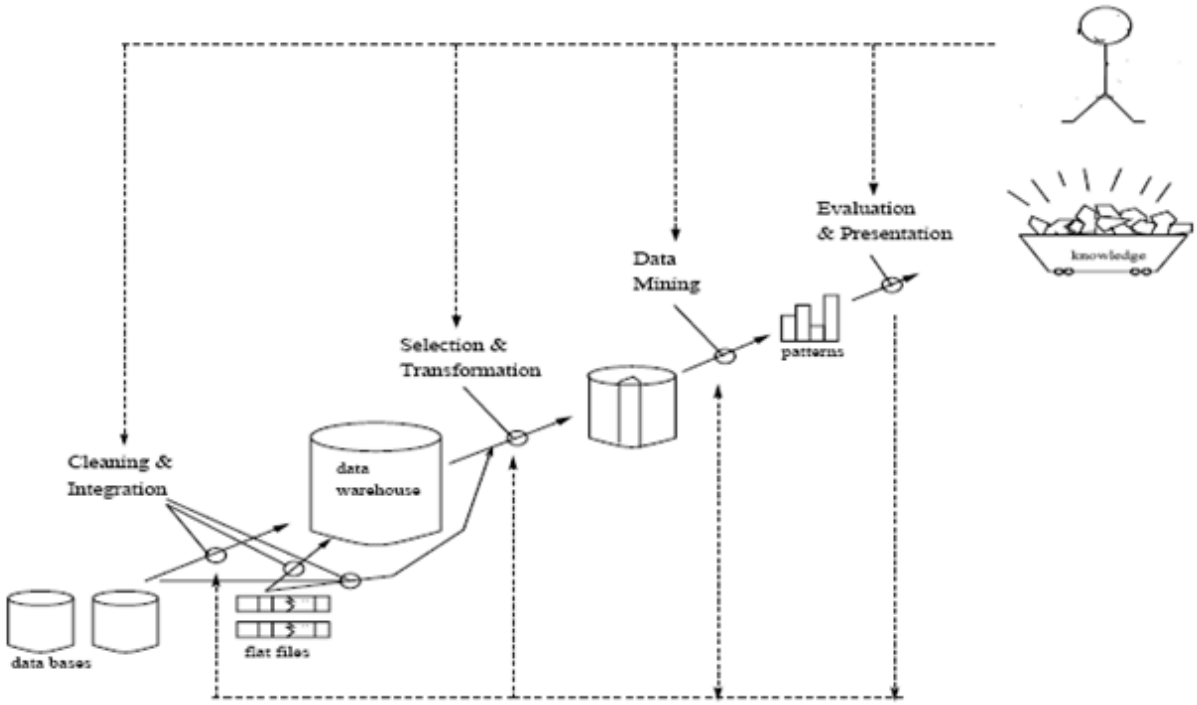
- Data mining refers to extracting or mining" knowledge from large amounts of data.
- There are many other terms related to data mining, such as knowledge mining, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

❖ Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases", or KDD

**Essential step in the process of knowledge discovery in databases**

Knowledge discovery as a process is depicted in following figure and consists of an iterative sequence of the following steps:

- data cleaning: to remove noise or irrelevant data
- data integration: where multiple data sources may be combined
- data selection: where data relevant to the analysis task are retrieved from the database
- data transformation: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations
- data mining :an essential process where intelligent methods are applied in order to extract data patterns
- pattern evaluation to identify the truly interesting patterns representing knowledge based on some interestingness measures
- knowledge presentation: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.



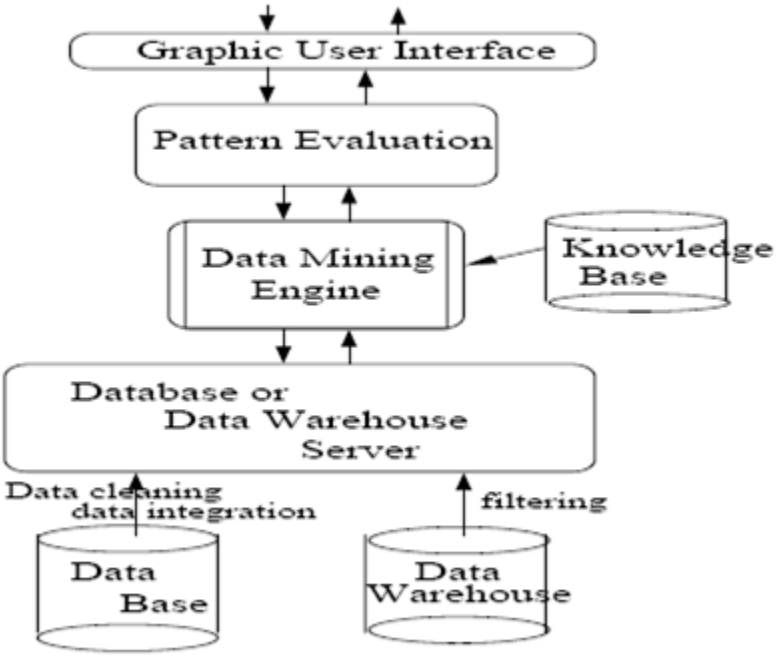

*Data mining as a step in the process of knowledge discovery.*

***Architecture of a typical data mining system/Major Components***

Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories.

❖ Based on this view, the architecture of a typical data mining system may have the following major components:

1. A database, data warehouse, or other information repository, which consists of the set of databases, data warehouses, spreadsheets, or other kinds of information repositories containing the student and course information.
2. A database or data warehouse server which fetches the relevant data based on users' data mining requests.
3. A knowledge base that contains the domain knowledge used to guide the search or to evaluate the interestingness of resulting patterns. For example, the knowledge base may contain metadata which describes data from multiple heterogeneous sources.
4. A data mining engine, which consists of a set of functional modules for tasks such as classification, association, classification, cluster analysis, and evolution and deviation analysis.
5. A pattern evaluation module that works in tandem with the data mining modules by employing interestingness measures to help focus the search towards interestingness patterns.
6. A graphical user interface that allows the user an interactive approach to the data mining system.




*Architecture of a typical Data mining system*

**How is a data warehouse different from a database? How are they similar?**

❖ Differences between a data warehouse and a database:

- A data warehouse is a repository of information collected from multiple sources, over a history of time, stored under a unified schema, and used for data analysis and decision support; whereas a database, is a collection of interrelated data that represents the current status of the stored data.

- There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another.
- A database system supports ad-hoc query and on-line transaction processing
- Similarities between a data warehouse and a database: Both are repositories of information, storing huge amounts of persistent data.

**1.2 Data mining: on what kind of data? / Describe the following advanced database systems and applications: object-relational databases, spatial databases, text databases, multimedia databases, the World Wide Web.**

- In principle, data mining should be applicable to any kind of information repository.
- This includes relational databases, data warehouses, transactional databases, advanced database systems, flat files, and the World-Wide Web.
- Advanced database systems include object-oriented and object-relational databases, and special c application-oriented databases, such as spatial databases, time-series databases, text databases, and multimedia databases.

***Flat files:***

Flat files are actually the most common data source for data mining algorithms, especially at the research level.

- Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied.
- The data in these files can be transactions, time-series data, scientific measurements, etc.

***Relational Databases:***

A relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships.

- Tables have columns and rows, where columns represent attributes and rows represent tuples.
- A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.
- In following figure it presents some relations Customer, Items, and Borrow representing business activity in a video store.
- These relations are just a subset of what could be a database for the video store and is given as an example.

Borrow

| customerID | date | itemID | # | … |
|---|---|---|---|---|
| C1234 | 99/09/06 | 98765 | 1 | … |
| … | | | | |

Customer

| customerID | name | address | password | birthdate | family_income | group | … |
|---|---|---|---|---|---|---|---|
| C1234 | John Smith | 120 main street | Marty | 1965/10/10 | $45000 | A | … |
| … | | | | | | | |

Items

| itemID | type | title | media | category | Value | # | … |
|---|---|---|---|---|---|---|---|
| 98765 | Video | Titanic | DVD | Drama | $15.00 | 2 | … |
| … | | | | | | | |

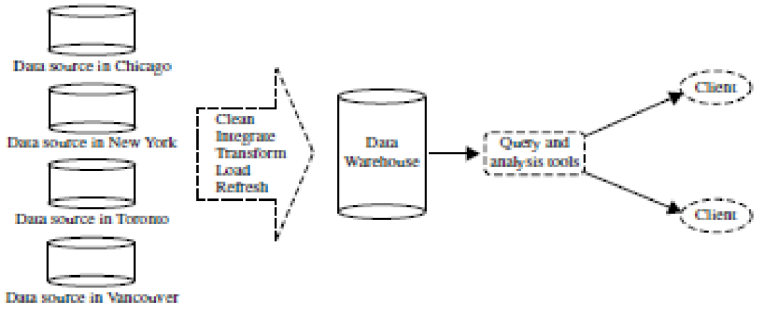
Fragments of some relations from a relational database for OurVideoStore.

- The most commonly used query language for relational database is SQL, which allows retrieval and manipulation of the data stored in the tables, as well as the calculation of aggregate functions such as average, sum, min, max and count.
- For instance, an SQL query to select the videos grouped by category would be:
  SELECT count(*) FROM Items WHERE type=video GROUP BY category.
- Data mining algorithms using relational databases can be more versatile than data mining algorithms specifically written for flat files, since they can take advantage of the structure inherent to relational databases.
- While data mining can benefit from SQL for data selection, transformation and consolidation, it goes beyond what SQL could provide, such as predicting, comparing, detecting deviations, etc.

## ***Data warehouses***

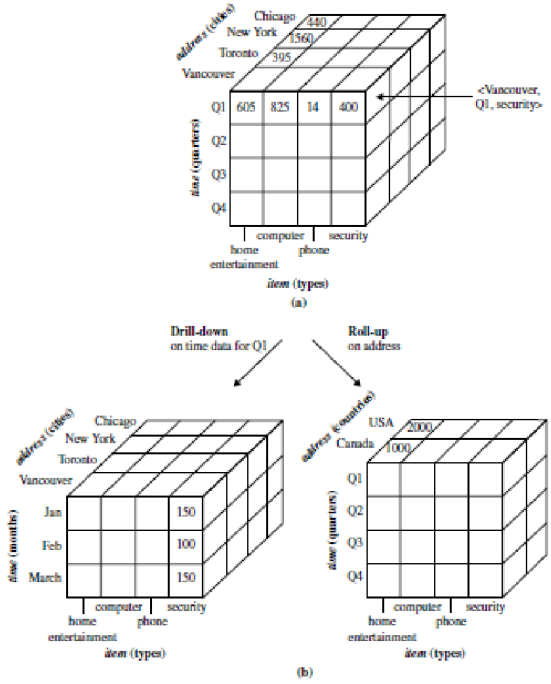
A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site.

- Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing.
- The figure shows the basic architecture of a data warehouse




*Typical framework of a data warehouse for AllElectronics*

- In order to facilitate decision making, the data in a data warehouse are organized around major subjects, such as customer, item, supplier, and activity.
- The data are stored to provide information from a historical perspective and are typically summarized.
- A data warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount.
- The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube.
- It provides a multidimensional view of data and allows the precomputation and fast accessing of summarized data.




*A multidimensional data cube, commonly used for data warehousing, (a) showing summarized data for AllElectronics and (b) showing summarized data resulting fromdrill-down and roll-up operations on the cube in (a). For improved readability, only some of the cube cell values are shown.*

- The data cube structure that stores the primitive or lowest level of information is called a base cuboid.
- Its corresponding higher level multidimensional (cube) structures are called (non-base) cuboids.
- A base cuboid together with all of its corresponding higher level cuboids form a data cube.
- By providing multidimensional data views and the precomputation of summarized data, data warehouse systems are well suited for On-Line Analytical Processing, or OLAP.
- OLAP operations make use of background knowledge regarding the domain of the data being studied in order to allow the presentation of data at different levels of abstraction. Such operations accommodate different user viewpoints.

- Examples of OLAP operations include drill-down and roll-up, which allow the user to view the data at differing degrees of summarization, as illustrated in above figure.

**Transactional databases**

In general, a transactional database consists of a flat file where each record represents a transaction.

- A transaction typically includes a unique transaction identity number (trans ID), and a list of the items making up the transaction (such as items purchased in a store) as shown below:

| *trans_ID* | *list_of_item_IDs* |
|---|---|
| T100 | I1, I3, I8, I16 |
| T200 | I2, I8 |
| ... | ... |

*Fragement of a transactional database for sales at AllElectronics*

**Advanced database systems and advanced database applications**

**An objected-oriented database** is designed based on the object-oriented programming paradigm where data are a large number of objects organized into classes and class hierarchies.

- Each entity in the database is considered as an object.
- The object contains a set of variables that describe the object, a set of messages that the object can use to communicate with other objects or with the rest of the database system and a set of methods where each method holds the code to implement a message.

**Spatial database**

**A spatial database** contains spatial-related data, which may be represented in the form of raster or vector data.

- Raster data consists of n-dimensional bit maps or pixel maps, and vector data are represented by lines, points, polygons or other kinds of processed primitives, Some examples of spatial databases include geographical (map) databases, VLSI chip designs, and medical and satellite images databases.
- Ex : 2-D satellite images may be represented as ***raster format*** .
  Maps can be represented in ***Vector format*** **.**

**Time – Series Databases :**

**Time-Series Databases:** Time-series databases contain time related data such stock market data or logged activities.

- These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis.
- Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.

**Text database**

**A text database** is a database that contains text documents or other word descriptions in the form of long sentences or paragraphs, such as product specifications, error or bug reports, warning messages, summary reports, notes, or other documents.

**Multimedia database**

**A multimedia database** stores images, audio, and video data, and is used in applications such as picture content-based retrieval, voice-mail systems, video-on-demand systems, the World Wide Web, and speech-based user interfaces.

**World- Wide Web**

**The World-Wide Web** provides rich, world-wide, on-line information services, where data objects are linked together to facilitate interactive access.

- Some examples of distributed information services associated with the World-Wide Web include America Online, Yahoo!, AltaVista, and Prodigy.

