

# Contents

Sl.No	Topic	Page No
1.1	Introduction	1
1.2	Necessity of learning DBMS	1
1.3	Applications of DBMS	1
1.4	Types of databases	2
1.5	Performance Measurement of Databases	10
1.6	Goals of parallel databases	11
1.7	Techniques of query Evaluation	12
1.8	Optimization of Parallel Query	12
1.9	Goals of Query optimization	12
1.10	Approaches of Query Optimization	13
1.11	Virtualization on Multicore Processor	13
	References	14
	MOOCS	15
	Quiz	16
	Video Links	17

## Database Management System

### 1.1 Introduction

A database management system (DBMS) is a software package designed to define, manipulate, retrieve and manage data in a database. A DBMS generally manipulates the data itself, the data format, field names, record structure and file structure. It also defines rules to validate and manipulate this data.

Database management systems are set up on specific data handling concepts, as the practice of administrating a database evolves. The earliest databases only handled individual single pieces of specially formatted data whereas the newer systems can handle different kinds of data and tie them together in more elaborate ways.

### 1.2 Necessity of learning DBMS

Traditionally, data was organized in file formats; DBMS overcomes the deficiencies in traditional style of data management and has the following characteristics:

- ❖ **Real-world entity:** It is realistic and uses real-world entities to design its architecture. **Example:** A University database where we represent students as an entity and their roll\_number as an attribute.

- ❖ **Relation-based tables:** It allows entities and the relations among them to form tables. We can understand the architecture of it by looking at the table names.
- ❖ **Isolation of data and application:** A database system is entirely different from its data where database is an active entity and data is said to be passive.
- ❖ **Less redundancy:** It follows the rules of normalization and splits a relation when any of its attributes is having redundancy in values.
- ❖ **Consistency:** It provides a greater consistency as compared to earlier forms of data storing applications.
- ❖ **Query Language:** It is equipped with query language and makes it more efficient to retrieve and manipulate data which was not possible in the earlier file-processing system.

### 1.3 Applications of DBMS

- ❖ **ACID Properties:** Atomicity, Consistency, Isolation, and Durability (ACID) properties help the database stay healthy in multi-transactional environments & in case of failure.
- ❖ **Multiuser and Concurrent Access:** Supports multi-user environment and allows them to access and manipulate data in parallel.
- ❖ **Multiple views:** Facilitates multiple views for different users. A user who is in the Accounts department has a different view of database than a user working in the Transport department. This feature enables the users to have a concentrate view of the database according to their requirements.
- ❖ **Security:** Features like multiple views offer security to some extent where users are unable to access data of other users and departments. It offers different levels of security features, enabling multiple users to have different views with different features. As DBMS is not saved on the disk as traditional file systems, it is hard for miscreants to break the code.

### 1.4 Types of databases

Some common **examples of relational databases** are: Sybase, Oracle, MySQL, Microsoft SQL Server and etc.

1. Centralized Database
2. Distributed Database
3. Relational Database
4. NoSQL Database
5. Cloud Database
6. Object-oriented Databases
7. Hierarchical Databases
8. Network Databases
9. Personal Database
10. Operational Database
11. Enterprise Database
12. Parallel Databases

## *1) Centralized Database*

It stores data at a centralized database system and comforts the users to access the stored data from varied locations through various applications. An example of a University database can be a Student which carries a central database of each student in the university.

Example: University\_of\_Mumbai

Advantages:

- Decreased the risk of data management
- Data consistency
- Enables organizations to establish data standards.
- Less costly

Disadvantages

- Large size increases the response time for fetching the data
- Complex to update
- Server failure leads to the entire data loss.

## *2) Distributed Database*

In distributed database data is distributed among different database systems of an organization and are connected via communication links helping the end-users to access the data easily. **Examples:** Oracle, Apache Cassandra, HBase, Ignite and etc.

Distributed database system can be further classified into:

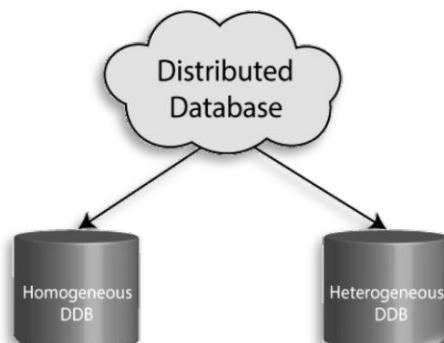


Figure 1: Architecture of a distributed database

- ❖ **Homogeneous DDB:** Executes on the same operating system using the same application process carrying same hardware devices.
- ❖ **Heterogeneous DDB:** Executes on different operating systems with different application procedures carrying different hardware devices.

Advantages of Distributed Database

- Modular development
- Server failure will not affect the entire data set.

### 3) Relational Database

It is based on the relational data model that stores data in the form of rows and columns forming a table. It uses SQL for storing, maintaining and manipulating of the data invented by E.F. Codd in the year 1970. Each table carries a key making the data unique from the others. **Examples:** Oracle, Sybase, MySQL and etc.

#### Properties of Relational Database

Four commonly known properties of a relational model are Atomicity, Consistency, Durability, Isolation (ACID):

**Atomicity:** It ensures the data operation will complete either with success or with failure following the 'all or nothing' strategy. Example: A transaction will either be committed or rolled back.

**Consistency:** Any operation over the data should be consistent in terms of its value either before or after the operation. Example: Account balance before and after the transaction should remain conserved.

**Isolation:** Data remains isolated even when numerous concurrent users are accessing data at the same time. Example: When multiple transactions are processed at the same instance, effect of a transaction should not be visible to the other transactions.

**Durability:** It ensures permanent data storage as it completes the specified operation and issues commit.

### 4) NoSQL Database

NoSQL is used for storing a wide range of data sets in different ways. Example: MongoDB.

Based on demand NoSQL is further classified into the following types:

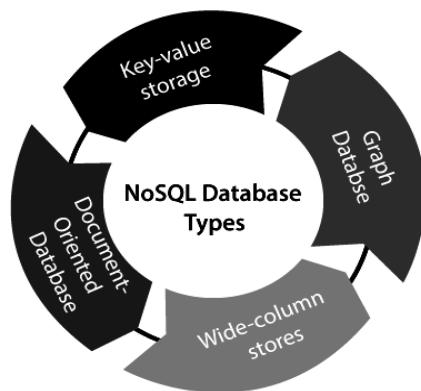


Figure 2: NoSQL database

- i. **Key-value storage:** Stores every single item as a key.
- ii. **Document-oriented Database:** It is used to store data as JSON-like document

- iii. **Graph Databases:** It is used for storing large amounts of data in a graph-like structure. Example: Social networking websites internally uses the graph database.
- iv. **Wide-column stores:** Data is stored in large columns together, in its place of storing in rows.

#### Advantages

- ❖ Good productivity in the application development
- ❖ Better option to handle large data sets
- ❖ Highly scalable
- ❖ Quicker access using the key field/value

#### 5) Cloud Database

Data is stored in a virtual environment getting executed over the cloud computing platform having numerous cloud platforms. Examples:

- ❖ Amazon Web Services
- ❖ Microsoft Azure
- ❖ Google Cloud SQL and etc.

#### 6) Object-oriented Databases

It uses the object-based data model approach for storing data where the data is represented and stored as objects.

Examples: Realm, ObjectBox

#### 7) Hierarchical Databases

Stores the data in the form of parent-children relationship and organizes the data in a tree-like structure. Data in the form of records are connected through links where each child record has only one parent whereas the parent record can have numerous child records.

Examples: IBM Information Management System (IMS) and the RDM Mobile

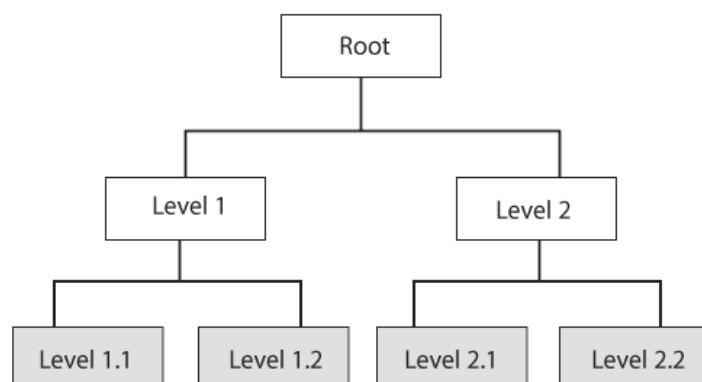


Figure 3: Hierarchical Database

*8) Network Databases*

Data is represented in the form of a network model where the data in the form of nodes is connected through the links between them.

Examples: Integrated Data Store (IDS), IDMS (Integrated Database Management System), Raima Database Manager

*9) Personal Database*

Designed for a single user where data collection and storing is on the user's system.

Examples: DSRao\_database

**Advantages**

- ❖ Simple.
- ❖ Less storage space.

*10) Operational Database*

Operational database is designed for executing the day-to-day operations in businesses.

Examples: Microsoft SQL Server, AWS Dynamo, Apache Cassandra, MongoDB

*11) Enterprise Database*

It is used for managing enormous amount of data allowing simultaneous access to the users with greater efficiency.

Examples: Microsoft SQL Server, IBM DB2, SAP Sybase ASE, MariaDB Enterprise

**Advantages**

- ❖ Multiprocessing
- ❖ Executing parallel queries.

*12) Parallel Databases*

Organizations are in a need to handle and maintain substantial amount of data with higher transfer rate and greater efficiency of a system. Parallel database system improves the performance through parallelization of varied operations like loading, manipulating, storing, building and evaluating. Processing speed and efficiency is increased by using multiple disks and CPUs in parallel. Figure 4, 5 and 6 shows the different architecture proposed and successfully implemented in the area of Parallel Database systems. In the figures, P represents Processors, M represents Memory, and D represents Disks/Disk setups.

Parallel database systems are classified into two groups:

- i. Multiprocessor architecture and

## ii. Hierarchical System or Non-Uniform Memory Architecture

### Multiprocessor architecture

It has the following alternatives:

- ❖ Shared memory architecture
- ❖ Shared disk architecture
- ❖ Shared nothing architecture

### Shared memory architecture

In shared memory architecture multiple processors share the same single primary/main memory and have its own hard disk for storage. As shown in the figure 4, several processors are connected through an interconnection network with Main memory and disk setup. Interconnection network is usually a high speed network making data sharing easy among the various components.

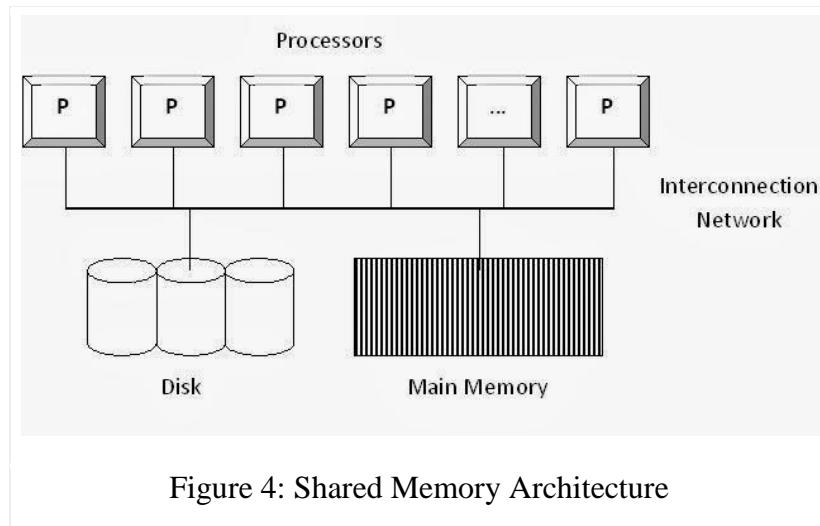


Figure 4: Shared Memory Architecture

Advantages:

- ❖ Simple to implement
- ❖ Effective communication among the processors
- ❖ Less communication overhead

Disadvantages:

- ❖ Limited degree of parallelism
- ❖ Addition of processor would slow down the existing processors.
- ❖ Cache-coherency need to be maintained
- ❖ Bandwidth issue

### Shared disk architecture

As shown in figure 5, in shared disk architecture each processor has its own private memory sharing the single mass storage in common.

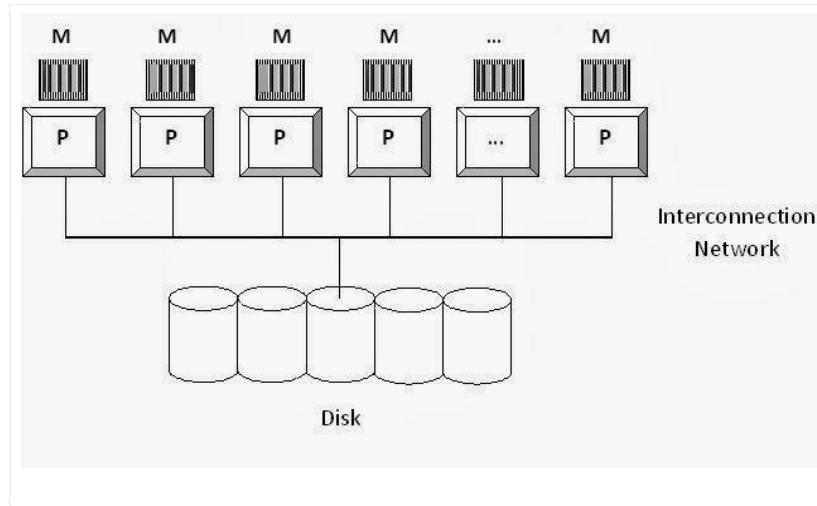


Figure 5: Shared Disk Architecture

**Advantages:**

- ❖ Fault tolerance is achieved
- ❖ Interconnection to the memory is not a bottleneck
- ❖ Supports large number of processors

**Disadvantages:**

- ❖ Limited scalability
- ❖ Inter-processor communication is slow

### **Applications:**

Digital Equipment Corporation(DEC).

### **Shared nothing architecture**

As shown in figure 6, in shared nothing architecture, each processor has its own main memory and mass storage device setup. The entire setup is a collection of individual computers connected via a high speed communication network.

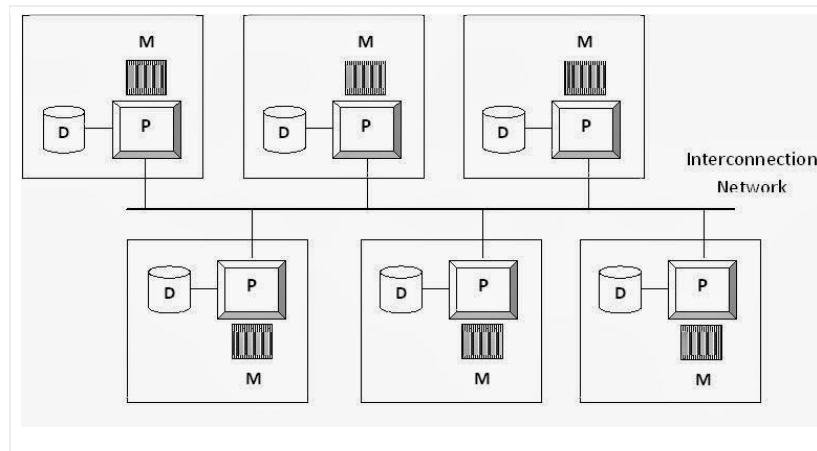


Figure 6: Shared Nothing Architecture

**Advantages:**

- ❖ Flexible to add any number of processors
- ❖ Data request can be forwarded via interconnection n/w

**Disadvantages:**

- ❖ Data partitioning is required
- ❖ Cost of communication is higher

### **Applications**

- ❖ Teradata
- ❖ OracleCUBE
- ❖ The Grace and Gamma research prototypes
- ❖ Tandem and etc.

## **Hierarchical System or Non-Uniform Memory Architecture**

- ❖ Non-Uniform Memory Architecture (NUMA), has the non-uniform memory access.
- ❖ Cluster is formed by a group of connected computers including shared nothing, shared disk and etc.
- ❖ NUMA takes longer time to communicate among each other as it uses local and remote memory.

**Advantages:**

- ❖ Improved performance
- ❖ High availability
- ❖ Proper resource utilization
- ❖ Highly Reliable

**Disadvantages**

- ❖ High cost
- ❖ Numerous Resources
- ❖ Complexity in managing the systems

## **1.5 Performance Measurement of Databases**

Performance measurement includes the factors like Speedup and Scale-up.

**Speedup:** Ability to execute tasks in lesser time by increasing the number of resources.

$$\text{Speedup} = (\text{Original time}) / (\text{Parallel time}) \quad (\text{Equation 1})$$

Original time = time required as to execute the task using a single or 1 processor

Parallel time = time required as to execute the task using numerous or 'n' processors

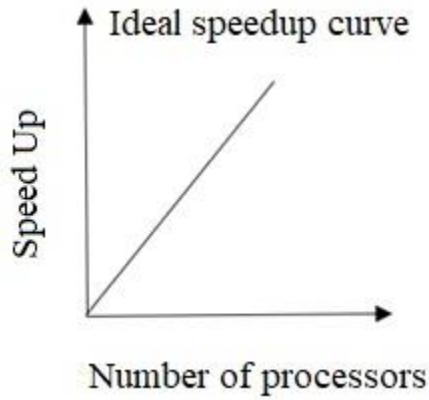


Figure 7: Speedup curve

**Example:**



Figure 8: A CPU requires 3 minutes to execute a process

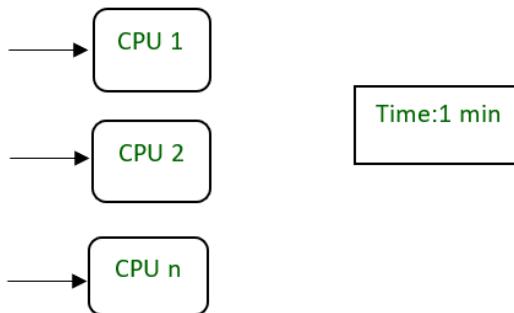


Figure 9: 'n' CPU requires 1 min to execute a process by dividing into smaller tasks

**Scale-up:** Ability to maintain the performance of the system when workload and resources increase proportionally.

$$\text{Scaleup} = (\text{Volume Parallel}) / (\text{Volume Original}) \quad (\text{Equation 2})$$

Where ,

Volume Parallel = volume executed in a given amount of time using 'n' processor

Volume Original = volume executed in a given amount of time using 1 processor

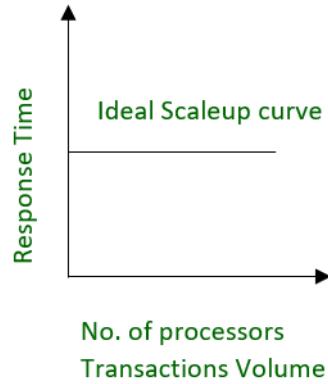


Figure 10: Ideal Scaleup curve

### Example

20 users are using a CPU at 100% efficiency, if we try to add some more users, it becomes difficult for a single processor to handle additional users instead a new processor can be added as to serve the users in parallel mode and provides 200% efficiency.

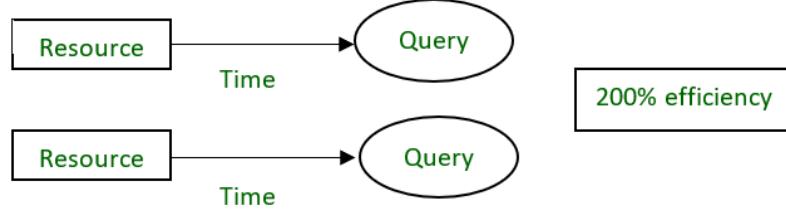


Figure 11: Increase in efficiency

### 1.6 Goals of parallel databases

- ❖ Improve performance:
- ❖ Improve availability of data:
- ❖ Improve reliability
- ❖ Provide distributed access of data:

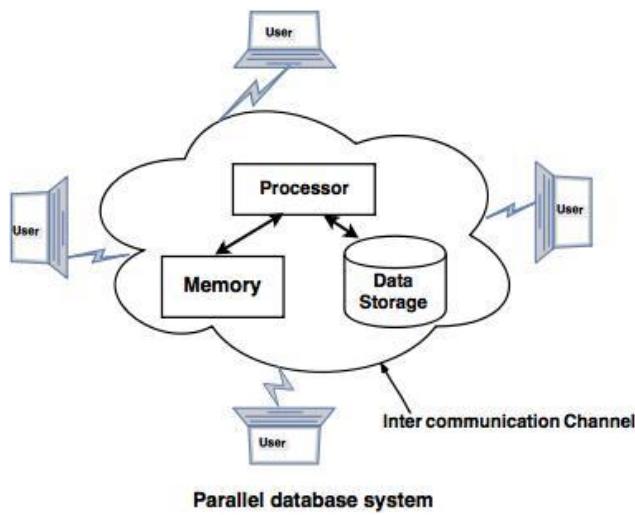


Figure 12: Parallel database system

## 1.7 Techniques of query Evaluation

1. Inter query parallelism
2. Intra Query Parallelism

**Inter query parallelism:** Allows running multiple queries on varied processors simultaneously achieving pipelined parallelism in improving the output of the system.

**Example:** If there are 18 queries, each taking 10 seconds for evaluation. The total time taken to complete evaluation process is 180 seconds. Inter query parallelism achieves this task only in 10 seconds.

**Intra Query Parallelism:** In intra query parallelism query is divided into sub queries and has the ability to run simultaneously on varied processors minimizing the query evaluation time. Intra query parallelism improves the response time of the system.

**Example:** If there are 18 queries, each taking 10 seconds for evaluation. The total time taken to complete evaluation process is 180 seconds. Inter query parallelism achieves this task only in 10 seconds. We can achieve in only 10 seconds by using intra query evaluation as each query is divided in sub-queries.

## 1.8 Optimization of Parallel Query

Parallel Query optimization is about selecting the efficient query evaluation plan and to minimize the cost of query evaluation.

### Factors playing an important role in parallel query optimization

- a) Time spent to find the best evaluation plan.
- b) Time required to execute the plan.

## 1.9 Goals of Query optimization

- ❖ Speed up the queries
- ❖ Increase the performance of the system.
- ❖ Select the best query evaluation plan.
- ❖ Avoid the unwanted plan.

## 1.10 Approaches of Query Optimization

1. Horizontal partitioning
2. Vertical partitioning
3. De-normalization

**Horizontal partitioning:** Tables are created vertically using columns.

**Vertical partitioning:** Tables are created with fewer columns and partition the table row wise.

**De-normalization:** Multiple tables are combined into one table.

## 1.11 Virtualization on Multicore Processor

Virtualization is a technique where the processing power of the computer is enhanced by adding multiple CPUs.

Multicore processors have the ability to solve the complicated processing and are used for heavy load process.

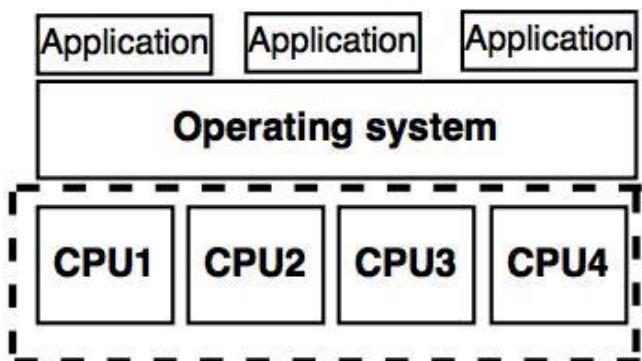


Figure 13: Virtualization on multicore processor

## **References**

1. C. J. Date, A. Kannan and S. Swamynathan, An Introduction to Database Systems, Pearson Education, Eighth Edition, 2009.
2. Abraham Silberschatz, Henry F. Korth and S. Sudarshan, Database System Concepts, McGraw-Hill Education (Asia), Fifth Edition, 2006.
3. Shio Kumar Singh, Database Systems Concepts, Designs and Application, Pearson Education, Second Edition, 2011.
4. Peter Rob and Carlos Coronel, Database Systems Design, Implementation and Management, Thomson Learning-Course Technology, Seventh Edition, 2007.
5. Patrick O'Neil and Elizabeth O'Neil, Database Principles, Programming and Performance, Harcourt Asia Pte. Ltd., First Edition, 2001.
6. Atul Kahate, Introduction to Database Management Systems, Pearson
7. Techopedia. <https://www.techopedia.com/definition/24361/database-management-systems-dbms> (Last accessed on 18.07.2021)
8. DBMS Tutorial. <https://www.javatpoint.com/dbms-tutorial> (Last accessed on 18.07.2021)
9. DBMS. <https://searchsqlserver.techtarget.com/definition/database-management-system>(Last accessed on 18.07.2021)

## MOOC List

1. Database Management Essentials (Coursera). <https://www.mooc-list.com/course/database-management-essentials-coursera>
2. Intro to relational database. <https://www.my-mooc.com/en/mooc/intro-to-relational-databases--ud197/>
3. Database systems Specialization. <https://www.coursera.org/specializations/database-systems>
4. Database Management System. [https://onlinecourses.swayam2.ac.in/cec19\\_cs05/preview](https://onlinecourses.swayam2.ac.in/cec19_cs05/preview)
5. Introduction to Databases. <https://www.edx.org/microbachelors/nyux-introduction-to-databases>

## Quiz

1. A Database Management System is a type of \_\_\_\_\_ software.
  2. The term "FAT" is stands for\_\_\_\_\_
  3. A huge collection of the information or data accumulated form several different sources is known as \_\_\_\_\_:
  4. \_\_\_\_\_ can be used to extract or filter the data & information from the data warehouse?
  5. \_\_\_\_\_ refers to the copies of the same data (or information) occupying the memory space at multiple places.
  6. \_\_\_\_\_refers to the "data about data"?
  7. \_\_\_\_\_ refers to the level of data abstraction that describes exactly how the data actually stored?
  8. In general, a file is basically a collection of all related\_\_\_\_\_.
  9. The term "Data" refers to \_\_\_\_\_:
  10. Rows of a relation are known as the \_\_\_\_\_.
  11. In a relation database, every tuples divided into the fields are known as the\_\_\_\_\_.
  12. In the relational table, \_\_\_\_\_ be represented by the term "attribute"?
  13. \_\_\_\_\_ is used in the application programs to request data from the database management system?
  14. The database management system can be considered as the collection of \_\_\_\_\_ that enables us to create and maintain the database.
  15. \_\_\_\_\_ refers collection of the information stored in a database at a specific time
  16. The term "ODBC" stands for\_\_\_\_\_
  17. The architecture of a database can be viewed as the \_\_\_\_\_
  18. The Database Management Query language is generally designed for the \_\_\_\_\_
  19. \_\_\_\_\_ is the collection of the interrelated data and set of the program to access them.
  20. A database is the complex type of the \_\_\_\_\_
  21. An advantage of the database management approach is \_\_\_\_\_
  22. \_\_\_\_\_ is the disadvantage of the file processing system
  23. Redundancy means \_\_\_\_\_
  24. Concurrent access means \_\_\_\_\_
  25. \_\_\_\_\_ refer to the correctness and completeness of the data in a database
  26. Either all of its operations are executed or none is called \_\_\_\_\_
  27. When data is processed, organized, structured or presented in a given context so as to make it useful, it is called \_\_\_\_\_
  28. \_\_\_\_\_ is an information repository which stores data.
  29. \_\_\_\_\_ level deals with physical storage of data.
  30. The process of hiding irrelevant details from user is called \_\_\_\_\_
  31. Example of Naive User is \_\_\_\_\_.
  32. A user who write software using tools such as Java, .Net, PHP etc. is \_\_\_\_\_
-

## **Video Links**

1. <https://www.youtube.com/watch?v=T7AxM7Vqvaw>
2. <https://www.youtube.com/watch?v=6Iu45VZGQDk>
3. <https://www.youtube.com/watch?v=wjfeGxqAQOY&list=PLrjkTql3jnm-CLxHftqLgkrZbM8fUt0vn>
4. <https://www.youtube.com/watch?v=ZaaSa1TtqXY>
5. <https://www.youtube.com/watch?v=lDpB9zF8LBw>
6. [https://www.youtube.com/watch?v=fSWAkJz\\_huQ](https://www.youtube.com/watch?v=fSWAkJz_huQ)
7. <https://www.youtube.com/watch?v=cMUQznvYZ6w>
8. <https://www.youtube.com/watch?v=mqprM5YUdpk>
9. <https://www.youtube.com/watch?v=3EJlovevfcA&list=PLxCzCOWd7aiFAN6I8CuViBuCdJgiOkT2Y&index=2>
10. <https://www.youtube.com/watch?v=ZtVw2iuFI2w&list=PLxCzCOWd7aiFAN6I8CuViBuCdJgiOkT2Y&index=3>
11. <https://www.youtube.com/watch?v=VyyTabQHevw&list=PLxCzCOWd7aiFAN6I8CuViBuCdJgiOkT2Y&index=4>

## **Contents**

<b>Sl.No</b>	<b>Topic</b>	<b>Page No</b>
2.1	<b>Types of distributed databases</b>	1
2.2	<b>Distributed DBMS (DDBMS) Architectures</b>	3
2.3	<b>Architectural Models</b>	3
2.4	<b>Design alternatives</b>	7
2.5	<b>Data Replication</b>	8
2.6	<b>Fragmentation</b>	9
	<b>References</b>	12
	<b>MOOCs</b>	13
	<b>Quiz</b>	14
	<b>Video Links</b>	15

### **2.1 Types of Distributed Databases**

As illustrated in figure 1, distributed databases are classified into homogeneous and heterogeneous, each with further sub-divisions.

Examples: Apache Ignite, Apache Cassandra, Apache HBase, Couchbase Server, Amazon SimpleDB, Clusterpoint, and FoundationDB

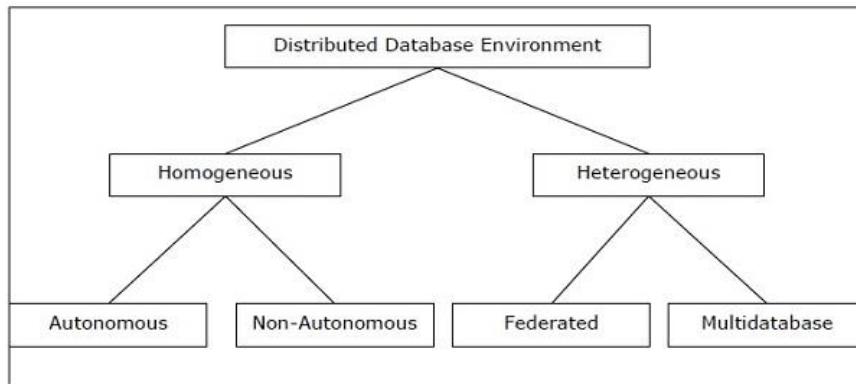


Figure 1: Distributed database environment

**2.1.1 Homogeneous Distributed Databases:** As illustrated in figure 2, all the sites use identical DBMS & operating systems and have the following properties:

- Similar software.
- Identical DBMS from the same vendor.
- Aware of all other neighboring sites cooperating with each other to process user requests.
- In case of a single database it is accessed through a single interface.

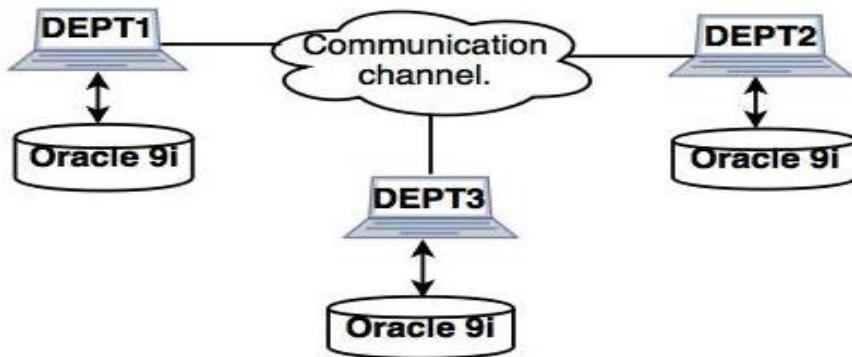


Figure 2: Homogenous distributed database

#### Types of Homogeneous Distributed Database

1. Autonomous
  2. Non-autonomous
- ❖ **Autonomous:** Each database is independent that functions on its personal and are incorporated with the aid of controlling software and use message passing to share data updates.
  - ❖ **Non-autonomous:** Data is distributed across the homogeneous nodes and a central or master DBMS co-ordinates data updates throughout the sites.

**2.1.2 Heterogeneous Distributed Databases:** As illustrated in figure3, different sites have different operating systems, DBMS products and data models and have the following properties are:

- ❖ Different sites use varied schemas and software.
- ❖ The system is composed of varied DBMSs.
- ❖ Complex query processing due to dissimilar schemas.
- ❖ Complex transaction processing due to dissimilar software.
- ❖ A site is not aware of the other sites leading to limited co-operation in processing user requests.

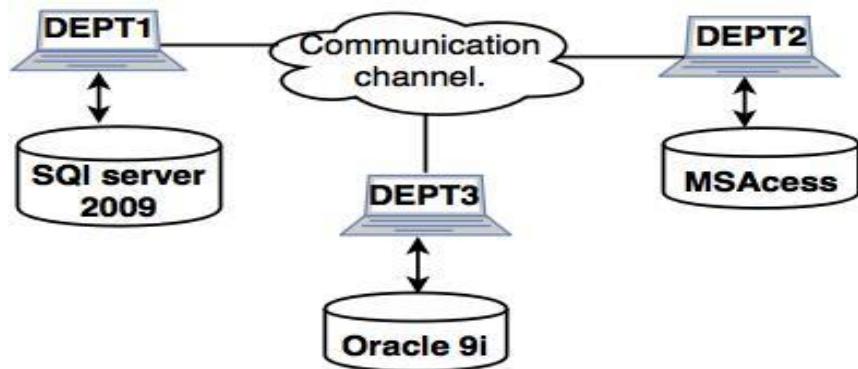


Figure 3: Heterogeneous distributed database

#### Types of Heterogeneous Distributed Databases

1. Federated
  2. Un-federated
- ❖ **Federated:** These systems are independent in nature and are integrated collectively as to feature as a single database system.
  - ❖ **Un-federated:** These systems employ a central coordinating module through which the databases are accessed.

