DATA WAREHOUSING

AND

DATA MINING

UNIT-1

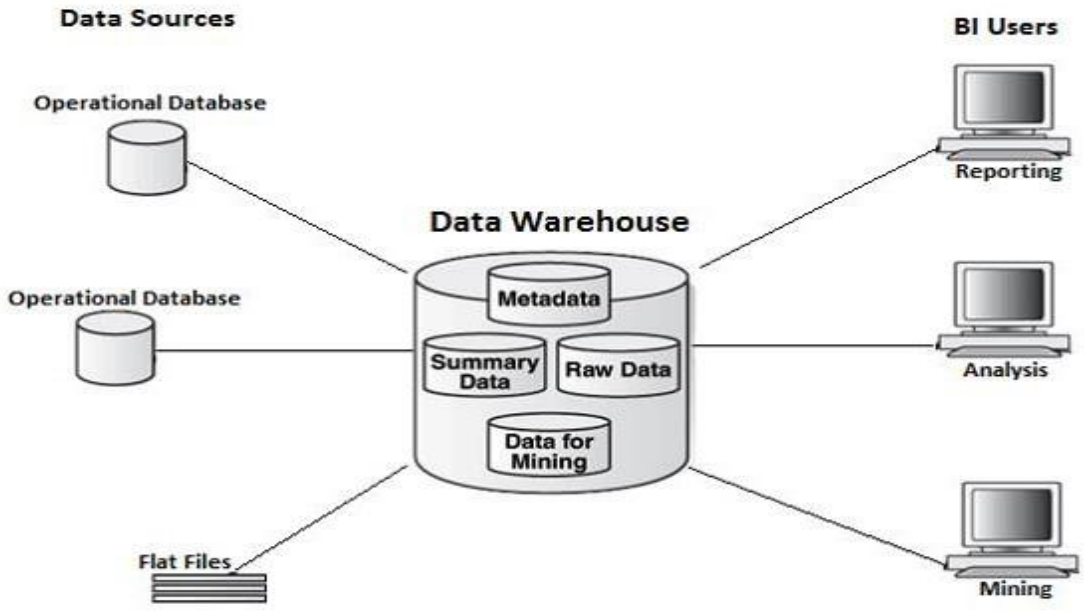
Introduction to data ware house Why Data Mining? What Is Data Mining? What Kinds of Data Can Be Mined?What Kinds of Patterns Can Be Mined? Which Technologies Are Used? Which Kinds of Applications Are Targeted? Major Issues in Data Mining. Data Objects and Attribute Types, Basic Statistical Descriptions of Data, Data Visualization, Measuring Data Similarity and Dissimilarity.

Introduction to Data warehouse:

A Data Warehouse consists of data from multiple heterogeneous data sources and is used for analytical reporting and decision making. Data Warehouse is a central place where data is stored from different data sources and applications.

The operational data store or other transformations before it is loaded to the DW system for information processing.

A Data Warehouse is used for reporting and analyzing of information and stores both historical and current data. The data in DW system is used for Analytical reporting, which is later used by Business Analysts, Sales Managers or Knowledge workers for decision-making.



In the above image, you can see that the data is coming from **multiple heterogeneous data** sources to a Data Warehouse. Common data sources for a data warehouse includes −

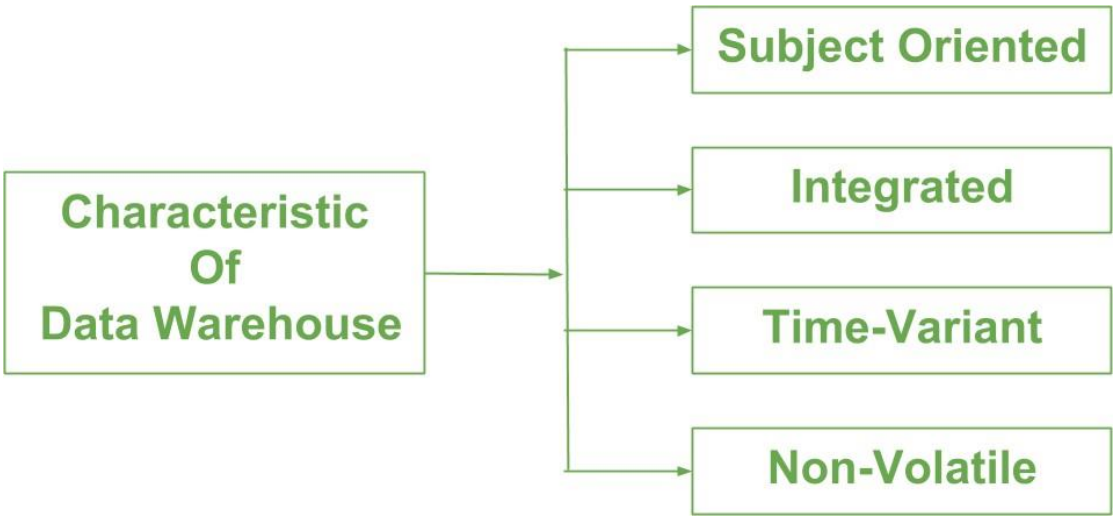
- Operational databases
- SAP and non-SAP Applications
- Flat Files (xls, csv, txt files)

Data in data warehouse is accessed by BI (Business Intelligence) users for Analytical Reporting, Data Mining and Analysis. This is used for decision making by Business Users, Sales Manager, Analysts to define future strategy.

Data Warehouse Definition:

A data warehouse is a subject-oriented, integrated, time-variant and non- volatile collection of data in support of management's decision-making process.

Characteristics/Features of Data Warehouse:

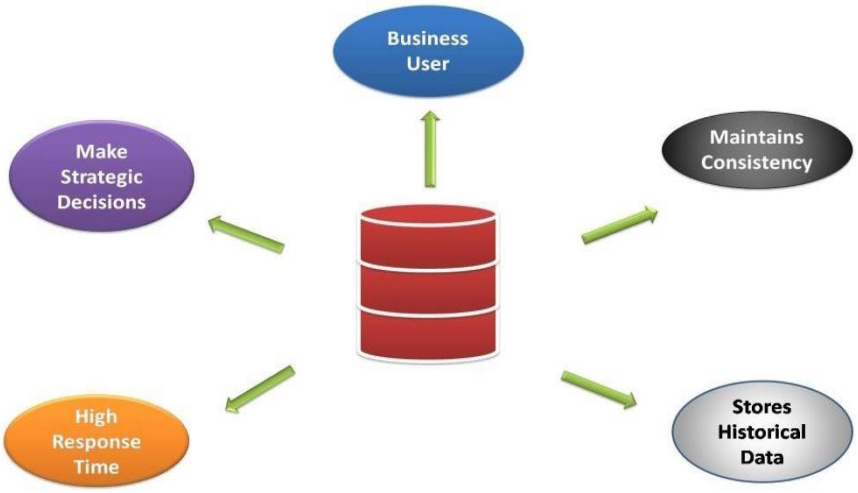


1. **Subject-Oriented :** A data warehouse target on the modeling and analysis of data for decision-makers. Therefore, data warehouses typically provide a concise and straightforward view around a particular subject, such as customer, product, or sales, instead of the global organization's ongoing operations. This is done by excluding data that are not useful concerning the subject and including all data needed by the users to understand the subject.
2. **Integrated:** A data warehouse integrates various heterogeneous data sources like RDBMS, flat files, and online transaction records. It requires performing data cleaning and integration during data warehousing to ensure consistency in naming conventions, attributes types, etc., among different data sources.
3. **Time-Variant:** Historical information is kept in a data warehouse. For example, one can retrieve files from 3 months, 6 months, 12 months, or even previous data from a data warehouse. These variations with a transactions system, where often only the most current file is kept.
4. **Non-Volatile :** The data warehouse is a physically separate data storage,

which is transformed from the source operational RDBMS. The operational updates of data do not occur in the data warehouse, i.e., update, insert, and delete operations are not performed.
It usually requires only two procedures in data accessing: Initial loading of data and access to data. Therefore, the DW does not require transaction processing, recovery, and concurrency capabilities, which allows for substantial speedup of data retrieval. Non-Volatile defines that once entered into the warehouse, and data should not change.

Need of Data Warehousing.

Data Warehousing is a progressively essential tool for business intelligence. It allows organizations to make quality business decisions. The data warehouse benefits by improving data analytics, it also helps to gain considerable revenue and the strength to compete more strategically in the market. By efficiently providing systematic, contextual data to the business intelligence tool of an organization, the data warehouses can find out more practical business strategies.



1. **Business User:** Business users or customers need a data warehouse to look at summarized data from the past. Since these people are coming from a non-technical background also, the data may be represented to them in an uncomplicated way.
2. **Maintains consistency:** Data warehouses are programmed in such a way that they can be applied in a regular format to all collected data from different sources, which makes it effortless for company decision-makers to analyze and share data insights with their colleagues around the globe. By standardizing the data, the risk of error in interpretation is also reduced and improves overall accuracy.
3. **Store historical data:** Data Warehouses are also used to store historical data that means, the time variable data from the past and this input can be used for various purposes.