**1.3 Data mining functionalities/Data mining tasks: what kinds of patterns can be mined?**

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories:

- Descriptive
- predictive

Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

Describe data mining functionalities, and the kinds of patterns they can discover

(or)

Define each of the following data mining functionalities: characterization, discrimination, association and correlation analysis, classification, prediction, clustering, and evolution analysis. Give examples of each data mining functionality, using a real-life database that you are familiar with.

**1 .Concept/class description: characterization and discrimination**

Data can be associated with classes or concepts. It describes a given set of data in a concise and summarative manner, presenting interesting general properties of the data. These descriptions can be derived via

1. data characterization, by summarizing the data of the class under study (often called the target class) in general terms .
2. data discrimination, by comparison of the target class with one or a set of comparative classes
3. both data characterization and discrimination

**Data characterization**

It is a summarization of the general characteristics or features of a target class of data.

Example:

A data mining system should be able to produce a description summarizing the characteristics of a student who has obtained more than 75% in every semester; the result could be a general profile of the student.

The output of data characterization can be presented in various forms like

- Pie charts
- Bar charts
- Curves
- Multimedimensional cubes
- Multimedimensional tables etc.

The resulting descriptions can be presented as generalized relations or in rule forms called characteristic rules.

**Data Discrimination** is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

Example

The general features of students with high GPA's may be compared with the general features of students with low GPA's. The resulting description could be a general comparative profile of the students such as 75% of the students with high GPA's are fourth-year computing science students while 65% of the students with low GPA's are not.

Discrimination descriptions expressed in rule form are referred to as discriminant rules.

**2 . Mining Frequent patterns , Associations and correlations .**

**Frequent patterns**, as the name suggests, are patterns that occur frequently in data.

- There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures.
- A *frequent itemset* typically refers to a set of items that often appear together in a transactional data set—for example, milk and bread, which are frequently bought together in grocery stores by many customers.
- A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card, is a (*frequent*) *sequential pattern*.
- A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (*frequent*) *structured pattern*.
- Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

It is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. For example, a data mining system may find association rules like

major(X, "computing science"") ⇒ owns(X, "personal computer")

[support = 12%, confidence = 98%]

where X is a variable representing a student. The rule indicates that of the students under study, 12% (support) major in computing science and own a personal computer. There is a 98% probability (confidence, or certainty) that a student in this group owns a personal computer.

Example:

**Correlation analysis**

Correlation analysis is a technique use to measure the association between two variables.

- Correlation is degree or type of relationship b/w two or more quantities ( variables).
- A **correlation coefficient** (**r**) is a statistic used for measuring the strength of a supposed linear association between two variables. Correlations range from -1.0 to +1.0 in value.
- A correlation coefficient of 1.0 indicates a perfect positive relationship in which two or more variables fluctuate together ( one increases/decreases and other one also increases/decreases).
- A correlation coefficient of 0.0 indicates no relationship between the two variables. That is, one cannot use the scores on one variable to tell anything about the scores on the second variable.
- A correlation coefficient of -1.0 indicates a perfect negative relationship in which high values of one variable increases and other decreases.

**3 . Classification and prediction**

**Classification:**

Classification:

- It predicts categorical class labels
- It classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Typical Applications
  - credit approval
  - target marketing
  - medical diagnosis
  - treatment effectiveness analysis

- **Classification** can be defined as the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
- The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

Example:

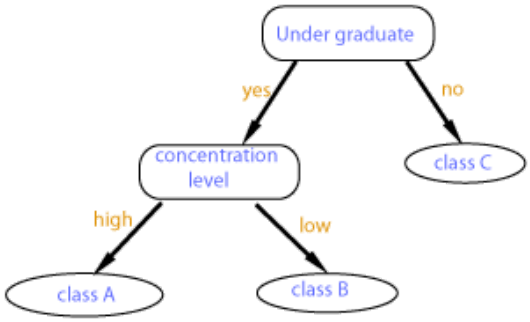
An airport security screening station is used to determine if passengers are potential terrorist or criminals. To do this, the face of each passenger is scanned and its basic pattern(distance between eyes, size, and shape of mouth, head etc) is identified. This pattern is compared to entries in a database to see if it matches any patterns that are associated with known offenders

- A classification model can be represented in various forms, such as

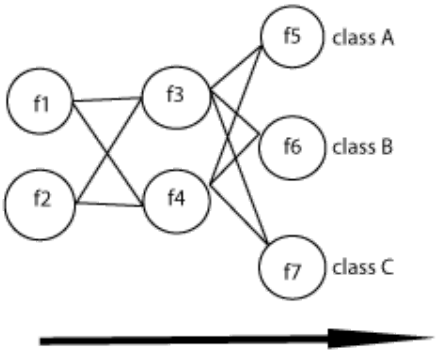
**1) IF-THEN rules,**

student ( class , "undergraduate") AND concentration ( level, "high") ==> class A
student (class ,"undergraduate") AND concentrtion (level,"low") ==> class B
student (class , "post graduate") ==> class C

**2) Decision tree**




**3) Neural network ( mathematical formulae )**




**Prediction:**

Find some missing or unavailable numerical data values rather than class labels referred to as prediction.

- Although prediction may refer to both numerical prediction and class label prediction, it is usually confined to numerical data value prediction and thus is distinct from classification.
- Prediction also encompasses the identification of distribution trends based on the available data.
- *Regression analysis* is a statistical methodology that is most often used for numerical prediction .
- Classification and prediction may need to be preceded by a ***relevance analysis*** , which attempts to identify attributes that do not contribute to the classification or prediction process. These attributes can then be excluded.

**Example:**

Predicting flooding is difficult problem. One approach is uses monitors placed at various points in the river. These monitors collect data relevant to flood prediction: water level, rain amount, time, humidity etc. These water levels at a potential flooding point in the river can be predicted based on the data collected by the sensors upriver from this point. The prediction must be made with respect to the time the data were collected.

**Classification vs. Prediction**

Classification differs from prediction in that the former is to construct a set of models (or functions) that describe and distinguish data class or concepts, whereas the latter is to predict some missing or unavailable, and often numerical, data values.

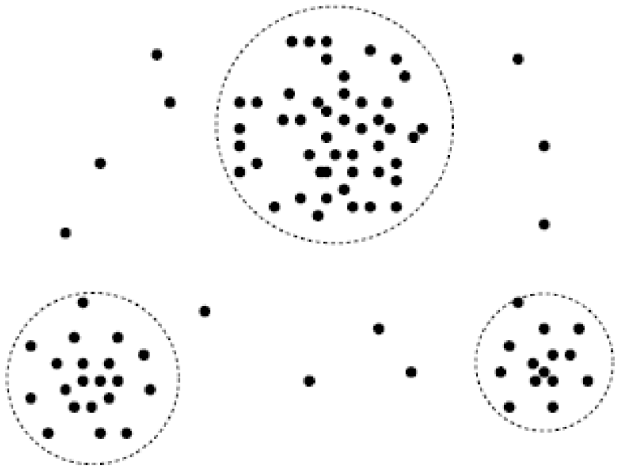
Their similarity is that they are both tools for prediction: Classification is used for predicting the class label of data objects and prediction is typically used for predicting missing numerical data values.