#### Advantages

- ❖ Organizational Structure
- ❖ Shareability and Local Autonomy
- ❖ Improved Availability
- ❖ Improved Reliability
- ❖ Improved Performance
- ❖ Economics
- ❖ Modular Growth

#### Disadvantages

- ❖ Complexity
- ❖ Cost
- ❖ Security
- ❖ Integrity Control More Difficult
- ❖ Lack of Standards

- ❖ Lack of Experience
- ❖ Database Design More Complex

## Rules for DDBMS

- i. Local Autonomy
- ii. No Reliance on a Central Site
- iii. Continuous Operation
- iv. Location Independence
- v. Fragmentation Independence
- vi. Replication Independence
- vii. Distributed Query Processing
- viii. Distributed Transaction Processing
- ix. Hardware Independence
- x. Operating System Independence
- xi. Network Independence
- xii. Database Independence

**2.2 Distributed DBMS (DDBMS) Architectures:** DDBMS architectures are developed on the following parameters:

1. Distribution
  2. Autonomy
  3. Heterogeneity
- ❖ **Distribution:** It states the physical distribution of data throughout the varied sites.
  - ❖ **Autonomy:** It indicates the distribution of control of the database system and the degree to which each constituent DBMS can operate independently.
  - ❖ **Heterogeneity:** It refers to the uniformity or dissimilarity of the data models, system components and databases.

## 2.3 Architectural Models

Common architectural models are:

1. Client - Server Architecture for DDBMS
2. Peer - to - Peer Architecture for DDBMS
3. Multi - DBMS Architecture

**Client - Server Architecture for DDBMS:** Is a two-level architecture in which the functionality is divided into servers and clients. Server functions include primarily encompass data management, query processing, optimization and transaction management whereas the client functions include particularly user interface with common functionalities like consistency checking and transaction management.

Client - server architectures are classified as:

- Single Server Multiple Client

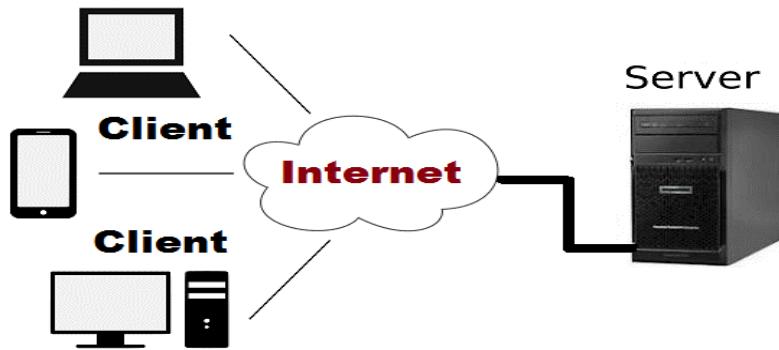


Figure 4: Single Server Multiple Client

- Multiple Server Multiple Client

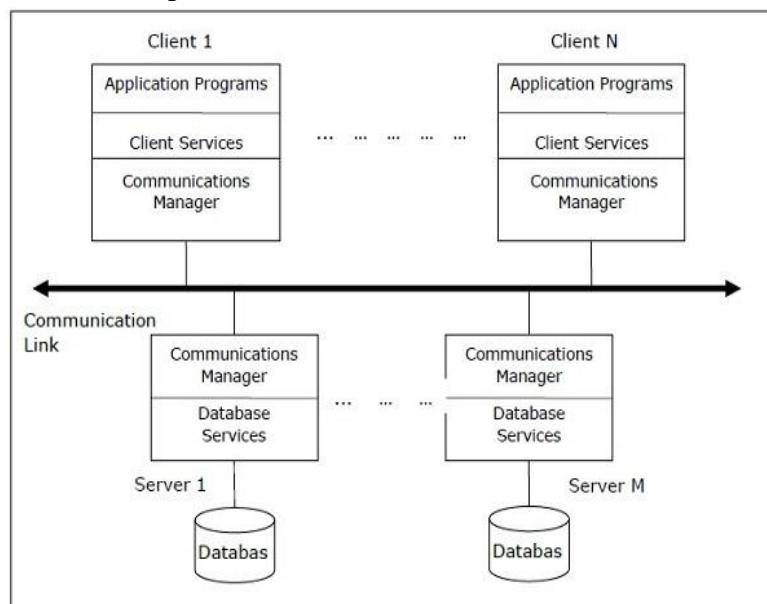


Figure 5: Multiple server multiple client

#### Peer- to-Peer Architecture for DDBMS

In Peer-to-Peer architecture, each peer acts both as a client and a server to impart database services and share their resource with other peers to co-ordinate their activities.

This architecture commonly has four levels of schemas –

- ❖ **Global Conceptual Schema:** Illustrates the global logical view of data.
- ❖ **Local Conceptual Schema:** Illustrates logical data organization at each site.
- ❖ **Local Internal Schema** – Illustrates physical data organization at each site.
- ❖ **External Schema** – Illustrates user view of data.

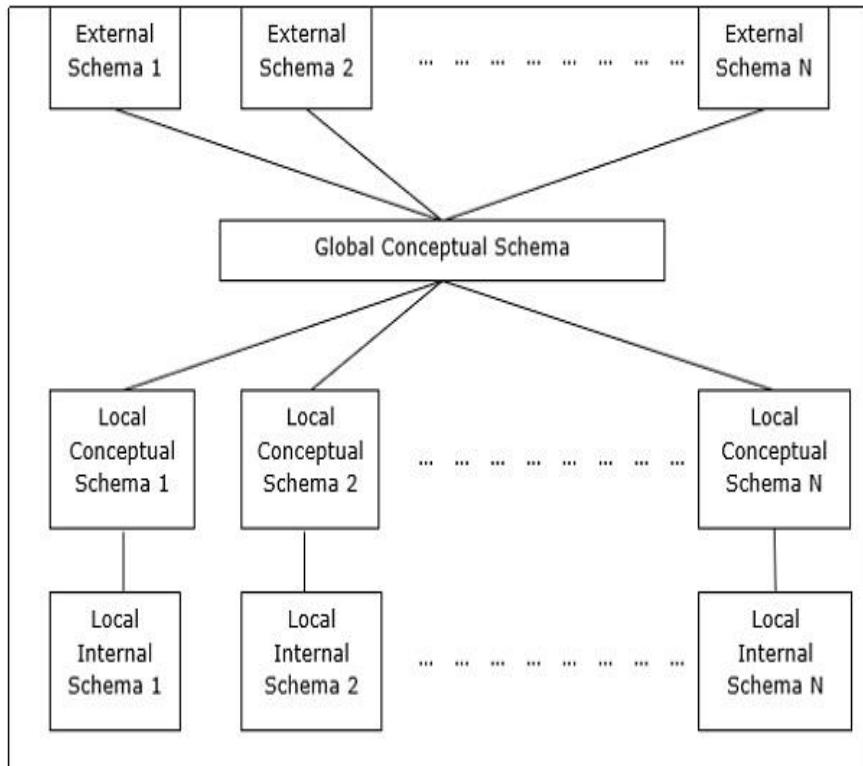


Figure 6: Peer- to-Peer Architecture

**Multi - DBMS Architectures:** Is an integrated database system formed by a collection of two or more autonomous database systems.

Multi-DBMS can be expressed through six levels of schemas –

- ❖ **Multi-database View Level** – Illustrates multiple user views comprising of subsets of the integrated distributed database.
- ❖ **Multi-database Conceptual Level** – Illustrates integrated multi-database that comprises of global logical multi-database structure definitions.
- ❖ **Multi-database Internal Level** – Illustrates the data distribution across different sites and multi-database to local data mapping.
- ❖ **Local database View Level** – Illustrates public view of local data.
- ❖ **Local database Conceptual Level** – Illustrates local data organization at each site.
- ❖ **Local database Internal Level** – Illustrates physical data organization at each site.

Two design alternatives for Multi - DBMS Architectures are:

- ❖ Model with multi-database conceptual level.
- ❖ Model without multi-database conceptual level.

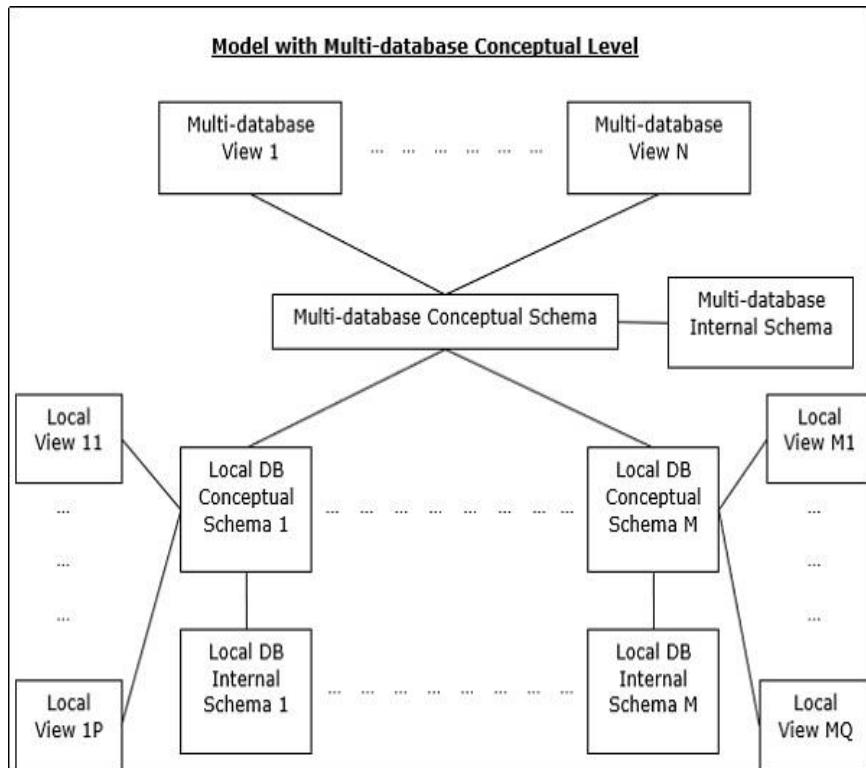


Figure 7: Model with multi-database conceptual level

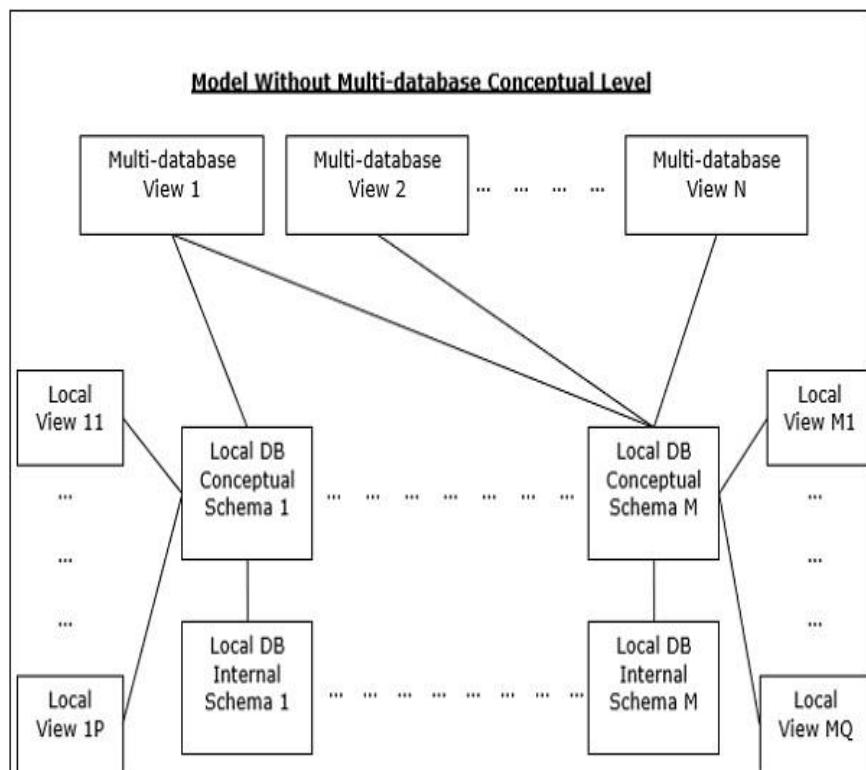


Figure 8: Model without multi-database conceptual level

## 2.4 Design Alternatives

The distribution design alternatives for the tables in a DDBMS are as follows:

- ❖ Non-replicated and non-fragmented
- ❖ Fully replicated
- ❖ Partially replicated
- ❖ Fragmented
- ❖ Mixed

#### Non-replicated & Non-fragmented

In this layout, different tables are located at varied sites. Data is placed in close proximity to the site where it is used maximum and is suitable where the percentage of queries need to join information in tables placed at varied sites is low. An appropriate distribution strategy reduces the communication cost during data processing.

#### Fully Replicated

In fully replicated layout, a copy of all the database tables is stored at each site due to which queries are executed in a fast manner with negligible communication cost. On the other side, massive redundancy in data incurs enormous cost during update operations and is appropriate for the systems where a large number of queries are to be handled with less number of database updates.

#### Partially Replicated

Copies of tables are stored at varied sites and the distribution is done in accord to the frequency of access. This takes into attention the truth that the frequency of having access to the tables range notably from site to site and the number of copies of the tables depends on how frequently the access queries execute and the site which generate the access queries.

#### Fragmented

In this layout, a table is split into two or extra pieces known as fragments or partitions with each fragment stored at varied sites providing increase in parallelism and better disaster recovery. Various fragmentation techniques are as follows:

- ❖ Vertical fragmentation
- ❖ Horizontal fragmentation
- ❖ Hybrid fragmentation

#### Mixed Distribution

This layout is a combination of fragmentation and partial replications. Tables are initially fragmented either in horizontal or vertical form and are replicated across the different sites in accord to the frequency of accessing the fragments.

### **2.5 Data Replication**

Is a manner of storing separate copies of the database at varied sites and is a popular fault tolerance technique of distributed databases.

Advantages:

- ❖ Reliability
- ❖ Reduction in Network Load
- ❖ Quicker Response
- ❖ Simpler Transactions

Disadvantages:

- ❖ Increased Storage Requirements
- ❖ Increased Cost and Complexity of Data Updating
- ❖ Undesirable Application – Database coupling

Commonly used replication techniques are:

- ❖ Snapshot replication
- ❖ Near-real-time replication
- ❖ Pull replication

## 2.6 Fragmentation

Fragmentation is process of dividing a table into a set of smaller tables called as **fragments**. Fragmentation is classified into three types: horizontal, vertical, and hybrid (combination of horizontal and vertical). Horizontal fragmentation can further be classified into two strategies: primary horizontal fragmentation and derived horizontal fragmentation.

Fragmentation should be done in a manner as the original table be reconstructed from the fragments as required and is referred as “reconstructiveness.”

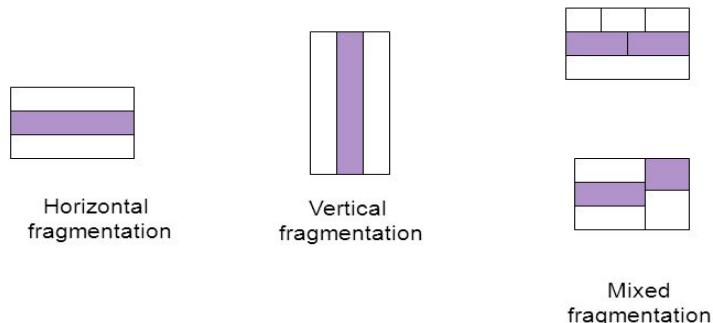


Figure 9: Fragmentation types

Advantages

- ❖ Increase in efficiency of the database system.
- ❖ Local query optimization techniques are enough for most queries.
- ❖ Security and privacy is maintained.

Disadvantages of Fragmentation

- ❖ Requirement of data from varied sites results in low access speed.
- ❖ Recursive fragmentations need expensive techniques.
- ❖ Lack of back-up copies renders the database ineffective.

### 2.6.1 Vertical Fragmentation

In this layout columns of a table are grouped into fragments and to maintain reconstructiveness, each fragment need to include the primary key field(s) of the table. Vertical fragmentation is primarily used to enforce privacy of data.

Example: A University database keeps records of all registered students in a Student table with the following schema.

STUDENT

Regd_No	Stu_Name	Course	Address	Semester	Fees	Marks

Now, the fees details are maintained in the accounts section. In this case, the designer will fragment the database as follows –

```
CREATE TABLE Stu_Fees AS
SELECT Regd_No, Fees
FROM STUDENT;
```

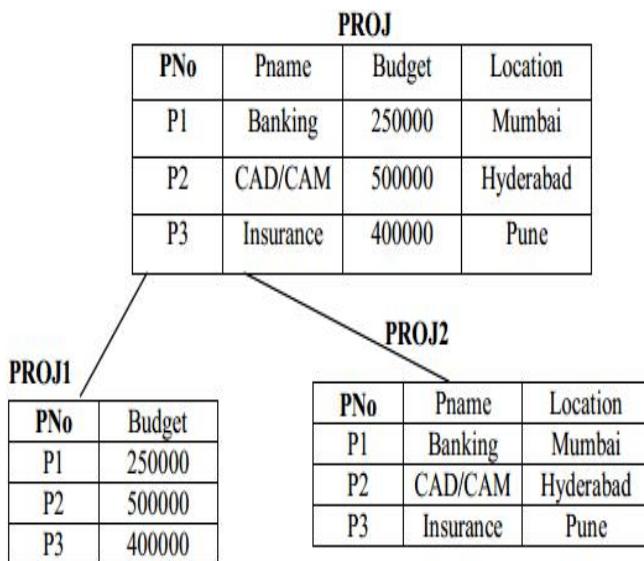


Figure 10: Vertical fragmentation

### 2.6.2 Horizontal Fragmentation

Groups the tuples of a table in accordance to the values of one or more fields. Horizontal fragmentation confirms the rule of “reconstructiveness” by having all the columns of the original base table.

Example: Details of all students of CS course needs to be maintained at the department of Computer Science. Horizontal fragment of the database is written as:

```

CREATE View COMP_STD AS
SELECT * FROM STUDENT
WHERE COURSE = "CS";

```

The diagram illustrates horizontal fragmentation. A large base table on the left is divided into two smaller fragments on the right by vertical lines. The base table has columns Attr1, Attr2, Attr3, Attr4, and Attr5, with data values ranging from 0 to 1. The top fragment has rows (0, 1, 1, 0, 1), (1, 0, 1, 1, 0), (0, 0, 0, 1, 0), (1, 1, 1, 1, 0), and (0, 0, 0, 0, 1). The bottom fragment has rows (1, 1, 1, 1, 0) and (0, 0, 0, 0, 1).

Attr1.	Attr2.	Attr3.	Attr4.	Attr5.
0	1	1	0	1
1	0	1	1	0
0	0	0	1	0
1	1	1	1	0
0	0	0	0	1

Attr1.	Attr2.	Attr3.	Attr4.	Attr5.
0	1	1	0	1
1	0	1	1	0

Attr1.	Attr2.	Attr3.	Attr4.	Attr5.
1	1	1	1	0
0	0	0	0	1

Figure 11: Horizontal fragmentation

### 2.6.3 Hybrid Fragmentation

A combination of horizontal and vertical fragmentation techniques are used in hybrid fragmentation and is the most flexible fragmentation technique as it generates fragments with minimal extraneous information, but reconstruction of the original base table is complex activity.

Hybrid fragmentation can be done in two alternative ways –

- ❖ Generate a set of horizontal fragments followed by the generation of vertical fragments from one or more of the horizontal fragments.
- ❖ Generate a set of vertical fragments followed by the generation of horizontal fragments from one or more of the vertical fragments.

The diagram illustrates hybrid fragmentation. A large base table on the left is divided into five fragments across three sites. Site 1 contains two fragments: one for columns Proj-name, Pno, and Location, and another for column Dept-no. Site 2 contains two fragments: one for columns Proj-name, Pno, and Location, and another for column Dept-no. Site 3 contains one fragment for column Proj-name. Site 4 contains one fragment for column Proj-name. Site 5 contains one fragment for column Dept-no.

Project			
Proj-name	Pno	Location	Dept-no
ProductX	1	Bellaire	5
ProductY	2	Sugarland	5
ProductZ	3	Houston	5
Computing	10	Stafford	4
Web sales	20	Houston	1
Benefits	30	Stafford	4

Proj-name	Pno	Location	Dept-no
ProductX	1	Bellaire	5
ProductY	2	Sugarland	5
ProductZ	3	Houston	5

Proj-name	Pno	Location	Dept-no
Computing	10	Stafford	4
Benefits	30	Stafford	4

Proj-name	Pno	Location	Dept-no
Web sales	20	Houston	1

Proj-name	Pno	Location	Dept-no
Computing	10	Stafford	4
Benefits	30	Stafford	4

Figure 12: Hybrid fragmentation

## References

1. David Bell and Jane Grimson. Distributed Database Systems (1st. ed.). Addison-Wesley Longman Publishing Co., Inc., USA. 1992.
2. Özsü MT, Valduriez P. Principles of distributed database systems. Englewood Cliffs: Prentice Hall; 1999 Feb.
3. Dye C. Oracle distributed systems. O'Reilly & Associates, Inc.; 1999 Apr 1.
4. Ozsu MT, Valduriez P. Distributed Databases: Principles and Systems. 1999.
5. Tuples S. Database Internals. 2002
6. Özsü, M. Tamer. . Distributed Database Systems. 2002.
7. Silberschatz A, Korth HF, Sudarshan S. Database system concepts. New York: McGraw-Hill; 1997 Apr.
8. Özsü MT, Valduriez P. Distributed and parallel database systems. ACM Computing Surveys (CSUR). 1996 Mar 1;28(1):125-8.
9. Van Alstyne MW, Brynjolfsson E, Madnick SE. Ownership principles for distributed database design. 1992.
10. Valduriez P, Jimenez-Peris R, Özsü MT. Distributed Database Systems: The Case for NewSQL. InTransactions on Large-Scale Data-and Knowledge-Centered Systems XLVIII 2021 (pp. 1-15). Springer, Berlin, Heidelberg.
11. Domaschka J, Hauser CB, Erb B. Reliability and availability properties of distributed database systems. In2014 IEEE 18th International Enterprise Distributed Object Computing Conference 2014 Sep 1 (pp. 226-233). IEEE.
12. Distributed databases. <https://www.db-book.com/db4/slide-dir/ch19-2.pdf> (Last accessed on 18.07.2021)
13. Distributed database management systems. <https://cs.uwaterloo.ca/~tozsuz/courses/cs856/F02/lecture-1-ho.pdf>. (Last accessed on 18.07.2021)

## MOOCs

1. DISTRIBUTED DATABASE SYSTEMS. <https://www.my-mooc.com/en/mooc/distributed-database-systems/>
2. Building Globally Distributed Databases with Cosmos DB. <https://www.coursera.org/projects/building-globally-distributed-databases-with-cosmos>
3. Distributed Database Systems. <https://www.classcentral.com/course/distributed-database-11170>
4. Database Systems Concepts & Design. <https://www.udacity.com/course/database-systems-concepts-design--ud150>
5. Database Systems. <https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-830-database-systems-fall-2010/>
6. Advanced Databases (saylor.org). <https://www.mooc-list.com/course/advanced-databases-saylororg>
7. Distributed Databases and Data Warehouses. <https://www.hse.ru/en/edu/courses/292696011>
8. Distributed Systems. [https://onlinecourses.nptel.ac.in/noc21\\_cs87/preview](https://onlinecourses.nptel.ac.in/noc21_cs87/preview)
9. Distributed Systems. <https://online.stanford.edu/courses/cs244b-distributed-systems>
10. DISTRIBUTED SYSTEMS. <https://www.distributedsystemscourse.com/>
11. Distributed Systems & Cloud Computing with Java. [https://www.udemy.com/course/distributed-systems-cloud-computing-with-java/?ranMID=39197&ranEAID=vedj0cWlu2Y&ranSiteID=vedj0cWlu2Y-Mp23N6jSsU4TZxvJaRgOrg&LSNPUBID=vedj0cWlu2Y&utm\\_source=aff-campaign&utm\\_medium=udemyads](https://www.udemy.com/course/distributed-systems-cloud-computing-with-java/?ranMID=39197&ranEAID=vedj0cWlu2Y&ranSiteID=vedj0cWlu2Y-Mp23N6jSsU4TZxvJaRgOrg&LSNPUBID=vedj0cWlu2Y&utm_source=aff-campaign&utm_medium=udemyads)

## Quiz

1. Storing a separate copy of the database at multiple locations is \_\_\_\_\_
2. \_\_\_\_\_ is the advantage of a distributed database over a centralized database
3. A distributed database is a collection of data which belong \_\_\_\_\_ to the same system but are spread over the \_\_\_\_\_ of the network.
4. \_\_\_\_\_ mean programs can be written as if a database is not distributed for its user.
5. In a distributed Database reduction of redundancy is obtained by \_\_\_\_\_
6. \_\_\_\_\_ are the main goals of a distributed database.
7. An autonomous homogenous environment is \_\_\_\_\_
8. A transaction manager is \_\_\_\_\_
9. Location transparency allows for \_\_\_\_\_
10. \_\_\_\_\_ is a heterogeneous distributed database
11. In \_\_\_\_\_ some of the columns of a relation are from different sites
12. \_\_\_\_\_ is a distributed database
13. \_\_\_\_\_ strategies is used by a distributed database
14. A sophisticated locking mechanism is known as 2-phase locking which includes the Growing phase and .....
15. A transaction processing system is also called as .....
16. The transactions are always ..... if it always locks a data item in shared mode before reading it.
17. .....is a server which is widely used in relational database systems.
18. \_\_\_\_\_ transaction property will check whether all the operation of a transaction completed or none
19. The total ordering of operations across groups ensures .....of transactions.
20. A distributed transaction can be ..... if queries are issued at one or more nodes.
21. \_\_\_\_\_ heterogeneous database systems is/are independent in nature and integrated together so that they function as a single database system
22. \_\_\_\_\_ is true for a heterogeneous database system
23. Global Wait-for graph is used for \_\_\_\_\_ in Distributed database
24. In Distributed database, \_\_\_\_\_ are the transactions for which a <ready T> log is found in the log file, but neither a <commit T> log nor an <abort T> log is found.
25. Read one, write all available protocol is used to increase \_\_\_\_\_ in a distributed database system.

## **Video links**

1. Distributed databases. <https://www.youtube.com/watch?v=QyR4TIbEJjo>
2. DBMS - Distributed Database System.  
<https://www.youtube.com/watch?v=aUyqZxn12sY>
3. Introduction to Distributed Databases.  
[https://www.youtube.com/watch?v=0\\_m5gPfzEYQ](https://www.youtube.com/watch?v=0_m5gPfzEYQ)
4. Introduction to Distributed Database System.  
[https://www.youtube.com/watch?v=RKmK\\_vKZsq8&list=PLduM7bkxBdOjbMXkTRdsSIWQKR43nSmd](https://www.youtube.com/watch?v=RKmK_vKZsq8&list=PLduM7bkxBdOjbMXkTRdsSIWQKR43nSmd)
5. Distributed Databases. <https://www.youtube.com/watch?v=J-sj3GUrq9k>
6. Centralised vs Distributed Databases.  
<https://www.youtube.com/watch?v=QjvjeQquon8>
7. Learn System design : Distributed datastores.  
<https://www.youtube.com/watch?v=l9JSK9OBzA4>
8. Architecture of Distributed database systems.  
<https://www.youtube.com/watch?v=vuApQk27Jus>
9. Distributed Database Introduction.  
<https://www.youtube.com/watch?v=Q1RIpXS7lPc>

## **Contents**

<b>Sl.No</b>	<b>Topic</b>	<b>Page No</b>
1.1	<b>Structured data types</b>	1
1.2	<b>Operations on structured data</b>	2
1.3	<b>Nested relations</b>	3
1.4	<b>Structured types in Oracle</b>	3
1.5	<b>Database objects in DBMS</b>	5
1.6	<b>Object datatype categories</b>	8
1.7	<b>Object tables</b>	10
1.8	<b>Object Identifiers</b>	10
1.9	<b>REFs</b>	11
1.10	<b>Collection types</b>	12
	<b>References</b>	16
	<b>MOOCs</b>	17
	<b>Quiz</b>	18
	<b>Video Links</b>	19

## Structured Data Types in DBMS

### 1.1 Structured data types

It is a user defined data type with elements which aren't atomic, are divisible and can be used separately or as a single unit as needed. The major advantage of using objects is the ability to define new data types (Abstract Data Types).

```
CREATE TYPE type_name AS  
(Attribute1_name data_type(size),  
Attribute2_name data_type(size),  
Attribute3_name data_type(size),  
.....  
AttributeN_name data_type(size));
```

Here, data\_type can be any of the following;

- ❖ It can be one of the valid data types like CHAR, VARCHAR2, NUMBER, INTEGER, etc.
- ❖ It can be another User Defined Type.

Example:

```
CREATE TYPE phone AS  
(Country_code NUMBER(4),  
STD_Code NUMBER(5),  
Phone_Number NUMBER(10))
```

```
CREATE TABLE contact  
(Contact_name VARCHAR2(20),  
Street VARCHAR2(20),  
City VARCHAR2(20),  
Ph PHONE);
```

**CREATE TYPE** is used for creating a structured data type whereas **DROP** is used for deleting.

Let us consider ‘Stud-Dept’ schema where table ‘Student’ is created with four columns namely Student\_No (a system generated column), Student\_Name (Name of the Student), Student\_Address (Address of the student) is a structured type column of type ‘Student\_ADDRESS-T’ and ProjNo (Project number of the Student) and StuImage (Images of the Student).

**CREATE TYPE Student\_ADDRESS-T as Row (street varchar2 (20), city varchar2 (20), state varchar2 (20), pin\_code varchar2 (10))**

```
CREATE TABLE Student (Student_No integer system generated, Student_Name varchar2 (20), Student_Address Student_ADDRESS-T,Proj-no varchar2 (20),Stuimage jpeg-image);
```

```
CREATE TABLE PROJECT  
(Projno integer, Pname varchar (20),  
Location REF (address-t) SCOPE LOCS,  
Student_No set of (integer));
```

```
CREATE TABLE DEPT  
(Dno integer, Dname varchar2 (20),  
Dlocation REF (address-t) SCOPE LOCS,  
Projno set of (integer));
```

PROJECT is a table specifying project number, location of the project and specifies the total no. of employees working on the project.

DEPT table specifies Department\_Name and department\_no (unique value) along with location of department. It also specifies the project completed or undertaken by the department.

Suppose , there is a need to find the names of Students along with their images who are living in ‘Namdevwada’ of ‘Nizamabad’.

```
SELECT Student_Name, Image FROM Student WHERE Address.street = 'Namdevwada'  
and address.city = 'Nizamabad';
```

### **1.2 Operations on Structured Data**

Structured data can be manipulated using built in methods for types defined using type constructor. These methods are similar to operations used for data types (atomic) of traditional RDBMS.

#### **i. Operations on Arrays**

Arrays are used in the same manner as in traditional RDBMS. ‘Array index’ method is used to return the number of elements in the array.

#### **ii. Operations on Rows**

Row type is a collection of fields values where each fields are accessed by traditional notation. Example: address-t.city specify the attribute ‘city’ of the type address-t.

Consider ‘Student-Dept’ schema in which we have to find the names of those employees who resides in ‘Namdevwada’ of ‘Nizamabad’.

```
SELECT S.Student_No,S.Student_Name  
FROM Student S  
WHERE S.Address.area ='Namdevwada' AND S.Address.city='Nizamabad'  
AND S.Address.city = 'Nizamabad'
```

#### **iii. Operations on Sets and Multi-sets**

Set and multisets are used in the conventional manner by using the use of  $=, <, >, >,<$  evaluation operators. An item of a hard and fast may be as compared with the aid of other objects the use of E (belongs to) relation. Two set objects can create a new item the usage of U, (Union Operation). They also can create a new object by way of subtracting using ‘-’ (set distinction operator ). Multi-set also uses the identical operations as utilized by the sets however the operations are implemented at the variety of copies of detail into account.

#### iv. Operations on Lists

List includes operations like ‘append’, ‘concatenate’, ‘head’, ‘tail’ and so forth. To govern the objects of list for example ‘concatenate’ or ‘append’ appends one listing to any other, ‘head’ returns the primary detail of list, ‘tail’ returns the list after getting rid of the primary element.

**1.3 Nested Relations :-**Attributes having complex kinds like setof (base), bagof (base) etc are known as ‘Nested Relation,. So ‘Unnesting’ is a manner or transforming a nested relation into 1NF relation. Let us consider ‘Student-dept’ schema wherein for each employee, we store the following information in Student table :-

1. Student\_No
2. Student\_Name
3. Student\_Address
4. Projno

The domains of some of the information stored for an Student are non atomic as, Projno, specifies the number of projects worked on by the Student . A Student may also have a set of projects to be worked on.

#### 1.4 Structured types in Oracle

Let us see some examples of defining and manipulating Structured types in Oracle.

```
CREATE TYPE Address AS OBJECT  
(Street VARCHAR2(20),  
City VARCHAR2(20),  
State VARCHAR2(20),  
Pincode NUMBER(10));
```

Execution of the above statement will create a new ABSTRACT datatype named ADDRESS and store the definition as part of the database.

This new type can be used to define an attribute in any TABLEs or TYPES as follows;

```
CREATE TABLE Student  
(Student_name VARCHAR2(20),  
Addr ADDRESS,  
Phone NUMBER(10));
```

This table Student will consist of 3 columns wherein the first one and the third one are of regular datatypes VARCHAR, and NUMBER respectively, and the second one is of the abstract type ADDRESS. The table PERSON will look like as follows;

Student_name	Addr				Phone
	Street	City	State	Pincode	

**Table 1: Student table**

### Advantages

1. Adopted by machine learning algorithms
2. Adopted by business users
3. Increased access to other tools

### Disadvantages

1. **Limited use**
2. **Limited storage**

### Examples

Common examples of machine-generated structured data are weblog statistics and point of sale data, such as barcodes and quantity.

### What is unstructured data?

**Unstructured data is data stored in its native format and not processed until it is used**, which is known as schema-on-read. It comes in a numerous of file formats, including email, social media posts, presentations, chats, IoT sensor data, and satellite imagery.