4. **Make strategic decisions:** Data warehouses contribute to making better strategic decisions. Some business strategies may be depending upon the data stored within the datawarehouses.
5. **High response time:** Data warehouse has got to be prepared forsomewhat sudden masses and type of queries that demands a major degree of flexibility and fast latency.

Applications of Data warehouse

1. **Airline:** In the Airline system, it is used for operation purpose like crew assignment, analyses of route profitability, frequent flyer program promotions, etc.
2. **Banking:** It is widely used in the banking sector to manage the resources available on desk effectively. Few banks also used for themarket research, performance analysis of the product and operations.
3. **Healthcare:** Healthcare sector also used Data warehouse to strategize and predict outcomes, generate patient's treatment reports, share data with tie-in insurance companies, medical aid services, etc.
4. **Public sector:** In the public sector, data warehouse is used for intelligence gathering. It helps government agencies to maintain and analyze tax records, health policy records, for every individual.
5. **Investment and Insurance sector:** In this sector, the warehouses are primarily used to analyze data patterns, customer trends, and to track market movements.
6. **Retain chain:** In retail chains, Data warehouse is widely used for distribution and marketing. It also helps to track items, customer buying pattern, promotions and also used for determining pricing policy.
7. **Telecommunication**: A data warehouse is used in this sector for product promotions, sales decisions and to make distribution decisions.
8. **Hospitality Industry:** This Industry utilizes warehouse services to design as well as estimate their advertising and promotion campaigns where they want to target clients based on their feedback and travel patterns.

Advantages of Data Warehouses

1. **Faster and more efficient data analysis** – With a data warehouse, data can be processed and analyzed much more quickly than with traditional methods. This means that companies can get the information they need faster, which can help them make better decisions.
2. **Better decision-making** – By using a data warehouse, companies can gather and analyze more data than they could before. This allows them to make more informed decisions based on a deeper understanding of their

business.

3. **Improved data quality** – Data warehouses are designed to ensure that data is consistent and accurate. This means that companies can trust the information they are using to make decisions.

4. **Increased data accessibility** – Because data warehouses are organized in a way that makes it easy to access information, companies can quickly find the data they need. This can save time and improve productivity.

5. **Cost savings** – By using a data warehouse, companies can save money on storage and processing costs. This is because datawarehouses are designed to be more efficient than traditional storagemethods.

Disadvantages of Data Warehouses

1. **Costly setup and maintenance** – Building a data warehouse can bea time-consuming and expensive process. It requires specialized knowledge and resources to design and maintain, which can be a significant investment for companies.

2. **Limited flexibility** – Once a data warehouse is set up, it can be challenging to make changes to the data structure. This can make it difficult to adapt to changes in the business environment, or to incorporate new types of data.

3. **Data silos** – Because data warehouses are designed to store specific types of data, it can be challenging to integrate data from different sources. This can lead to data silos, where different teams or departments have their own sets of data that are not shared with others.

4. **Data latency** – Depending on the size and complexity of a data warehouse, it can take some time for data to be processed and

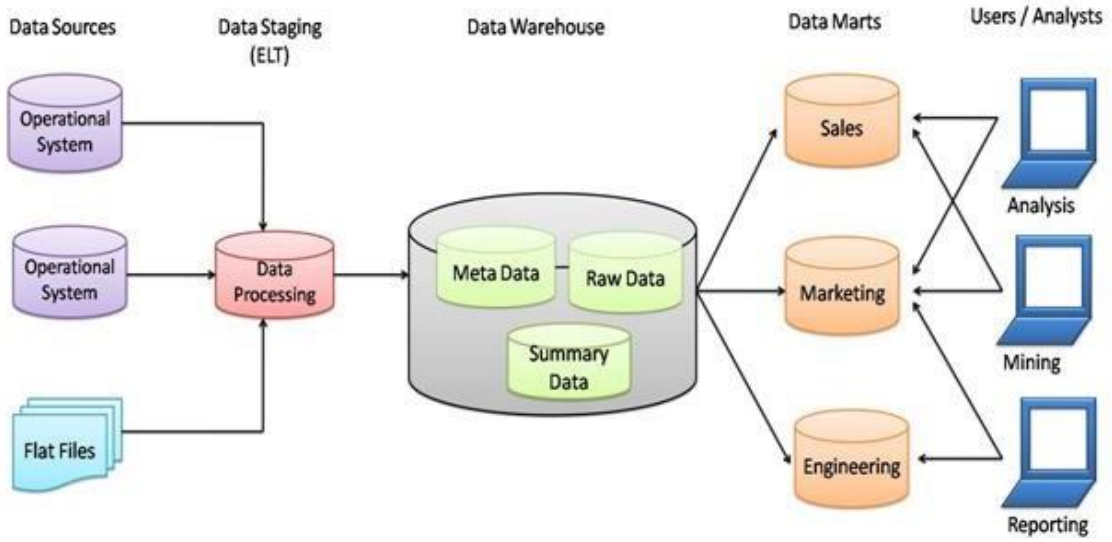
analyzed. This can create delays in decision-making, which can be adisadvantage in fast-paced business environments.

5. **Data security** – Data warehouses contain large amounts of sensitive information, which can make them a target for cyber-attacks or data breaches. Ensuring the security of a data warehouse requires ongoing vigilance and investment in cyber security measures.

Difference between Data Base vs Data warehouse:

Operational Database Systems	Data Warehouses
Operational systems are generally designed to support high-volume transaction processing.	Data warehousing systems are generally designed to support high-volume analytical processing. (i.e. OLAP).
Operational systems focuses on Data in.	Data warehousing systems focuses on Information out.
In Operational systems data is stored with a functional or process orientation.	In Data warehousing systems data is stored with a subject orientation.
Performance is low for analysis queries.	Performance is high for analysis queries.
It is used for Online Transactional Processing (OLTP)	It is used for Online Analytical Processing (OLAP).
Operational systems represent current transactions.	Data warehousing systems reads the historical data.
Data within operational systems are generally updated regularly.	Data within a data warehouse is non-volatile, meaning when new data is added old data is not erased so rarely updates.
Complex data structures.	Multi dimensional data structures.

Building Blocks/Components of Data Warehouse:



1. Source Data Component:

In the Data Warehouse, the source data comes from different places. Theyare group into four categories:

- **External Data:** For data gathering, most of the executives and data analysts rely on information coming from external sources for a numerous amount of the information they use. They use statistical features associated with their organization that is brought out by some external sources and department.
- **Internal Data:** In every organization, the consumer keeps their "private" spreadsheets, reports, client profiles, and generally even department databases. This is often the interior information, a part that might be helpful in every data warehouse.
- **Operational System data:** Operational systems are principally meant to run the business. In each operation system, we periodically take the old data and store it in achieved files.
- **Flat files:** A flat file is nothing but a text database that stores data in aplain text format. Flat files generally are text files that have all data processing and structure markup removed. A flat file contains a table with a single record per line.

2. Data Staging:

After the data is extracted from various sources, now it's time to prepare the data files for storing in the data warehouse. The extracted data collected from various sources must be transformed and made ready in a format that is suitable to be saved in the data warehouse for querying and analysis. The data staging contains three primary functions

- **Data Extraction:** This stage handles various data sources. Data analysts should employ suitable techniques for every data source.
- **Data Transformation:** As we all know, information for a knowledge warehouse comes from many alternative sources. If information extraction for a data warehouse posture huge challenges, informationtransformation gifts even important challenges. We tend to perform many individual tasks as a part of information transformation. First, we tend to clean the info extracted from every source of data. Standardization of information elements forms an outsized part of data transformation. Data transformation contains several kinds of combining items of information from totally different sources. Information transformation additionally contains purging supply information that's not helpful and separating outsourced records into new mixtures. Once the data transformation performs ends, we've got a set of integrated information that's clean, standardized, and summarized.
- **Data Loading:** When we complete the structure and construction of the data warehouse and go live for the first time, we do the initial loading of the data into the data warehouse storage. The initial load

moves high volumes of data consuming a considerable amount oftime.

3. Data Storage in Warehouse:

Data storage for data warehousing is split into multiple repositories. These data repositories contain structured data in a very highly normalized form for fast and efficient processing.