**4 . Clustering analysis**

Clustering analyzes data objects without consulting a known class label.

- The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.
- Cluster of objects are formed so that objects within a cluster have high similarity to one another , but are very dissimilar to objects in other clusters.
- Each cluster that is formed can be viewed as a class of objects.

Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together as shown below:



*A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.*

**Classification vs. Clustering**

- In general, in classification you have a set of predefined classes and want to know which class a new object belongs to.
- Clustering tries to group a set of objects and find whether there is *some* relationship between the objects.

In the context of machine learning, classification is *supervised learning* and clustering is *unsupervised learning*

**5 . Outlier analysis:**

A database may contain data objects that do not comply with general behavior or model of data.

- These data objects are outliers. In other words, the data objects which do not fall within the cluster will be called as outlier data objects.
- Noisy data or exceptional data are also called as outlier data. The analysis of outlier data is referred to as outlier mining.

Example

Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the location and type of purchase, or the purchase frequency.

**6 . Data evolution analysis**

It describes and models regularities or trends for objects whose behavior changes over time.

- Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Example:

The data of result the last several years of a college would give an idea if quality of graduated produced by it.

## Technologies used in data mining

Several techniques used in the development of data mining methods. **Some of them are mentioned below:**

***1. Statistics:***

- It uses the mathematical analysis to express representations, model and summarize empirical data or real world observations.
- Statistical analysis involves the collection of methods, applicable to large amount of data to conclude and report the trend.

***2. Machine learning***

- **Arthur Samuel** defined machine learning as a field of study that gives computers the ability to learn without being programmed.
- When the new data is entered in the computer, algorithms help the data to grow or change due to machine learning.
- In machine learning, an algorithm is constructed to predict the data from the available database **(Predictive analysis).**
- It is related to computational statistics.

**The four types of machine learning are:**

**1. Supervised learning**

- It is based on the classification.
- It is also called as **inductive learning**. In this method, the desired outputs are included in the training dataset.

**2. Unsupervised learning**

Unsupervised learning is based on clustering. Clusters are formed on the basis of similarity measures and desired outputs are not included in the training dataset.

**3. Semi-supervised learning**

Semi-supervised learning includes some desired outputs to the training dataset to generate the appropriate functions. This method generally avoids the large number of labeled examples (i.e. desired outputs) .

**4. Active learning**

- Active learning is a powerful approach in analyzing the data efficiently.
- The algorithm is designed in such a way that, the desired output should be decided by the algorithm itself (the user plays important role in this type).

***3. Information retrieval***

Information deals with uncertain representations of the semantics of objects (text, images).
**For example:** Finding relevant information from a large document.

**4. Database systems and data warehouse**

- Databases are used for the purpose of recording the data as well as data warehousing.
- Online Transactional Processing (OLTP) uses databases for day to day transaction purpose.
- To remove the redundant data and save the storage space, data is normalized and stored in the form of tables.
- **E**ntity-**R**elational modeling techniques are used for relational database management system design.
- Data warehouses are used to store historical data which helps to take strategical decision for business.
- It is used for online analytical processing (OALP), which helps to analyze the data.

***5. Decision support system***

- Decision support system is a category of information system. It is very useful in decision making for organizations.
- It is an interactive software based system which helps decision makers to extract useful information from the data, documents to make the decision.

## Data Mining Applications

Here is the list of areas where data mining is widely used −

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

**Financial Data Analysis**

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

**Retail Industry**

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry −

- Design and Construction of data warehouses based on the benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

**Telecommunication Industry**

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

**Biological Data Analysis**

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very

important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

**Other Scientific Applications**

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

**Intrusion Detection**

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

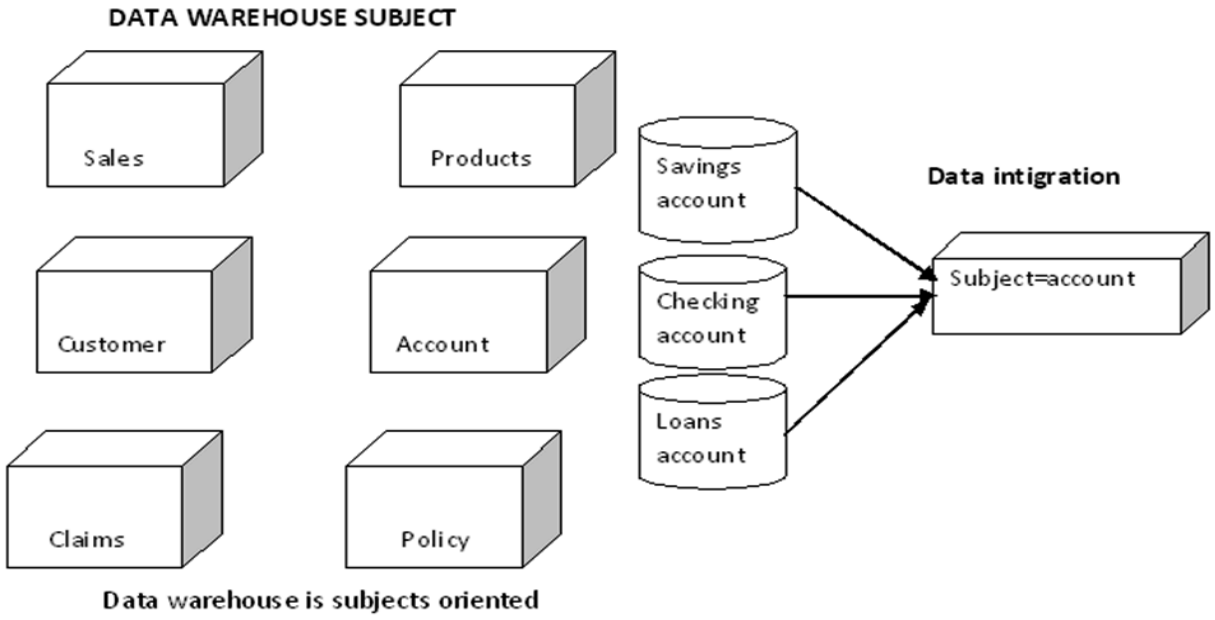
- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

**What is a data warehouse?**

A data warehouse is a subject oriented, integrate, time-variant, and nonvolatile collection of data in support of management's decision making process "Data warehousing: The process of constructing and using data warehouses".