### Advantages

1. Freedom of the native format
2. Faster accumulation rates
3. Data lake storage

### Disadvantages

1. Requires data science expertise
2. Specialized tools

### Examples

It lends itself well to determining how effective a marketing campaign is, or to uncovering potential buying trends through social media and review websites.

### Differences between structured and unstructured data:

1. Defined vs Undefined Data
2. Qualitative vs Quantitative Data
3. Storage in Data Houses vs Data Lakes
4. Easy vs Hard to Analyze

## 5. Predefined format vs a variety of formats

What is semi-structured data?

Semi-structured data refers to what might commonly be considered unstructured data; however that still has metadata that identifies certain characteristics. The metadata incorporates enough information to enable the information to be extra efficiently catalogued, searched, and analyzed than strictly unstructured data. Think of semi-structured data as the move-between of structured and unstructured data.

A good example of semi-structured data vs. structured data might be a tab delimited document containing customer data versus a database incorporating CRM tables. On the other side of the coin, semi-structured has more hierarchy than unstructured data; the tab delimited file is more specific than a list of remarks from a customer's instagram.

## 1.5 Database Objects in DBMS

A **database object** is any described object in a database that is used to store or reference data. Anything which we make from **create command** is called as Database Object. It may be used to keep and manage the data. Examples: view, sequence, indexes and etc.

Database Object	Advantage
Table	Basic unit of storage; composed rows and columns
View	Logically represents subsets of data from one or more tables
Sequence	Generates primary key values
Index	Improves the performance of some queries
Synonym	Alternative name for an object

Different database Objects:

1. **Table** – This database object is used to create a table in database.

**Syntax :**

CREATE TABLE [schema.]table

(column datatype [DEFAULT expr][, ...]);

**Example :**

CREATE TABLE dept

(Deptno NUMBER(2),  
Dname VARCHAR2(20),  
Location VARCHAR2(20));

**Output :**

DESCRIBE dept;

Name	Null?	Type
DeptNo		Number(2)
DName		Varchar2(20)
Location		Varchar2(20)

2. **View** – A view is a logical table. A view contains no data of its own and is like a window through which data from tables can be viewed or changed. Table / tables on which a view is based are called base table/ tables.

**Syntax :**

```
CREATE [OR REPLACE] [FORCE|NOFORCE] VIEW view
    [(alias[, alias]...)]
    AS subquery
    [WITH CHECK OPTION [CONSTRAINT constraint]]
    [WITH READ ONLY [CONSTRAINT constraint]];
```

**Example :**

```
CREATE VIEW dsrao
    AS SELECT Student_id ID_NUMBER, last_name Last_Name,
        salary*12 Annual_Salary
    FROM Student
    WHERE department_id = 111;
```

**Output :**

```
SELECT *
FROM dsrao;
```

3. **Sequence** – This database object is used to create a sequence in database.

**Use:** It is used to create a primary key value by which we can identify the record uniquely. It is generated and incremented by an internal Oracle routine.

**Syntax :**

```
CREATE SEQUENCE sequence
    [INCREMENT BY n]
    [START WITH n]
    [{MAXVALUE n | NOMAXVALUE}]
    [{MINVALUE n | NOMINVALUE}]
    [{CYCLE | NOCYCLE}]
    [{CACHE n | NOCACHE}];
```

**Example :**

```
CREATE SEQUENCE dept_deptid_seq
    INCREMENT BY 10
    START WITH 120
    MAXVALUE 9999
    NOCACHE
    NOCYCLE;
```

**Check if sequence is created by :**

```
SELECT sequence_name, min_value, max_value,
    increment_by, last_number
    FROM user_sequences;
```

4. **Index** – Are used to create indexes in database and speed up the rows with the aid of retrieval the usage of a pointer. Indexes can be created explicitly or routinely and calls for a complete scan in the absence of an index on a column. Indexes are logically and physically impartial of the table they index. They can be created or dropped at any time and haven't any effect on the base tables or different indexes.

**Syntax :**

CREATE INDEX index

    ON table (column[, column]...);

**Example :**

CREATE INDEX emp\_last\_name\_idx

    ON employees(last\_name);

5. **Synonym** – This database item is used to create indexes in database. It simplifies access to objects with the aid of developing a synonym. Creating a synonym eliminates the need to qualify the object name with the schema and provides you with an alternative name for a table, view, sequence, procedure, or other objects.

**Syntax:**

PUBLIC : creates a synonym accessible to all users

synonym : is the name of the synonym to be created

object : identifies the object for which the synonym is created

**Syntax :**

CREATE [PUBLIC] SYNONYM synonym FOR object;

**Example :**

CREATE SYNONYM d\_sum FOR dept\_sum\_vu;

Oracle is an **object-relational** database management system (ORDBMS), this means that users can outline additional forms of information--specifying both the structure of the data and the ways of operating on it--and use those types within the relational model. This approach provides value to the facts saved in a database. Object datatypes make it easier for application developers to work with complex data such as images, audio, and video.

### Complex Data Models

The Oracle server allows us to go for complicated enterprise models in SQL and cause them to a part of your database schema.

### Multimedia Datatypes

Much efficiency of database systems arises from their optimized management of fundamental data types like numbers, dates, and characters. Facilities exist for comparing values, determining their distributions, constructing efficient indexes, and performing other optimizations. Text, video, sound, graphics, and spatial data are examples of vital business entities that don't suit neatly into those basic kinds. Oracle Enterprise Edition supports modeling and implementation of these complex data types commonly known as multimedia data types.

## 1.6 Object Datatype Categories

There are two categories of object data types:

- ❖ Object types
- ❖ Collection types

Object datatypes use the built-in datatypes and other user-defined datatypes as the constructing blocks for datatypes that model the structure and behavior of data in applications.

### Object Types

Object types are abstractions of the real-global entities and are a schema object with three types of components specifically name, attributes and methods. A structured data unit that matches the template is termed to be an **object**.

### Purchase Order Example

Examples: `external_student`, `lineitem`, and `purchase_order`.

The attributes of `purchase_order` are `id`, `contact`, and `lineitems`. The attribute `contact` is an object, and the attribute `lineitems` is a nested table.

```
CREATE TYPE external_student AS OBJECT (
```

```
    name    VARCHAR2(30),
    phone   VARCHAR2(20) );
```

```
CREATE TYPE lineitem AS OBJECT (
```

```
    item_name  VARCHAR2(30),
    quantity   NUMBER,
    unit_price NUMBER(12,2) );
```

```
CREATE TYPE lineitem_table AS TABLE OF lineitem;
```

```
CREATE TYPE purchase_order AS OBJECT (
```

```
    id      NUMBER,
    contact external_student,
    lineitems lineitem_table,
    MEMBER FUNCTION
    get_value RETURN NUMBER );
```

For example, you can define a relational table to keep track of your contacts:

```
CREATE TABLE contacts (
    contact  external_person
    date     DATE );
```

The `contacts` table is a relational table with an object type defining certainly one of its columns. Objects that occupy columns of relational tables are called **column objects**.

### Types of Methods

Methods of an item kind version the behavior of items and are extensively categorised into member, static and comparison.

In the example, `purchase_order` has a method named `get_value`. Each purchase order object has its own `get_value` method.

Example: `x` and `y` are PL/SQL variables that hold purchase order objects and `w` and `z` are variables that hold numbers, the following two statements can leave `w` and `z` with distinct values:

```
w = x.get_value();
z = y.get_value();
```

The term `x.get_value()` is an invocation of the method `get_value`.

### Object Type Constructor Methods

Every object type has a system-defined **constructor method**; and is, a method that makes a new object according to the object type's specification.

For example, the expression:

```
purchase_order(
    1000376,
    external_student ("John Smith","1-800-555-1212"),
    NULL )
```

represents a purchase order object with the following attributes:

```
id      1000376
contact  external_student ("John Smith","1-800-555-1212")
lineitems NULL
```

### Comparison Methods

Oracle has amenities for comparing two data items and determines which is greater. Oracle affords two ways to define an order relationship among objects of a given object type: map methods and order strategies.

**Map** strategies use Oracle's capacity to examine/compare built-in types.

**Order** methods are more general and are used to compare two objects of a given object type. It returns -1 if the first is smaller, 0 if they are equal, and 1 if the first is greater.

## 1.7 Object Tables

An **object table** is a unique form of table that holds objects and provides a relational view of the attributes of those items.

For example, the following statement defines an object table for objects of the `external_person` type defined earlier:

```
CREATE TABLE external_student_table OF external_student;
```

Example:

```
INSERT INTO external_student_table VALUES ("John Smith", "1-800-555-1212");
```

```
SELECT VALUE(p) FROM external_student_table p
```

```
WHERE p.name = "John Smith";
```

The first instruction inserts an `external_person` object into `external_person_table` as a multicolumn table. Second selects from `external_person_table` as a single column table.

## Row Objects and Column Objects

Objects that appear in object tables are called **row objects** and the objects that emerge in table columns or as attributes of other objects are known as **column objects**.

### 1.8 Object Identifiers

Every row object in an object table has an associated logical object identifier (OID). Oracle assigns a completely unique system-generated identifier of length 16 bytes as the OID for every row object with the aid of default. The OID column of an object table is a hidden column. Oracle uses this value to construct object references to the row objects that might be used for fetching and navigating objects.

The purpose of the OID for a row object is to uniquely identify it in an object table by creating and maintaining an index on the OID column of an object table.

## Primary-Key Based Object Identifiers

Oracle lets in the choice of specifying the primary key value of a row object as the object identifier for the row object.

### Object Views Description

An object view is a virtual object table and its rows are row objects.

### 1.9 REFs

Oracle presents a built-in datatype called `REF` to encapsulate references to row objects of a specified object type. From a modeling perspective, `REFs` provide the ability to confine an association among two row objects. Oracle makes use of object identifiers to build such `REFs`.

## Scoped REFs

In declaring a column type, collection element, or object type attribute to be a `REF`, you can constrain it to contain only references to a specified object table. Such a `REF` is called a **scoped REF**. Scoped `REFs` require less storage space and permit more efficient access than unscoped `REFs`.

## Dangling REFs

It is feasible for the object identified with the aid of a `REF` to become unavailable through either deletion of the object or a change in privileges. Such a `REF` is referred to as **dangling**.

## Dereference REFs

Accessing the object stated by a `REF` is called **dereferencing** the `REF`. Dereferencing a dangling `REF` consequences in a null object.

Oracle provides **implicit dereferencing** of `REFs`. For example, consider the following:

```
CREATE TYPE person AS OBJECT (
    name  VARCHAR2(30),
    manager REF person );
```

If `x` represents an object of type `PERSON`, then the expression:

`x.manager.name`

represents a string containing the `name` attribute of the `person` object referred to by the `manager` attribute of `x`. The previous expression is a shortened form of:

`y.name`, where `y = DEREF(x.manager)`

### Obtain REFs

You can obtain a `REF` to a row object by selecting the object from its object table and applying the `REF` operator. For example, you can obtain a `REF` to the purchase order with identification number 1000376 as follows:

```
DECLARE OrderRef REF TO purchase_order;
```

```
SELECT REF(po) INTO OrderRef
  FROM purchase_order_table po
 WHERE po.id = 1000376;
```

## 1.10 Collection Types

Each collection type describes a data unit made up of an indefinite variety of elements, all of the identical datatype. The collection types include **array types** and **table types**.

Array types and table types are schema objects. The corresponding data units are referred to as **VARRAYs** and **nested tables**.

Collection types have constructor strategies. The call of the constructor method is the call of the type, and its argument is a comma separated listing of the new collection's elements. The constructor approach is a feature returning the new collection as its value.

### VARRAYs

An **array** is an ordered set of data **elements**. Each element has an **index**, which is a number corresponding to the element's position in the array.

The number of elements in an array is the **size** of the array. Oracle permits arrays to be of variable size, that's why they may be called VARRAYS. You have to specify a maximum size while you declare the array type.

Example: `CREATE TYPE prices AS VARRAY(10) OF NUMBER(12,2);`

The VARRAYS of type `prices` have no more than 10 elements, each of datatype `NUMBER(12, 2)`.

Creating an array type does not allocate space. It defines a datatype, which you can use as:

- ❖ The datatype of a column of a relational table
- ❖ An object type attribute
- ❖ A PL/SQL variable, parameter, or function return type.

A VARRAY is normally stored in line; that is, in the same tablespace as the other data in its row. If it is sufficiently large, however, Oracle stores it as a BLOB.

### Nested Tables Description

A **nested table** is an unordered set of data **elements**, all of the identical datatype. It has a single column, and the kind of that column is a built-in type or an object type. If an object type, the table can also be viewed as a multicolumn table, with a column for each characteristic of the object type. For example, in the purchase order example, the following statement declares the table type used for the nested tables of line items:

```
CREATE TYPE lineitem_table AS TABLE OF lineitem;
```

A table type definition does not allocate space. It defines a type, which you can use as:

- The datatype of a column of a relational table
- An object kind attribute
- A PL/SQL variable, parameter, or function return type

When a table type seems as the kind of a column in a relational table or as an characteristic of the underlying object type of an object table, Oracle stores all of the nested table data in a single table, which it buddies with the enclosing relational or object table. For example, the following declaration defines an object table for the object type `purchase_order`:

```
CREATE TABLE purchase_order_table OF purchase_order
  NESTED TABLE lineitems STORE AS lineitems_table;
```

The second line specifies `lineitems_table` as the storage table for the `lineitems` attributes of all of the `purchase_order` objects in `purchase_order_table`.

A convenient way to get entry to the elements of a nested table individually is to apply a nested cursor.

### FINAL and NOT FINAL Types

A type declaration ought to have `NOT FINAL` keyword, if you want it to have subtypes. The default is that the kind is `FINAL`; that is, no subtypes may be created for the kind. This lets in for backward compatibility.

#### Example of Creating a NOT FINAL Object Type

```
CREATE TYPE Person_t AS OBJECT
  ( ssn NUMBER,
    name VARCHAR2(30),
    address VARCHAR2(100)) NOT FINAL;
```

`Person_t` is declared to be a `NOT FINAL` type. This enables definition of subtypes of `Person_t`.

`FINAL` types can be altered to be `NOT FINAL`. In addition, `NOT FINAL` types with no subtypes can be altered to be `FINAL`.

## NOT INSTANTIABLE Types and Methods

A type can be declared to be NOT INSTANTIABLE. This implies that there is no constructor (default or user-defined) for the type. Thus, it is not possible to construct instances of this type.

The typical use would be define instantiable subtypes for such a type, as follows:

```
CREATE TYPE Address_t AS OBJECT(...) NOT INSTANTIABLE NOT FINAL;  
CREATE TYPE USAAddress_t UNDER Address_t(...);  
CREATE TYPE IntlAddress_t UNDER Address_t(...);
```

For example:

```
CREATE TYPE T AS OBJECT  
(  
    x NUMBER,  
    NOT INSTANTIABLE MEMBER FUNCTION func1() RETURN NUMBER  
) NOT INSTANTIABLE;
```

A subtype of a NOT INSTANTIABLE type can override any of the non-instantiable methods of the supertype and provide concrete implementations. If there are any non-instantiable methods remaining, the subtype must also necessarily be declared NOT INSTANTIABLE.

A non-instantiable subtype can be defined under an instantiable supertype. Declaring a non-instantiable type to be FINAL is not allowed.

## Difference Between Object Oriented Database and Object Relational Database

The most important difference among Object Oriented Database and Object Relational Database is that Object Oriented Database is a database that represents data in the form of objects like in Object Oriented Programming at the same time Object Relational Database is a database that is based totally on the relational model and object-oriented database model.

## Object Oriented Database

Object-oriented databases constitute data in the form of objects and classes. As per the object-oriented paradigm, an object is a real-world entity. In addition, a class enables to create objects. Moreover, object-oriented databases follow the principles of object-oriented programming.

In addition, object-oriented databases support OOP concepts such as inheritance, encapsulation and so forth. It additionally helps complex objects such as maps, sets, lists, tuples or collections of multiple primitive objects. Furthermore, Object-oriented database permits the user to create persistent objects which assist to overcome the database issues like concurrency and recovery. These objects stay in computer memory even after completing the execution.

## Object Relational Database

Object-relational database is an advanced version of the object-oriented database. It gives a approach to the issues users face in object-oriented databases; some of these issues include cost for computing resources, possibilities of design errors and data inconsistency.

Furthermore, these databases help objects and inheritance and offer a higher interface for many object-oriented languages. Users can also use data model extensions with custom data types and strategies. Moreover, companies like Microsoft, Oracle, and Sybase have object-relational versions of their products.

### **Difference between Object Oriented Database and Object Relational Database**

An object-oriented database is a database that represents information in the form of objects as used in object-oriented Programming. An object-relational database, on the other hand, is a database that depends on the relational model and the object-oriented database model. Thus, this is the main difference among object oriented database and object relational database.

Answer the Following:

1. What is an OID?
2. What are the strategies for obtaining a legitimate OID?
3. Which association maintains an OID registry?

## References:

1. "Object Database." Wikipedia, Wikimedia Foundation, 16 Mar. 2019, [Available here](#).
2. "What Is an Object-Relational Database (ORD)? – Definition from Techopedia." Techopedia.com, [Available here](#).
3. "Object-Relational Database." Wikipedia, Wikimedia Foundation, 8 July 2018, [Available here](#).
4. "What is an Object-Oriented Database?", Study.com, [Available here](#).
5. <https://www.geeksforgeeks.org/database-objects-in-dbms/>(14.08.21)
6. <https://www.ibm.com/docs/en/db2/11.5?topic=tables-reference-types>(14.08.21)
7. <https://www.ibm.com/docs/en/i/7.2?topic=schema-object-identifier-oid>(14.08.21)
8. <https://pediaaa.com/difference-between-object-oriented-database-and-object-relational-database/>(14.08.21)

## MOOCs

1. Introduction to Structured Query Language (SQL). University of Michigan. Coursera.  
<https://www.coursera.org/learn/intro-sql>
2. Intermediate Relational Database and SQL. Coursera.  
<https://www.coursera.org/projects/intermediate-rdb-sql>
3. Introduction to Relational Database and SQL. Coursera.  
<https://www.coursera.org/projects/introduction-to-relational-database-and-sql>
4. Oracle SQL - A Complete Introduction. Udemy.  
[https://www.udemy.com/course/introduction-to-oracle-sql/?LSNPUBID=JVFXdTr9V80&ranEAID=JVFXdTr9V80&ranMID=39197&ranSiteID=JVFXdTr9V80-d.tBI6h6Ou\\_r6Fk7THfQ7Q&utm\\_medium=udemyads&utm\\_source=aff-campaign](https://www.udemy.com/course/introduction-to-oracle-sql/?LSNPUBID=JVFXdTr9V80&ranEAID=JVFXdTr9V80&ranMID=39197&ranSiteID=JVFXdTr9V80-d.tBI6h6Ou_r6Fk7THfQ7Q&utm_medium=udemyads&utm_source=aff-campaign)
5. Oracle SQL: An Introduction to the most popular database. Udemy.  
[https://www.udemy.com/course/oracle-sql-an-introduction-to-the-most-popular-database/?ranMID=39197&ranEAID=JVFXdTr9V80&ranSiteID=JVFXdTr9V80-gPUoUHGA.bk7GEc2CHkc5g&LSNPUBID=JVFXdTr9V80&utm\\_source=aff-campaign&utm\\_medium=udemyads](https://www.udemy.com/course/oracle-sql-an-introduction-to-the-most-popular-database/?ranMID=39197&ranEAID=JVFXdTr9V80&ranSiteID=JVFXdTr9V80-gPUoUHGA.bk7GEc2CHkc5g&LSNPUBID=JVFXdTr9V80&utm_source=aff-campaign&utm_medium=udemyads)
6. Oracle SQL Developer: Mastering its Features + Tips & Tricks. Udemy.  
[https://www.udemy.com/course/oracle-sql-developer-tips-and-tricks/?LSNPUBID=JVFXdTr9V80&ranEAID=JVFXdTr9V80&ranMID=39197&ranSiteID=JVFXdTr9V80-BJvtSlb2eHT3z05lbG2Tow&utm\\_medium=udemyads&utm\\_source=aff-campaign](https://www.udemy.com/course/oracle-sql-developer-tips-and-tricks/?LSNPUBID=JVFXdTr9V80&ranEAID=JVFXdTr9V80&ranMID=39197&ranSiteID=JVFXdTr9V80-BJvtSlb2eHT3z05lbG2Tow&utm_medium=udemyads&utm_source=aff-campaign)
7. Oracle Database 12c Fundamentals. Pluralsight.  
[https://www.pluralsight.com/courses/oracle-database-12c-fundamentals?clickid=Wrd1mUSpBxyLWCdRIKxBMx0uUkBtkN3JqS-kwM0&irgwc=1&mpid=1193463&aid=7010a000001xAKZAA2&utm\\_medium=digital\\_affiliate&utm\\_campaign=1193463&utm\\_source=impactradius](https://www.pluralsight.com/courses/oracle-database-12c-fundamentals?clickid=Wrd1mUSpBxyLWCdRIKxBMx0uUkBtkN3JqS-kwM0&irgwc=1&mpid=1193463&aid=7010a000001xAKZAA2&utm_medium=digital_affiliate&utm_campaign=1193463&utm_source=impactradius)
8. Step by Step Practical Oracle SQL with real life exercises. Udemy.  
[https://www.udemy.com/course/oracle-and-sql-step-by-step-learning/?LSNPUBID=JVFXdTr9V80&ranEAID=JVFXdTr9V80&ranMID=39197&ranSiteID=JVFXdTr9V80-Qclzu0fxjk7S80GoaFVfw&utm\\_medium=udemyads&utm\\_source=aff-campaign](https://www.udemy.com/course/oracle-and-sql-step-by-step-learning/?LSNPUBID=JVFXdTr9V80&ranEAID=JVFXdTr9V80&ranMID=39197&ranSiteID=JVFXdTr9V80-Qclzu0fxjk7S80GoaFVfw&utm_medium=udemyads&utm_source=aff-campaign)

## **QUIZ**

1. Varrays are a good choice when \_\_\_\_\_
2. The constructs of a procedure, function or a package are \_\_\_\_\_ .
3. \_\_\_\_\_ sorts rows in SQL
4. The \_\_\_\_\_ is a statement that queries or reads data from a table
5. The SQL keyword(s) \_\_\_\_\_ is used with wildcards.
6. The advantage of pl/sql is \_\_\_\_\_
7. \_\_\_\_\_ are the features of pl/sql
8. \_\_\_\_\_ clause creates temporary relation for the query on which it is defined.
9. \_\_\_\_\_ command makes the updates performed by the transaction permanent in the database
10. \_\_\_\_\_ command is used to change the definition of a table in SQL
11. A CASE SQL statement is \_\_\_\_\_
12. Shared locks are applied while performing \_\_\_\_\_
13. Sequence can generate \_\_\_\_\_
14. A sequence in SQL can generate a maximum number \_\_\_\_\_
15. \_\_\_\_\_ is NOT a type of constraint in SQL language
16. \_\_\_\_\_ data dictionary table can be used to show the object privileges granted to the user on specific columns
17. \_\_\_\_\_ is a constraint that can be defined only at the column level
18. \_\_\_\_\_ is a view
19. SQL Server has mainly \_\_\_\_\_ many types of views
20. Dynamic Management View is a type of \_\_\_\_\_
21. You can delete a view with \_\_\_\_\_ command.
22. \_\_\_\_\_ is stored only in the Master database.

## **Video Links**

1. Oracle - SQL - Creating Synonyms.  
<https://www.youtube.com/watch?v=uKKLgpsIkCY>
2. Oracle SQL Programming - Sequences, Indexes & Synonyms.  
<https://www.youtube.com/watch?v=s4HGIcYKtUw>
3. How to Create and Use Indexes in Oracle Database.  
<https://blogs.oracle.com/sql/post/how-to-create-and-use-indexes-in-oracle-database>
4. Oracle - SQL - Creating Index. <https://www.youtube.com/watch?v=fkBdIroDWQs>
5. Oracle - SQL - Creating Views. [https://www.youtube.com/watch?v=MfvrQH\\_DG8s](https://www.youtube.com/watch?v=MfvrQH_DG8s)
6. SQL tutorial 61: SEQUENCE in Oracle Database By Manish Sharma RebellionRider.  
<https://www.youtube.com/watch?v=RrajmYKzIVQ>
7. Oracle - SQL - Creating Sequences.  
<https://www.youtube.com/watch?v=H5FvZjPsxV4>
8. SQL tutorial 62: Indexes In Oracle Database By Manish Sharma RebellionRider.  
<https://www.youtube.com/watch?v=F5NrQYD4a9g>
9. How to Use Create Table, Alter Table, and Drop Table in Oracle Database.  
<https://blogs.oracle.com/sql/post/how-to-use-create-table-alter-table-and-drop-table-in-oracle-database>

## **UNIT 2 – Chapter 2**

### **DIMENSIONAL MODELLING**

#### **Unit Structure**

2.0 Objectives

2.1-Dimensional Modelling

    2.1.1 Objectives of Dimensional Modelling

    2.1.2 Advantages of Dimensional Modelling

    2.1.3 Disadvantages of Dimensional Modelling

2.2 Elements of Dimensional Data Model

2.3 Steps of Dimensional Modelling

    2.2.1 Fact Table

    2.2.2 Dimension Tables

2.4 Benefits of Dimensional Modelling

2.5 Dimensional Models

2.6 Types of Data Warehouse Schema:

- 2.6.1 Star Schema
- 2.6.2 Snowflake Schema
- 2.6.3 Galaxy Schema
- 2.6.4 Star Cluster Schema

## 2.7 Star Schema Vs Snowflake Schema: Key Differences

## 2.8 Summary

# 2.0 Objectives

This chapter will enable the readers to understand the following concepts:

- Meaning of Dimensional Modelling including its objectives, advantages, and disadvantages
- The steps in Dimensional Modelling
- Understanding of Fact Tables and Dimension Tables
- Benefits of Dimensional Modelling
- Understanding of different schemas – Star, Snowflake, Galaxy and Start Cluster schema
- Key differences between the Star Schema and the Snowflake Schema

# 2.1-Dimensional Modelling

**Dimensional Modelling (DM)** is a data structure technique optimized for data storage in a Data warehouse. The purpose of dimensional modelling is to optimize the database for faster retrieval of data. The concept of Dimensional Modelling was developed by Ralph Kimball and consists of “fact” and

A dimensional model in data warehouse is designed to read, summarize, analyse numeric information like values, balances, counts, weights, etc. in a data warehouse. In contrast, relation models are optimized for addition, updating and deletion of data in a real-time Online Transaction System.

These dimensional and relational models have their unique way of data storage that has specific advantages.

### 2.1.1 Objectives of Dimensional Modelling

The purposes of dimensional modelling are:

1. To produce database architecture that is easy for end-clients to understand and write queries.

2. To maximize the efficiency of queries. It achieves these goals by minimizing the number of tables and relationships between them.

#### 2.1.2 Advantages of Dimensional Modelling

Following are the benefits of dimensional modelling are:

- **Dimensional modelling is simple:** Dimensional modelling methods make it possible for warehouse designers to create database schemas that business customers can easily hold and comprehend. There is no need for vast training on how to read diagrams, and there is no complicated relationship between different data elements.
- **Dimensional modelling promotes data quality:** The star schema enable warehouse administrators to enforce referential integrity checks on the data warehouse. Since the fact information key is a concatenation of the essentials of its associated dimensions, a factual record is actively loaded if the corresponding dimensions records are duly described and also exist in the database. By enforcing foreign key constraints as a form of referential integrity check, data warehouse DBAs add a line of defence against corrupted warehouses data.
- **Performance optimization is possible through aggregates:** As the size of the data warehouse increases, performance optimization develops into a pressing concern. Customers who have to wait for hours to get a response to a query will quickly become discouraged with the warehouses. Aggregates are one of the easiest methods by which query performance can be optimized.

#### 2.1.3 Disadvantages of Dimensional Modelling

- To maintain the integrity of fact and dimensions, loading the data warehouses with a record from various operational systems is complicated.
- It is severe to modify the data warehouse operation if the organization adopting the dimensional technique changes the method in which it does business.

## 2.2 Elements of Dimensional Data Model

### Fact

- Facts are business measurements. Facts are normally but not always numeric values that could be aggregated. e.g., number of products sold per quarter.
- Facts are the measurements/metrics or facts from your business process. For a Sales business process, a measurement would be quarterly sales number

## **Dimension**

- Dimensions are called contexts. Dimensions are business descriptors that specify the facts, for example, product name, brand, quarter, etc.
- Dimension provides the context surrounding a business process event. In simple terms, they give who, what, where of a fact. In the Sales business process, for the fact quarterly sales number, dimensions would be
- Who – Customer Names
- Where – Location
- What – Product Name

In other words, a dimension is a window to view information in the facts.

## **Attributes**

The Attributes are the various characteristics of the dimension in dimensional data modelling.

In the Location dimension, the attributes can be

- State
- Country
- Zipcode etc.

Attributes are used to search, filter, or classify facts. Dimension Tables contain Attributes

## **Fact Table**

A fact table is a primary table in dimension modelling.

A Fact Table contains

1. Measurements/facts
2. Foreign key to dimension table

## **Dimension Table**

- A dimension table contains dimensions of a fact.
- They are joined to fact table via a foreign key.
- Dimension tables are de-normalized tables.
- The Dimension Attributes are the various columns in a dimension table
- Dimensions offers descriptive characteristics of the facts with the help of their attributes
- No set limit set for given for number of dimensions
- The dimension can also contain one or more hierarchical relationships

## 2.3 Steps of Dimensional Modelling

The accuracy in creating your Dimensional modelling determines the success of your data warehouse implementation. Here are the steps to create Dimension Model

1. Identify Business Process
2. Identify Grain (level of detail)
3. Identify Dimensions
4. Identify Facts
5. Build Schema

The model should describe the Why, How much, When/Where/Who and What of your business process

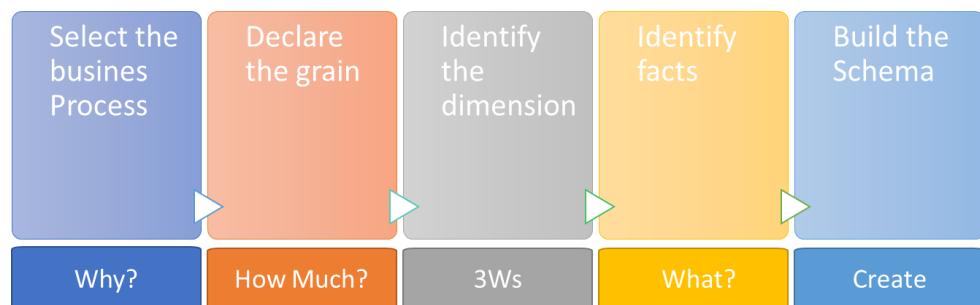


Figure 1- Steps of Dimensional Modelling

### Step 1) Identify the Business Process

Identifying the actual business process a data Warehouse should cover. This could be Marketing, Sales, HR, etc. as per the data analysis needs of the organization. The selection of the Business process also depends on the quality of data available for that process. It is the most important step of the Data Modelling process, and a failure here would have cascading and irreparable defects.

To describe the business process, you can use plain text or use basic Business Process Modelling Notation (BPMN) or Unified Modelling Language (UML).

### Step 2) Identify the Grain

The Grain describes the level of detail for the business problem/solution. It is the process of identifying the lowest level of information for any table in your data warehouse. If a table contains sales data for every day, then it should be daily granularity. If a table contains total sales data for each month, then it has monthly granularity.

During this stage, you answer questions like

- Do we need to store all the available products or just a few types of products? This decision is based on the business processes selected for Datawarehouse

- Do we store the product sale information on a monthly, weekly, daily or hourly basis? This decision depends on the nature of reports requested by executives
- How do the above two choices affect the database size?

For example, the CEO at an MNC wants to find the sales for specific products in different locations on a daily basis. So, the grain is "product sale information by location by the day."

### **Step 3) Identify the Dimensions**

Dimensions are nouns like date, store, inventory, etc. These dimensions are where all the data should be stored. For example, the date dimension may contain data like a year, month and weekday.

For example, the CEO at an MNC wants to find the sales for specific products in different locations on a daily basis.

- Dimensions: Product, Location and Time
- Attributes: For Product: Product key (Foreign Key), Name, Type, Specifications
- Hierarchies: For Location: Country, State, City, Street Address, Name

### **Step 4) Identify the Fact**

This step is co-associated with the business users of the system because this is where they get access to data stored in the data warehouse. Most of the fact table rows are numerical values like price or cost per unit, etc.

For example, the CEO at an MNC wants to find the sales for specific products in different locations on a daily basis. The fact here is Sum of Sales by product by location by time.

### **Step 5) Build Schema**

In this step, you implement the Dimension Model. A schema is nothing but the database structure (arrangement of tables). There are two popular schemas

- STAR SCHEMA
- SNOWFLAKE SCHEMA

For example, a city and state can view a store summary in a fact table. Item summary can be viewed by brand, color, etc. Customer information can be viewed by name and address.

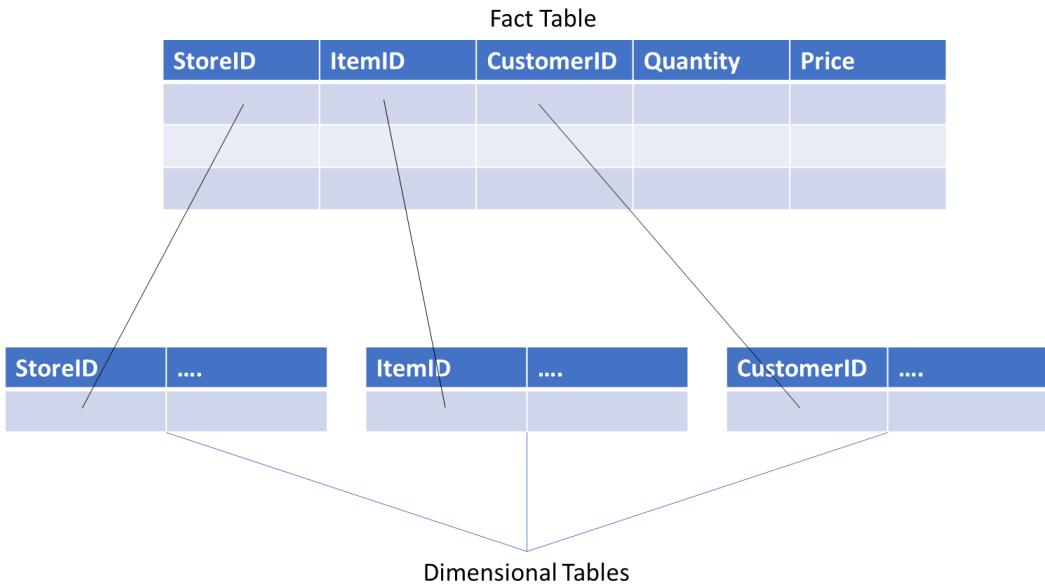


Figure 2 - Fact Tables and Dimension Tables

### 2.2.1 Fact Table

Location ID	Product Code	Customer ID	Unit Sold
44	172321	22345623	2
82	212121	31211324	1
58	434543	10034213	3

In this example, Customer ID column in the facts table is the foreign keys that join with the dimension table. By following the links, we can see that row 2 of the fact table records the fact that customer 31211324, Gaurav, bought one items at Location 82.