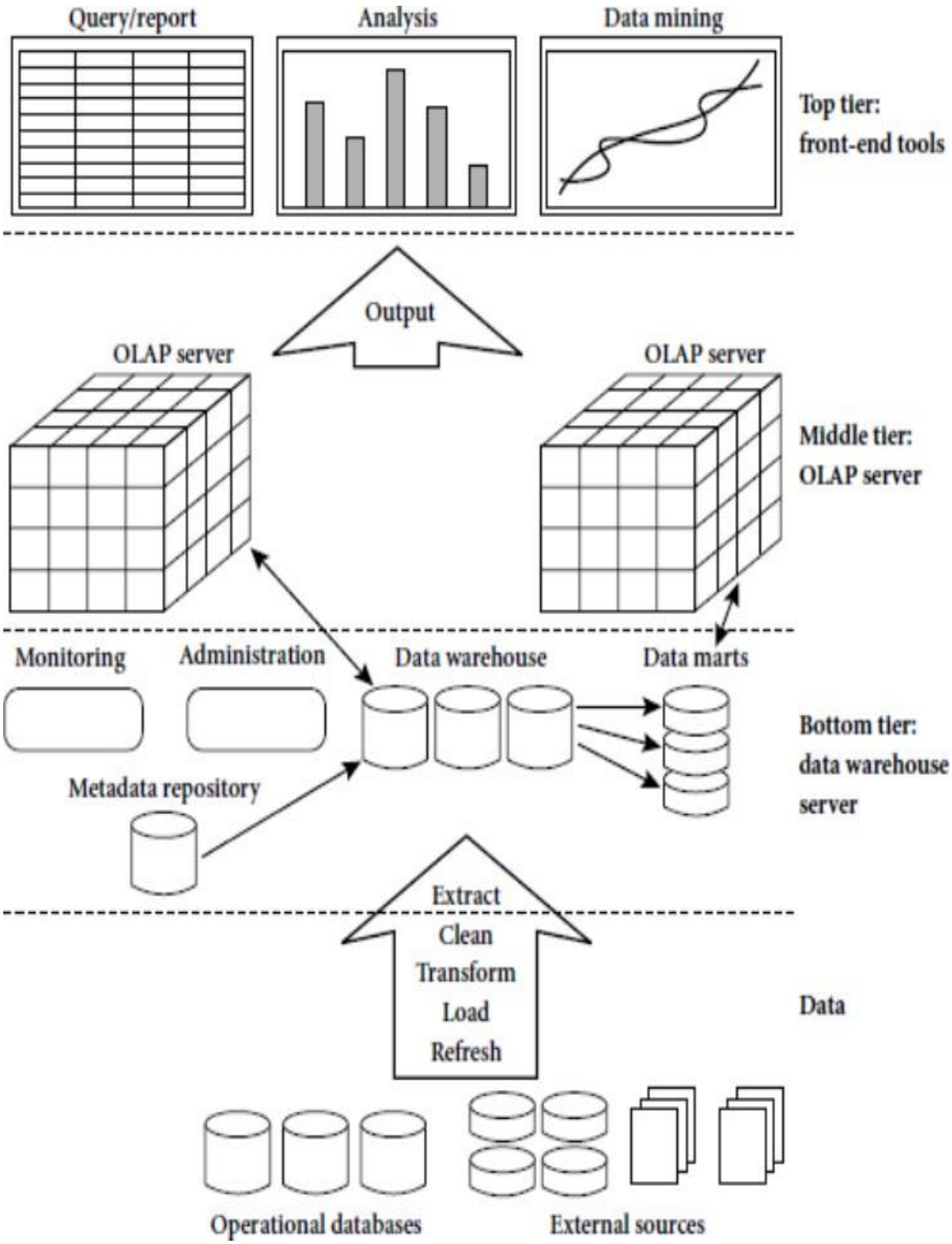
- **Metadata:** Metadata means data about data i.e. it summarizes basic details regarding data, creating findings & operating with explicit instances of data. Metadata is generated by an additional correction or automatically and can contain basic information about data.
- **Raw Data:** Raw data is a set of data and information that has not yet been processed and was delivered from a particular data entity to the data supplier and hasn't been processed nonetheless by machine or human. This data is gathered out from online sources to deliver deep insight into users' online behavior.
- **Summary Data or Data summary:** Data summary is an easy termfor a brief conclusion of an enormous theory or a paragraph. This is often one thing where analysts write the code and, in the end, they declare the ultimate end in the form of summarizing data. Data summary is the most essential thing in data mining and processing.

4. Data Marts:

Data marts are also the part of storage component in a data warehouse. It can store the information of a specific function of an organization that is handled by a single authority. There may be any number of data marts in a particular organization depending upon the functions. In short, data marts contain subsets of the data stored in data warehouses.

Now, the users and analysts can use data for various applications like reporting, analyzing, mining, etc. The data is made available to them whenever required.

A Three Tier Data Warehouse Architecture:



Tier-1: The bottom tier is a warehouse database server that is almostalways 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, and transformation (e.g., to merge similar data from differentsources into a unified format), as well as load and refresh functions to update the data warehouse. Example gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection). This tier also contains a metadata repository, which stores information about the data warehouse and its contents.

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

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

Difference between OLTP vs OLAP:

S.No.	OLAP	OLTP
1	OLAP stands for Online analytical processing.	OLTP stands for online transaction processing.
2	It includes software tools that help in analyzing data mainly for business decisions.	It helps in managing online database modification.
3	It utilizes the data warehouse.	It utilizes traditional approaches of DBMS.
4	It is popular as an online database query management system.	It is popular as an online database modifying system.
5	OLAP employs the data warehouse.	OLTP employs traditional DBMS.
6	It holds old data from various Databases.	It holds current operational data.
7	Here the tables are not normalized.	Here, the tables are normalized.
8	It allows only read and hardly write operations.	It allows both read and write operations.
9	Here, the complex queries are involved.	Here, the queries are simple.

Data Mining

INTRODUCTION: Data mining is nothing but discovery of *knowledge data* from large database. Generally, the term mining refers to mining of gold from rocks or sand is called gold mining.

1.1. Why Data Mining?

- The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and 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.
- Data mining can be viewed as a result of the natural evolution of information technology. It means, providing a path to extract the required data of an industry from warehousing machine. This is the witness of developing knowledge of an industry.
- It includes *data collection, database creation, data management (i.e* data storage and retrieval, and database transaction processing) and *data analysis and understanding (*involving data warehousing and data mining).

1.1.1. Evolution of data mining and data warehousing: In the development of data mining, we should know the evolution of database. This includes,

Data collection and Database creation*:* In the 1960's, database and information technology began with file processing system. It is powerful database system. But it is providing inconsistency of data. It means, a user needs to maintain duplicate data of an industry.

Database Management System*:* In b/w 1970 – 1980, the progress of database is
→ Hierarchical and network database systems were developed.
→ Relational database systems were developed
→ Data modeling tools were developed in early 1980s (such as E-R model etc.
→ Indexing and data organization techniques were developed. (such as B+ tree, hashing etc).
→ Query languages were developed. (such as SQL, PL/SQL)
→ User interfaces, forms and reports, query processing.
→ On-line transaction processing (OLTP)

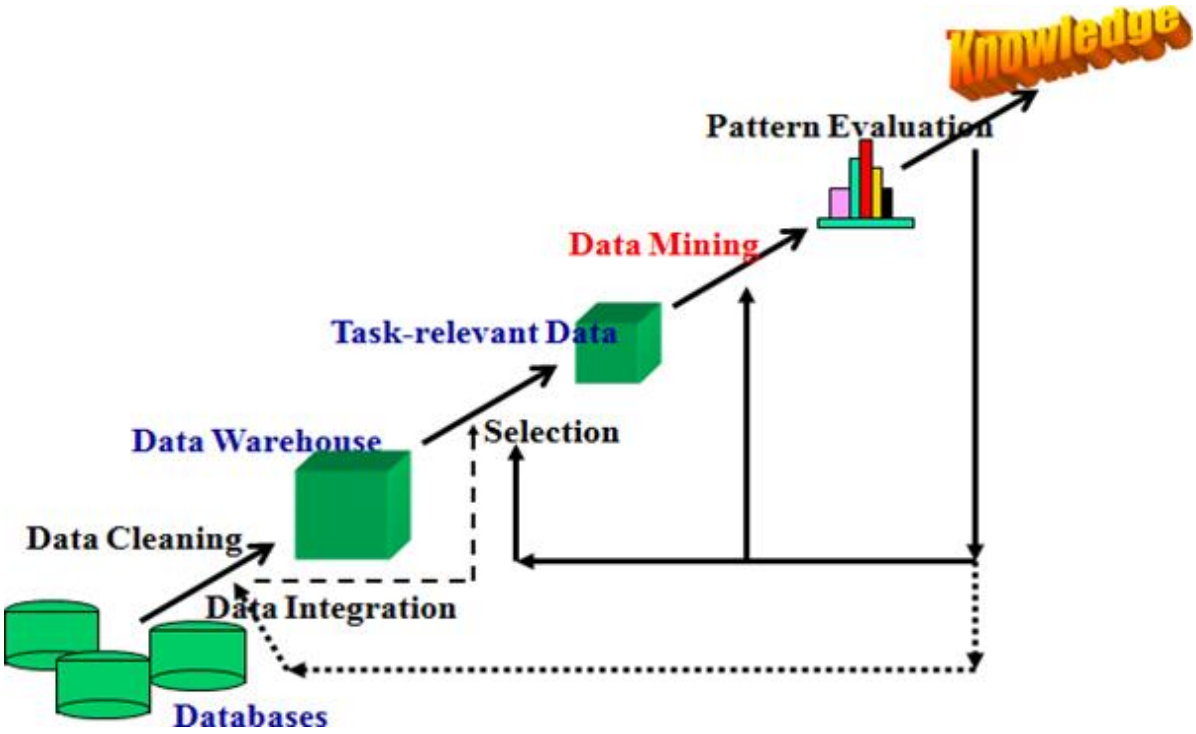
Advanced Database Systems: In mid 1980s to till date,
→ Advanced data models were developed. (such as extended relational, object-oriented, object-relational, spatial, temporal, multimedia, scientific databases etc.