**Subject-oriented**: A data warehouse is organized around major subjects, such as cus- tomer, supplier, product, and sales. Rather than concentrating on the day-to-day oper- ations and

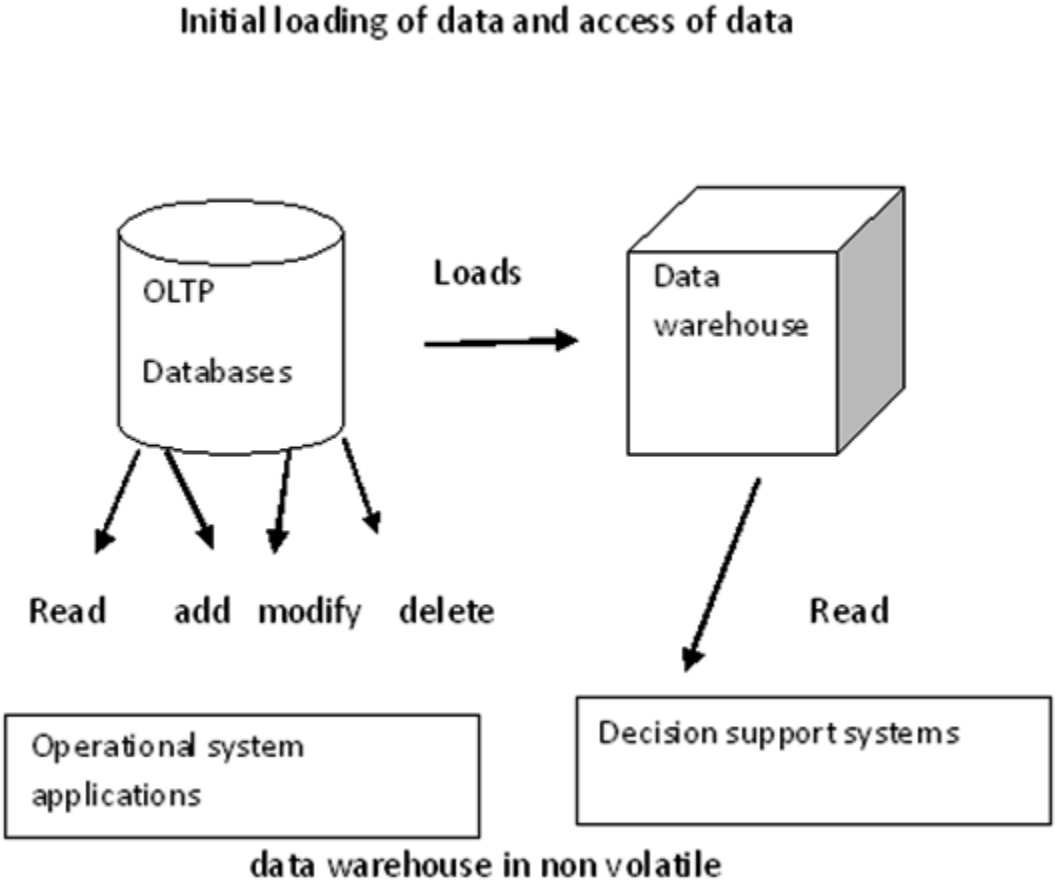
transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers.




**Integrate**d: A data warehouse is usually constructed by integrating multiple heteroge- neous sources, such as relational databases, flat files, and on-line transaction records. Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.

**Time-variant**: Data are stored to provide information from a historical perspective (e.g., the past 5–10 years). Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

**Nonvolatile:** A data warehouse is always a physically separate store of data trans- formed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: *initial loading of data* and *access of data*.




*3.1.1 Differences between Operational Database Systems and Data Warehouses*

What a data warehouse is by comparing these two kinds of systems.

- The major task of on-line operational database systems is to perform on-line transaction and query processing. These systems are called on-line transaction processing (OLTP) systems.

- They cover most of the day-to-day operations of an organization, such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.

- Data warehouse systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. These systems are known as on-line analytical processing (OLAP) systems.

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

| *Feature* | *OLTP* | *OLAP* |
|---|---|---|
| Characteristic | operational processing | informational processing |
| Orientation | transaction | analysis |
| User | clerk, DBA, database professional | knowledge worker (e.g., manager, executive, analyst) |
| Function | day-to-day operations | long-term informational requirements, decision support |
| DB design | ER based, application-oriented | star/snowflake, subject-oriented |
| Data | current; guaranteed up-to-date | historical; accuracy maintained over time |
| Summarization | primitive, highly detailed | summarized, consolidated |
| View | detailed, flat relational | summarized, multidimensional |
| Unit of work | short, simple transaction | complex query |
| Access | read/write | mostly read |
| Focus | data in | information out |
| Operations | index/hash on primary key | lots of scans |
| Number of records accessed | tens | millions |
| Number of users | thousands | hundreds |
| DB size | 100 MB to GB | 100 GB to TB |
| Priority | high performance,high availability | highflexibility,end-user autonomy |
| Metric | transaction throughput | query throughput, response time |

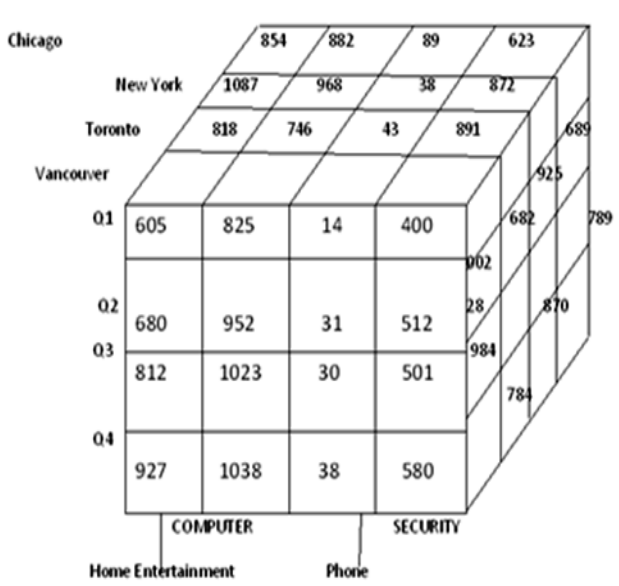
## *MULTIDIMENSIONAL DATA MODEL*

A data warehouse is based on multidimensional data model which views data in the form of data cube.

*From tables and spreadsheets to data cubes*

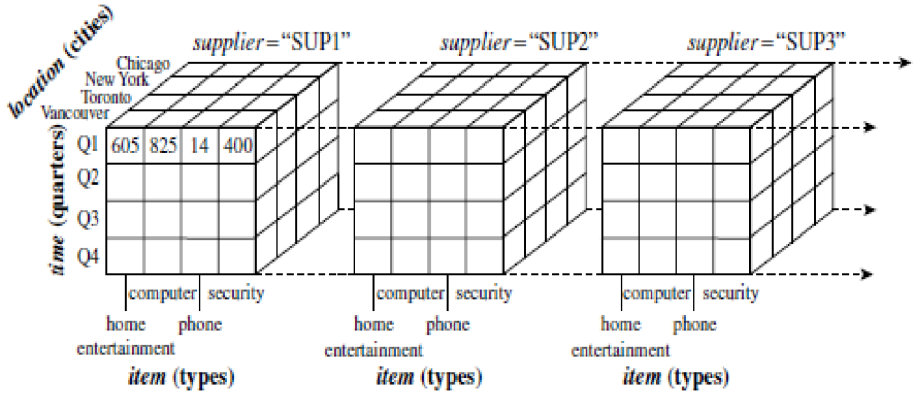
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions.
- Dimensions are perspectives or entities with respect to which an organization wants to keep records such as time, item, branch, location etc.
- Dimension table, such as item (item name, brand, type), or time (day, week, month, quarter, year) gives further descriptions about dimensions
- 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 base cuboids. The top most o-D cuboids, which hold the highest-level of summarization, called the apex cuboids. The lattice of cuboids forms a data cube.
- A 3-D view of sales data warehouse, according to the dimensions time, item, and location. The measure displayed is dollars sold (in thousands).