### 2.2.2 Dimension Tables

Customer ID	Name	Gender	Income	Education	Region
12232232	Rohan	Male	23000	3	4
22432253	Sandeep	Male	35000	5	1
31211324	Gaurav	Male	120000	1	3

## 2.4 Benefits of Dimensional Modelling

- Standardization of dimensions allows easy reporting across areas of the business.
- Dimension tables store the history of the dimensional information.
- It allows to introduce entirely new dimension without major disruptions to the fact table.
- Dimensional also to store data in such a fashion that it is easier to retrieve the information from the data once the data is stored in the database.
- Compared to the normalized model dimensional table are easier to understand.
- Information is grouped into clear and simple business categories.

- The dimensional model is very understandable by the business. This model is based on business terms, so that the business knows what each fact, dimension, or attribute means.
- Dimensional models are de-normalized and optimized for fast data querying. Many relational database platforms recognize this model and optimize query execution plans to aid in performance.
- Dimensional modelling in data warehouse creates a schema which is optimized for high performance. It means fewer joins and helps with minimized data redundancy.

## 2.5 Dimensional Models

A multidimensional model views data in the form of a data-cube. A data cube enables data to be modelled and viewed in multiple dimensions. It is defined by dimensions and facts.

The dimensions are the perspectives or entities concerning which an organization keeps records. For example, a shop may create a sales data warehouse to keep records of the store's sales for the dimension time, item, and location. These dimensions allow the user to keep track of things, for example, monthly sales of items and the locations at which the items were sold. Each dimension has a table related to it, called a dimensional table, which describes the dimension further. For example, a dimensional table for an item may contain the attributes item\_name, brand, and type.

A multidimensional data model is organized around a central theme, for example, sales. This theme is represented by a fact table. Facts are numerical measures. The fact table contains the names of the facts or measures of the related dimensional table.

**Multidimensional Schema** is especially designed to model data warehouse systems. The schemas are designed to address the unique needs of very large databases designed for the analytical purpose (OLAP).

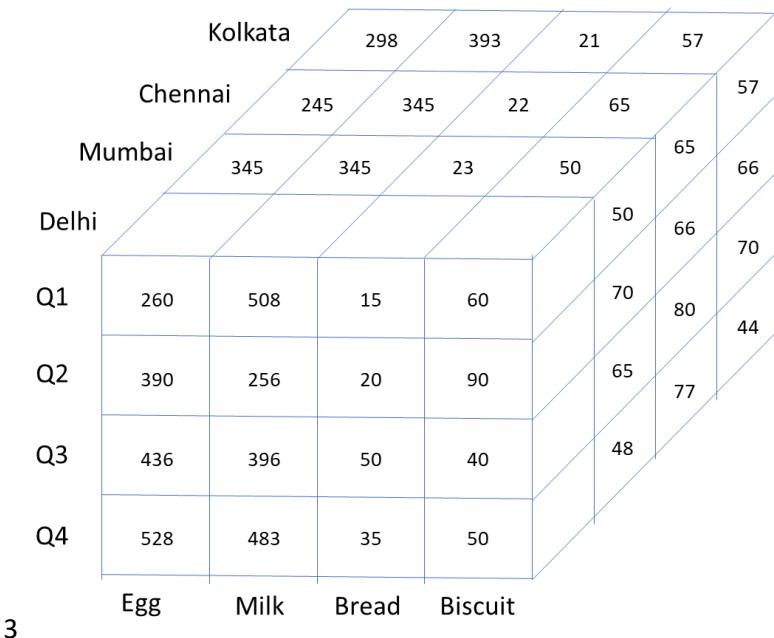
For example, consider the data of a shop for items sold per quarter in the city of Delhi. The data is shown in the table. In this 2D representation, the sales for Delhi are shown for the time dimension (organized in quarters) and the item dimension (classified according to the types of an item sold). The fact or measure displayed in rupee\_sold (in thousands).

Location = "Delhi"				
	Item (Type)			
Quarter	Egg	Milk	Bred	Biscuit
Q1	260	508	15	60
Q2	390	256	20	90
Q3	436	396	50	40
Q4	528	483	35	50

Now, if we want to view the sales data with a third dimension, For example, suppose the data according to time and item, as well as the location is considered for the cities Chennai, Kolkata, Mumbai, and Delhi. These 3D data are shown in the table. The 3D data of the table are represented as a series of 2D tables.

Location ="Delhi"					Location ="Mumbai"					Location ="Chennai"					Location ="Kolkata"				
	Item (Type)				Item (Type)				Item (Type)				Item (Type)						
Quarter	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit			
Q1	260	508	15	60	345	345	23	50	245	345	22	65	298	393	21	57			
Q2	390	256	20	90	234	243	31	70	333	325	27	66	303	367	18	66			
Q3	436	396	50	40	342	346	24	65	348	300	25	80	400	350	27	70			
Q4	528	483	35	50	345	432	24	48	380	321	30	77	377	365	30	44			

Conceptually, it may also be represented by the same data in the form of a 3D data cube, as shown in fig:



3

## 2.6 Types of Data Warehouse Schema:

Following are 3 chief types of multidimensional schemas each having its unique advantages.

- Star Schema
- Snowflake Schema
- Galaxy Schema

### 2.6.1 Star Schema

**Star Schema** in data warehouse, in which the center of the star can have one fact table and a number of associated dimension tables. It is known as star schema as its structure resembles a star. The Star Schema data model is the simplest type of Data Warehouse schema. It is also known as Star Join Schema and is optimized for querying large data sets.

In the following Star Schema example, the fact table is at the center which contains keys to every dimension table like Dealer\_ID, Model ID, Date\_ID, Product\_ID, Branch\_ID & other attributes like Units sold and revenue.

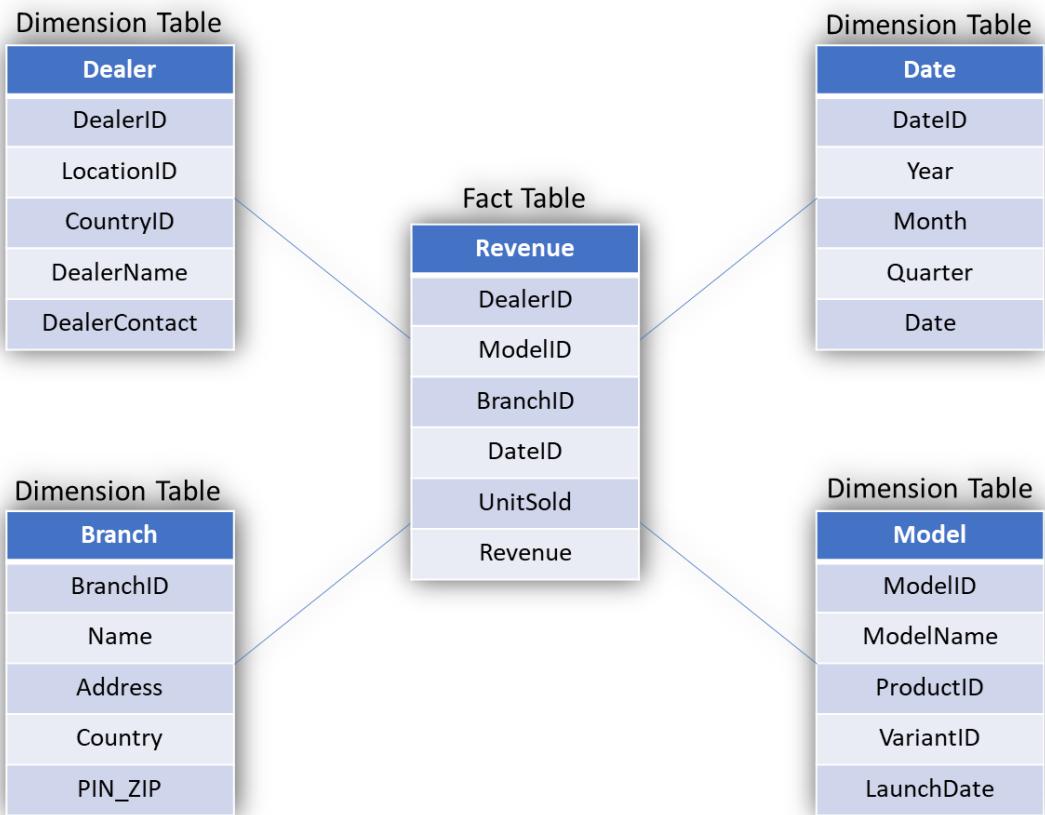


Figure 3 - Star Schema

#### Characteristics of Star Schema:

- Every dimension in a star schema is represented with the only one-dimension table.
- The dimension table should contain the set of attributes.
- The dimension table is joined to the fact table using a foreign key
- The dimension table are not joined to each other
- Fact table would contain key and measure
- The Star schema is easy to understand and provides optimal disk usage.
- The dimension tables are not normalized. For instance, in the above figure, Country\_ID does not have Country lookup table as an OLTP design would have.
- The schema is widely supported by BI Tools

#### 2.6.2 Snowflake Schema

**Snowflake Schema** in data warehouse is a logical arrangement of tables in a multidimensional database such that the ER diagram resembles a snowflake shape. A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. The dimension tables are normalized which splits data into additional tables.

In the following Snowflake Schema example, Country is further normalized into an individual table.

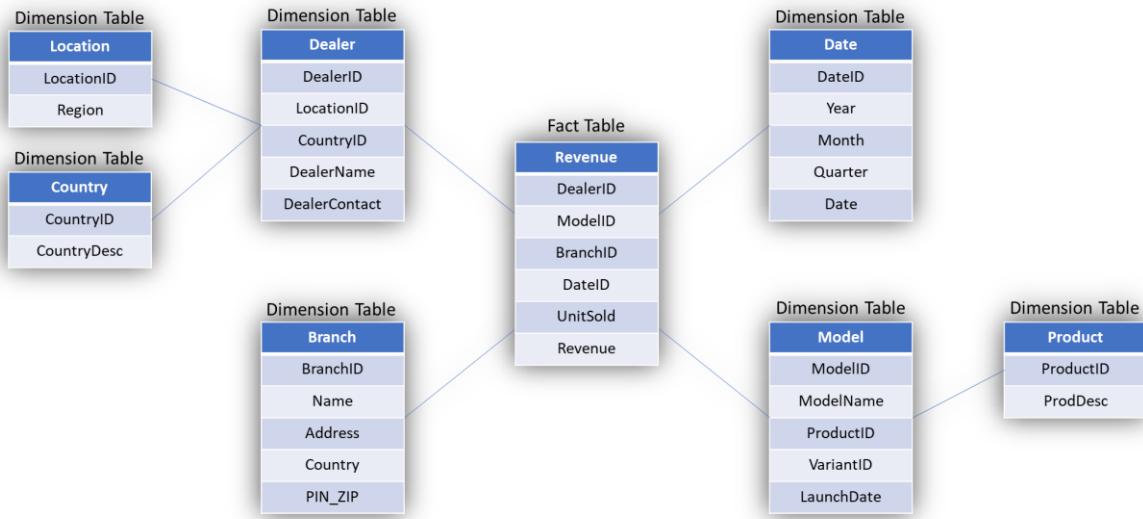


Figure 4 - Snowflake Schema

#### Characteristics of Snowflake Schema:

- The main benefit of the snowflake schema it uses smaller disk space.
- Easier to implement a dimension is added to the Schema
- Due to multiple tables query performance is reduced
- The primary challenge that you will face while using the snowflake Schema is that you need to perform more maintenance efforts because of the more lookup tables.

#### 2.6.3 Galaxy Schema

A **Galaxy Schema** contains two fact table that share dimension tables between them. It is also called Fact Constellation Schema. The schema is viewed as a collection of stars hence the name Galaxy Schema.

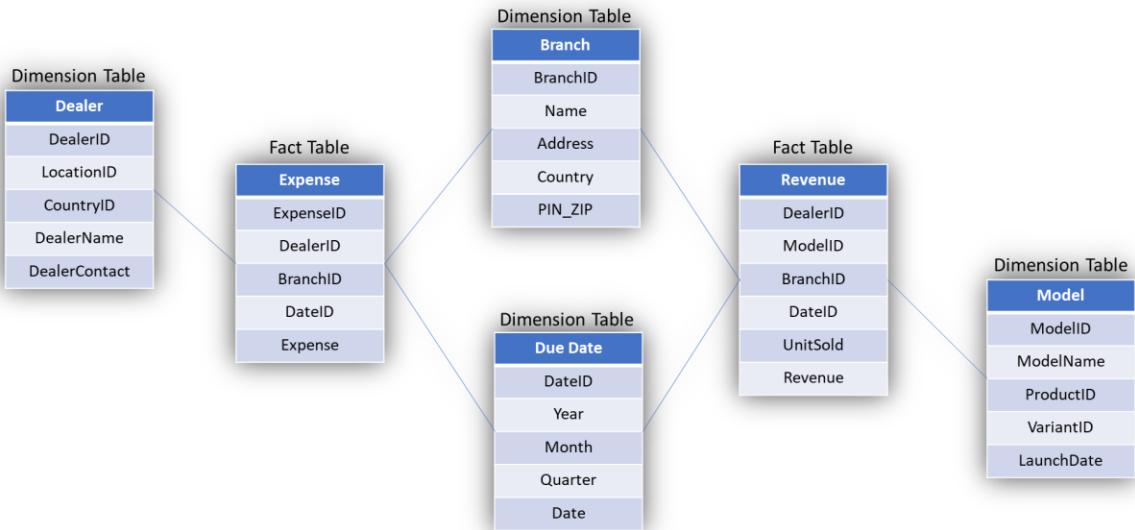


Figure 5 - Galaxy Schema

As you can see in above example, there are two facts table

1. Expense
2. Revenue.

In Galaxy schema shared dimensions are called Conformed Dimensions.

#### Characteristics of Galaxy Schema:

- The dimensions in this schema are separated into separate dimensions based on the various levels of hierarchy. For example, if geography has four levels of hierarchy like region, country, state, and city then Galaxy schema should have four dimensions.
- Moreover, it is possible to build this type of schema by splitting the one-star schema into more Star schemes.
- The dimensions are large in this schema which is needed to build based on the levels of hierarchy.
- This schema is helpful for aggregating fact tables for better understanding.

#### 2.6.4 Star Cluster Schema

Snowflake schema contains fully expanded hierarchies. However, this can add complexity to the Schema and requires extra joins. On the other hand, star schema contains fully collapsed hierarchies, which may lead to redundancy. So, the best solution may be a balance between these two schemas which is Star Cluster Schema design.

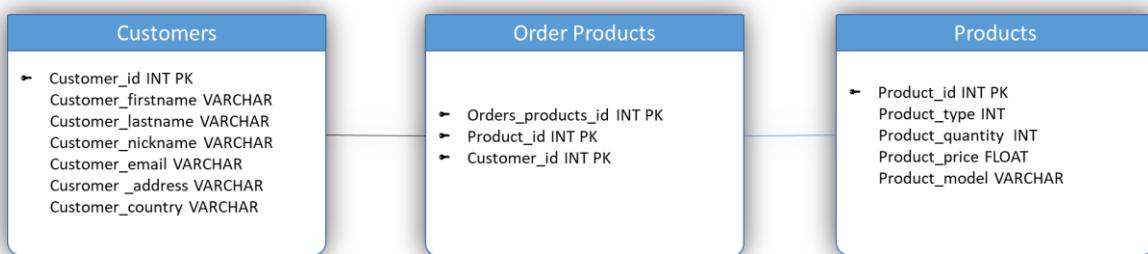


Figure 6 - Star Cluster Schema

Overlapping dimensions can be found as forks in hierarchies. A fork happens when an entity acts as a parent in two different dimensional hierarchies. Fork entities then identified as classification with one-to-many relationships.

## 2.7 Star Schema Vs Snowflake Schema: Key Differences

Following is a key difference between Star Schema and Snowflake Schema:

Star Schema	Snowflake Schema
Hierarchies for the dimensions are stored in the dimensional table.	Hierarchies are divided into separate tables.
It contains a fact table surrounded by dimension tables.	One fact table surrounded by dimension table which are in turn surrounded by dimension table
In a star schema, only single join creates the relationship between the fact table and any dimension tables.	A snowflake schema requires many joins to fetch the data.
Simple DB Design.	Very Complex DB Design.
Denormalized Data structure and query also run faster.	Normalized Data Structure.
High level of Data redundancy	Very low-level data redundancy
Single Dimension table contains aggregated data.	Data Split into different Dimension Tables.
Cube processing is faster.	Cube processing might be slow because of the complex join.
Offers higher performing queries using Star Join Query Optimization. Tables may be connected with multiple dimensions.	The Snowflake schema is represented by centralized fact table which unlikely connected with multiple dime

## 2.8 Summary

- Multidimensional schema is especially designed to model data warehouse systems
- The star schema is the simplest type of Data Warehouse schema. It is known as star schema as its structure resembles a star.
- A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. It is called snowflake because its diagram resembles a Snowflake.
- In a star schema, only single join defines the relationship between the fact table and any dimension tables.
- Star schema contains a fact table surrounded by dimension tables.
- Snowflake schema is surrounded by dimension table which are in turn surrounded by dimension table
- A snowflake schema requires many joins to fetch the data.
- A Galaxy Schema contains two fact table that shares dimension tables. It is also called Fact Constellation Schema.
- Star cluster schema contains attributes of Star and Snowflake Schema.

## **UNIT 2 : CHAPTER 1**

### **Data Warehouse**

#### **Unit Structure**

##### **1.0 Objectives:**

###### **1.1 Introduction to Data Warehouse**

###### **1.2 Evolution of Data Warehouse**

###### **1.3 Benefits of Data Warehouse**

###### **1.4 Data Warehouse Architecture**

###### **1.4.1 Basic Single-Tier Architecture**

###### **1.4.2 Two-Tier Architecture**

###### **1.4.3 Three-Tier Architecture**

###### **1.5 Properties of Data Warehouse Architectures**

###### **1.6 ETL Process in Data Warehouse**

###### **1.7 Cloud-based ETL Tools vs. Open Source ETL Tools**

###### **1.8 ETL and OLAP Data Warehouses**

### 1.8.1 The Technical Aspects of ETL

## 1.9 Data Warehouse Design Approaches

### 1.9.1 Bill Inmon – Top-down Data Warehouse Design Approach

### 1.9.2 Ralph Kimball – Bottom-up Data Warehouse Design Approach

## 1.10 Data Mart

### 1.10.1 Reasons for creating a data mart

## 1.11 Types of Data Marts

### 1.11.1 Dependent Data Marts

### 1.11.2 Independent Data Marts

### 1.11.3 Hybrid Data Marts

## 1.12 Characteristics of Data Mart

## 1.13 Summary

## 1.14 References for further reading

## **1.0 Objectives:**

This chapter will make the readers understand the following concepts:

- Meaning of data warehouse
- Concept behind Data Warehouse
- History and Evolution of Data Warehouse
- Different types of Data Warehouse Architectures
- Properties of data warehouse
- Concept of Data Staging
- ETL process
- Design approaches to Data Warehouse
- Data Marts and their types

## **1.1 Introduction to Data Warehouse**

As organizations grow, they usually have multiple data sources that store different kinds of information. However, for reporting purposes, the organization needs to have a single view of

the data from these different sources. This is where the role of a Data Warehouse comes in. **A Data Warehouse helps to connect and analyse data that is stored in various heterogeneous sources.** The process by which this data is collected, processed, loaded, and analysed to derive business insights is called Data Warehousing.

The data that is present within various sources in the organization can provide meaningful insights to the business users if analysed in a proper way and can assist in making data as a strategic tool leading to improvement of processes. Most of the databases that are attached to the source systems are transactional in nature. This means that these databases are used typically for storing transactional data and running operational reports on it. The data is not organized in a way where it can provide strategic insights. A data warehouse is designed for generating insights from the data and hence, helps to convert data into meaningful information that can make a difference.

Data from various operational source systems is loaded onto the Data Warehouse and is therefore a central repository of data from various sources that can provide cross functional intelligence based on historic data. Since the Data Warehouse is separated from the operational databases, it removes the dependency of working with transactional data for intelligent business decisions.

While the primary function of the Data Warehouse is to store data for running intelligent analytics on the same, it can also be used as a central repository where historic data from various sources is stored.

In order to be able to provide actionable intelligence to the end users, it is important that the Data Warehouse consists of information from different sources that can be analysed as one for deriving business intelligence for the organization as a whole. For example, in case of an insurance company, to be able to find out the customers who have more propensity to provide a fraud claim, the insurance company must be able to analyse data from the various sources like the policy system, claims systems, CRM systems, etc.

In most cases, the data from these disparate systems is stored in different ways and hence cannot be taken as it is and loaded onto the data warehouse. Also, the purpose for which a data warehouse is built is different from the one for which the source system was built. In the case of our insurance company above, the policy system was built to store information with

regards to the policies that are held by a customer. The CRM system would have been designed to store the customer information and the claims system was built to store information related to all the claims made by the customers over the years. For us to be able to determine which customers could potentially provide fraud claims, we need to be able to cross reference information from all these source systems and then make intelligent decisions based on the historic data.

Hence, the data has to come from various sources and has to be stored in a way that makes it easy for the organization to run business intelligence tools over it. There is a specific process to extract data from various source systems, translate this into the format that can be uploaded onto the data warehouse and then load the data on the data warehouse. This process for extraction, translation and loading of data is explained in detail subsequently in the chapter.

Besides the process of ensuring availability of the data in the right format on the data warehouse, it is also important to have the right business intelligence tools in place to be able to mine data and then make intelligent predictions based on this data. This is done with the help of business intelligence and data visualization tools that enable converting data into meaningful information and then display this information in a way that is easy for the end users to understand.

With the improvement in technology and the advent of new tools, an enormous amount of data is being collected from various sources. This could be data collected from social media sites where every click of the user is recorded for further analysis. Such enormous amount of data creates a big data situation that is even more complex to store and analyse. Specialised tools are required to analyse such amounts of data.

The kind of analysis that is done on the data can vary from high level aggregated dashboards that provide a cockpit view to a more detailed analysis that can provide as much drill down of information as possible. Hence, it is important to ensure that the design of the data warehouse takes into consideration the various uses of the data and the amount of granularity that is needed for making business decisions.

Most times, the kind of analysis that is done using the data that is stored in the data warehouse is time-related. This could mean trends around sales numbers, inventory holding, profit from products or specific segments of customers, etc. These trends can then be utilized

to forecast the future with the use of predictive tools and algorithms. The Data Ware house provides the basic infrastructure and data that is need by such tools to be able to help the end-users in their quest for information.

In order to understand Data Warehouse and the related concepts in more details, it is important for us to understand a few more related terms:

1. An Operational Data Store (ODS)
2. Data Marts
3. Data Lakes

### **Operational Data Store (ODS)**

As the name suggests, an Operational Data Store or ODS is primarily meant to store data that near current operational data from various systems. The advantage of such a data store is that it allows querying of data which is more real-time as compared to a data warehouse. However, the disadvantage is that the data cannot be used to do complex and more time-consuming queries that can be run on a data warehouse. This is because the data on the operational data store has not yet gone through the process of transformation and is not structured for the purpose of complex queries. It provides a way to query data without having to burden the actual transactional system.

### **Data Marts**

Data marts are like a mini data warehouse consisting of data that is more homogenous in nature rather than a varied and heterogeneous nature of a data warehouse. Data marts are typically built for the use within an department or business unit level rather than at the overall organizational level. It could aggregate data from various systems within the same department or business unit. Hence, data marts are typically smaller in size than data warehouses.

### **Data Lakes**

A concept that has emerged more recently is the concept of data lakes that store data in a raw format as opposed to a more structures format in the case of a data warehouse. Typically, a data lake will not need much transformation of data without loading onto the data lake. It is

generally used to store bulk data like social media feeds, clicks, etc. One of the reasons as to why the data is not usually transformed before loading onto a data lake is because it is not usually known what kind of analysis would be carried out on the data. More often than not, a data scientist would be required to make sense of the data and to derive meaningful information by applying various models on the data.

*Table 1 - Data Mart v/s Data Lake v/s Data Warehouse*

<b>Data Store</b>	<b>Primary Use</b>	<b>Amount of Data and Cost of Setup</b>
<i>Data Mart</i>	Meant for use within a business unit or function	Lesser than Data Warehouse and Data Lake
<i>Data Warehouse</i>	Meant for use at organizational level across business units	More than Data Mart but less than Data Lake
<i>Data Lake</i>	Meant for advanced and predictive analytics	Greater than Data Mart and Data Warehouse

Some of the other names of a Data Warehouse system are Decision Support System, Management Information System, Business Intelligence System or Executive Information System.

## **1.2 Evolution of Data Warehouse**

As the information systems within the organizations grew more and more complex and evolved over time, the systems started to develop and handle more and more amount of information. The need for an ability to analyze the data coming out from the various systems became more evident over time.

The initial concept of a “Business Data Warehouse” was developed by IBM researchers Barry Devlin and Paul Murphy in late 1980s. It was intended to provide an architectural model as to how the data would flow from an operational system to an environment that could support decision making for the business. The evolution of Data Warehouse can be traced back to the 1960 when Dartmouth and General Mills developed the terms like dimension and facts in a joint research paper. In 1970, A. Nielsen and IRI used this concept to introduce the dimensional data marts for retail sales. It was much later in the year 1983, that Tera Data Corporation introduced a Database Management System that was designed specifically for the decision support process.

Later, in the late 1980s, IBM researchers developed the Business Data Warehouse. Inmon Bill was considered as a father of data warehouse. He had written about a variety of topics for building, usage, and maintenance of the warehouse & the Corporate Information Factory.

### **1.3 Benefits of Data Warehouse**

There are numerous benefits that a data warehouse can provide organizations. Some of these benefits are listed below:

1. Enhancing business intelligence within the organization:

A data warehouse is able to bring data from various source systems into a single platform. This allows the users to make better business decisions that are based on data which cuts across different system and can provide an integrated view rather than an isolated view of the data.

This is made possible since the data has been extracted, translated, and then loaded onto the data warehouse platform from various cross-functional and cross-departmental source systems. Information that provides such an integrated view of the data is extremely useful for the senior management in making decisions at the organizational level.

2. Right information at the right time

Given the ability of the Data Warehouse to be able to provide information requested on demand, it is able to provide the right information to the organizational users at the time when it is required the most. Time is usually of essence when it comes to business decisions. Organizations not only need to spend valuable time and effort in collating information from various sources. Manual collation of such data not only takes time but is also error prone and cannot be completely trusted.

A Data warehouse platform can take care of all such issues since the data is already loaded and can be queried upon as desired. Thereby saving precious time and effort for the organizational users.

3. Improving the quality of data

A data warehouse platform consists of data that is extracted from various systems and has been translated to the required format for the data warehouse. Second disk significantly

improves the quality of the data second thereby increases the quality of decisions that are made based on such data.

Given that the data in the data warehouse is usually automatically uploaded, the chances of errors to creep into the process are quite minimal. This is not a manual process which is prone to errors.

#### 4. Return on investment

Building a data warehouse is usually an upfront cost for the organization. However, the return that it provides in terms of information and the ability to make right decisions at the right time provides a return on investment that is usually manyfold with respect to the amount that has been invested upfront. In the long run, a data warehouse helps the organization in multiple ways to generate new revenue and save costs.

#### 5. Competitive edge

A data warehouse is able to provide the top management within the organization a capability to make business decisions that are based on data that cuts across the organizational silos. It is therefore more reliable and the decisions that are based on such data are able to provide a competitive edge to the organization viz-a-viz their competition

#### 6. Better decision-making process

Use of a data warehouse that provides the capability to integrate information from various systems across the organization can lead to better decision-making process within the organization. The senior management will have an integrated view of information coming from video source systems and therefore will be able to make decisions that are not limited by a siloed view.

#### 7. Predict with more confidence

The data that is stored within the data warehouse provides better quality and consistency than any manual process. This can give more confidence to the users that any predictions that are driven from this data warehouse would be more accurate and can be trusted with more confidence than manual processes.

#### 8. Streamlined data flow within the organization

A data warehouse is able to integrate data from multiple sources within the organization and therefore streamlines and provides a consistent view of data that is stored in various systems – bringing them into a single repository.

#### **1.4 Data Warehouse Architecture**

As we seek to understand what the data warehouse is, it is important for us to understand the different types of deployment architectures by way of which a data warehouse can be implemented within an organization. Every data warehouse implementation is different from each other However, there are certain elements that can common between all of them.

The data warehouse architecture defines the way in which information is processed, transformed, loaded and then presented to the end users for the purpose of generating business insights. In order to understand the data warehouse architecture, we need to understand the some of the terminologies associated with it.

Day-to-day operations of an organization are typically run by production systems such as payroll, HR, finance, etc. that generate data and transactions on a daily basis and are usually called Online Transaction Processing (OLTP) systems. Such applications are usually the sources of data for a data warehousing platform. On the other hand, a data warehouse is primarily designed to support analytical capabilities on top of data that comes from all of these various source systems and is therefore termed as an Online Analytical Processing (OLAP) system. The online analytical processing system provides users with the capability to produce ad hoc reports as required and on demand.

As can be seen that the online transaction processing systems are usually updated regularly based on the data and transactions that happen daily on that system. In contrast, an online analytical processing system or the data warehouse is usually updated through an ETL process that extracts the data from the source systems on a regular basis, transforms the data into a format that will be required for the data warehouse and then loads the data onto the data warehouse as per the pre-defined processes.

It may be noticed that the data in the data warehouse is typically not real time data and there is usually a delay in moving the data from these source systems to the data warehouse . However, this is something that most businesses are fine with as long as they get an

integrated view of data from across different functions of the organization and as long as the data is automatically uploaded on the data warehouse for generation of these insights on demand.

A data warehouse architecture may be implemented in many different ways . Some of the common ways of implementing the data warehouse architecture are listed below.

- Basic architecture for a Data Warehouse or a single tier architecture
- Staging-area based architecture for a Data Warehouse or a two tier architecture
- Staging area and data-mart based architecture for a Data Warehouse or a three-tier architecture

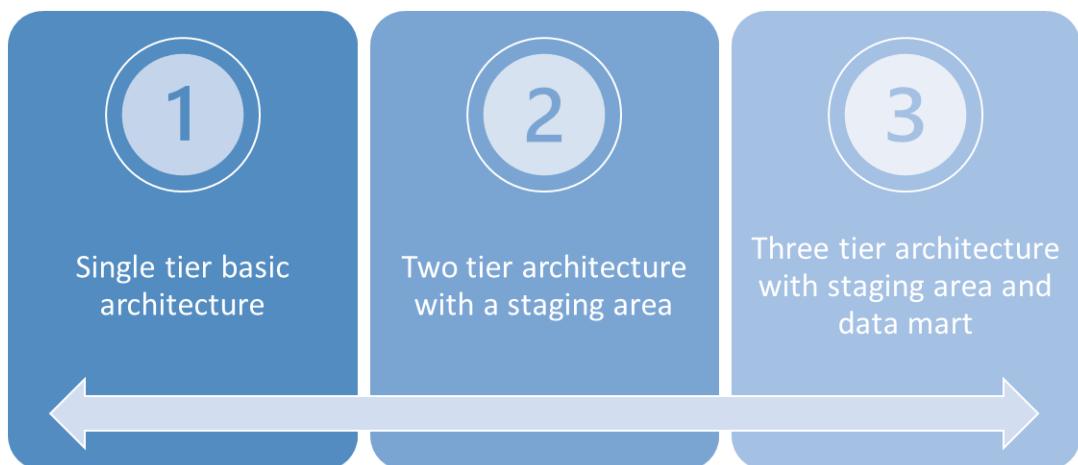


Figure 7 - Various Implementation Architectures for Data Warehouse

#### 1.4.1 Basic Single-Tier Architecture

This type of architecture is not used much but it does provide a good idea of how basic data warehouse can be implemented. It aims to remove data redundancy. In this basic architecture, the only physical layer available is the source systems.

This means that the data warehouse is implemented as a multidimensional view of operational data created by specific middleware, or an intermediate processing layer.

The figure below shows the implementation of a basic data warehouse architecture which has the sources systems abstracted by a middleware that aims to remove the provide a separation between transaction and analytical capabilities.