Data Warehousing and Data mining*:* In late 1980 to till date
→ Developed Data warehouse and OLAP technology
→ Data mining and knowledge discovery were introduced.

Web-based Databases Systems: In 1990 – till date
→ XML based database systems and web mining were developed.
New Generation of Integrated Information Systems**:** From 2000 onwards developed an integrated information system.

What is Data Mining: The term Data Mining refers to *extracting or "mining" knowledge*

from large amounts of data. The term mining is actually a misnomer (i.e. unstructured data). For example, mining of gold from rocks or sand is referred to as gold mining.
Data mining is the process ofdiscovering meaningful new trends bystoring the large amount of data inrepository of database. It also uses pattern recognition techniques as well as statistical techniques.



1.1.2. Data mining steps in the knowledge discovery process (KDD):

1. **Data cleaning**: To remove noise and inconsistent data
2. **Data integration**: where multiple data sources may be combined
3. **Data selection**: where data relevant to the analysis task are retrieved from the database
4. **Data transformation**: where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations
5. **Data mining**: an essential process where intelligent methods are applied to extract data patterns
6. **Pattern evaluation**: To identify the truly interesting patterns representing knowledge based on interestingness measures
7. **Knowledge presentation**: where visualization and knowledge representation techniques are used to present mined knowledge to users.

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledgebase.

1.2. What Kind of Data Can Be Mined?

Data mining can be applied to any kind of information repositories such as Databases data, data warehouse, transactional data bases, advanced systems, flat files and the *World Wide Web*. Advanced databases systems include object-oriented, object-relation databases, time series databases, text databases and multimedia databases.

→**1.3.1.** ***Databases Data:*** A database system is also called a ***database management system*** **(DBMS).** It consists of a collection of interrelated data, known as a database, and set of software programs to manage and access the data. The software programs provide mechanisms for defining database structures and data storage. These also provide data consistency and security, concurrency, shared or distributed data access etc.

A ***relational database*** is a collection of tables, each of which is assigned a unique name. Each table consists of a set of **attributes** (columns or fields) and a set of **tupples** (records or rows). Each **tupple** is identified by a unique key and is described by a set of attribute values. For this, ER models are constructed for relational databases. For example, *AllElectronics* Industry illustrated with following information. *custer, item, employee, branch.*

customer table

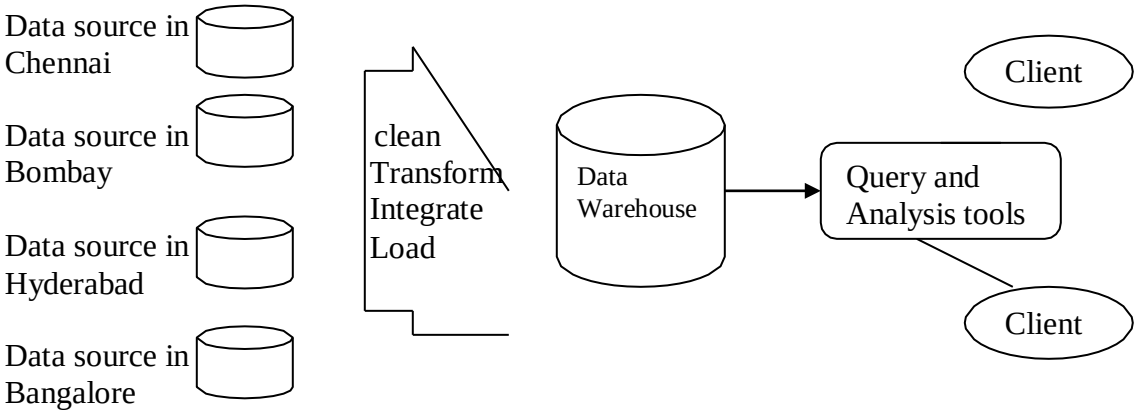
Cust-id	cust_name	Gender	Address Place	item-id

item table

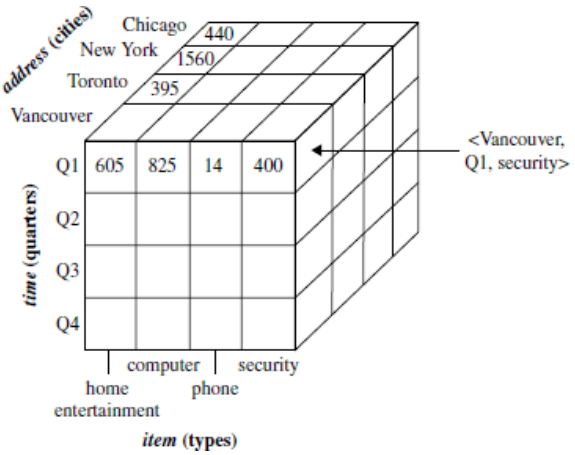
item-id	item_name	Price	Manufacturing

AllElectronics company sales his products (such as computers and printers) to the customers. Here providing the relation b/w **custormer** table (file) and **product** table. By this relation can identify what types of products are taken the customer.

→**1.3.2.** ***Data Warehouses*****:** A data warehouse is a repository of information collected from multiple sources, stored under a schema and resides at a single site. The data warehouses are constructed by a process of data cleaning, data transformation, data integration, data loading and periodic data refreshing.



A data warehouse is mainly 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 sales amount. The 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 preprocess and fast accessing of summarized data.



A data cube for summarized sales data of AllElectronics is presented in fig. The cube has three dimensions such as address (Chennai, Bombay, Hyd, Bang), time with Q1,Q2,Q3,Q4 and item with home needs, computer, phone and security. In this, aggregate value stored in each cell of the cube.

By providing multidimensional data views, performed the OLAP operations. Such as **drill-down,** and **roll-up.**

→**1.3.3.** ***Transactional Databases*****:** A transactional database consists of a file where each record represents a transaction. A transaction includes a unique transaction such as data of the transaction, the customer id number, the ID number of the sales person and so on.

AllElectronics transactions can be stored in a table with one record per transaction. This is shown in fig.

Transaction_id	List of items	Transaction dates
T100	I1, I3, I8, I16	18-12-2018
T200	I2, I8	18-12-2018

1.4 What Kinds of Patterns Can Be Mined? (or) Data Mining Functionalities:

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining tasks are classified into two categories ***descriptive*** and ***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.

1.4.1. Concept/Class Description: Descriptions of a individual classes or a concepts in summarized, concise and precise terms called class or concept descriptions. These descriptions can be divided into 1. Data Characterization 2. Data Discrimination.

Data Characterization:

- It is summarization of the general characteristics of a target class of data (forms).
- The data corresponding to the user specified class are collected by a database query.

The output of data characterization can be presented in various forms like *pie charts, bar charts curves, multidimensional cubes, multidimensional tables* etc. The resulting descriptions can be presented as generalized relations are called *characteristic rules.*

Data Discriminations: Comparison of two target class data objects from one or set of contrasting (distinct) classes. The target and contrasting classes can be specified by the user, and the corresponding data objects are retrieved through database queries.

For example, comparison of products whose sales increased by 10% in the last year with those whose sales decreased by 30% during the same period. This is called data discrimination.

1.4.2. Mining Frequent Patterns, Associations and Correlations:

1.4.2.1. Frequent Patterns: A *frequent itemset* typically refers to a set of items that often appear in a transactional data. For example, milk, and bread are frequently purchased by many customers. *AllElectronics industry occurring the products which are frequently purchased by the customers.* Generally, home needs are frequently used by the more customers.

1.4.2.2. **Association Analysis**: "*What is association analysis ?"*

Association analysis is the discovery of *association rules* showing attribute with value conditions that occur frequently together the given set of data. It is used for transaction data analysis. The Association rule of the form X ==> Y.

For example, In AllElectronics relational database, data mining system may find association rules like

buys(X, "computer") ==> buys(X, "software")

Here, who buys "computer", they buys "software".

age (X, "20 .. 29") & income (X, "20k .. 29k") ==> buys(X,

In this, the Association rule indicate that that indicates who employee of AllElectronics have the age b/w 20 to 29 and earning income b/w 20000 to 29000 are purchased CD player at AllElectronics Company.