| | Location ="Chicago" Item | | | | Location="New York" item | | | | location="Toronto" item | | | | location="Vancouver" item | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Home Ent | comp. phone | sec | | Home ENT | comps. phone | sec | | Home ENT | comps. Phone | phone | sec | Home | sec | ent comp. | |
| Q1 | 854 | 882 | 89 | 623 | 1087 | 968 | 38 | 872 | 818 | 746 | 43 | 591 | 605 | 825 | 14 | 00 |
| Q2 | 943 | 890 | 64 | 698 | 1130 | 1124 | 41 | 925 | 894 | 795 | 52 | 682 | 680 | 952 | 31 | 512 |
| Q3 | 1032 | 924 | 59 | 789 | 1034 | 1048 | 45 | 002 | 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 |

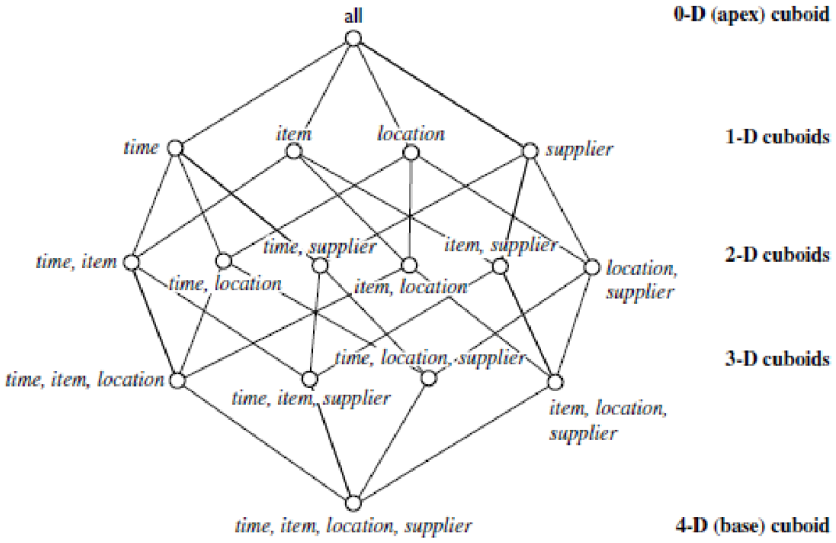



*A 3-D data cube representation of the data in above Table*

If we continue in this way, we may display any *n*-D data as a series of (*n* − 1)-D "cubes." The data cube is a metaphor for multidimensional data storage. The actual physical storage of such data may differ from its logical representation. The important thing to remember is that data cubes are *n*-dimensional and do not confine data to 3-D.




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




*Lattice of cuboids, making up a 4-D data cube for time, item, location, and supplier. Each cuboid represents a different degree of summarization.*

*Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Data Models*

- The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them. Such a data model is appropriate for on-line transaction processing.
- A data warehouse, however, requires a concise, subject-oriented schema that facilitates on-line data analysis.
- The most popular data model for a data warehouse is a **multidimensional model**, which can exist in the formof a **star schema**, a **snowflake schema**, or a **fact constellation schema**.

**Star schema:**

- The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (**fact table**) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (**dimension tables**), one for each dimension.
- The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

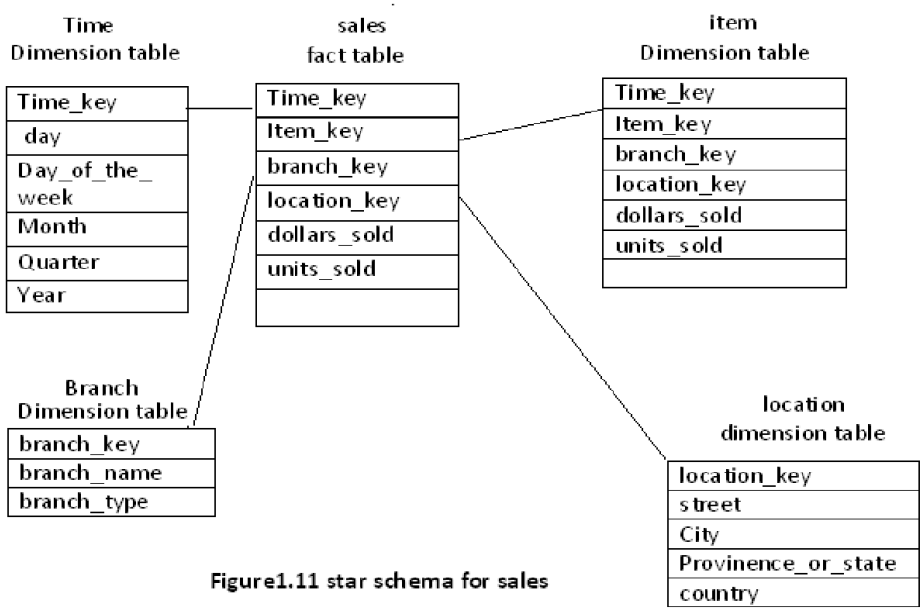



Figure1.11 star schema for sales

**Snowflake schema:**

- The snowflake schema is a variant of the star schema model, where some dimension tables are *normalized*, thereby further splitting the data into additional tables.
- The resulting schema graph forms a shape similar to a snowflake.
- The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies.