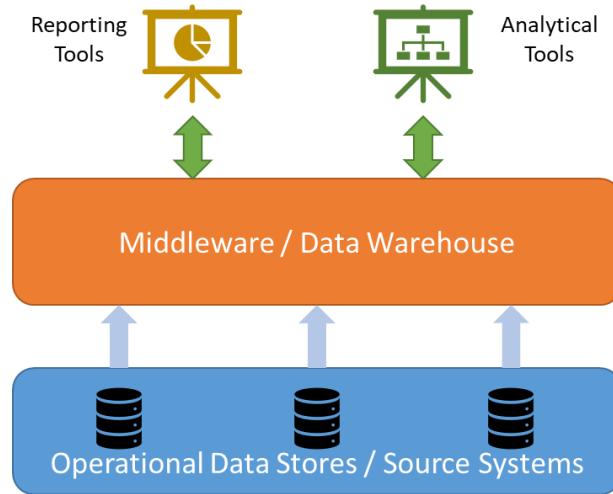


Figure 8 - Basic Data Warehouse Architecture

#### 1.4.2 Two-Tier Architecture

The need for separation plays a crucial role in defining the two-tier architecture for a data warehouse system, as shown in the figure below:

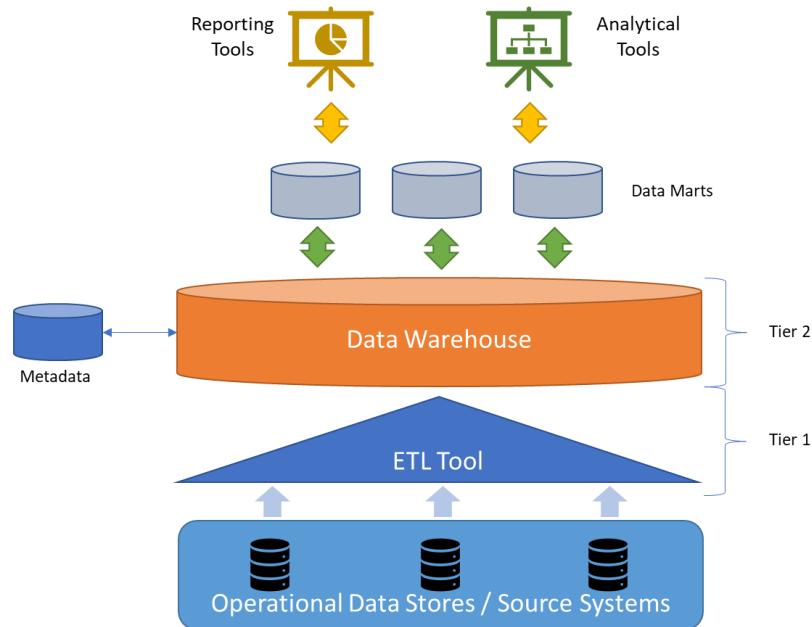


Figure 9 - Two Tier architecture for Data Warehouse

The two the two-layer architectures highlights a separation between physically available resources and data warehouse thus it is divided into four different four different stages which are according to the dataflow. these different stages are mentioned as below.

1. **Source Layer:** as discussed earlier the data warehouse uses heterogeneous sources of data the data which is initially stored in a corporate relational databases or legacy databases or it may come from any source within the organization or outside the organization.
2. **Data Staging:** The data which we are going to store should be extracted, cleans to remove any inconsistencies integrity to merge heterogeneous sources into one standard schema. Thus extraction, transformation, loading tools ETL can combine heterogeneous schema by extracting, cleaning, transforming, validating and load data into data warehouse.
3. **Data Warehouse layer:** Information is saved to one logically centralized individual repository: a data warehouse. The data warehouses can be directly accessed, but it can also be used as a source for creating data marts, which partially replicate data warehouse contents and are designed for specific enterprise departments. Meta-data repositories store information on sources, access procedures, data staging, users, data mart schema, and so on.
4. **Analysis:** In this layer, integrated data is efficiently, and flexible accessed to issue reports, analyse information, and simulate business scenarios. It should feature information navigators, complex query optimizers, and customer-friendly GUIs.

#### 1.4.3 Three-Tier Architecture

The three-tier architecture consists of the source layer (containing multiple source system), the reconciled layer and the data warehouse layer (containing both data warehouses and data marts).

The reconciled layer is between the source data and data warehouse. It creates a standard reference model for the whole enterprise. And, at the same time it separates the problem of data extraction and integration from data warehouse. This layer is also directly used to perform better operational tasks e.g. producing daily reports or generating data flows periodically to benefit from cleaning and integration.

While this architecture is useful for extensive global enterprise systems, the major disadvantage is the use of extra file storage space because of redundant reconciled layer that makes the analytical tools little further away from being real time.

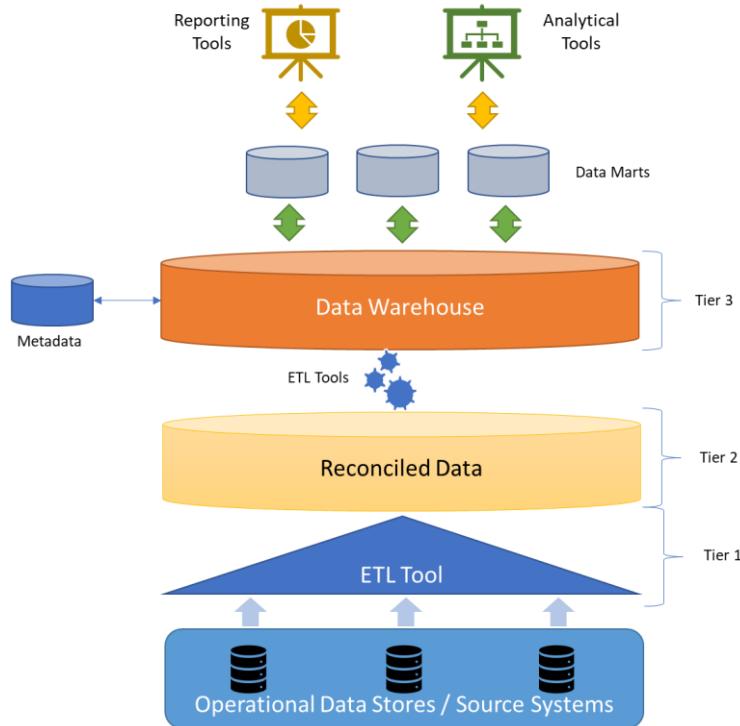


Figure 10 - Three Tier architecture for Data Warehouse

## 1.5 Properties of Data Warehouse Architectures

The following architecture properties are necessary for a data warehouse system:

- 1. Separation:** There should be separation between analytical and transactional processing as much as possible.
- 2. Scalability:** To upgrade the data volumes which has to be managed and processed and number of user requirements which have to be met we need hardware and software architectures that should be simple to upgrade.
- 3. Extensibility:** The architecture should be able to perform new operations and technologies without redesigning the whole system.
- 4. Security:** Security plays a very important role in information technology .Monitoring accesses providing passwords are necessary because of the strategic data stored in the data warehouses.
- 5. Administrability:** Data Warehouse management should not be complicated.



*Figure 11 – Properties of Data Warehouse Architecture*

## 1.6 ETL Process in Data Warehouse

ETL (or Extract, Transform, Load) is a process of data integration that encompasses three steps — extraction, transformation, and loading. In a nutshell, ETL systems take large volumes of raw data from multiple sources, converts it for analysis, and loads that data into your warehouse

It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area and then finally, loads it into the Data Warehouse system.

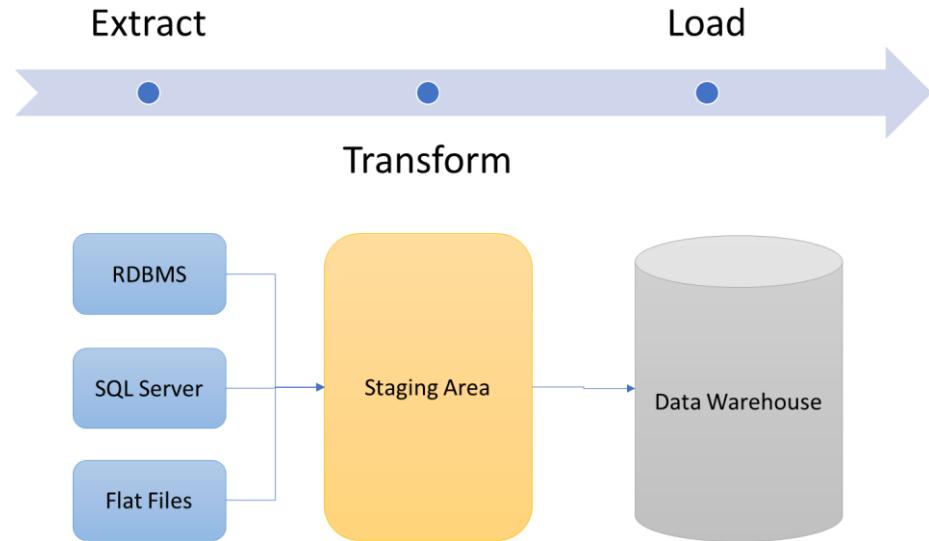


Figure 12 - High Level ETL Process flow

### **Extraction:**

In this step data is extracted from various sources into a staging area. This area acts as a buffer between the data warehouse and source systems. As we know the data comes from various sources, hence the data will be in different formats and we cannot directly transfer this data into data warehouse. The staging area is used by companies for data cleaning.

A major challenge during this extraction process is how ETL tool differentiates structured and unstructured data. All unstructured items such as emails web pages etc can be difficult to extract without the right tool. It is important to extract the data from various source systems and store it into staging area first and not indirectly into data warehouse because of their various formats. It is, therefore, one of the major steps of ETL process.

### **Transformation:**

The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format. All the data from multiple source systems is normalized and converted to a single system format — improving data quality and compliance. ETL yields transformed data through these methods:

- Filtering – loading only certain attributes into the data warehouse.
- Cleaning – filling up the NULL values with some default values, mapping U.S.A, United States and America into USA, etc.
- Joining – joining multiple attributes into one.
- Splitting – splitting a single attribute into multiple attributes.

- Sorting – sorting tuples on the basis of some attribute (generally key-attribute).

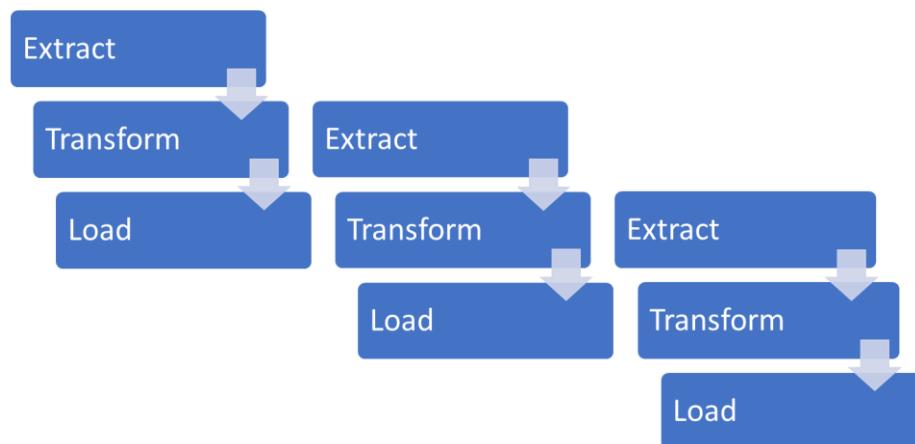
### **Loading:**

The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse. Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals. The rate and period of loading solely depends on the requirements and varies from system to system.

ETL process can also use the pipelining concept i.e. as soon as some data is extracted, it can be transformed and during that period some new data can be extracted. And while the transformed data is being loaded into the data warehouse, the already extracted data can be transformed.

Finally, data that has been extracted to a staging area and transformed is loaded into your data warehouse. Depending upon your business needs, data can be loaded in batches or all at once. The exact nature of the loading will depend upon the data source, ETL tools, and various other factors.

The block diagram of the pipelining of ETL process is shown below:



*Figure 13 - ETL Pipeline View*

**ETL Tools:** Most used ETL tools are

- Sybase
- Oracle Warehouse builder,
- CloverETL
- MarkLogic.

## **1.7 Cloud-based ETL Tools vs. Open Source ETL Tools**

ETL is a critical component off overall data warehouse architecture so choosing the right one is very crucial there are various different options available and one can choose depending upon the overall ETL needs, data schemas Ant operational structure

Cloud-based ETL tools like Xplenty offer rapid, real-time streaming, quick integrations, and easy pipeline creation. The primary benefit of cloud-based ETL tools is that they work immediately out-of-the-box. Plus, they're hyper-useful for a variety of ETL needs, especially if most of your warehouse exists in the cloud (i.e., Redshift, Snowflake, or Big Query).

Open source ETL tools come in a variety of shapes and sizes. There are ETL frameworks and libraries that you can use to build ETL pipelines in Python. There are tools and frameworks you can leverage for GO and Hadoop. Really, there is an open-source ETL tool out there for almost any unique ETL need.

## **1.8 ETL and OLAP Data Warehouses**

Data engineers have been using ETL for over two decades to integrate diverse types of data into online analytical processing (OLAP) data warehouses. The reason for doing this is simple: to make data analysis easier.

Normally, business applications use online transactional processing (OLTP) database systems. These are optimized for writing, updating, and editing the information inside them. They're not good at reading and analysis. However, online analytical processing database systems are excellent at high-speed reading and analysis. That's why ETL is necessary to transform OLTP information, so it can work with an OLAP data warehouse.

During the ETL process, information is:

- Extracted from various relational database systems (OLTP or RDBMS) and other sources.
- Transformed within a staging area, into a compatible relational format, and integrated with other data sources.
- Loaded into the online analytical processing (OLAP) data warehouse server.

### 1.8.1 The Technical Aspects of ETL

It's important to pay close attention to the following when designing your ETL and ELT processes:

- **Ensure accurate logging:** It's vital to make sure your data system provides "accurate logging" of new information. To ensure accurate logging, you'll need to audit data after loading to check for lost or corrupt files. With proper auditing procedures, you can debug your ETL/ELT process when data integrity challenges arise (as they invariably do).
- **Flexibility to work with diverse sources of structured and unstructured data:** Your data warehouse may need to integrate information from a lot of incompatible sources like PostgreSQL, Salesforce, Cassandra, and in-house financial applications. Some of this information could lack the data structures required for analysis. You need to design your ETL/ELT process to deal with all forms of data—structured and unstructured alike.
- **Stability and reliability:** ETL/ELT pipelines get overloaded, crash, and run into problems. Your goal should be to build a fault-tolerant system that can recover after a shutdown so your data can move without getting lost or corrupted even in the face of unexpected issues.
- **Designing an alert system:** To ensure the accuracy of your business insights, an alert system that notifies you of potential problems with the ETL/ELT process is essential. For example, you'll want to receive notifications and reports for expired API credentials, bugs related to third-party APIs, connector errors, general database errors, and more.
- **Strategies to speed up the flow of data:** When data warehouses and BI platforms have access to information that is up-to-date, they offer better, more accurate insights at a moment's notice. Therefore, it's important to focus on reducing data latency, i.e., the time it takes for a data packet to move from one area of the system to the next.
- **Growth flexibility:** Your ETL/ELT solution should be flexible to scale up and down according to your organization's changing data needs. This will save money on cloud-server processing and storage fees, while providing the ability to scale up as required.

- **Support for incremental loading:** Using change data capture (CDC) speeds up the ETL process by permitting incremental loading. This lets you update only a small part of your data warehouse while ensuring data synchronicity.

## 1.9 Data Warehouse Design Approaches

Very important aspect of building data warehouses is the design of data warehouse. Selection of right data Warehouse saves lot of time, efforts, and project cost.

The two different approaches are normally followed when designing a data warehouse solution and based on the requirement of the project we can choose one that suits the particular scenario.

These methodologies are a result of research from Bill Inmon and Ralph Kimball.

### 1.9.1 Bill Inmon – Top-down Data Warehouse Design Approach

“Bill Inmon” is sometimes also referred to as the “father of data warehousing”; his design methodology is based on a top-down approach. In the top-down approach, the data warehouse is designed first and then data mart are built on top of data warehouse.

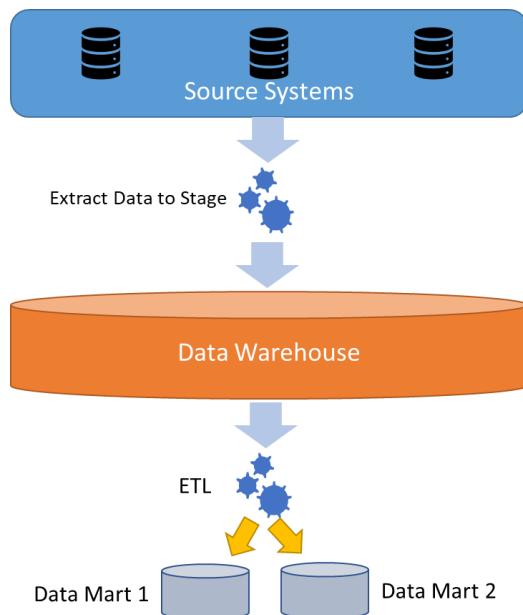


Figure 14 - Top Down Approach

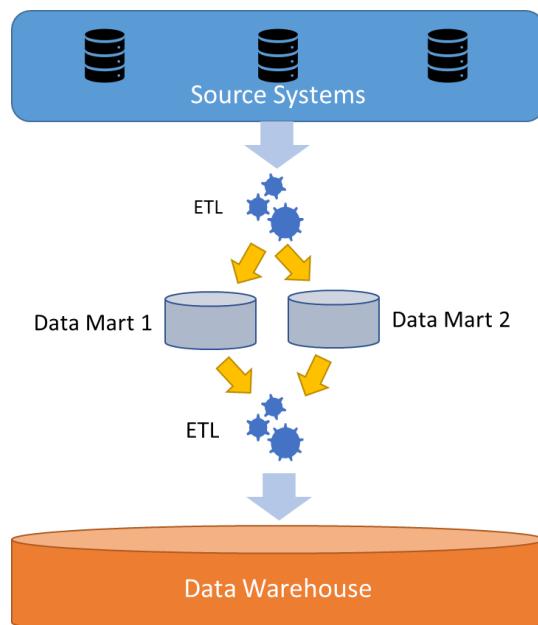
Below are the steps that are involved in top-down approach:

- Data is extracted from the various source systems. The extracts are loaded and validated in the stage area. Validation is required to make sure the extracted data is accurate and correct. You can use the ETL tools or approach to extract and push to the data warehouse.
- Data is extracted from the data warehouse in regular basis in stage area. At this step, you will apply various aggregation, summarization techniques on extracted data and loaded back to the data warehouse.
- Once the aggregation and summarization is completed, various data marts extract that data and apply the some more transformation to make the data structure as defined by the data marts.

#### 1.9.2 Ralph Kimball – Bottom-up Data Warehouse Design Approach

Ralph Kimball is a renowned author on the subject of data warehousing. His data warehouse design approach is called dimensional modelling or the Kimball methodology. This methodology follows the bottom-up approach. As per this method, data marts are first created to provide the reporting and analytics capability for specific business process, later with these data marts enterprise data warehouse is created.

Basically, Kimball model reverses the Inmon model i.e. Data marts are directly loaded with the data from the source systems and then ETL process is used to load in to Data Warehouse.



*Figure 15 - Bottom up Approach*

Below are the steps that are involved in bottom-up approach:

- The data flow in the bottom-up approach starts from extraction of data from various source system into the stage area where it is processed and loaded into the data marts that are handling specific business process.
- After data marts are refreshed the current data is once again extracted in stage area and transformations are applied to create data into the data mart structure. The data is then extracted from Data Mart to the staging area is aggregated, summarized and so on loaded into EDW and then made available for the end user for analysis and enables critical business decisions.

## 1.10 Data Mart

Data marts is the access layer of a data warehouse that is used to provide users with data. Data warehouses typically house enterprise-wide data, and information stored in a data mart usually belongs to a specific department or team.

The key objective for data marts is to provide the business user with the data that is most relevant, in the shortest possible amount of time. This allows users to develop and follow a project, without needing to wait long periods for queries to complete. Data marts are designed to meet the demands of a specific group and have a comparatively narrow subject area.. Data marts may contain millions of records and require gigabytes of storage.

The fundamental use of a data mart is Business Intelligence (BI) applications. BI is used to gather, store, access, and analyze record. It can be used by smaller businesses to utilize the data they have accumulated since it is less expensive than implementing a data warehouse.

### 1.10.1 Reasons for creating a data mart

- Creates collective data by a group of users
- Easy access to frequently needed data
- Ease of creation
- Improves end-user response time
- Lower cost than implementing a complete data warehouse
- Potential clients are more clearly defined than in a comprehensive data warehouse
- It contains only essential business data and is less cluttered.

## 1.11 Types of Data Marts

There are mainly two approaches to designing data marts. These approaches are

- Dependent Data Marts
- Independent Data Marts

### 1.11.1 Dependent Data Marts

A dependent data mart is a logical subset or a physical subset of a higher data warehouse. According to this technique, the data marts are treated as the subsets of a data warehouse. In this technique, firstly a data warehouse is created from which further various data marts can be created. These data marts are dependent on the data warehouse and extract the essential records from it. In this technique, as the data warehouse creates the data mart; therefore, there is no need for data mart integration. It is also known as a **top-down approach**.

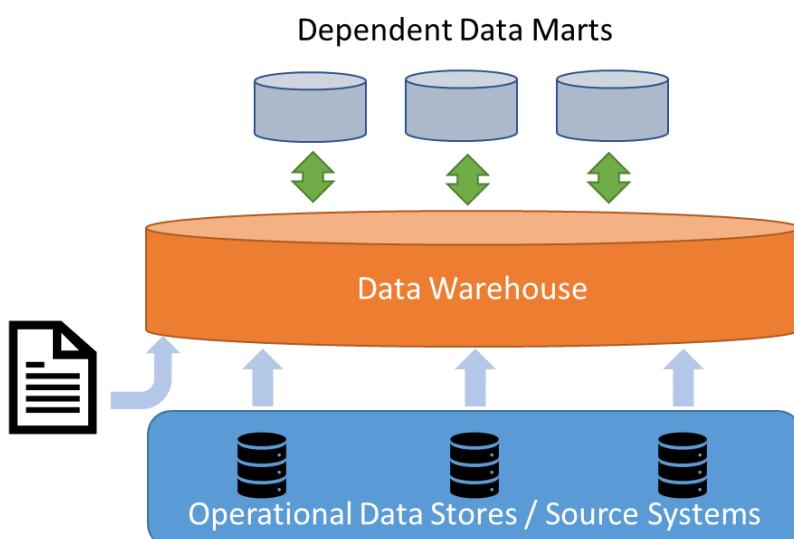
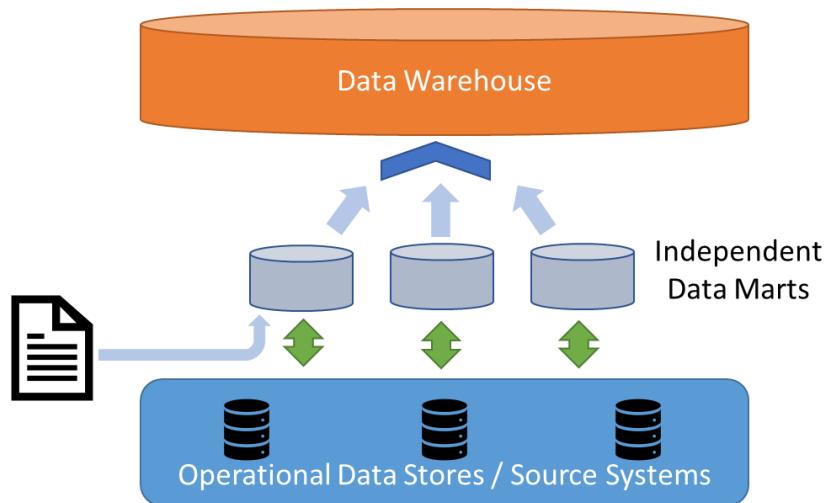


Figure 16 - Dependent Data Mart

### 1.11.2 Independent Data Marts

The second approach is Independent data marts (IDM). Here, firstly independent data marts are created, and then a data warehouse is designed using these independent multiple data marts. In this approach, as all the data marts are designed independently; therefore, the integration of data marts is required. It is also termed as a **bottom-up approach** as the data marts are integrated to develop a data warehouse.



*Figure 17 - Independent Data Warehouse*

Other than these two categories, one more type exists that is called "**Hybrid Data Marts**".

#### 1.11.3 Hybrid Data Marts

It allows us to combine input from sources other than a data warehouse. This could be helpful for many situations; especially when Adhoc integrations are needed, such as after a new group or product is added to the organizations

### 1.12 Characteristics of Data Mart

Below are some of the features of a data mart:

- Since the source of the data is concentrated to subject the user response time is enhanced by using it.
- For frequently required data, using data marts will be beneficial since it is subset to central Data Warehouse and hence data size will be lesser.
- Also, since the volume of the data is limited the processing time will be quite reduced compared to central Data Warehouse.
- These are basically agile and can accommodate the changes in the model quite quickly and efficiently compared to the data warehouse.
- Datamart requires a single subject expert to handle, in contrast to warehouse data, the expertise we require in multiple subject warehouses. Because of this, we say that data mart is more agile.

- We can segregate access categories to a low level with partitioned data and with data mart, it is a lot easy.
- Infrastructure dependency is quite limited, and data can be stored in on different hardware platforms upon segmentation.

### **1.13 Summary**

- A data warehouse is a data base which is kept separate from organizations operational database.
- It possesses consolidated historical data which helps organizations to analyse its business
- Data warehouse helps in consolidated historical data analysis
- An operational database query allows to read and modify operations while an OLAP query needs read only access of stored data
- An operational database second maintains current data while on the other hand a data warehouse maintains historical data
- OLAP systems are used by knowledge workers such as executives, managers, and analysts
- ETL stands for extract, transform and load
- ETL provides a method of moving the data from various sources into data warehouse
- In the first step, extraction, data is extracted from the source system into the staging area.
- In the transformation step, the data extracted from source is cleaned and transformed.
- In the third step, loading, data into the target data warehouse
- A data mart is defined as a subset of data warehouse time is focused on a single functional area of an organization
- Datamart helps to enhance user experience by reducing response time due to reduction in the volume of data
- There are three types of data marts – dependent, independent and hybrid

### **1.14 References for further reading**

Reference books:

1. Ponniah, Paulraj, Data warehousing fundamentals: a comprehensive guide for IT professionals, John Wiley & Sons, 2004.
2. Dunham, Margaret H, Data mining: Introductory and advanced topics, Pearson Education India, 2006.
3. Gupta, Gopal K, Introduction to data mining with case studies, PHI Learning Pvt. Ltd., 2014.
4. Han, Jiawei, Jian Pei, and Micheline Kamber, Data mining: concepts and techniques, Second Edition, Elsevier, Morgan Kaufmann, 2011.
5. Ramakrishnan, Raghu, Johannes Gehrke, and Johannes Gehrke, Database management systems, Vol. 3, McGraw-Hill, 2003
6. Elmasri, Ramez, and Shamkant B. Navathe, Fundamentals of Database Systems, Pearson Education, 2008, (2015)
7. Silberschatz, Abraham, Henry F. Korth, and Shashank Sudarshan, Database system concepts, Vol. 5, McGraw-Hill, 1997.

#### Web References:

1. <https://www.guru99.com/data-mining-vs-datawarehouse.html>
2. [https://www.tutorialspoint.com/dwh/dwh\\_overview](https://www.tutorialspoint.com/dwh/dwh_overview)
3. <https://www.geeksforgeeks.org/>
4. <https://blog.eduonix.com/internet-of-things/web-mining-text-mining-depth-mining-guide>

#### UNIT 2 – Chapter 3

##### OLAP in the Data Warehouse

###### Unit Structure

- 3.0 Objectives
- 3.1 What is OLAP

- 3.2 OLAP Cube
- 3.3 Basic analytical operations of OLAP
  - 3.3.1 Roll-up
  - 3.3.2 Drill-down
  - 3.3.3 Slice
  - 3.3.4 Pivot
- 3.4 Characteristics of OLAP Systems
- 3.5 Benefits of OLAP
  - 3.5.1 Motivations for using OLAP
- 3.6 Types of OLAP Models
  - 3.6.1 Relational OLAP
  - 3.6.2 Multidimensional OLAP (MOLAP) Server
  - 3.6.3 Hybrid OLAP (HOLAP) Server
  - 3.6.4 Other Types
- 3.7 Difference between ROLAP, MOLAP, and HOLAP
- 3.8 Difference Between ROLAP and MOLAP
- 3.9 Summary

## 3.0 Objectives

This chapter will enable the readers to understand the following concepts:

- An overview of what OLAP is
- Meaning of OLAP cubes
- The basic analytical operations of OLAP including Rollup, Drill down, Slide & Dice and Pivot
- Characteristics of an OLAP Systems
- Types of OLAP systems that consist of Relational OLAP, Multi-dimensional OLAP and Hybrid OLAP
- Other types of OLAP systems
- Advantages and disadvantages of each of the OLAP systems
- Differences between the three major OLAP systems

## 3.1 What is OLAP

**Online Analytical Processing (OLAP)** is a category of software that allows users to analyse information from multiple database systems at the same time. It is a technology that enables analysts to extract and view business data from different points of view.

Analysts frequently need to group, aggregate, and join data. These operations in relational databases are resource intensive. With OLAP data can be pre-calculated and pre-aggregated, making analysis faster.

**OLAP (Online Analytical Processing)** is the technology behind many Business Intelligence (BI) applications. OLAP is a powerful technology for data discovery, including capabilities for limitless report viewing, complex analytical calculations, and predictive “what if” scenario (budget, forecast)

planning. OLAP performs multidimensional analysis of business data and provides the capability for complex calculations, trend analysis, and sophisticated data modelling.

It is the foundation for many kinds of business applications for Business Performance Management, Planning, Budgeting, Forecasting, Financial Reporting, Analysis, Simulation Models, Knowledge Discovery, and Data Warehouse Reporting. OLAP enables end-users to perform ad hoc analysis of data in multiple dimensions, thereby providing the insight and understanding they need for better decision making.

OLAP databases are divided into one or more cubes. The cubes are designed in such a way that creating and viewing reports become easy. OLAP stands for Online Analytical Processing.

## 3.2 OLAP Cube

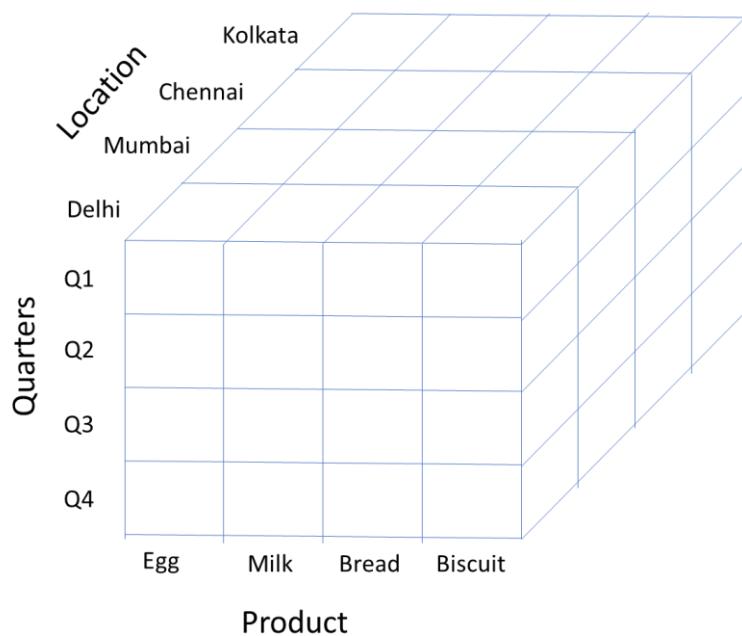


Figure 18 - OLAP Cube

At the core of the OLAP concept, is an OLAP Cube. The OLAP cube is a data structure optimized for very quick data analysis.

The OLAP Cube consists of numeric facts called measures which are categorized by dimensions. OLAP Cube is also called the **hypercube**.

Usually, data operations and analysis are performed using the simple spreadsheet, where data values are arranged in row and column format. This is ideal for two-dimensional data. However, OLAP contains multidimensional data, with data usually obtained from a different and unrelated source. Using a spreadsheet is not an optimal option. The cube can store and analyse multidimensional data in a logical and orderly manner.

### How does it work?

A Data warehouse would extract information from multiple data sources and formats like text files, excel sheet, multimedia files, etc.

The extracted data is cleaned and transformed. Data is loaded into an OLAP server (or OLAP cube) where information is pre-calculated in advance for further analysis.

### **3.3 Basic analytical operations of OLAP**

Four types of analytical operations in OLAP are:

1. Roll-up
2. Drill-down
3. Slice and dice
4. Pivot (rotate)

#### **3.3.1 Roll-up**

Roll-up is also known as "consolidation" or "aggregation." The Roll-up operation can be performed in 2 ways

- Reducing dimensions
- Climbing up concept hierarchy. Concept hierarchy is a system of grouping things based on their order or level.

Consider the following diagram

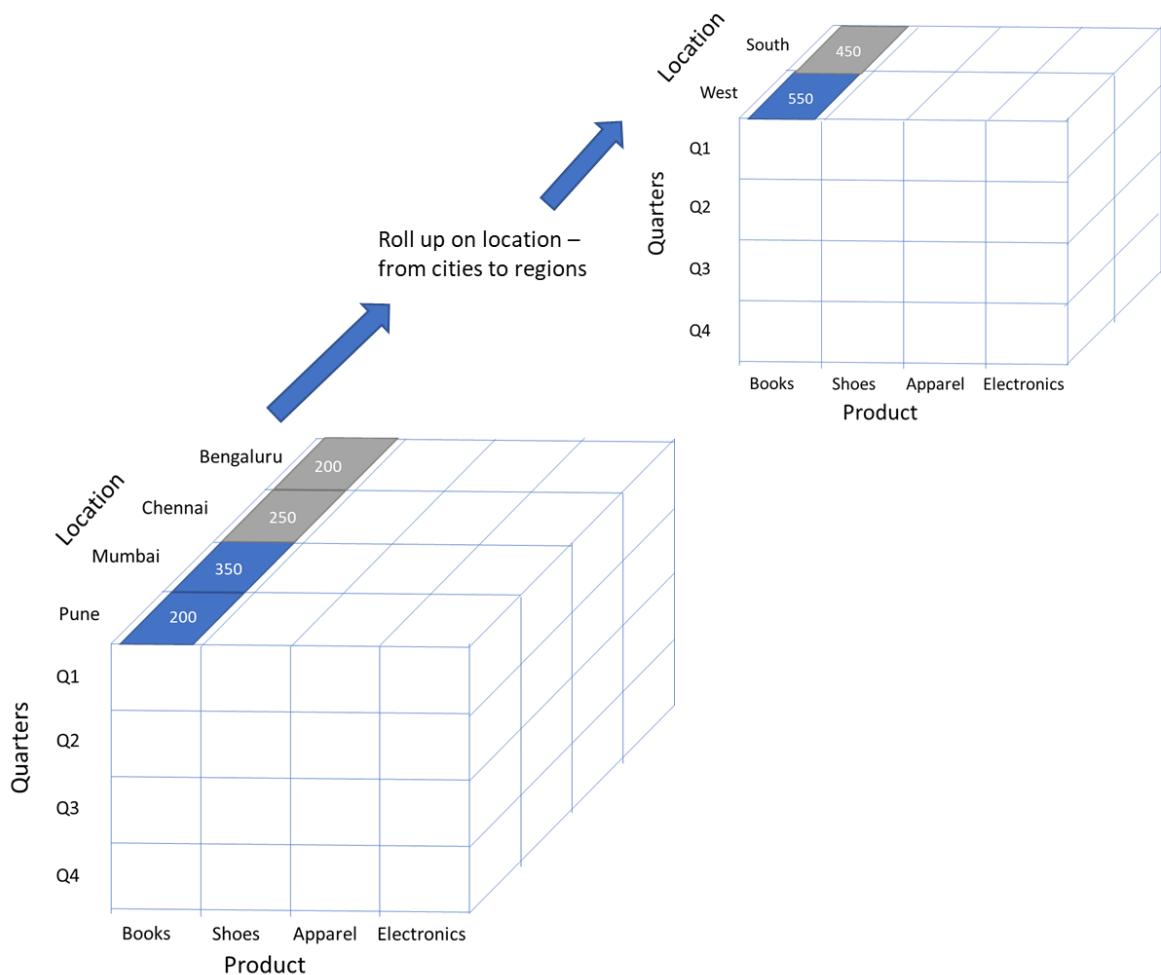


Figure 19 - Example of Roll-up

In this example, cities Pune and Mumbai are rolled up as West; and the sales figures of Chennai and Bengaluru are rolled up as South

- The sales figure of Pune and Mumbai are 200 and 350 respectively. They become 550 after roll-up
- In this aggregation process, data is location hierarchy moves up from city to the region.
- In the roll-up process at least one or more dimensions need to be removed. In this example, Quarter dimension is removed.