1.4.2.3. Classification and Regressive prediction:

Classification is the process of finding a set of models that describes and distinguishes data classes or concepts.

- The derived model may be represented in various forms such as *classification (IF-THEN) rules, decision trees, mathematical formulae or neural networks.*
- A *decision tree* is a **flow-chart** like tree structure. The decision trees can easily converted to classification rule. The neural networks are used for classification to provide connection b/w computers.

age(X, "youth") AND income(X, "high") ⟶ *class(X, "A")*
age(X, "youth") AND income(X, "low") ⟶ *class(X, "B")*
age(X, "middle_aged") ⟶ *class(X, "C")*
age(X, "senior") ⟶ *class(X, "C")*

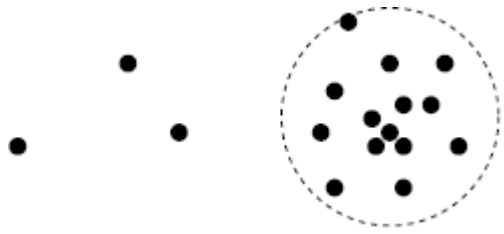
Regression for Predication is used to predict missing or unavailable data values rather than class labels. Prediction refers to both *data value prediction* and class *label prediction*. The predicted values are numerical data and are often referred to as ***prediction.***

1.4.2.4. **Cluster Analysis:** ("*What is cluster analysis?")*

Clustering is a method of grouping data into different groups, so that in each group share similar trends and patterns. The objectives of clustering are

- To uncover natural groupings
- To initiate hypothesis about the data
- To find consistent and valid organization of data.

For example, Cluster analysis can be performed on AllElectronics customers. It means, to identify homogeneous (same group) customers. By this cluster may represent target groups for marketing to increase the sales.

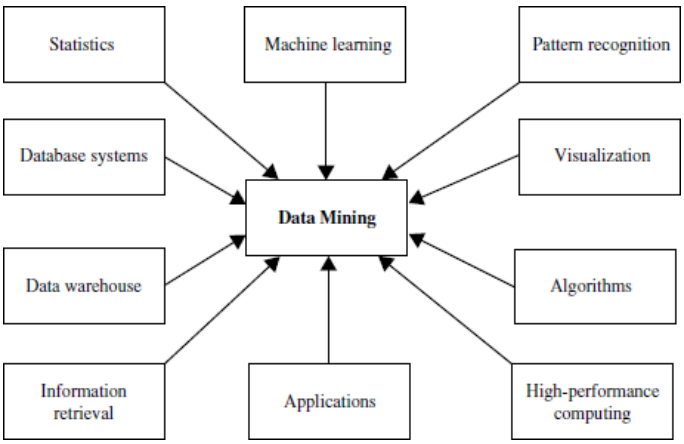


1.4.2.5. Outlier Analysis: In this analysis, a database may contain data objects that do not do what someone wants. Most data mining methods discard outliers as noise or exceptions. Finding such type of applications are fraud detection is referred as **outlier mining.**

For example, Outlier analysis may uncover usage of credit cards by detecting purchases of large number of products when comparing with regular purchase of large product customers.

1.5. Which Technologies Are Used? (or) Classification of Data Mining Systems:

Data mining is classified with many techniques. Such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, high performance computing, and many application domains (Shown in Figure).Data mining system can be categorized according to various criteria.



Statistics: A *statistical model* is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions. Statistical models are widely used to model data and data classes. For example, in data mining tasks like data characterization and classification, statistical models of target classes can be built.

Machine Learning: *Machine learning* investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to *automatically* learn to recognize complex patterns and make intelligent decisions based on data. Machine learning is a fast-growing discipline.

- **Supervised learning** is basically a synonym for classification. The supervision in the learning comes from the labeled examples in the training data set. For example, in the postal code recognition problem, a set of handwritten postal code images and their corresponding machine-readable translations are used as the training examples, which supervise the learning of the classification model.
- **Unsupervised learning** is essentially a synonym for clustering. The learning processis unsupervised since the input examples are not class labeled. For example, an unsupervised learning method can take, as input, a set of images of handwritten digits. Suppose that it finds 10 clusters of data. These clusters may correspond to the 10 distinct digits of 0 to 9, respectively.
- **Semi-supervised** learning is a class of machine learning techniques that make use of both labeled and unlabeled examples when learning a model. For a two-class problem, one class as the *positive examples* and the other class as the *negative examples.*
- **Active learning** is a machine learning approach that lets users play an active role in the learning process.

Database Systems and Data Warehouses:

- **Database systems** can focus on the creation, maintenance, and use of databases for organizations and end-users. Particularly, database systems principles in data models, query languages, query processing and optimization methods, data storage, and indexing and accessing methods. Many data mining tasks need to handle large data sets or even real-time,

fast streaming data. Recent database systems have built systematic data analysis capabilities on database data using data warehousing and data mining facilities.

- **A data warehouse** integrates data from multiple sources and various timeframes. It provides OLAP facilities in multidimensional databases to promote multidimensional data mining. It maintains recent data, previous data and historical data in database.

Information Retrieval:

- **Information retrieval** (IR) is the science of searching for documents or information in documents. The typical approaches in information retrieval adopt probabilistic models. For example, a text document can be observing as a container of words, that is, a multi set of words appearing in the document.

Pattern recognition is the process of recognizing patterns by using machine learning algorithm. Pattern recognition can be defined as the classification of data based on knowledge already gained or on statistical information extracted from patterns and/or their representation. One of the important aspects of the pattern recognition is its application potential. **Examples:** Speech recognition, speaker identification, multimedia document recognition (MDR), automatic medical diagnosis.

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.

An ***algorithm*** in data mining (or machine learning) is a set of heuristics and calculations that creates a model from data. To create a model, the algorithm first analyzes the data you provide, looking for specific types of patterns or trends.

High Performance Computing (HPC) framework which can abstract the increased complexity in current computing systems and at the same time provide performance benefits by exploiting multiple forms of parallelism in Data Mining algorithms.

Data Mining Applications*:* 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.*

1.6. Which Kinds of Applications Are Targeted?

Data mining has seen great successes in many applications. Presentations of data mining in knowledge-intensive application domains, such as bioinformatics and software engineering,

- **Business intelligence** (**BI**) technologies provide historical, current, and predictive views of business operations. Examples include reporting, online analytical processing, business performance management, competitive intelligence, benchmarking, and predictive analytics.
    - Data mining is the core of business intelligence. Online analytical processing tools in business intelligence depend on data warehousing and multidimensional data mining. Classification and prediction techniques are the core of predictive analytics in business intelligence, for which there are many applications in analyzing markets, supplies, and sales.
- A **Web search engine** is a specialized computer server that searches for information on the Web. The search results of a user query are often returned as a list (sometimes called hits). The hits may consist of web pages, images, and other types of files.

    - Web search engines are essentially very large data mining applications. Various data mining techniques are used in all aspects of search engines, ranging from *crawling*

(e.g., deciding which pages should be crawled and the crawling frequencies), indexing (e.g., selecting pages to be indexed and deciding to which extent the index should be constructed), and searching (e.g., deciding how pages should be ranked, which advertisements should be added, and how the search results can be personalized or made "context aware").

1.7. Major issues in Data Mining: Data mining is a dynamic and fast-expanding field with great strengths. Major issues in data mining research, partitioning them into five groups: *mining methodology, user interaction, efficiency and scalability, diversity of data types,* and *data mining and society*.

→ **Mining methodology**: In this methodology the user interaction on different issues such as

- *Mining various and new kinds of knowledge.*
- *Mining knowledge in multidimensional space.*
- *Data mining—an interdisciplinary effort.*
- *Boosting the power of discovery in a networked environment.*
- *Handling uncertainty, noise, or incompleteness of data.*
- *Pattern evaluation and pattern- or constraint-guided mining.*

→**User Interaction:** Interesting areas of research include *how to interact with a data mining system, how to incorporate a user's background knowledge in mining,* and *how to visualize and comprehend data mining results.*

- *Interactive mining.*
- *Incorporation of background knowledge.*
- *Ad hoc data mining and data mining query language.*
- *Presentation and visualization of data mining results.*

→**Efficiency and Scalability:**

- *Efficiency and scalability of data mining algorithms.*
- *Parallel, distributed, and incremental mining algorithms.*
- *Cloud computing* and *cluster computing.*

→**Diversity of Database Types:**

- *Handling complex types of data*
- *Mining dynamic, networked, and global data repositories*

→**Data Mining and Society:**

- *Social impacts of data mining.*
- *Privacy-preserving data mining.*
- *Invisible data mining.*

1.8. Data Objects and Attribute Types:

A **data object** represents an entity.

- In a sales database, the objects may be customers, store items, and sales;
- in a medical database, the objects may be patients;
- in a university database, the objects may be students, professors, and courses.
- Data objects are typically described by attributes. Data objects can also be referred to as *samples, examples, instances, data points,* or *objects*.
- The data objects are stored in a database, they are *data tuples*. That is, the rows of a database correspond to the data objects, and the columns correspond to the attributes.