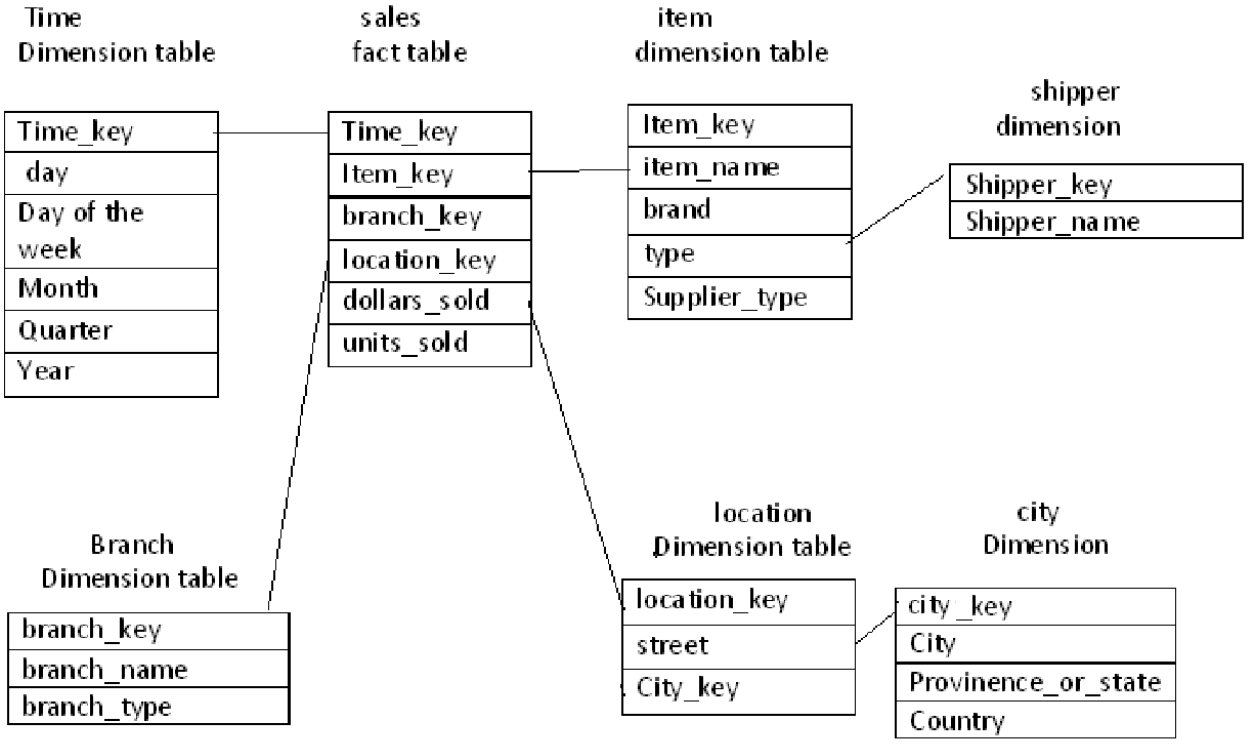
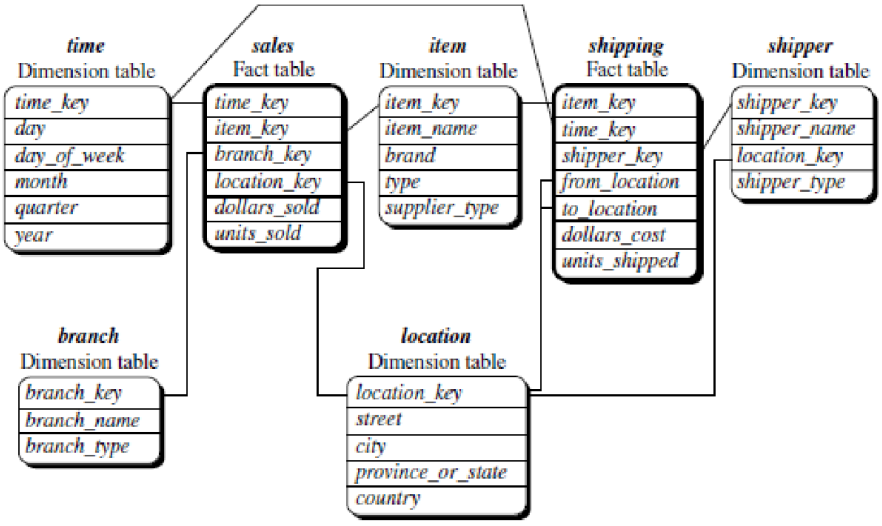



Figure 1.12 Snowflake schemas for sales

**Fact constellation:**

- Sophisticated applications may require multiple fact tables to *share* dimension tables.
- This kind of schema can be viewed as a collection of stars, and hence is called a **galaxy schema** or a **fact constellation .**




*Fact constellation schema of a sales and shipping data warehouse.*

*OLAP Operations in the Multidimensional Data Model*

- In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies.
- This organization provides users with the flexibility to view data from different perspectives.
- A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand.
- Hence, OLAP provides a user-friendly environment for interactive data analysis.

**Roll-up:**

- The roll-up operation (also called the *drill-up* operation by some vendors) performs aggregation on a data cube, either by *climbing up a concept hierarchy* for a dimension or by *dimension reduction*.
- Example : The result of a roll-up operation performed on the central cube by climbing up the concept hierarchy for *location* given in Figure.
- This hierarchy was defined as the total order "*street <city < province or state < country*."
- The roll-up operation shown aggregates the data by ascending the *location* hierarchy from the level of *city* to the level of *country*.

**Drill-down:**

- Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data.
- Drill-down can be realized by either *stepping down a concept hierarchy* for a dimension or *introducing additional dimensions*.
- Figure 4.12 shows the result of a drill-down operation performed on the central cube by stepping down a concept hierarchy for *time* defined as "*day* < *month* < *quarter* < *year*."
- Drill-down occurs by descending the *time* hierarchy fromthe level of *quarter* to the more detailed level of *month*.