### 3.3.2 Drill-down

In drill-down data is fragmented into smaller parts. It is the opposite of the rollup process. It can be done via

- Moving down the concept hierarchy
- Increasing a dimension

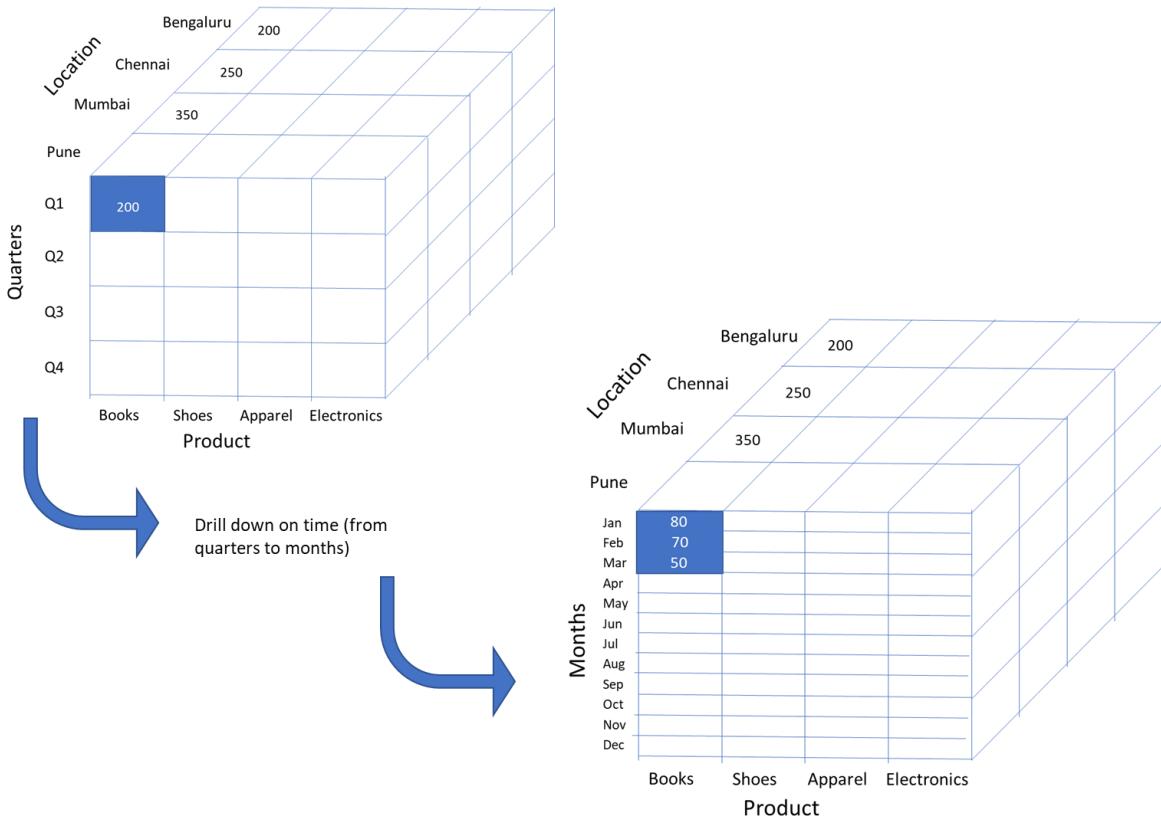


Figure 20 - Drill down Example

Consider the diagram above

- Quarter Q1 is drilled down to months January, February, and March. Corresponding sales are also registered
- In this example, dimension months are added.

### 3.3.3 Slice

Here, one dimension is selected, and a new sub-cube is created. Following diagram explain how slice operation performed:

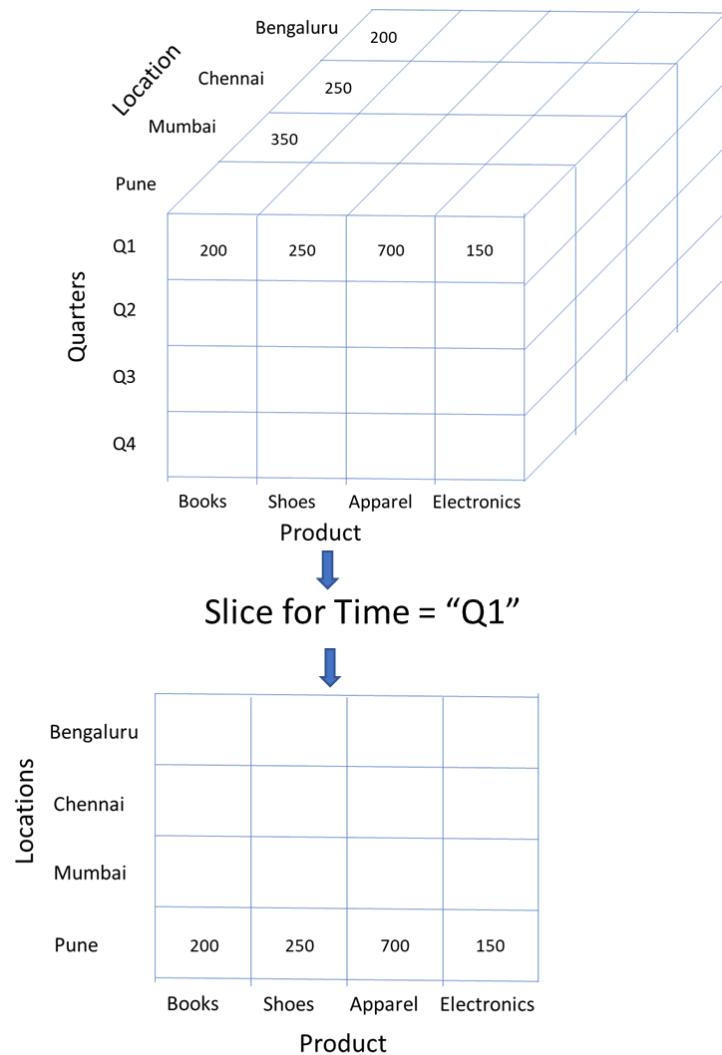


Figure 21 - Example of Slice

- Dimension Time is Sliced with Q1 as the filter.
- A new cube is created altogether.

## Dice

This operation is similar to a slice. The difference in dice is you select 2 or more dimensions that result in the creation of a sub-cube.

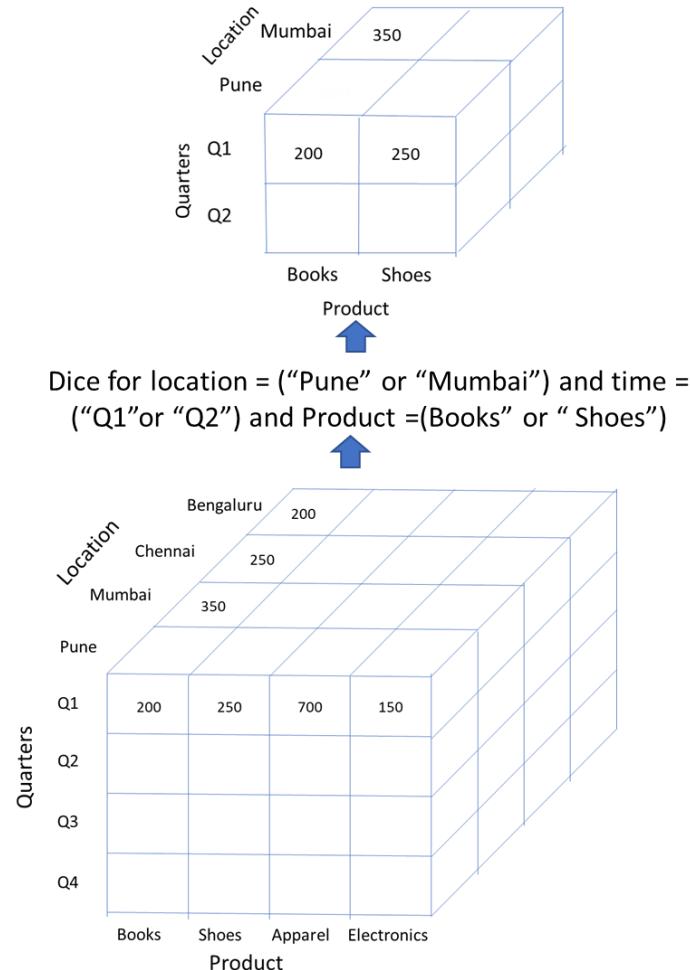


Figure 22 - Example of Dice

### 3.3.4 Pivot

In Pivot, you rotate the data axes to provide a substitute presentation of data.

In the following example, the pivot is based on item types.

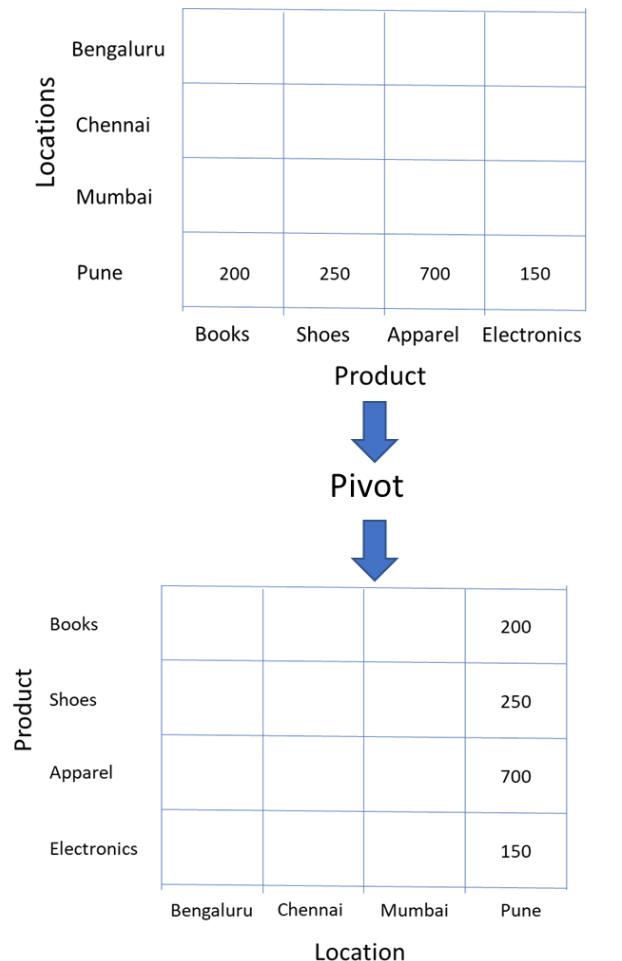


Figure 23 - Example of Pivot

## 3.4 Characteristics of OLAP Systems

In the **FASMI characteristics of OLAP methods**, the term derived from the first letters of the characteristics are:

### Fast

It defines which the system targeted to deliver the most feedback to the client within about five seconds, with the elementary analysis taking no more than one second and very few taking more than 20 seconds.

### Analysis

It defines which the method can cope with any business logic and statistical analysis that is relevant for the function and the user, keep it easy enough for the target client. Although some pre-programming may be needed we do not think it acceptable if all application definitions have to be allow the user to define new Adhoc calculations as part of the analysis and to document on the data in any desired method, without having to program so we excludes products (like Oracle Discoverer) that do not allow the user to define new Adhoc calculation as part of the analysis and to document

on the data in any desired product that do not allow adequate end user-oriented calculation flexibility.

### Share

It defines which the system tools all the security requirements for understanding and, if multiple write connection is needed, concurrent update location at an appropriated level, not all functions need customer to write data back, but for the increasing number which does, the system should be able to manage multiple updates in a timely, secure manner.

### Multidimensional

This is the basic requirement. OLAP system must provide a multidimensional conceptual view of the data, including full support for hierarchies, as this is certainly the most logical method to analyze business and organizations.

### Information

The system should be able to hold all the data needed by the applications. Data sparsity should be handled in an efficient manner.

The main characteristics of OLAP are as follows:

1. **Multidimensional conceptual view:** OLAP systems let business users have a dimensional and logical view of the data in the data warehouse. It helps in carrying slice and dice operations.
2. **Multi-User Support:** Since the OLAP techniques are shared, the OLAP operation should provide normal database operations, containing retrieval, update, adequacy control, integrity, and security.
3. **Accessibility:** OLAP acts as a mediator between data warehouses and front-end. The OLAP operations should be sitting between data sources (e.g., data warehouses) and an OLAP front-end.
4. **Storing OLAP results:** OLAP results are kept separate from data sources.
5. **Uniform documenting performance:** Increasing the number of dimensions or database size should not significantly degrade the reporting performance of the OLAP system.
6. OLAP provides for distinguishing between zero values and missing values so that aggregates are computed correctly.
7. OLAP system should ignore all missing values and compute correct aggregate values.
8. OLAP facilitate interactive query and complex analysis for the users.
9. OLAP allows users to drill down for greater details or roll up for aggregations of metrics along a single business dimension or across multiple dimensions.
10. OLAP provides the ability to perform intricate calculations and comparisons.
11. OLAP presents results in several meaningful ways, including charts and graphs.

## **3.5 Benefits of OLAP**

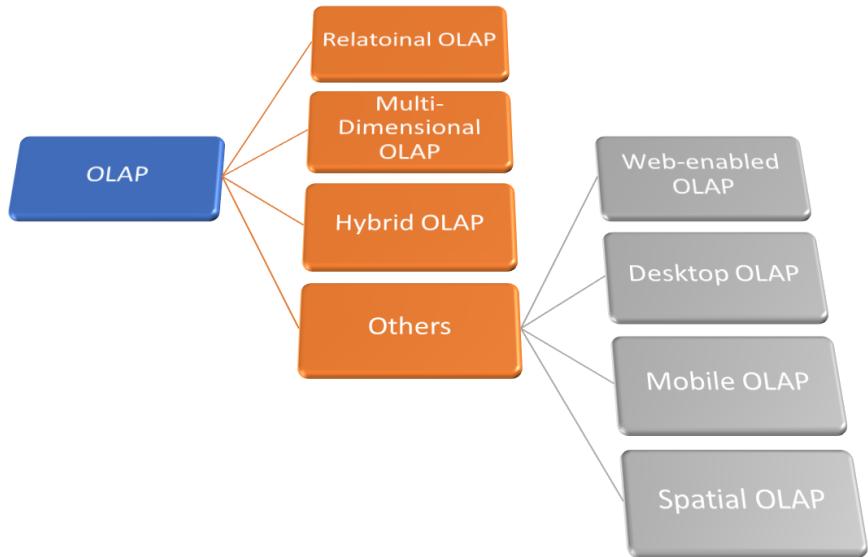
OLAP holds several benefits for businesses: -

- OLAP helps managers in decision-making through the multidimensional record views that it is efficient in providing, thus increasing their productivity.
- OLAP functions are self-sufficient owing to the inherent flexibility support to the organized databases.
- It facilitates simulation of business models and problems, through extensive management of analysis-capabilities.
- In conjunction with data warehouse, OLAP can be used to support a reduction in the application backlog, faster data retrieval, and reduction in query drag.

### **3.5.1 Motivations for using OLAP**

1. **Understanding and improving sales:** For enterprises that have much products and benefit a number of channels for selling the product, OLAP can help in finding the most suitable products and the most famous channels. In some methods, it may be feasible to find the most profitable users. **For example**, considering the telecommunication industry and considering only one product, communication minutes, there is a high amount of record if a company want to analyze the sales of products for every hour of the day (24 hours), difference between weekdays and weekends (2 values) and split regions to which calls are made into 50 region.
2. **Understanding and decreasing costs of doing business:** Improving sales is one method of improving a business, the other method is to analyze cost and to control them as much as suitable without affecting sales. OLAP can assist in analyzing the costs related to sales. In some methods, it may also be feasible to identify expenditures which produce a high return on investments (ROI). **For example**, recruiting a top salesperson may contain high costs, but the revenue generated by the salesperson may justify the investment.

## **3.6 Types of OLAP Models**



*Figure 24 - OLAP Hierarchical Structure*

OLAP (Online Analytical Processing) was introduced into the business intelligence (BI) space over 20 years ago, in a time where computer hardware and software technology weren't nearly as powerful as they are today. OLAP introduced a ground-breaking way for business users (typically analysts) to easily perform multidimensional analysis of large volumes of business data.

Aggregating, grouping, and joining data are the most difficult types of queries for a relational database to process. The magic behind OLAP derives from its ability to pre-calculate and pre-aggregate data. Otherwise, end users would be spending most of their time waiting for query results to be returned by the database. However, it is also what causes OLAP-based solutions to be extremely rigid and IT-intensive.

### 3.6.1 Relational OLAP

These are intermediate servers which stand in between a relational back-end server and user frontend tools. They use a relational or extended-relational DBMS to save and handle warehouse data, and OLAP middleware to provide missing pieces. ROLAP servers contain optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services. ROLAP technology tends to have higher scalability than MOLAP technology.

ROLAP systems work primarily from the data that resides in a relational database, where the base data and dimension tables are stored as relational tables. This model permits the multidimensional analysis of data.

This technique relies on manipulating the data stored in the relational database to give the presence of traditional OLAP's slicing and dicing functionality. In essence, each method of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement.

ROLAP stands for Relational Online Analytical Processing. ROLAP stores data in columns and rows (also known as relational tables) and retrieves the information on demand through user submitted

queries. A ROLAP database can be accessed through complex SQL queries to calculate information. ROLAP can handle large data volumes, but the larger the data, the slower the processing times.

Because queries are made on-demand, ROLAP does not require the storage and pre-computation of information. However, the disadvantage of ROLAP implementations are the potential performance constraints and scalability limitations that result from large and inefficient join operations between large tables. Examples of popular ROLAP products include Meta cube by Stanford Technology Group, Red Brick Warehouse by Red Brick Systems, and AXSYS Suite by Information Advantage.

### Relational OLAP Architecture

ROLAP Architecture includes the following components

- Database server.
- ROLAP server.
- Front-end tool.

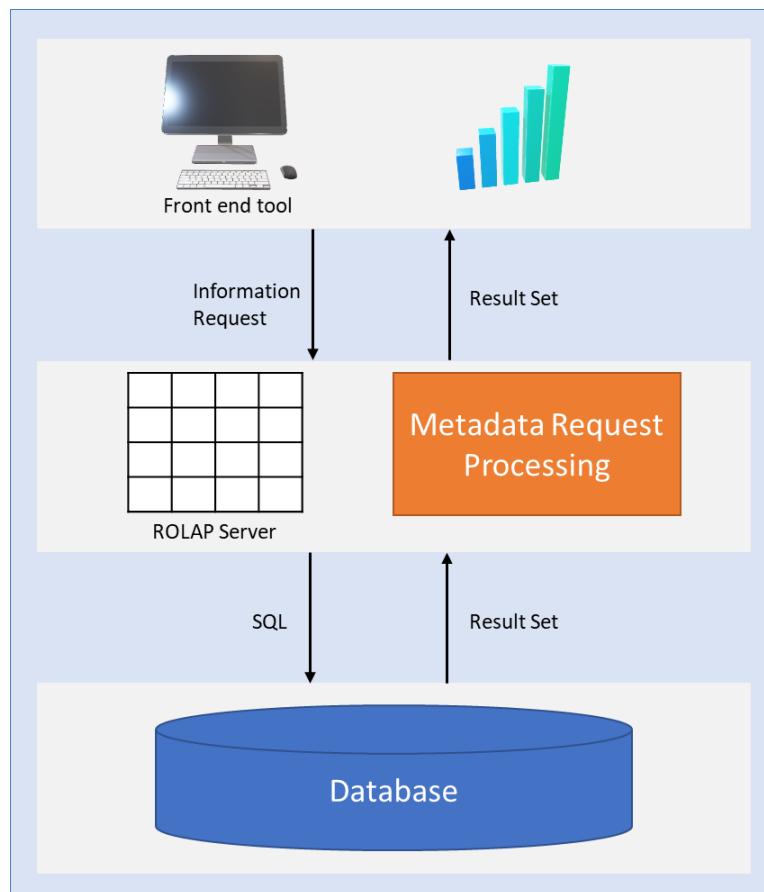


Figure 25 - ROLAP Architecture

**Relational OLAP (ROLAP)** is the latest and fastest-growing OLAP technology segment in the market. This method allows multiple multidimensional views of two-dimensional relational tables to be created, avoiding structuring record around the desired view.

Some products in this segment have supported reliable SQL engines to help the complexity of multidimensional analysis. This includes creating multiple SQL statements to handle user requests,

being 'RDBMS' aware and also being capable of generating the SQL statements based on the optimizer of the DBMS engine.

### Advantages

- **Can handle large amounts of information:** The data size limitation of ROLAP technology is depends on the data size of the underlying RDBMS. So, ROLAP itself does not restrict the data amount.
- RDBMS already comes with a lot of features. So ROLAP technologies, (works on top of the RDBMS) can control these functionalities.

### Disadvantages

- **Performance can be slow:** Each ROLAP report is a SQL query (or multiple SQL queries) in the relational database, the query time can be prolonged if the underlying data size is large.
- **Limited by SQL functionalities:** ROLAP technology relies on upon developing SQL statements to query the relational database, and SQL statements do not suit all needs.

### 3.6.2 Multidimensional OLAP (MOLAP) Server

A MOLAP system is based on a native logical model that directly supports multidimensional data and operations. Data are stored physically into multidimensional arrays, and positional techniques are used to access them.

One of the significant distinctions of **MOLAP** against a **ROLAP** is that data are summarized and are stored in an optimized format in a multidimensional cube, instead of in a relational database. In MOLAP model, data are structured into proprietary formats by client's reporting requirements with the calculations pre-generated on the cubes.

### MOLAP Architecture

MOLAP Architecture includes the following components

- Database server.
- MOLAP server.
- Front-end tool.

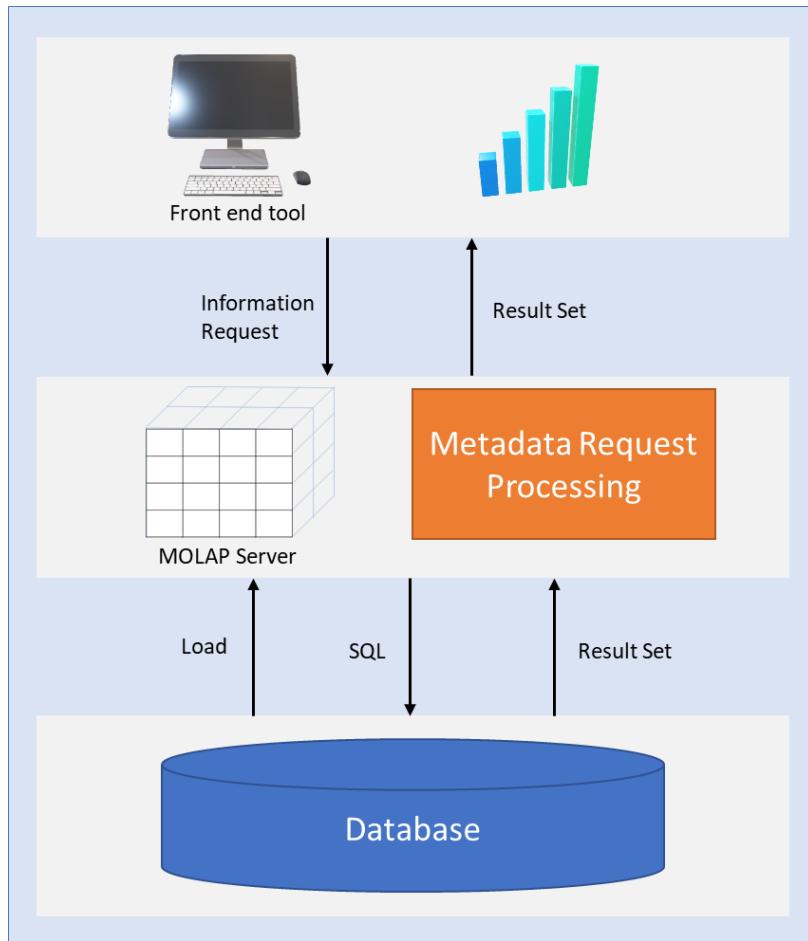


Figure 26 - MOLAP Architecture

**MOLAP** structure primarily reads the precompiled data. MOLAP structure has limited capabilities to dynamically create aggregations or to evaluate results which have not been pre-calculated and stored.

Applications requiring iterative and comprehensive time-series analysis of trends are well suited for MOLAP technology (e.g., financial analysis and budgeting).

Examples include Arbor Software's Essbase, Oracle's Express Server, Pilot Software's Lightship Server, Sniper's TM/1, Planning Science's Gentium and Kenan Technology's Multiway.

Some of the problems faced by clients are related to maintaining support to multiple subject areas in an RDBMS. Some vendors can solve these problems by continuing access from MOLAP tools to detailed data in and RDBMS.

This can be very useful for organizations with performance-sensitive multidimensional analysis requirements and that have built or are in the process of building a data warehouse architecture that contains multiple subject areas.

An example would be the creation of sales data measured by several dimensions (e.g., product and sales region) to be stored and maintained in a persistent structure. This structure would be provided to reduce the application overhead of performing calculations and building aggregation during initialization. These structures can be automatically refreshed at predetermined intervals established by an administrator.

## Advantages

- **Excellent Performance:** A MOLAP cube is built for fast information retrieval and is optimal for slicing and dicing operations.
- **Can perform complex calculations:** All evaluation have been pre-generated when the cube is created. Hence, complex calculations are not only possible, but they return quickly.

## Disadvantages

- **Limited in the amount of information it can handle:** Because all calculations are performed when the cube is built, it is not possible to contain a large amount of data in the cube itself.
- **Requires additional investment:** Cube technology is generally proprietary and does not already exist in the organization. Therefore, to adopt MOLAP technology, chances are other investments in human and capital resources are needed.

### 3.6.3 Hybrid OLAP (HOLAP) Server

HOLAP incorporates the best features of **MOLAP** and **ROLAP** into a single architecture. HOLAP systems save more substantial quantities of detailed data in the relational tables while the aggregations are stored in the pre-calculated cubes. HOLAP also can drill through from the cube down to the relational tables for delineated data. The **Microsoft SQL Server 2000** provides a hybrid OLAP server.

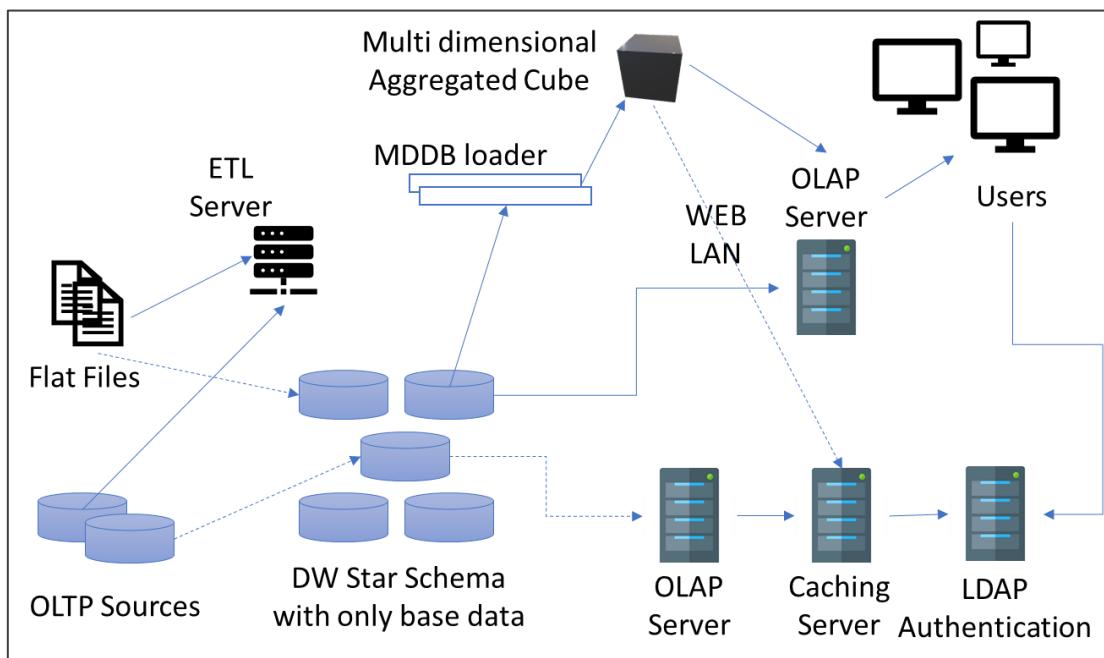


Figure 27 - HOLAP Architecture

## Advantages of HOLAP

- HOLAP provide benefits of both MOLAP and ROLAP.
- It provides fast access at all levels of aggregation.
- HOLAP balances the disk space requirement, as it only stores the aggregate information on the OLAP server and the detail record remains in the relational database. So no duplicate copy of the detail record is maintained.

### **Disadvantages of HOLAP**

HOLAP architecture is very complicated because it supports both MOLAP and ROLAP servers.

#### **3.6.4 Other Types**

There are also less popular types of OLAP styles upon which one could stumble upon every so often. We have listed some of the less popular brands existing in the OLAP industry.

#### **Web-Enabled OLAP (WOLAP) Server**

WOLAP pertains to OLAP application which is accessible via the web browser. Unlike traditional client/server OLAP applications, WOLAP is considered to have a three-tiered architecture which consists of three components: a client, a middleware, and a database server.

#### **Desktop OLAP (DOLAP) Server**

DOLAP permits a user to download a section of the data from the database or source, and work with that dataset locally, or on their desktop.

#### **Mobile OLAP (MOLAP) Server**

Mobile OLAP enables users to access and work on OLAP data and applications remotely through the use of their mobile devices.

#### **Spatial OLAP (SOLAP) Server**

SOLAP includes the capabilities of both Geographic Information Systems (GIS) and OLAP into a single user interface. It facilitates the management of both spatial and non-spatial data.

## **3.7 Difference between ROLAP, MOLAP, and HOLAP**

<b>ROLAP</b>	<b>MOLAP</b>	<b>HOLAP</b>
ROLAP stands for Relational Online Analytical Processing.	MOLAP stands for Multidimensional Online Analytical Processing.	HOLAP stands for Hybrid Online Analytical Processing.
The ROLAP storage mode causes the aggregation of the division to be stored in indexed views in the relational database that was specified in	The MOLAP storage mode principle the aggregations of the division and a copy of its source information to be saved in a multidimensional	The HOLAP storage mode connects attributes of both MOLAP and ROLAP. Like MOLAP, HOLAP causes the aggregation of the division to

the partition's data source.	operation in analysis services when the separation is processed.	be stored in a multidimensional operation in an SQL Server analysis services instance.
ROLAP does not because a copy of the source information to be stored in the Analysis services data folders. Instead, when the outcome cannot be derived from the query cache, the indexed views in the record source are accessed to answer queries.	This MOLAP operation is highly optimize to maximize query performance. The storage area can be on the computer where the partition is described or on another computer running Analysis services. Because a copy of the source information resides in the multidimensional operation, queries can be resolved without accessing the partition's source record.	HOLAP does not causes a copy of the source information to be stored. For queries that access the only summary record in the aggregations of a division, HOLAP is the equivalent of MOLAP.
Query response is frequently slower with ROLAP storage than with the MOLAP or HOLAP storage mode. Processing time is also frequently slower with ROLAP.	Query response times can be reduced substantially by using aggregations. The record in the partition's MOLAP operation is only as current as of the most recent processing of the separation.	Queries that access source record for example, if we want to drill down to an atomic cube cell for which there is no aggregation information must retrieve data from the relational database and will not be as fast as they would be if the source information were stored in the MOLAP architecture.