1.8.1. What Is an Attribute?

- An **attribute** is a data field, representing a characteristic or feature of a data object.
- The attribute may also call, dimension, feature, and variable. The term *dimension* is commonly used in data warehousing. The term *feature* is commonly used in Machine learning, while statisticians prefer the term *variable.* Data mining and database professionalscommonly use the term *attribute*.
- For example, Attributes is described for a customer object is as *customer ID*, *name*, and *address*.

Types of Attribute: The type of an attribute is determined by the set of possible values. They are nominal, binary, ordinal, or numeric.

- **Nominal Attributes:** Nominal means "relating to names." The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical. The values are also known as enumerations.
    - Example: *hair color* and *marital status* are two attributes describing person objects. In our application, possible values for *hair color* are black, brown, blond, red, auburn, gray, and white. The attribute *marital status* can take on the values single, married, divorced, and widowed.
- **Ordinal Attributes:** An **ordinal attribute** is an attribute with possible values that have a meaningful order or *ranking* among them, but the magnitude between successive values is not known.
    - **Example:** *drink size* corresponds to the size of drinks available at a fast-food restaurant. This ordinal attribute has three possible values: *small*, *medium*, and *large*. The values have a meaningful sequence (which corresponds to increasing drink size.
    - Other examples of ordinal attributes include *grade* (e.g., *A+*, *A*, *A-*, *B+* and so on).
    - Professional ranks can be enumerated in a sequential order: for example, *assistant*, *associate*, and *professors*.
- **Binary:** Nominal attribute with only 2 states (0 and 1). Eg: true or false, yes or no.
    - Symmetric binary: both outcomes equally important e.g., gender
    - Asymmetric binary: outcomes not equally important. e.g., medical test (positive vs. negative).
- **Numeric Attributes:** A **numeric attribute** is *quantitative*; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be *interval-scaled* or *ratio-scaled*.
    - ***Interval-Scaled Attributes:*** Interval-scaled attributes are measured on a scale of equal-size units.
        - **Example**: A temperature attribute is interval-scaled. Suppose that we have the outdoor temperature value for a number of different days, where each dayis an object. For example, a temperature of 20_C is five degrees higher than a temperature of 15_C. Calendar dates are another example. For instance, the years 2002 and 2010 are eight years apart.
- **Ratio-Scaled Attributes:**
    - A **ratio-scaled attribute** is a numeric attribute with an inherent zero-point. e.g., *temperature in Kelvin, length, counts.*

1.9. Basic Statistical Descriptions of Data:

Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

1.9.1. Measures of central tendency:

It means, measure the location of the middle or center of a data distribution. It includes mean, median, mode, and midrange.

- **Mean**: measure of the "center" of a set of data is the *(arithmetic) mean*.
  - Let $x1, x2, : : : , xN$ be a set of N values or *observations*, such as for some numeric attribute X, like *salary*. The **mean** of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12} = \frac{696}{12} = 58.$$

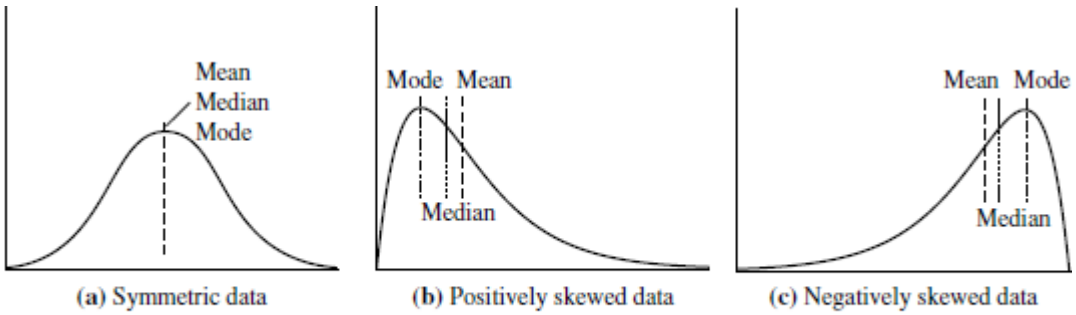
- **Median.** Let's find the median of the data from the above example. The data are already sorted in increasing order. The median can be any value within the two middlemost values of 52 and 56. $\frac{52+56}{2} = \frac{108}{2} = 54.$ Thus, the median is $54,000.

- **Mode:** In the example, the data are bimodal. The two modes are $52,000 and $70,000.

Step 1: The number that occurs most frequently in a data set is called the mode.
Step 2: 30, 36, 47,50, 52, 52, 56, 60, 63, 70, 70, 110.
Step 3: Since the number 52 and 70 appears two times. So, the mode of the data set are 52 and 70.

- **Midrange.** The midrange of the data of above Example is $\frac{30,000+110,000}{2} = \$70,000.$

symmetric data distribution: In a unimodal frequency curve with perfect **symmetric** data distribution, the mean, median, and mode are all at the same center value, as shown in Figure (a).
Data in most real applications are not symmetric. They may instead be either **positively skewed**, where the mode occurs at a value that is smaller than the median (Figure b), or **negatively skewed**, where the mode occurs at a value greater than the median (Figure c).



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

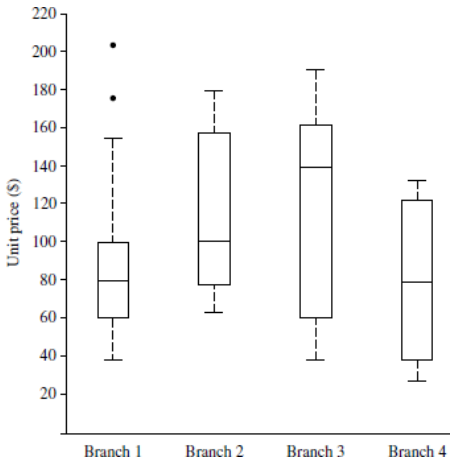
Mean, median, and mode of symmetric versus positively and negatively skewed data.

1.9.2. Measuring the Dispersion of Data: The measures include range, quartiles, quartiles, percentiles, and the inter-quartile range. The five-number summary can be displayed as a boxplot, outliers, variance and standard deviation. also indicate the spread of a data distribution.

- The **range** of the set is the difference between the largest (max()) and smallest (min()) values.
- **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially *equal size* consecutive sets.
- The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the **inter-quartile range** (**IQR**) and is defined as $IQR = Q3 - Q1$.
    - For example, the quartiles are the three values that split the sorted data set into four equal parts. The data of above example contain 12 values, already sorted in increasing order. Thus, the qartiles for this data are the third, sixth, and ninth values, respectively, in the sorted list. Therefore, $Q1$ is \$47,000 and $Q3$ is \$63,000. Thus, theinter-quartile range is $IQR = 63 - 47 = \$16{,}000$.

1.9.2.1. Boxplot: Boxplot incorporates the five-number summary as follows:

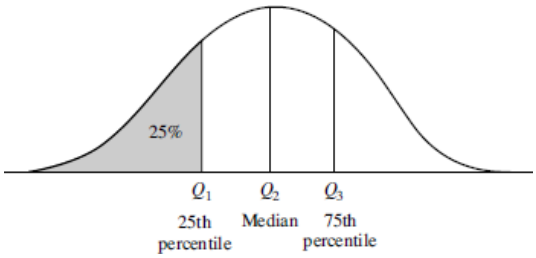
- *Data is represented with a box*
- *The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR*
- *The median is marked by a line within the box*
- *Two lines(whiskers) outside the box extended to Minimum and Maximu.*



The figure shows Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period. For branch 1, the median price of items sold is \$80, $Q1$ is \$60, and $Q3$ is \$100. Notice that two outlying observations for this branch were plotted individually, as their values of 175 and 202 are more than 1.5 times the IQR here of 40.

Note: Boxplots can be computed in $O.n\log n/$ time.

- **Outliers:** points beyond a specified outlier threshold, plotted individually.
- **Percentile:** The three Quartiles are shown as : Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)



***1.9.3.* Variance and Standard Deviation:** Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

The **variance** of N observations, $x_1, x_2, \ldots, x_N$, for a numeric attribute X is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2,$$

where x^- is the mean value of the observations, as defined in Eq. mean formula. The **standard deviation**,σ, of the observations is the square root of the variance, σ^2 .
In the example 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. we found $\bar{x} = \$58{,}000$
Using mean value i.e.