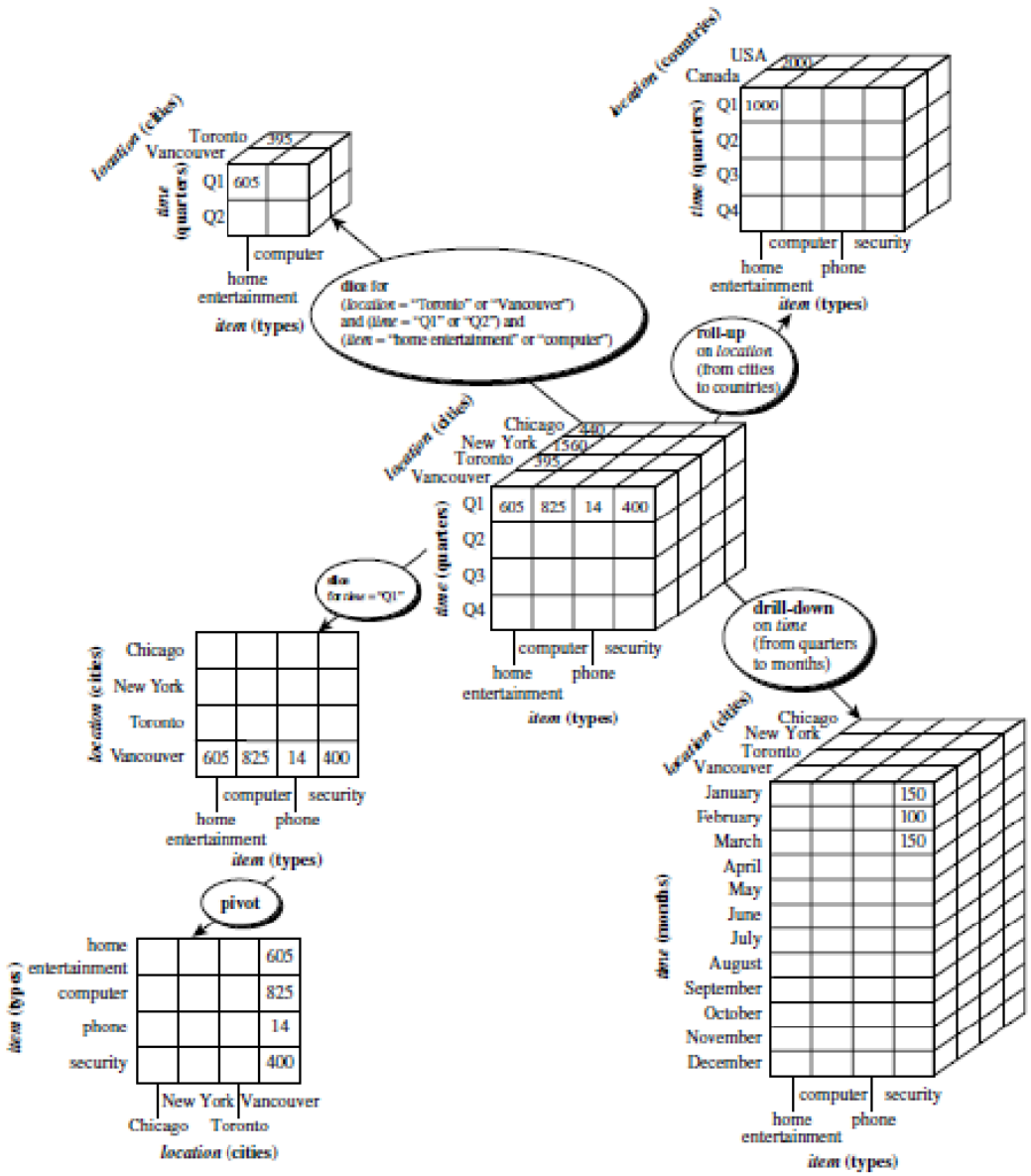
**Slice and dice:**

- The *slice* operation performs a selection on one dimension of the given cube, resulting in a subcube.
- Figure 4.12 shows a slice operation where the sales data are selected from the central cube for the dimension *time* using the criterion *time* = "Q1."
- The *dice* operation defines a subcube by performing a selection on two or more dimensions.\
- Figure 4.12 shows a dice operation on the central cube based on the following selection criteria that involve three dimensions: (*location* = "Toronto" or "Vancouver") and (*time* D="Q1" or "Q2") and (item D "home entertainment" or "computer").

**Pivot (rotate):**

- *Pivot* (also called *rotate*) is a visualization operation that rotates the data axes in view to provide an alternative data presentation.
- Figure 4.12 shows a pivot operation where the *item* and *location* axes in a 2-D slice are rotated. Other examples include rotating the axes in a 3-D cube, or transforming a 3-D cube into a series of 2-D planes.

**Other OLAP operations:** Some OLAP systems offer additional drilling operations. For example, **drill-across** executes queries involving (i.e., across) more than one fact table. The **drill-through** operation uses relational SQL facilities to drill through the bottom level of a data cube down to its back-end relational tables.




***Data warehouse Architecture***

*Steps for the Design and construction of Data warehouse are:*

1. Design of Data warehouse : A business analysis framework
2. Data warehouse design process

*1.Design of a data warehouse: a Business Analysis Framework*

- To design an effective data warehouse we need to understand and analyze business needs and construct a business framework.
- The construction of a large and complex information system can be viewed as the construction of a large and complex building, for which the owner, architect, a builder have different views.
- These views are combined to form a complex framework that represents the top-down, business-driven, or owner's perspective, as well as the bottom-up, builder-driven, or implementer's view of the information system.
- 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 an dimensional tables
  - Business query view - Sees the perspective of data in the warehouse from the view of end-user

*2 Data warehouse Design process*

- Top-down, Bottom-up approaches or a combination of both
- Top-down: Starts with overall Design an planning (mature)
- Bottom-up: Starts with experiments an prototypes(rapid)
- From software engineering point of view
  - Water fall: Structured an systemically analysis at each step before proceeding to the next
  - Spiral: Rapid generation of increasingly functional systems, short turn around time, quick turn around
- Typical data warehouse design process
  1. Choose a business process to model, e.g., orders, invoices, etc.
  2. Choose the grain (atomic level of data) of the business process
  3. Choose the dimensions that will apply to each fact table record.
  4. Dimension the measure that will populate each fact table record.

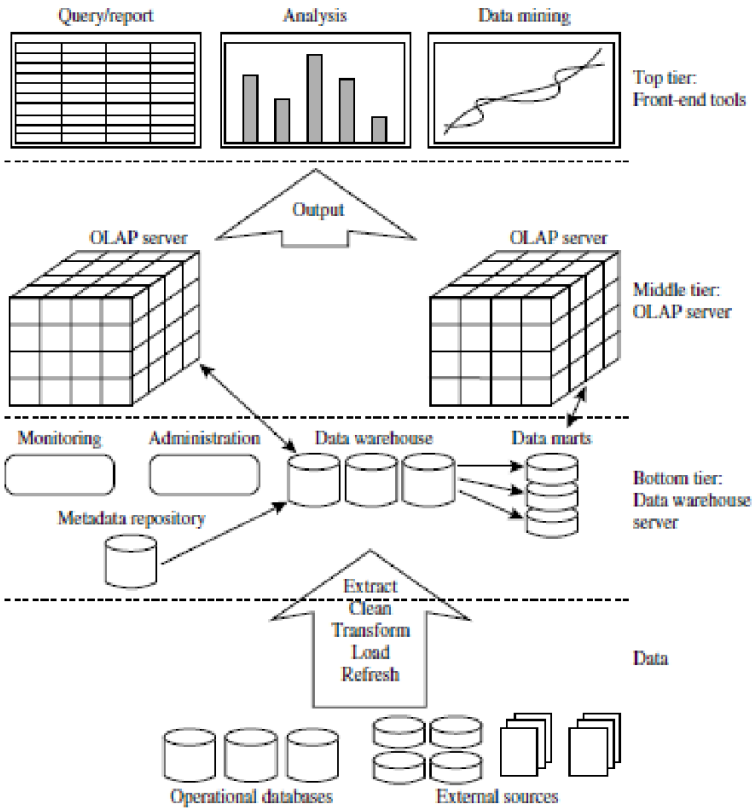
*Multi-Tiered Architecture- Three tier Data warehouse Architecture*

Data warehouse often adopt a three-tier architecture.

1. The bottom tier is a warehouse database server that us almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants). These tools and utilities perform data extraction, cleaning an transformation. This tier also contains a metadata respiratory, which stores information about the data warehouse and its contents

2. The middle tier is an OLAP server that is typically implemented using either(1) a relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations; or (2) a multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implemented multidimensional data and operations.

3. The top tier is a front-end client layer, which contains query and reporting, analysis tools, and/or data mining tools (e.g. trend analysis, prediction, and so on).




*A three-tier data warehousing architecture.*

Three Data Warehouse Models

1. Enterprise warehouse

Collects all of the information about subjects spanning in the enterprise the entire organization.

2. Data Mart

   A subset of corporate-wide data that is of value to a specific groups, such as marketing

3. Virtual Warehouse

   A set of views over operational database. Only some of the possible summary views may be materialized.

*3.3.5 OLAP Server Architectures*

1. Relational OLAP (ROLAP)
   - Use relational or extended-relational DBMS to store and manage warehouse data
   - And OLAP middle ware to support missing pieces
   - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
   - Greater scalability
2. Multidimensional OLAP (MOLAP)
   - Array-based multidimensional storage engine(sparse matrix techniques)
   - Fast indexing to pre-computed summarized data
3. Hybrid OLAP (HOLAP)
   - User flexibility, e.g., low level: relational, high-level: array
4. Specialized SQL servers
   - Specialized support for SQL queries over star/snowflake schemas