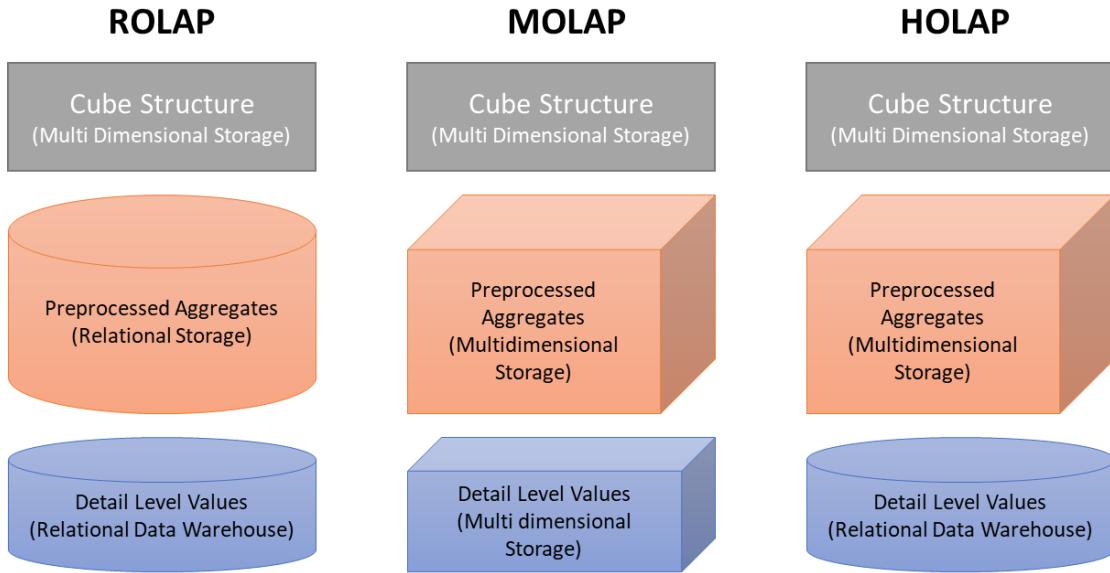


Figure 28 - Difference between ROLAP, MOLAP and HOLAP

### 3.8 Difference Between ROLAP and MOLAP

ROLAP	MOLAP
ROLAP stands for Relational Online Analytical Processing.	MOLAP stands for Multidimensional Online Analytical Processing.
It usually used when data warehouse contains relational data.	It used when data warehouse contains relational as well as non-relational data.
It contains Analytical server.	It contains the MDDB server.
It creates a multidimensional view of data dynamically.	It contains prefabricated data cubes.
It is very easy to implement	It is difficult to implement.
It has a high response time	It has less response time due to prefabricated cubes.

### 3.9 Summary

- Online Analytical Processing (OLAP) is a category of software that allows users to analyse information from multiple database systems at the same time.
- OLAP enables end-users to perform ad hoc analysis of data in multiple dimensions, thereby providing the insight and understanding they need for better decision making.
- At the core of the OLAP concept, is an OLAP Cube. The OLAP cube is a data structure optimized for very quick data analysis.

- The OLAP Cube consists of numeric facts called measures which are categorized by dimensions. OLAP Cube is also called the hypercube.
- Four types of analytical operations in OLAP are Roll-up, Drill-down, Slice & dice and Pivot (rotate)
- The Roll-up operation can be performed by Reducing dimensions or climbing up concept hierarchy
- The Drill down operation can be performed by moving down the concept hierarchy or Increasing a dimension
- Slice is when one dimension is selected, and a new sub-cube is created. Dice is where two or more dimensions are selected as a new sub-cube is created
- In Pivot, you rotate the data axes to provide a substitute presentation of data.
- FASMI characteristics of OLAP methods - Fast, Analysis, Share, Multi-dimensional and Information
- OLAP helps in understanding and improving sales. It also helps in understanding and improving the cost of doing business
- Three major types of OLAP models are Relational OLAP, Multi-dimensional OLAP and Hybrid OLAP
- Relational OLAP systems are intermediate servers which stand in between a relational backend server and user frontend tools. They use a relational or extended-relational DBMS to save and handle warehouse data, and OLAP middleware to provide missing pieces
- MOLAP structure primarily reads the precompiled data. MOLAP structure has limited capabilities to dynamically create aggregations or to evaluate results which have not been pre-calculated and stored.
- HOLAP incorporates the best features of MOLAP and ROLAP into a single architecture. HOLAP systems save more substantial quantities of detailed data in the relational tables while the aggregations are stored in the pre-calculated cubes.

## **Module 3 - DATA MINING AND PREPROCESSING**

### **Chapter 6 Introduction to Data Mining**

#### **6.0 Objectives**

This chapter provides an overview of the following –

- What is Data mining?

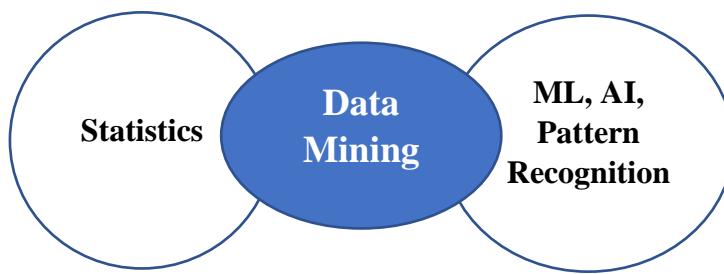
- Data Mining Applications
  - Knowledge Discovery in Data (KDD) Process in detail
  - Basic architecture of Data Mining system
  - Issues and Challenges in Data Mining
- 

## 6.1 Introduction

---

We say that today is the age of Big Data. The sheer volume of data being generated today is exploding. The rate of data creation or generation is mind boggling. Mobile phones, social media, imaging technologies which are used for medical diagnosis, non-traditional IT devices like RFID readers, GPS navigation systems —all these are among the fastest growing sources of data. Now keeping up with this huge influx of data is difficult, but what is more challenging is analysing vast amounts of this generated data, to identify meaningful patterns and extract useful information. Data in its original form is crude, unrefined so it must be broken down, analysed to have some value. So, Data Mining is finding insightful information which is hidden in the data.

**Data Mining**, (sometimes also known as Knowledge Discovery in Data (KDD)), is an automatic or semi-automatic ‘mining’ process used for extracting useful data from a large set of raw data. It analyses large amount of scattered information to find meaningful constructs from it and turns it into knowledge. It checks for anomalies or irregularities in the data, identifies patterns or correlations among millions of records and then converts it into knowledge about future trends and predictions. It covers a wide variety of domains and techniques including Database Technology, Multivariate Statistics, Engineering and Economics (provides methods for Pattern recognition and predictive modelling), ML (Machine Learning), Artificial Intelligence, Information Science, Neural Networks, Data Visualization many more.



Data Mining – Confluence of Multiple Domains

---

## 6.2 Data Mining Applications

---

Data Mining and big data are used almost everywhere. **Data Mining is increasingly used by companies having solid consumer focus like in retail sales, advertising and marketing, financial institutions, bioinformatics etc.** Almost all Commercial companies use data mining and big data to gain insights into their customers, processes, staff, and products. Many companies use mining to offer customers a better user experience, as well as to cross-sell, increase the sale, and customize their products.

Data Mining is helping the **consumer-oriented companies** in determining the relation between product price and product positioning. It also helps to determine the relation between consumer demographics and competition. It aids in determining the impact of relationships on consumer satisfaction, sales and profits. The retail companies conduct sales campaigns including discount coupons, bonuses for loyal customers, advertisements to attract and retain their customers. Careful analysis of the effectiveness of such sales campaigns is essential to improve profits and reduce business cost.



**Financial institutions** use data mining and analysis to predict stock markets, determine the risk of lending money, and learn how to attract new clients for their services.

**Credit card companies** monitor the spending habits of the customer and can easily recognize duplicitous purchases with a certain degree of accuracy using rules derived by processing billions of transactions.

**Mobile phone companies** study their subscribers' calling patterns to determine, for example, whether a caller's frequent contacts are on a rival network. If that rival network is offering an attractive promotion which might cause the subscriber to defect, the mobile phone company can proactively offer the subscriber an attractive offer to avoid defection.

**Social media** companies such as LinkedIn and Facebook, data itself is their primary product. The valuations of these companies are heavily derived from the data they gather and host, which contains more and more intrinsic value as the data grows. For example, a small update on your social media account bombards you with related information and advertisements. Facebook constructs social graphs to analyze which users are connected to each other as an interconnected network. These graphs also help them to search people with similar interest, hobbies and shared locations.

Another domain where data mining is gaining rapid grounds is **biological data analysis**. Genetic sequencing and human genome mapping provide a detailed genetic makeup and lineage. The healthcare industry is using this information to predict illnesses a person is likely to get in his lifetime and take steps to avoid these diseases or reduce its impact through the use of customized medications and treatment.

The process of gathering data is fairly well established. In fact, forward-thinking organizations started collecting data even before they knew how they were going to use it. They recognized that data had great value, even if they did not yet know how to extract that value. The challenge is now how to use that data to gain valuable insights for the business. Efforts among data mining professionals are now focused on how to leverage all of the data. Companies that effectively employ data mining tools and techniques can translate their collected data into valuable insights about their business processes and strategies. Such insights can then be used to make better business decisions that increase productivity and revenue, leading to accelerated business growth and higher profits.

---

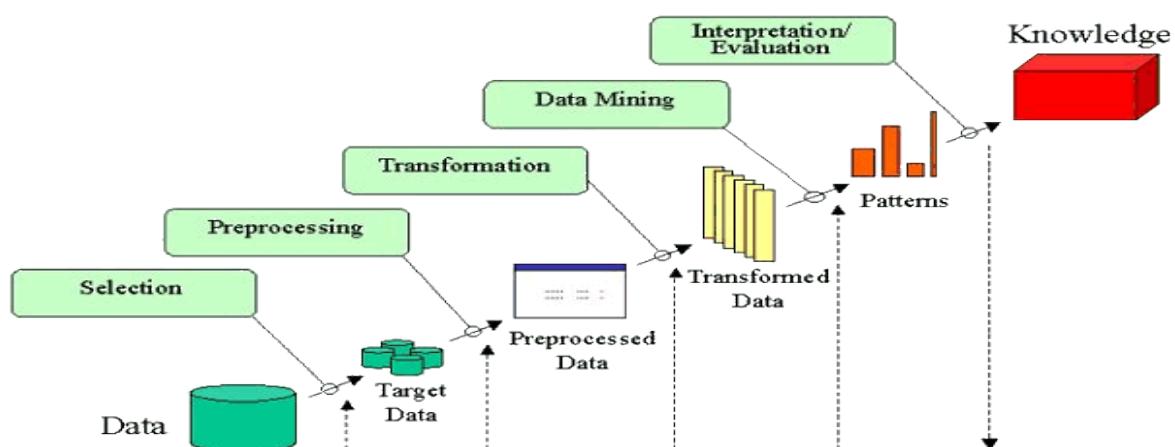
### 6.3 Knowledge Discovery in Data (KDD) Process

---

It is an interactive and iterative sequence comprising of 9 phases. Teams commonly learn new things in a phase that cause them to go back and refine the work done in prior phases based on new insights and information that have been uncovered. The diagram given below depicts the iterative movement between phases until the team members have sufficient information to move to the next phase. The process begins with finding the KDD goals and ends with the successful implementation of the discovered knowledge.

1. **Domain Understanding** – In this preliminary step the team needs to understand and define the goals of the end-user and the environment in which the KDD process will take place.
2. **Selection & Addition** – In this phase it is important to determine the dataset which will be utilized for the KDD process. So, the team needs to first find the relevant data which is accessible. Data from multiple sources can be integrated in this phase. Note that this is the data which is going to lead us to Knowledge. So, if some attributes from the data are missing then it will lead to half-cooked Knowledge. Therefore, the objective of this phase is determining the suitable and complete dataset on which the discovery will be performed.

3. **Pre-processing & Cleansing** – The data received from the earlier phase is like a rough diamond. Now in this phase you need to polish the diamond so that everyone can know its beauty. So now the main task in this phase is to sanitize and prepare the data for use. Data cleansing is a subprocess that focuses on removing errors in your data so your data becomes true and consistent. Sanity checks are performed to check that the data does not contain physically or theoretically impossible values such as people taller than 3 meters or someone with an age of 299 years.
4. **Data Transformation** – Once your team has cleansed and integrated the data, now you may have to transform your data so it becomes suitable for the next phase of data mining. In this phase, the data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations. The aggregation operators perform mathematical operations like Average, Count, Max, Min, Sum etc. on the numeric property of the elements in the collection. This phase is very project-specific.
5. **Data Mining** – In this phase methods like Association, Classification, Clustering and/or Regression are applied in order to extract patterns. We may need to use the Data Mining Algorithm several times until the desired output is obtained.
6. **Evaluation** – In this phase we evaluate and understand the mined patterns, rules and reliability to the goal set in the first phase. Here we assess the pre-processing steps for their impact on the Data Mining Algorithm outcomes. For example, we can assess the outcome of the algorithm by adding an extra feature in phase 4 and repeating from there. This phase focuses on comprehensibility and efficacy of the newly developed model.
7. **Discovered Knowledge Presentation** – The last phase is all about the use and overall feedback and discovery results acquired by Data Mining. The interesting discovered patterns are presented to the end-user and may be stored as new knowledge in the knowledge base. The success of this phase decides the effectiveness of the entire KDD process.



#### 6.4 Architecture of Data Mining System / Components of Data Mining System

Data Mining architecture consists of various components which make up the entire process of data mining.

- **Sources of Data** – There is varied sources of data available today. Many a times data is present only in the form of text files or spreadsheets, excel worksheets. Today major chunk of data is gathered from internet or WWW which forms the information repository units. Data from these various sources are in different formats. So, this data cannot be used directly for processing. The main challenge is to clean and transform the data so that it can be used for analysis. Various tools and techniques are implemented to ensure reliability and completeness of the data.
- **Database or Data Warehouse Server** – The data received from various data sources is stored in the database server. This data is ready to be processed. It is fetched depending on the user's request.
- **Data Mining Engine** – This is the most crucial component of this architecture. It usually contains a lot of modules which perform a variety of tasks. The tasks like Association, Characterization, Prediction, Clustering, Classification, Regression, Time series analysis, naïve Bayes, Support Vector machines, Random Forests, Decision Trees etc can be performed.
- **Pattern Evaluation Modules** – This module measures how interesting the pattern that has been devised is actually. Usually, a threshold value is used for evaluation. This module has a direct link of interaction with the data mining engine. The main aim of this module is to determine the interesting and useful patterns that could make the data of better quality.
- **Graphical User Interface (GUI)** – This module interacts with the user. GUI helps the user to access and use the system efficiently. GUI hides the complexity of the system thereby displaying only the relevant components to the user. When the user submits a query, the module interacts with the overall set of data mining system, producing a relevant output which is displayed in an appropriate manner. This component also allows the user to browse database and database warehouse schemas, evaluate mined patterns and visualize the patterns in different forms.
- **Knowledge Base** – Knowledge Base contains the domain knowledge which is required to search or evaluate the patterns. It may also contain data from user experiences and beliefs. The data mining engine interacts with the knowledge base thereby providing more efficient, accurate and reliable results. The pattern evaluation module also interacts with the knowledge base to get various inputs and updates from it.

---

## 6.5 Issues and Challenges in Data Mining

---

Efficient and effective data mining in large databases poses number of challenges to researchers and developers. The issues in data mining include the following –

- 1. Mining methodology and user interaction issues** – These include the types of knowledge mined, the ability to mine the knowledge at multiple granularities, the use of domain knowledge, ad hoc mining and knowledge visualization.
  - **Mining different types of knowledge in databases** – The need of different users is never the same. Different users may be interested in different kinds of knowledge. So, it is necessary for data mining to cover broad range of knowledge discovery tasks like association, correlation analysis, prediction, clustering, outlier analysis etc. These tasks require the development of numerous data mining techniques.
  - **Interactive mining of knowledge at multiple levels of abstraction** – Data mining process should be interactive. User should be able to interact with the data mining system to view the data and discovered patterns at multiple granularities and from different angles. Interactive mining is very essential as it allows the user to concentrate on discovered patterns and to refine the data mining requests based on the discovery.
  - **Incorporation of background knowledge** – To guide the discovery process and to express the discovered patterns in a proper way at multiple levels of abstraction, background knowledge of the domain being studied is essential.
  - **Data mining query languages and ad hoc data mining** – Data mining query language should be developed for giving access to the user and describing ad hoc mining tasks. Also, it should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
  - **Presentation and Visualization of data mining results** – Once the patterns are discovered it needs to be expressed in high-level languages, visual presentations or in any other appropriate form so that it is easily understood by the stakeholders. Various visualization techniques like trees, graphs, charts, matrices etc need to be used.
  - **Handling noisy or incomplete data** – Data cleaning methods need to be applied to deal with the noisy data. If such techniques are not applied then the accuracy of the discovered patterns will be poor.
  - **Pattern evaluation** – A data mining system may discover thousands of patterns, many of which are uninteresting to the user as they represent the common knowledge or lack any freshness. So, the challenge faced by data mining systems are development of techniques to assess the ‘interestingness’ of discovered patterns.
- 2. Performance Issues** – There are performance issues related to the data mining systems. Few are given below –
  - **Efficiency and scalability of data mining algorithms** – In order to effectively extract the information from huge databases, data mining algorithms must be efficient and scalable. The running time of data mining algorithm must be predictable and acceptable in huge databases.

- **Parallel, distributed and incremental mining algorithms** – Wide distribution of data, huge size of databases and the computational complexity of some data mining methods are the factors responsible for development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions, which are processed in parallel. The results are then merged. The high cost of some data mining processes promotes the need for incremental data mining algorithms.

### **3. Diversity of database types –**

- **Handling relation and complex types of data** – As relational databases and data warehouses are the most commonly used, there is a need for developing efficient and effective data mining systems for handling such data. There are diverse data types so it must not be expected that one system should mine all kinds of data. So, specific data mining systems must be developed for mining specific type of data.
- **Mining information from heterogeneous databases and global information systems** – Today we have huge, distributed and heterogeneous databases which are connected with each other using computer networks. Data mining systems may help us reveal high-level data regularities in multiple heterogeneous databases which was not possible with simple query systems. Web mining which is one of the most challenging and evolving data mining field, reveals interesting knowledge about web usage, web dynamics, web contents etc.

### **4. Social impact of data mining** – Today we have powerful data mining tools being developed and put into use. But these tools and techniques pose a threat to our privacy and data security. Data mining applications derive large set of demographic information about the consumers that was previously not known or hidden in the data. The unauthorized use of such data could result in the disclosure of information that is supposed to be confidential. The focus of data mining technology is on the discovery of useful patterns and not on specific information regarding an individual. But due to the exploratory nature of data mining, it is impossible to know what patterns may be discovered and so, there is no certainty over how they may be used. So, data security-enhancing techniques and privacy-preserving data mining are the two new areas where research is being carried out. We must not definitely lose sight of all the benefits that data mining research brings (ranging from insights gained from medical and scientific applications to increased customer satisfaction) but we definitely expect our researchers and scientists to build solutions which ensure data privacy protection and security.

All these issues and challenges stimulate and motivate the researchers and experts to investigate this data mining field deeper.

## **6.6 Summary**

Data Mining is a task of discovering meaningful patterns and insights from huge datasets. It is an interdisciplinary field linked with areas like machine learning, artificial intelligence, statistics, data warehousing, neural networks, pattern recognition etc. It is widely used in telecom industry, retail industry, financial organizations, biomedical analysis, intrusion detection etc.

Knowledge Discovery of Data (KDD) Process consists of data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation.

The various components of a data mining system include database, data warehouse, servers, data mining engine, pattern evaluation module and GUI.

Performance issues, mining methodology and user interface issues, diversity of databases and social impact of data mining are the few challenges faced in this domain. The issues and challenges in data mining motivate the researchers and scientists to explore this field further.

---

## 6.7 Exercises

1. What is data mining?
  2. What are the steps involved in data mining when viewed as a process of knowledge discovery?
  3. How data mining helps in businesses? Explain with few examples.
  4. Explain in detail the purpose of each component of the data mining system.
  5. Explain the challenges in data mining field w.r.t to social implications of mining.
- 

## 6.8 References

- Dunham, Margaret H, Data mining: Introductory and advanced topics, Pearson Education India, 2006.
- Han, Jiawei, Jian Pei, and Micheline Kamber, Data mining: concepts and techniques, Second Edition, Elsevier, Morgan Kaufmann, 2011.

## **Module 3 - DATA MINING AND PREPROCESSING**

### **Chapter 7 Data Preprocessing**

---

## 7.0 Objectives

This chapter provides an overview of the following –

- What is the need of Data Preprocessing?
- What is Data Cleaning?

- How to deal with missing values, noisy data?
  - What is Data Integration?
  - What is Data Transformation?
- 

## 7.1 Introduction to Data Preprocessing

---

In real world data comes from various heterogeneous sources. Data directly taken from such diverse sources may have inconsistencies, errors, missing values or redundancy. Since redundancies, missing values and inconsistencies compromise the integrity of the dataset, we need to fix these issues otherwise the final outcome would be plagued with faulty insights and misleading conclusions. Chances are that with faulty dataset the system will develop biases and deviations which will result in poor user experience. So, **Data Preprocessing is a technique to turn raw and crude information gathered from diverse sources into clean and consistent dataset.**

**Data Preprocessing is one of the most vital steps in the data mining process. Data Preprocessing involves Data Cleaning, Data Integration, Data Transformation, Data Reduction etc.** These techniques are useful for removing the noisy data and preparing the quality data which gives efficient result of the analysis. These techniques, when applied before the data mining process will definitely improve the quality of the patterns mined as well as the time needed for actual mining process. Note that these techniques are not mutually exclusive.

Data Preprocessing is required because in real world collected data is generally –

- **Incomplete** – Data collected from various sources may be missing some important attributes, or it may have only aggregate data. Reasons for incomplete data lies in data collection process itself. The data or few of the important attributes were not collected in the first place as they were not relevant at the time of data collection. Data collection and data entry are error-prone processes. They often require human intervention, and because humans are only human, they make typos or lose their concentration for a second and introduce an error into the chain. The data collected by machines or computers isn't free from errors either. Incomplete data can be collected because of malfunctioning machine or sensor.
- **Noisy** – Data may be erroneous or may have outliers. An outlier is an observation that seems to be distant from other observations or an observation that follows a different logic or generative process than all the other observations. Outliers can gravely influence your data modelling so there is a need to investigate them first. Data errors may also point to defective equipment such as broken transmission lines or defective sensors. Data errors can also point to software bugs. So, fixing the data as soon as its captured is very important.
- **Inconsistent** – Data may have discrepancies like physically or theoretically impossible values, different codes or names. Inconsistent data may also be a result of

deviation from codebook. A codebook is a description of your data, a form of metadata.

Before using the data for analysis, we need it to be organized and sanitized properly. In the following section we will learn the various techniques which are used to get a clean dataset.

---

## 7.2 Introduction to Data Cleaning

---

**Data Cleaning**, also called as Data Cleansing or Scrubbing, is required because dataset is dirty. It is a part of data preprocessing. Inconsistent, incorrect, inaccurate, incomplete data is identified as a part of the data cleansing process.

**Data Cleaning can be implemented in many ways by adjusting the missing values, identifying and removing outliers, removing redundant rows or deleting irrelevant records.**

### 7.2.1 Missing Values

You may notice that many tuples from your dataset have no recorded values for certain attributes. Following are the methods for handling missing values. Which technique to use at what time is dependent on your particular case. Following techniques can be used to handle missing values –

Sr. No.	Technique	Advantage	Disadvantage
1	Ignore the tuple or omit the values	Easy to perform	You lose information for the tuple. It is especially poor when the percentage of missing values per attribute varies considerably.
2	Set value to null	Easy to perform	Not every modelling technique can handle null values
3	Fill the missing value manually		Approach is tedious and time consuming (in some cases not feasible) if there is a large set of data with missing values.
4	Input a static value such as 0 or a global constant in place of missing value	Easy to perform. You don't lose information from the other attributes in the tuple	Can lead to false estimations.
5	Fill the missing value with the attribute mean	Does not disturb the model as much	Harder to execute. You make data assumptions.
Sr.	Technique	Advantage	Disadvantage

No.			
6	Fill the missing value with the most probable value	Widely popular	Not easy to perform. Probable value can be estimated with Regression, Bayesian inference or decision trees.

Techniques from 4 to 6 bias the data. There is a high chance that the filled-in value is incorrect. However, technique 6 is used heavily as it uses the large percentage of information from the available data to predict missing values. Due to inference-based tools or decision tree induction there is a greater probability that the relationship between the missing value and the other attributes is preserved.

### **7.2.2 Noisy Data**

Noisy data is the data with a large amount of additional meaningless information in it. In general, noise is a random error or variance which may include faulty data gathering equipment, technology limitations, resource limitations and data entry problems. Due to noise, algorithms often miss out data patterns. Noisy data can be handled by the following methods –

- Binning** – In this method, the data values are sorted in an order, then grouped into ‘bins’ or buckets. Then each value in a particular bin is smoothed using its neighbourhood i.e., its surrounding values. It is said that binning method performs local smoothing as it looks up at its surrounding values to smooth the values of the attribute.

**Consider an example** – Suppose we have a set of following values which are sorted: [4, 8, 9, 15, 21, 21, 21, 24, 25, 26, 28, 29, 34]

Now, we will divide this dataset into sets of equal frequency –

Bin1: 4, 8, 9, 15

Bin2: 21, 21, 24, 25

Bin3: 26, 28, 29, 34

There are several ways of binning the values –

**Smoothing by bin means** – Here, all the values of a bin are replaced by the mean of the values from that bin.

Mean of 4, 8, 9, 15 = 9

Mean of 21, 21, 24, 25 = 23

Mean of 26, 28, 28, 34 = 29

Therefore, this way results in the following bins -

Bin1: 9, 9, 9, 9

Bin2: 23, 23, 23, 23

Bin3: 29, 29, 29, 29

**Smoothing by bin medians** – Here, all the values of a bin are replaced by the median of the values from that bin.

Median of 4, 8, 9, 15 = 9

Median of 21, 21, 24, 25 = 23

Median of 26, 28, 28, 34 = 28

Therefore, this way results in the following bins -

Bin1: 9, 9, 9, 9

Bin2: 23, 23, 23, 23

Bin3: 28, 28, 28, 28

**Smoothing by bin boundaries** – Here, all the values of a bin are replaced by the closest boundary of the values from that bin. Therefore, this way results in the following bins -

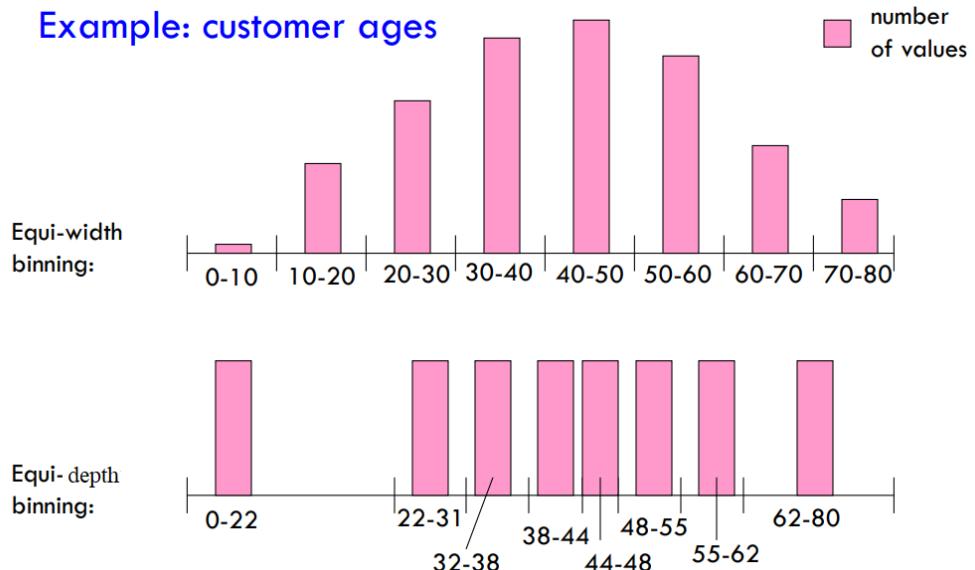
Bin1: 4, 4, 4, 15

Bin2: 21, 21, 25, 25

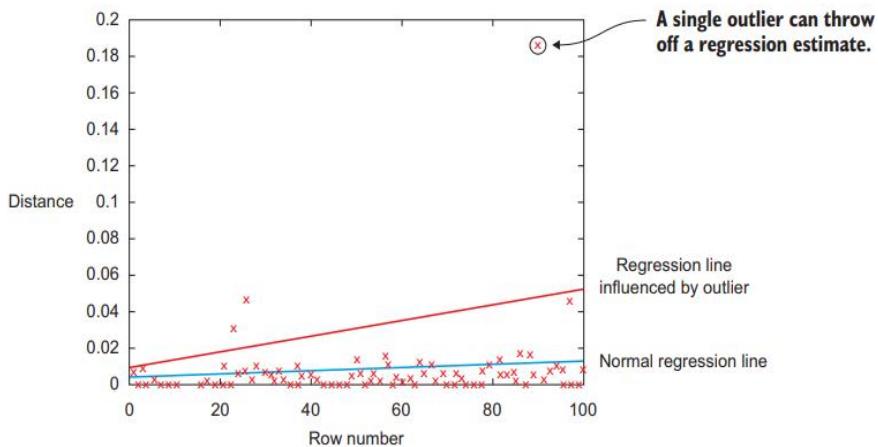
Bin3: 26, 26, 26, 34

Alternatively, bins may be equal-width or of equal-depth. Equal-width binning divides the range into N intervals of equal size. Here, outliers may dominate the result. Equal-depth binning divides the range into N intervals, each containing approximately same number of records. Here skewed data is also handled well.

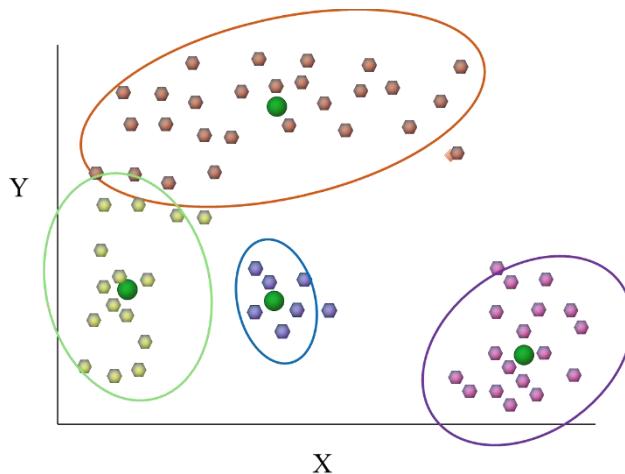
### Example: customer ages



2. **Regression** – Diagnostic plots can be insightful to find and identify data errors. We use a measure to identify data points which are outliers. We do a regression to get accustomed with the data and detect the impact of individual observations on the regression line. When a single observation has too much impact, this can point to an error in the data but it can also be a valid point.



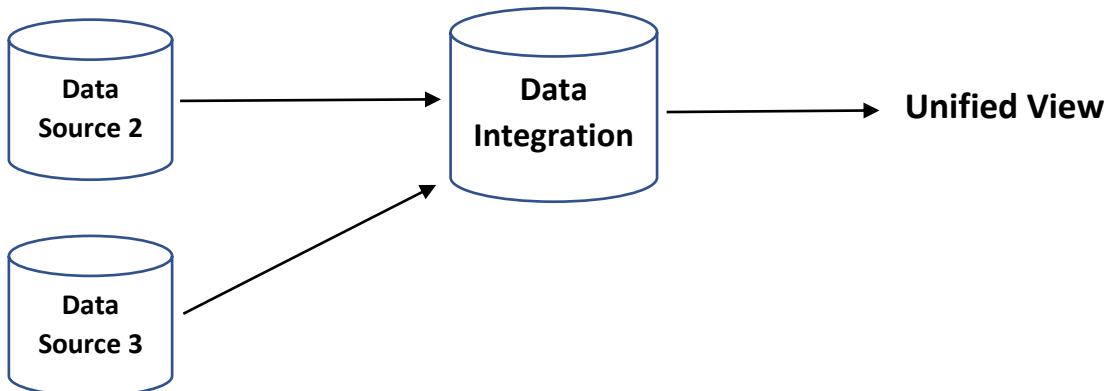
3. **Clustering** – Clustering involves grouping datapoints which exhibit similar characteristics. Datapoints which fall outside the set of clusters can be considered as outliers. The clusters should be optimized in such a way that the distance between the data points inside the cluster should be minimal and the distance among the different clusters should be as far as possible.



### 7.3 Data Integration

Data comes from diverse sources which we need to integrate. Data varies in size, type, format and structure, ranging from databases, spreadsheets, Excel files to text documents. **Data Integration technique combines data from diverse sources into a coherent data store and provides a unified view of that data.**





While performing the data integration you have to deal with several issues. Major issues faced during Data Integration are listed below -

- **Entity Identification Problem** – As the data is unified from diverse sources then how to match equivalent real-world entities. For instance, we have student data from two different sources. An entity from one source has student\_ID and the entity from other source has student\_PRN. Now, it's very difficult for a data analyst or the system to understand that these two entities actually refer to the same attribute. Here Schema Integration can be achieved using metadata of each attribute. Metadata is the data about data. Analyzing the metadata information can prevent error in schema integration. Structural integration can be achieved by ensuring that the functional dependency of an attribute in the source system and its referential constraints matches the functional dependency and referential constraint of the same attribute in the target system. For instance, suppose in one dataset discount is applied to an entire order whereas in another dataset discount is applied to every single product in the order. This discrepancy must be identified before the data from these two sources is integrated into a target system.
- **Redundancy and Correlation Analysis** – Redundant data can arise when attributes can be derived using another attribute in the data set. For instance, one data set has student age attribute and another dataset has student's date of birth attribute. Now, the age attribute is redundant as it could be derived using the date of birth. Inconsistencies in the attribute can also be one of the reasons for redundancy. Correlation Assessment can be used to determine redundancy. The attributes are analyzed to detect their interdependency on each other thereby detecting the correlation between them. Correlation is a statistical analysis method used to measure and describe the relationship between two attributes. It can also be used to determine how strong the relationship is. Correlation between attributes A and B is computed by Pearson's formula known as Correlational Coefficient.

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2} \times \sqrt{\sum(Y - \bar{Y})^2}}$$

Here  $\bar{X}$  is the mean of X attribute and  $\bar{Y}$  is the mean of Y attribute. Higher the correlation coefficient r, more strongly the attributes are correlated and one of them (either X or Y) can be discarded. If the correlation constant is 0 then the attributes are independent and if it is negative then one attribute discourages the other i.e., if value of one attribute increases then value of the other decreases.

Note that Correlation does not imply causation. It means that if there is a correlation between two attributes that does not necessarily imply that one is the cause of the other.

For discrete data, correlation between two attributes can be discovered using chi-square test  $\chi^2$ . Formula for chi-square is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Here,  $\chi^2$  is value of chi-square, O is the observed value of an attribute, E is the expected value of attributes which needs to be calculated. E represents how the attributes would be distributed if there would be NO relationship between the attributes.