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12} = \frac{696}{12} = 58.$$

To determine the variance and standard deviation of the data from that example, we set N = 12 and use Eq. **variance** to obtain

$$\sigma^2 = \frac{1}{12}(30^2 + 36^2 + 47^2 \ldots + 110^2) - 58^2$$
$$\approx 379.17$$
$$\sigma \approx \sqrt{379.17} \approx 19.47.$$

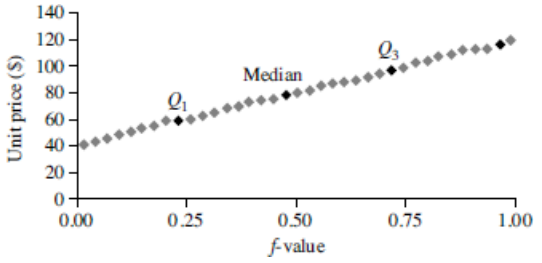
The basic properties of the standard deviation, σ, as a measure of spread are as follows:

- σ measures spread about the mean and should be considered only when the mean is chosen as the measure of center.
- $\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise, $\sigma > 0$.

1.9.4. Graphic Displays of Basic Statistical Descriptions of Data

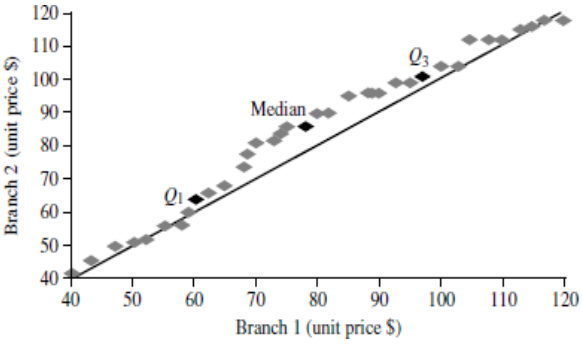
Graphic displays of basic statistical descriptions of data include *quantile plots*, *quantile–quantile plots*, *histograms*, and *scatter plots*. Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing.

1. Quantile Plot: A **quantile plot** is a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute. Second, it plots quantile information. Each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$.



For example, given the quantile plots of sales data for two different time periods, we can compare their *Q*1, median, *Q*3, and other *fi* values at a glance. This is shown in fig.

2. Quantile–Quantile Plot: A **quantile–quantile plot**, or **q-q plot**, graphs the quantiles of one univariate distribution against the corresponding quantiles of another. It is a powerful visualization tool in that it allows the user to view whether there isa shift in going from one distribution to another.



- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.
- A q-q plot for unit price data from two *AllElectronics* branches.

3. Histograms: "Histos" means pole or mast, and "gram" means chart, so a histogram is a chart of poles. Below figure shows a histogram for the data set of Table 2.1, where buckets (or bins) are defined by equal-width ranges representing $20 increments and the frequency is the count of items sold.

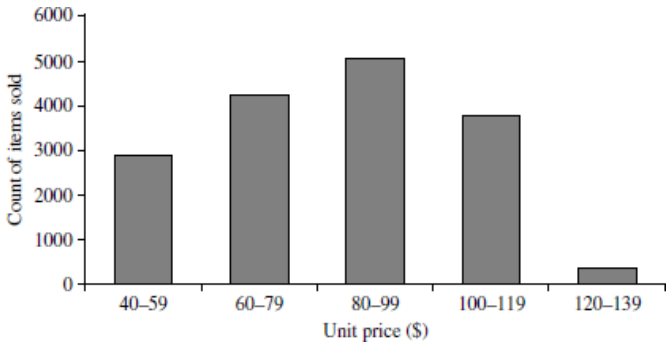
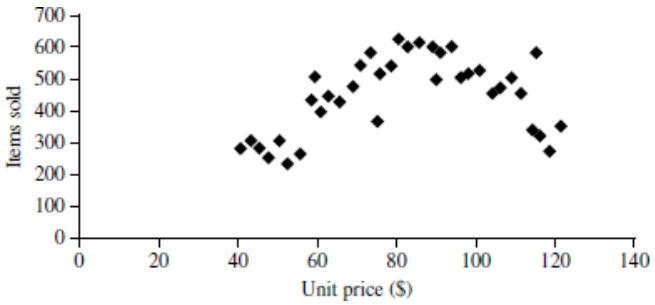
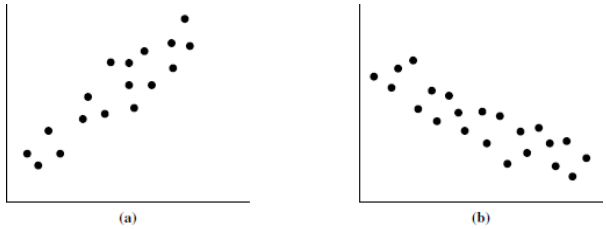


Table 2.1 A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350

4. Scatter Plots and Data Correlation: A **scatter plot** is one of the most effective graphical methods for determining a relationship, pattern, or trend between two numeric attributes. To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane. Figure shows a scatter plot for the set of data in Table 2.1.





Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

Data Correlation Two attributes, *X*, and *Y*, are **correlated** if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated). Figure (a) shows examples of positive and negative correlations between two attributes. If the plotted point's pattern slopes from lower left to upper right, this means that the values of *X* increase as the values of *Y* increase, suggesting a *positive correlation* (Figure b). If the pattern of plotted points slopes from upper left to lower right, the values of *X* increase as the values of *Y* decrease, suggesting a *negative correlation* (Figure c).

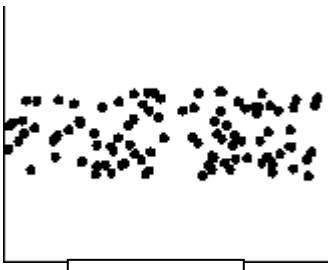


figure (a)

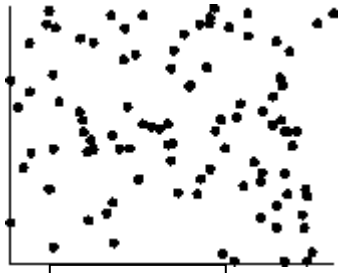


figure (b)

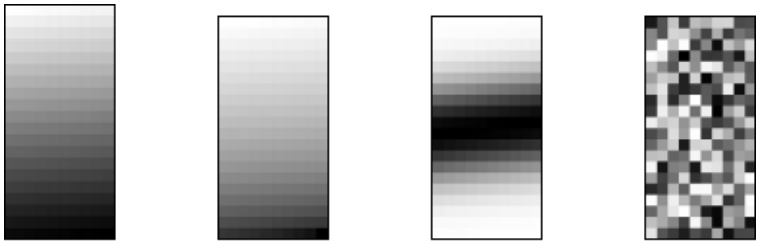
figure (c)

1.10. Data Visualization:

- **Data visualization** aims to communicate data clearly and effectively through graphical representation. Data visualization has been used extensively in many applications—for example, at work for reporting, managing business operations, and tracking progress of tasks.
- The basic concept of data visualization has several representative approaches, including pixel-oriented techniques, geometric projection techniques, icon-based techniques, and hierarchical and graph-based techniques.

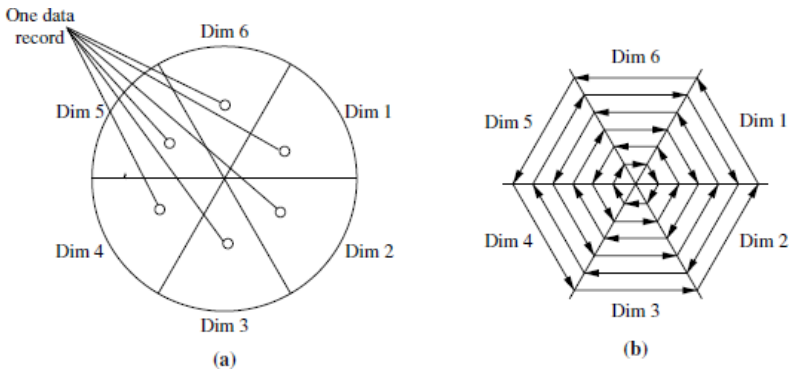
1. Pixel-oriented visualization. *AllElectronics* maintains a customer information table, which consists of four dimensions: *income*, *credit limit*, *transaction volume*, and *age*. This shown in figure.

- For a data set of 'm' dimensions, create 'm' windows on the screen, one for each dimension. The 'm' dimension values of a record are mapped to 'm' pixels at the corresponding positions in the windows. The colors of the pixels reflect the corresponding values.



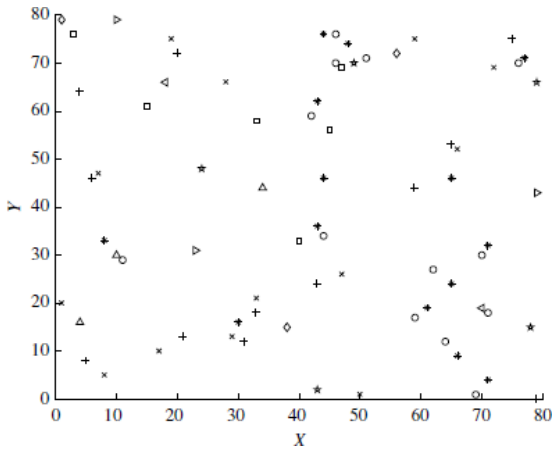
(a) *income* (b) *credit_limit* (c) *transaction_volume* (d) *age*

The ***circle segment technique*** uses windows in the shape of segments of a circle, as illustrated in Figure. This technique can ease the comparison of dimensions because the dimension windows are located side by side and form a circle.



2. <u>Geometric Projection Visualization Techniques:</u> Geometric projection techniques help users find interesting projections of multidimensional data sets. The central challenge the geometric projection techniques try to address is how to visualize a high-dimensional space on a 2-D display.

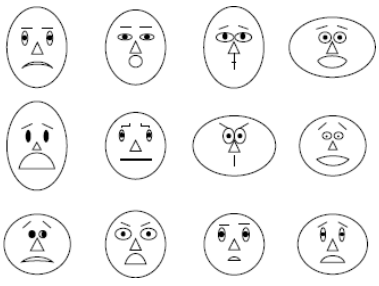
Visualization of a 2-D data set using a scatter plot.



A **scatter plot** displays 2-D data points using Cartesian coordinates. A third dimension can be added using different colors or shapes to represent different data points. Figure 2.13 shows an example, where *X* and *Y* are two spatial attributes and the third dimension is represented by different shapes. Through this visualization, we can see that points of types "+" and "_" tend to be colorcated.

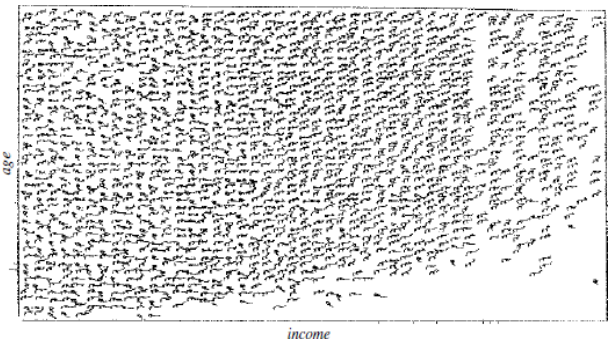
<u>2. Icon-Based Visualization Techniques:</u> It uses small icons to represent multidimensional data values. Two popular icon-based techniques: *Chernoff faces* and *stick figures*.

<u>Chernoff faces</u> were introduced in 1973 by statistician Herman Chernoff. They display multidimensional data of up to 18 variables (or dimensions) as a cartoon human face (shown in figure). Chernoff faces help reveal (make known) trends in the data. Components of the face, such as the *eyes*, *ears*, *mouth*, and *nose*, represent values of the dimensions by their shape, size, placement,



and orientation. For example, dimensions can be mapped to the following facial characteristics: *eye size, eye spacing, nose length, nose width, mouth curvature, mouth width, mouth openness, pupil size, eyebrow slant, eye eccentricity,* and *head eccentricity.*

The **stick figure** visualization technique maps multidimensional data to five-piece stick figures, where each figure has four limbs and a body.Two dimensions are mapped to the display (*x* and*y*) axes and the remaining dimensions are mappedto the angle and/or length of the limbs. **Figure shows** census data, where *age* and *income* are mapped to the display axes, and the remaining dimensions (*gender, education,* and so on) aremapped to stick figures.



3. Hierarchical Visualization Techniques: Visualization of the data using a hierarchical partitioning into subspaces.

- Methods
  - **Dimensional Stacking**
  - **Worlds-within-Worlds**
  - **Tree-Map**
  - **Cone Trees**
  - **InfoCube**

1. **Dimensional Stacking:** Partitioning of the n-dimensional attribute space in 2-D subspaces, which are 'stacked' into each other.
2. **"Worlds-within-Worlds,"** also known as *n*-Vision, is a representative hierarchical visualization method.
3. **Tree-maps** display hierarchical data as a set of nested rectangles.
4. **Cone Trees:** *3D cone tree* visualization technique works well for up to a thousand nodes or so First build a *2D circle tree* that arranges its nodes in concentric circles centered on the root node.
5. **InfoCube:** A 3-D visualization technique where hierarchical information is displayed as nested semi-transparent cubes. The outermost cubes correspond to the top level data, while the subnodes or the lower level data are represented as smaller cubes inside the outermost cubes, and so on.

1.11. Measuring Data Similarity and Dissimilarity:

A **cluster** is a collection of data objects such that the objects within a cluster are *similar* to one another and *dissimilar* to the objects in other clusters. The term *proximity* is used to refer to either similarity or dissimilarity.

Similarity and Dissimilarity

Distance or similarity measures are essential to solve many pattern recognition problems such as classification and clustering. Various distance/similarity measures are available in literature to compare two data distributions. As the names suggest, a similarity measures how close two distributions are. For multivariate data complex summary methods are developed to answer this question.

Similarity Measure

- Numerical measure of how alike two data objects are.
- Often falls between 0 (no similarity) and 1 (complete similarity).

Dissimilarity Measure

- Numerical measure of how different two data objects are.
- Range from 0 (objects are alike) to ∞ (objects are different).

Proximity refers to a similarity or dissimilarity.

Similarity/Dissimilarity for Simple Attributes

Here, p and q are the attribute values for two data objects.

Attribute Type	Similarity	Dissimilarity
Nominal	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$
Ordinal	$s = 1 - \frac{\|p-q\|}{n-1}$ (values mapped to integer 0 to n-1, where n is the number of values)	$d = \frac{\|p-q\|}{n-1}$
Interval or Ratio	$s = 1 - \|p-q\|$, $s = \frac{1}{1+\|p-q\|}$	$d = \|p-q\|$

1.12.

Data Matrix: In the central tendency, dispersion and variance & Standard deviation, the data is handled through a single attribute (i.e. one dimensional array). In *Data Matrix*, attribute handle multiple data (i.e. multi dimensional array).

For example: Suppose that we have *n* objects (e.g., persons, items, or courses) described by *p* attributes (also called *measurements* or *features*, such as age, height, weight, or gender).

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}.$$

The objects are $x1 = .x_{11}, x_{12}, \ldots, x_{1p})$, $x2 = .x_{21}, x_{22}, \ldots, x_{2p})$ and so on, where *xij* is the value for object *xi* of the *j*th attribute.

- The Data Matrix (or *object-by-attribute structure*) structure stores the *n* data objects in the form of a relational table, or *n*-by-*p* matrix (*n* objects _*p* attributes):

Dissimilarity matrix: (or *object-by-object structure*): This structure stores a collection of proximities that are available for all pairs of *n* objects. It is often represented by an *n*-by-*n* table:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

where $d.(i, j)$ is the measured **dissimilarity** or "difference" between objects *i* and *j*. In general, $d.(i, j)$ is a non-negative number that is close to 0 when objects *i* and *j* are highly similar or "near" each other, and becomes larger the more they differ. Note that $d(.i, i) = 0$; that is, the difference between an object and itself is 0.

Measures of similarity can often be expressed as a function of measures of dissimilarity. For example, for nominal data, $sim(.i, j) = 1 - d(.i, j)$, where $sim(.i, j)$ is the similarity between objects *i* and *j*.

"How is dissimilarity computed between objects described by nominal attributes?"

: The dissimilarity between two objects *i* and *j* can be computed based on the ratio of mismatches:

$$d(i, j) = \frac{p - m}{p},$$

where *m* is the number of *matches* (i.e., the number of attributes for which *i* and *j* are in the same state), and *p* is the total number of attributes describing the objects. Weights can be assigned to increase the effect of *m* or to assign greater weight to the matches in attributes having a larger number of states.

Dissimilarity between nominal attributes. Suppose that we have the sample data of **Table,** except that only the *object-identifier* and the attribute *test-1* are available, where *test-1* is nominal. (We will use *test-2* and *test-3* in later examples.) Let's compute the dissimilarity matrix (Equation), that is, Since here we have one nominal attribute, *test-1*, we set $p = 1$ in Eq. So that $d(i, j)$ evaluates to 0 if objects *i* and *j* match, and 1 if the objects differ. Thus, we get

A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

From this, we see that all objects are dissimilar except objects 1 and 4 (i.e., $d(.4,1) = 0$).

Dissimilarity between binary attributes: Suppose that a patient record table (Table 2.4) contains the attributes *name, gender, fever, cough, test-1, test-2, test-3,* and *test-4*, where *name* is an object identifier, *gender* is a symmetric attribute, and the remaining attributes are asymmetric binary.

Relational Table Where Patients Are Described by Binary Attributes

name	gender	fever	cough	test-1	test-2	test-3	test-4
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

For asymmetric attribute values, let the values *Y* (*yes*) and (*positive*) be set to 1, and the value *N* (*no* or *negative*) be set to 0. Suppose that the distance between objects is computed based

$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67,$$

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33,$$

$$d(Jim, Mary) = \frac{1+2}{1+1+2} = 0.75.$$