For calculating the expected value E, use the following formula

$$E = \frac{M_R * M_c}{n}$$

Here, E is the expected value,  $M_R$  is the row marginal for the cell you are calculating an expected value for,  $M_C$  is the column marginal and n is the sample size.

**Example** – Suppose we have surveyed 100 people asking whether they voted during State Assembly Elections. Respondents are categorized based on their age – one category is 18 – 35 years and second is 36 to 50 years. There are also two categories of voting – Voted and Not Voted.

Below is the 2x2 contingency table which depicts two rows for age categories and two columns for voting behaviour.

	Voted	Not Voted	Total
Age 18-35	24	31	55
Age 36-50	36	9	45
Total	60	40	100

Now, let us determine the expected value for each cell,

Expected value for Age 18-35, Voted:  $E=55*60/100=33$

Expected value for Age 18-35, Not Voted:  $E=55*40/100=22$

Expected value for Age 36-50, Voted:  $E=45*60/100=27$

Expected value for Age 36-50, Not Voted:  $E=45*40/100=18$

	Voted, Expected Value	Not Voted, Expected Value	Total
--	-----------------------	---------------------------	-------

<b>Age 18-35</b>	33	22	55
<b>Age 36-50</b>	27	18	45
<b>Total</b>	60	40	100

Now, as we have the observed value and the expected value, we can easily calculate chi-square.

$$\chi^2 = [(24-33)^2 / 33] + [(31-22)^2 / 22] + [(36-27)^2 / 27] + [(9-18)^2 / 18] = \\ 2.45 + 3.68 + 3 + 4.5 = 13.63$$

Now, let's calculate the degrees of freedom

$$df = (\text{rows} - 1) * (\text{columns} - 1) = (2 - 1) * (2 - 1) = 1$$

Now, we need to use the chi-square distribution table and find the critical value at the intersection of the degrees of freedom ( $df = 1$ ) and the level of significance which is 0.01. Our critical value is 6.63 which is smaller than  $\chi^2$  value 13.63. Therefore, we can conclude that voter age and voter turnout are related. However, we cannot determine how much they are related using this test.

- **Tuple Duplication** – Information unification may also lead to duplicate tuples. These duplicate tuples are the result of denormalized tables which are used as a source for data integration.
- **Detection and resolution of data value conflicts** – During data fusion process we need to pay attention to the units of measurement of the datasets. This may be due to differences in representation, scaling or encoding. For instance, if we are to study prices of fuel in different parts of the world then some datasets may contain price per gallon and others may contain price per litre. An attribute in one system may be recorded at a lower level of abstraction than the same attribute in another. Having different levels of aggregation is similar to having different types of measurements. For instance, one dataset may contain data per week whereas the other dataset may contain data per work-week. These types of errors are easy to detect and fix.

The diversity of the data sources poses a real challenge in the data integration process. Intelligent and vigilant integration of data will definitely lead to correct insights and speedy data mining process.

## 7.4 Data Transformation

The next task after cleansing and integration is transforming your data so it takes a suitable form for data mining. When data is homogeneous and well-structured, it is easier to analyze and look for patterns.

Data transformation involves the following –

- **Smoothing** – It is a process which is used to remove noise from the dataset by applying certain algorithms. It helps in predicting patterns from the dataset. Smoothing techniques include binning, regression and clustering.
- **Aggregation** – Aggregation is the method of storing, analyzing and presenting the data in a report or summary format. Telecom companies collect data about their customers. This gives them an idea about customer demographics and calling patterns. This aggregated data assists them in designing customized offers and discounts.
- **Generalization** – Here, low-level attributes are transformed into high-level attributes using concept hierarchies. This transformation from a lower level to a higher level is beneficial in getting a clear picture of the data. For instance, age data can be in the form of (20, 30, 40...) in a dataset. It can be transformed into a higher level into a categorical value (young, middle-aged, old). Data generalization can be divided into two approaches – Data cube process (OLAP) and Attribute oriented induction approach (AOI).
- **Normalization** – Data is transformed so that it falls under a given range. **The popular normalization methods are Min-Max normalization, Z-Score normalization and Decimal Scaling.** Note that Z-Score normalization and Decimal Scaling can change the original data slightly.

1. **Min-Max Normalization** – This method transforms the original data linearly. Suppose  $\min_F$  and  $\max_F$  are the minimum and maximum values of an attribute F. This method maps a value v of F to  $v'$  in the range [new\_min<sub>F</sub>, new\_max<sub>F</sub>] using the following formula

$$v' = \frac{v - \min_F}{\max_F - \min_F} (\text{new\_max}_F - \text{new\_min}_F) + \text{new\_min}_F$$

**Example** - Suppose the minimum and maximum value for an attribute profit are Rs10,000 and Rs1,00,000. We want the profit in the range of [0,1]. Using the above formula value Rs20,000 for attribute profit can be plotted to

$$v' = \frac{(20000 - 10000)}{(100000 - 10000)} (1 - 0) + 0 = 0.11$$

Min-Max Normalization preserves the relationships among the original data values. It can easily flag a ‘out-of-bounds’ error if a future input case falls outside of the original data range for the attribute.

2. **Z-Score Normalization** – This is used when actual minimum and maximum value of an attribute are unknown or when there are outliers. Here, the values for an attribute are normalized based on mean ad standard deviation.

$$z = \frac{x - \text{mean}(x)}{\text{stddev}(x)}$$

**Example** – Suppose mean of an attribute = 60000 and standard deviation = 10000. A value of 85000 for attribute can be transformed to

$$z = \frac{85000 - 60000}{10000} = 2.50$$

- 3. Decimal Scaling** – It normalizes the values of an attribute by changing the position of their decimal points. The number of points by which the decimal point is moved can be determined by the absolute maximum value of attribute. Decimal Scaling formula is

$$v'_i = \frac{v_i}{10^j}$$

Here,  $v_i$  value is normalized to  $v'_i$ ,  $j$  is the smallest integer such that  $\text{Max}(|v'|) < 1$

**For instance,** Values of an attribute varies from -99 to 99. The maximum absolute value is 99. For normalizing the values, we divide the numbers by 100 (i.e.,  $j=2$ ) so that values come out to be as 0.98, 0.97 and so on.

- **Attribute Construction** – Here, new attributes are created from an existing set of attributes. Attribute construction can discover missing information about relationships between data attributes which can be important for knowledge discovery.

## 7.5 Summary

Data Preprocessing is an extremely important step in mining as real-world data is usually noisy and inconsistent. Data Preprocessing involves data cleaning, data integration, data transformation and data reduction.

Data Cleaning deals with missing values in the dataset, sanitizes the data by removing the noise, identifies the outliers and remedies the inconsistencies.

Data Integration combines data from varied sources. Metadata, correlation analysis, data conflict detection contribute towards smooth data integration.

Data Transformation methods transforms the data into required forms for mining.

## 7.6 Exercises

1. Mostly in real-world data, tuples have missing values. What are the various methods to deal with missing values?
  2. What are the various issues to be considered w.r.t data integration?
  3. What are the normalization methods? Also explain the value ranges for each of them.
  4. Suppose a group of 12 sales price records has been sorted as follows – 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215  
Partition them into 3 bins by using the following methods –
    - a) Equal-width partitioning
    - b) Equal-frequency partitioning
    - c) Clustering
- 

## 7.7 References

- Dunham, Margaret H, Data mining: Introductory and advanced topics, Pearson Education India, 2006.
- Han, Jiawei, Jian Pei, and Micheline Kamber, Data mining: concepts and techniques, Second Edition, Elsevier, Morgan Kaufmann, 2011.

## **Module 3 - DATA MINING AND PREPROCESSING**

### **Chapter 8 Data Reduction**

---

#### **8.0 Objectives**

This chapter provides an overview of the following –

- What is Data Reduction?
- Data Cube Aggregation
- Dimensionality Reduction
- What is Data Compression?
- Numerosity reduction
- Data discretization and Concept hierarchy

---

#### **8.1 Introduction to Data Reduction**

In real world we usually deal with big data. So, it takes a long time for analyzing and mining this big data. In some case it may not be practically feasible to analyze such a huge amount of data. **Data reduction method results in a simplified and condensed description of the original data that is much smaller in size/quantity but retains the quality of the original data.** The strategy of data reduction decreases the sheer volume of data but retains the integrity of the data. Analysis and mining on such smaller data is practically feasible and mostly results in same analytical outcomes.

Data Reduction methods are given below –

1. **Data Cube Aggregation** – Here, data is grouped in a more manageable format. Mostly it is a summary of the data. Aggregation operations are applied to the data in the construction of a data cube.

**For instance**, suppose we have quarterly average of company sales from year 2015 to 2020. Now, instead of quarterly average we require yearly sales. So, we can summarize the data in such a way that we get cumulative sales each year instead of quarterly average. The resulting dataset is reduced in volume but its quality or integrity is not compromised as no data is lost.

Data cube is a multi-dimensional architecture. For simple data analysis, data cubes are widely used. Data cubes provide fast access to precomputed, summarized data, thereby benefitting on-line analytical processing (OLAP) as well as data mining.

2. **Attribute Subset Selection** – Here, irrelevant and redundant attributes are detected and removed to form the core attribute set. This process reduces the data volume and dimensionality. The main objective of this process is to determine minimum set of

attributes such that removing irrelevant attributes will not affect usability of the data. Moreover, it also reduces the cost of data analysis.

Mining on a reduced data set may also make the discovered pattern easier to understand. The question arises about how to select the attributes to be removed. Statistical significance tests are used so that such attributes can be selected.

Following methods are used for attribute subset selection –

- **Stepwise Forward Selection** – This method starts with an empty set of attributes as the minimal set. The most relevant attribute is chosen and added to the minimal set. At each iteration, the most relevant attribute from the remaining original attributes is selected and added to the minimal set.
- **Stepwise Backward Elimination** – Here, all the original attributes are considered in the initial set of attributes. In each iteration, the worst or irrelevant attribute is removed from the set.
- **Combination of Forward Selection and Backward Elimination** – This is the most common method used for attribute selection. As the name itself suggests this method combines Forward Selection and Backward Selection to select the most relevant attributes efficiently.
- **Decision Tree Induction** – Here, a flowchart like structure with nodes is constructed. Nodes denote a test on an attribute. Each branch corresponds to the result of the test and leaf nodes is a class prediction. The attribute which does not become a part of the tree is considered to be irrelevant and therefore discarded.

3. **Dimensionality Reduction** – In this method number of dimensions over which the data is spread across is reduced. With reduced dimensionality, it is easy to visualize and manipulate the data. Handling the high-dimensional data is very difficult in practice. There are two types of Dimensionality reduction –

- **Wavelet Transform** – Here, data vector X is transformed into another vector Y, in such a way that both the vectors are of same length. Now, the question arises that if both the vectors are of same length then how it can be useful for data reduction? The result of the wavelet transform can be truncated thus accomplishing dimensionality reduction. A small compressed approximation of the data can be retained by storing only a small fraction of the strongest wavelet coefficients.

**For example**, retain all wavelet coefficients larger than some particular threshold and the remaining coefficients are set to 0. The general procedure for applying a discrete wavelet transform uses a hierarchical pyramid algorithm that halves the data in each iteration, resulting in fast computational speed.

The method is as follows –

- a) The length, L, of the input data vector must be a power of 2. This condition can be satisfied by adding extra zeros to the data vector if required.

- b) Each transform involves applying two functions. The first applies some data smoothing such as sum or weighted average. The second performs a weighted difference which brings out the detailed features of the data.
- c) The two functions are recursively applied to the sets of data obtained in the previous iteration, until the resulting dataset obtained is of length 2.
- d) Selected values from the data sets obtained in the above iterations are designated the wavelet coefficients of the transformed data.

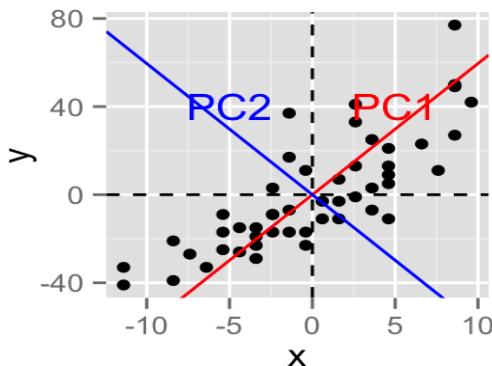
Wavelet transforms are well suited for data cube, sparse data or data which is highly skewed. Wavelet transform is often used in image compression, computer vision, analysis of time-series data and data cleansing.

- **Principal Components Analysis (PCA)** – It is a statistical process which transforms the observations of the correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called **Principal Components**.

PCA combines the essence of attributes by creating an alternative, smaller set of variables. The original data can now be projected onto this smaller set. So many a times PCA reveals relationships that were not previously detected or identified.

The basic process of PCA is as follows –

- a) For each attribute to fall within the same range, the input data is normalized. So, attributes with large domains do not dominate attributes with smaller domains.
- b) PCA computes  $k$  orthonormal vectors that provide a base for the normalized input data. These unit vectors are called as Principal Components. The input data is a linear combination of the principal components.
- c) Principal components are sorted in order of decreasing strength. The principal components essentially serve as a new set of axes for the data, by providing important information about variance. Observe in the below figure the direction in which the data varies the most actually falls along the red line. This is the direction with the most variation in the data. So, it's the first principal component (PC1). The direction along which the data varies the most out of all directions that are uncorrelated with the first direction is shown using the blue line. That's the second principal component (PC2).



- d) As the components are sorted based on decreasing order of strength, the size of data can be reduced by eliminating the weaker components. The weaker components will be with low variance. It is possible to reconstruct a good approximation of the original data with the help of strongest principal components.

PCA can also be used for finding hidden patterns if data has high dimensions. Some fields where PCA is used is AI, Computer vision and image compression.

- **Numerosity Reduction** – Numerosity Reduction is a data reduction technique which replaces the original data by smaller form of data representation to reduce the volume of data. These techniques can be **parametric** or **nonparametric**. For Parametric methods, data is represented using some model. The model is used to estimate the data so that only parameters of data are required to be stored instead of actual original data. Regression and Log-Linear methods are used for creating such models. Non-Parametric methods are used to store reduced representations of the data. These methods include histograms, clustering, sampling and data cube aggregation.

#### **Parametric Methods –**

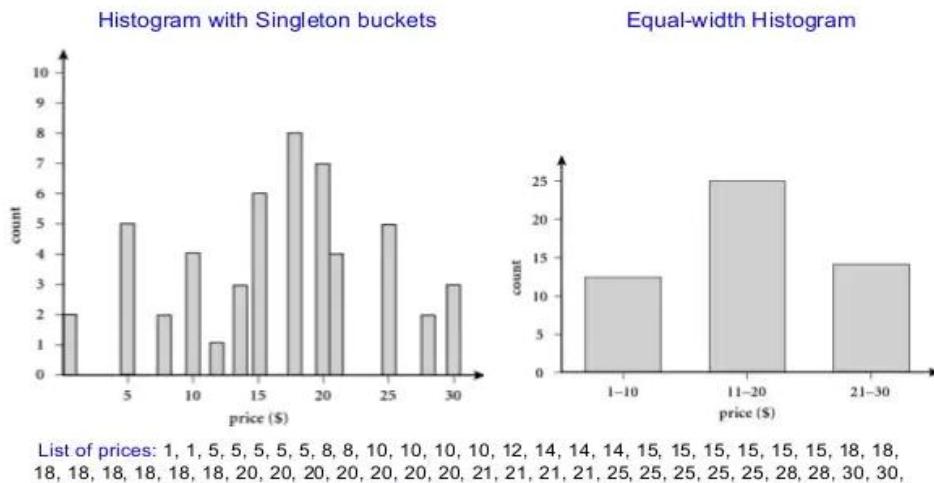
- a) **Regression** – Regression is of two types – Simple Linear regression and Multiple Linear regression. Regression model is Simple Linear regression when there is only single independent attribute, however, if there are multiple independent attributes then the model is Multiple Linear Regression. In Simple Linear regression the data are modelled to fit a straight line.

**For example,** a random variable ‘y’ can be modelled as a linear function of another random variable ‘x’ with the equation  $y=ax+b$  where  $a$  represents the slop of the line and  $b$  represents the y-intercept of the line. With reference to data mining,  $x$  and  $y$  are numerical database attributes. In Multiple Linear regression,  $y$  will be modelled as a linear function of two or more independent variables.

- b) **Log-Linear Model** – Log-Linear models approximate discrete multidimensional probability distributions. The method can be used to estimate the probability of each cell in a base cuboid for a set of discretized attributes, based on smaller cuboids making up the data cube lattice.

## Non-Parametric Methods

- a) **Histograms** – It is popular method of data reduction. Here, data is represented in terms of frequency. It uses binning to approximate data distribution. A Histogram for an attribute partitions the data distribution of the attribute into disjoint subsets or buckets. The buckets are displayed on a horizontal axis while the height and area represent the average frequency of the values depicted by the bucket. Many a times buckets represent continuous ranges for the given attribute.



There are several partitioning rules for buckets –

- **Equal-width** – Here, the width of each bucket range is same.
- **Equal-frequency** – Here, the bucket is created in such a way so that the number of contiguous data samples in each bucket are roughly the same.
- **V-optimal** – It is based on the concept of minimizing a quantity called as weighted variance. This method of partitioning does a better job of estimating the bucket contents. V-optimal Histogram attempts to have the smallest variance possible among the buckets.
- **Max-Diff** – Here, we consider the difference between each pair of adjacent values. A bucket boundary is established between

each pair for pairs having the  $\beta - 1$  largest differences, where  $\beta$  is the user specified number of buckets.

- b) **Clustering** – Clustering divides the data tuples into groups/clusters.

Partitioning is done in such way that objects within a cluster are similar to one another and are different to objects from other clusters. IN many fields it is beneficial to group together similar objects.

**For instance**, in a financial application we might be interested to find clusters of organizations who have similar financial performance. In medical application we might be interested to find clusters of patients with similar symptoms. It is easier to visualize clusters in two dimensions. Centroid of a cluster is the point (sometimes imaginary) for which each attribute value is the average of the values of the corresponding attribute for all the points in the cluster.

So, the centroid of four points with 6 attributes

8.0	6.2	0.5	24	11.1	-6.2
2.0	-3.5	0.9	24.2	17.3	-5.1
-3.6	8.1	0.8	20.6	10.3	-7.2
-6.0	6.7	0.7	12.7	9.2	-8

would be

0.1	4.38	0.73	20.38	11.98	-6.63
-----	------	------	-------	-------	-------

The Diameter of a cluster is the maximum distance between any two points of the cluster. The quality of the cluster depends upon the diameter of the cluster. We may merge those clusters whose resulting cluster has the lowest diameter. In data reduction, the cluster representations of the data are used to replace the original data.

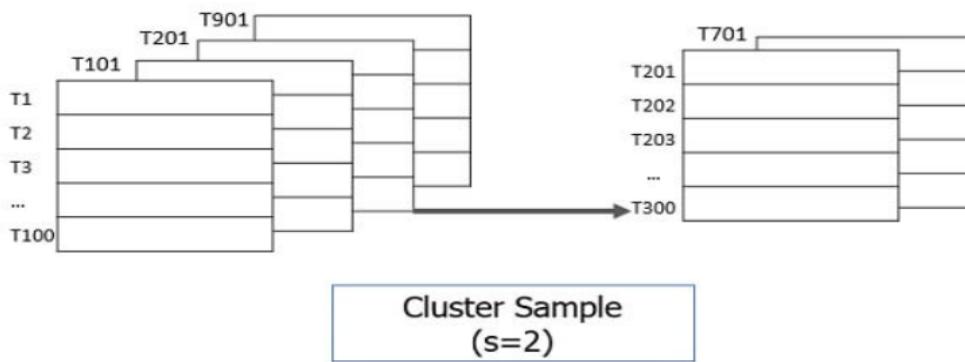
- c) **Sampling** – Sampling is one of the data reduction method which reduces the large dataset into a much smaller data sample.

There are many methods using which we can sample a large data set D containing N tuples –

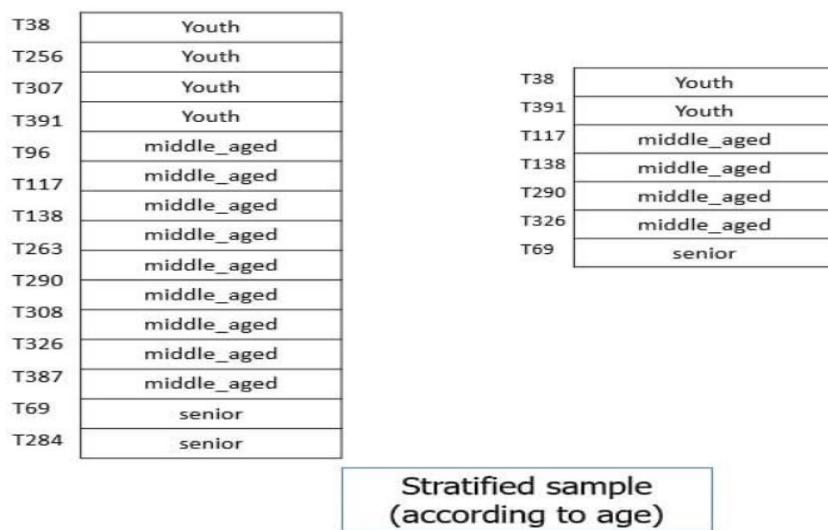
- **Simple Random Sample Without Replacement of size (SRSWOR)** – Here, ‘S’ number of tuples are drawn from N tuples such that  $S < N$  in the dataset D. The probability of drawing any tuple from the dataset D is  $1/N$  which means all the tuples have an equal chance of getting selected in the sample.
- **Simple Random Sample with Replacement of size (SRSWR)** – It is similar to SRSWOR but the tuple is drawn from dataset D, is recorded and then replaces back into the dataset D so that it can be drawn again.



- **Cluster Sample** – The tuples in the dataset D are clustered into M mutually disjoint subsets. From these clusters, a simple random sample of size S could be drawn where S<M. Data reduction can be implemented by using SRSWOR on these clusters.



- **Stratified Sample** – Here, the large dataset D is partitioned into mutually disjoint sets called ‘strata’. Now a simple random sample is taken from each stratum to get stratified data. This method is useful for skewed data.



## 8.2 Introduction to Data Discretization and Concept Hierarchy Generation

---

**Data Discretization is a method of translating attribute values of continuous data into a finite set of intervals with minimal information loss.** It is process of translating continuous data into intervals and then assigning specific label to each interval. Interval labels can then be used to replace actual data values. Replacing large number of continuous data values with small number of labels, reduces and simplifies the actual data.

For instance, suppose we have an attribute age with following data values

### Before Discretization

<b>Age</b>	10,11,13,14,17,19,30,31,32,38,40,42,70,72,73,75
------------	---

### After Discretization

Attribute	Age	Age	Age
	10,11,13,14,17,19	30,31,32,38,40,42	70,72,73,75
Interval Label	Young	Middle-Aged	Senior

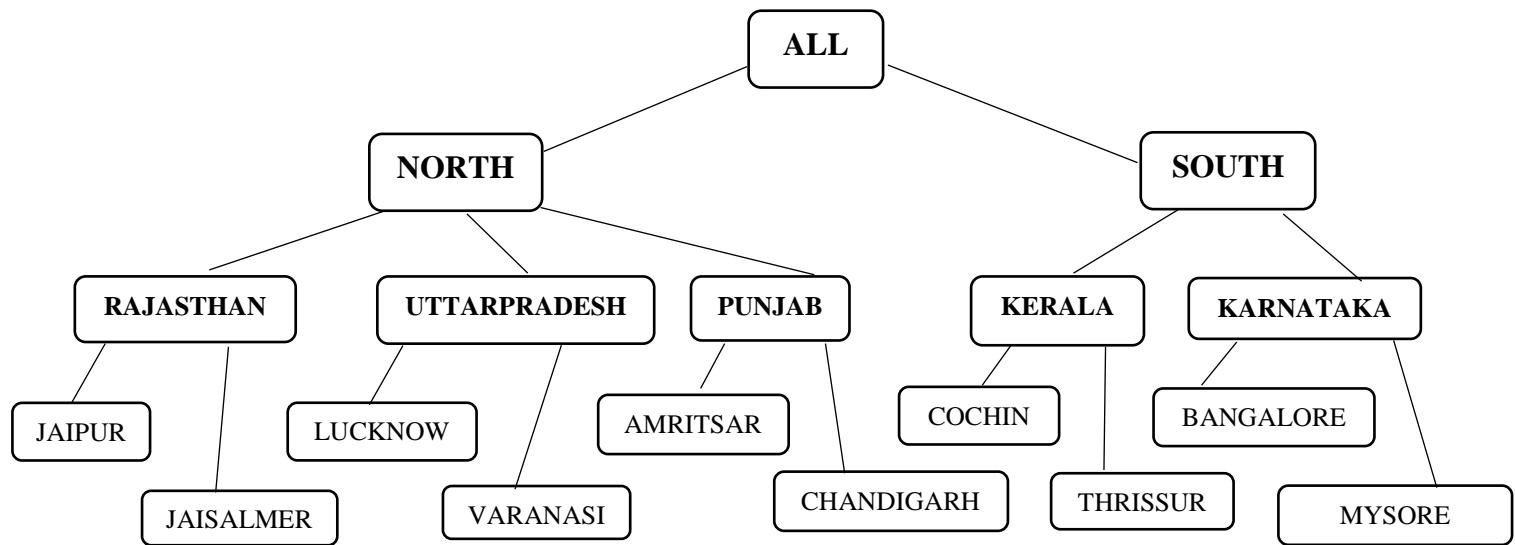
Discretizing techniques can be categorized as follows –

- **Discretization of the top-down** – Here, the procedure begins by first determining one or few points to divide the whole set of attributes, called as split points and then performs this recursively at the resulting intervals.
- **Discretization from Bottom-up** – Here, the procedure begins by considering all the continuous values as possible split points. Then it discards few by combining neighbourhood values to form intervals. This process is then recursively applied to resulting intervals.

**Discretization can be performed recursively on an attribute to provide a hierarchical or multi resolution partitioning of the attribute values known as Concept Hierarchy.**

For instance, Consider a concept hierarchy for the dimension ‘location’. City values for location include Jaipur, Jaisalmer, Amritsar, Chandigarh, Cochin, Thrissur, Lucknow, Varanasi, Bangalore, Mysore. Each city can be mapped to the state to which it belongs. Each state can then be mapped to province (North or South) to which they belong. These mappings form a concept hierarchy for the dimension ‘location’ mapping a set of low-level concepts (i.e. cities) to higher level, more general concepts (i.e. provinces).

The below given diagram shows this concept hierarchy.



There may be more than one concept hierarchy for a given attribute or dimension, based on different user viewpoints.

### 8.3 Data Discretization and Concept Hierarchy Generation for Numerical Data

Following are the methods which can be used –

- **Binning** – We have already discussed about this method as a part of Data smoothing. Binning can be used for data discretization and also for creation of idea hierarchy. Attribute values are grouped together into a number of equal-frequency bins or equal-width bins. Then bin mean or bin median is used for smoothing the values. This process can be recursively used for generating concept hierarchy. As binning does not use class information, it is an unsupervised discretization.
- **Histogram Analysis** – We have already discussed about Histograms as a part of Data Reduction techniques. Histograms partitions the values for an attribute into disjoint ranges called as buckets. We have already discussed Equal-width histogram and Equal-frequency Histogram the partitioning techniques. To automatically generate multilevel concept hierarchy, histogram analysis algorithm is applied recursively to each partition. The process is curtailed once the desired number of concept levels are reached.
- **Entropy-based Discretization** – Entropy is one of the most popularly used discretization technique. We always want to make meaningful splits or partitions in our continuous data. Entropy-base discretization helps to split our data at points where

we will gain the most insights when we give it to our data mining systems. Entropy describes how consistently a potential split will match up with a classifier. Lower entropy is better and a Zero value entropy is the best.

**For instance**, suppose we have data about income of people below 25 years of age.

	Income<=50000	Income>50000
Age<25	4	6

The above data will result in a high entropy value, almost closer to 1. Based on the above data we cannot be sure that if a person is below 25 years of age, then he will have income greater than 50000. Because data indicates that only 6/10 make more than 50000 and the rest makes below it. Now, let's change our data values.

	Income<=50000	Income>50000
Age<25	9	1

Now, this data will give a lower entropy value as it provides us more information on relation between age and income.

Now, let's see how we calculate the entropy value.

$$Entropy(D_1) = - \sum_{i=1}^m p_i \log_2 p_i$$

Here, m is the number of classifier values. In our example value of m=2 as we have 2 options: income<=50000 and income>50000. 'p' is the probability of getting specific classifier given the bin you are looking at. Now, lets just calculate entropy values for the above two table values.

In the first table we had 4 samples below 50000 income and 6 above 50000.

$$\text{Entropy (Age}<\text{25)} = - \left( \frac{4}{10} \log_2 \left( \frac{4}{10} \right) + \frac{6}{10} \log_2 \left( \frac{6}{10} \right) \right) = 0.529 + 0.442 = 0.971$$

Now, let's calculate for second table.

$$\text{Entropy (Age}<\text{25)} = - \left( \frac{9}{10} \log_2 \left( \frac{9}{10} \right) + \frac{1}{10} \log_2 \left( \frac{1}{10} \right) \right) = 0.137 + 0.332 = 0.469$$

So, if you observe entropy value moved from 0.971 to 0.469. We would have 0 entropy value if we had 10 in one category and 0 in the other category.

**Entropy-based discretization performs the following algorithm –**

1. Calculate Entropy value for your data.
2. For each potential split in your data
  - o Calculate Entropy in each potential bin
  - o Find the net entropy for your split
  - o Calculate entropy gain
3. Select the split with highest entropy gain

4. Recursively perform the partition on each split until a termination criterion is met.  
 Terminate once you have reached a specified number of bins or terminate once the entropy gain falls below a certain limit.

We want to perform splits which improve the insights we get from our data. So, we want to perform splits that maximize the insights we get from our data. Entropy gain measures that. So, we need to find and maximize entropy gain to perform splits.

We can calculate net entropy using the following equation.

$$Info_A(D) = \frac{|D_1|}{|D|} Entropy(D_1) + \frac{|D_2|}{|D|} Entropy(D_2)$$

The formula indicates that our information across two bins is equal to the ratio of the bin's size multiplied by that bin's entropy.

**Consider the following example.** Here, we have data regarding number of hours studied (continuous variable) and scored grade A achieved in test (classifier).

Hours Studied	Scored Grade A in test
4	N
5	Y
8	N
12	Y
15	Y

We will discretize the above given data by first calculating entropy of the data set.

	Scored Grade A	Scored Grade lower than A
Overall students	3	2

$$\text{Entropy } (D) = - \left( \frac{3}{5} \log_2 \left( \frac{3}{5} \right) + \frac{2}{5} \log_2 \left( \frac{2}{5} \right) \right) = 0.529 + 0.442 = 0.971$$

Now, we will iterate through and see which splits give us the maximum entropy gain. To find a split, we average two neighbouring values in the list.

### Split 1: 4.5

4 and 5 are the neighbouring values in the list. Suppose we split at  $(4+5)/2 = 4.5$

Now we get 2 bins as follows –

	Scored grade A	Scored lower than A
$\leq 4.5$	0	1
$> 4.5$	3	1

Now, we need to calculate entropy for each bin and find the information gain of this split.

$$\text{Entropy } (D_{\leq 4.5}) = - \left( \frac{1}{1} \log_2 \left( \frac{1}{1} \right) + \frac{0}{1} \log_2 \left( \frac{0}{1} \right) \right) = 0 + 0 = 0$$

$$\text{Entropy } (D_{>4.5}) = - \left( \frac{3}{4} \log_2 \left( \frac{3}{4} \right) + \frac{1}{4} \log_2 \left( \frac{1}{4} \right) \right) = 0.311 + 0.5 = 0.811$$

$$\text{Net entropy is } \text{Info}_A(D_{\text{new}}) = \frac{1}{5}(0) + \frac{4}{5}(0.811) = 0.6488$$

$$\text{Entropy gain is } \text{Gain}(D_{\text{new}}) = 0.971 - 0.6488 = 0.322$$

**Split 2: 6.5** – Average of next 2 values is  $(5+8)/2 = 6.5$ . Now, we will repeat the above procedure which we carried for Split1.

	<b>Scored grade A</b>	<b>Scored lower than A</b>
<b><math>\leq 6.5</math></b>	1	1
<b><math>&gt; 6.5</math></b>	2	1

$$\text{Entropy } (D_{\leq 6.5}) = 1$$

$$\text{Entropy } (D_{>6.5}) = 0.917$$

$$\text{Net entropy is } \text{Info}_A(D_{\text{new}}) = 0.944$$

$$\text{Entropy Gain is } \text{Gain}(D_{\text{new}}) = 0.971 - 0.944 = 0.27$$

This is less gain than we had in the earlier split (0.322) so our best split is still at 4.5. Let's check the next split at 10.

**Split3: 10** – Average of next 2 values is  $(8+12)/2 = 10$ . Now we will repeat the procedure.

	<b>Scored grade A</b>	<b>Scored lower than A</b>
<b><math>\leq 10</math></b>	1	2
<b><math>&gt; 10</math></b>	2	0

$$\text{Entropy } (D_{\leq 10}) = 0.917$$

$$\text{Entropy } (D_{>10}) = 0$$

$$\text{Net entropy is } \text{Info}_A(D_{\text{new}}) = 0.55$$

$$\text{Entropy Gain is } \text{Gain}(D_{\text{new}}) = 0.971 - 0.55 = 0.421$$

This is the maximum gain value we have as compared to the earlier splits (0.322 and 0.27).

**Split 4: 13.5** – This split will also result in lower entropy gain.

**Conclusion** – So, now after calculating entropy gains for various splits, we conclude that the best split is Split3. So, we will partition the data at 10.

The entropy and information gain measures are also used for decision tree induction.

- **Interval Merging by  $\chi^2$  Analysis – Chi merge is a simple algorithm which uses  $\chi^2$  (chi-square) statistic to discretize numeric attributes.** It is a supervised bottom-up data discretization technique. Here, we find the best neighbouring intervals and merge them to form larger intervals. This process is recursive in nature. The basic idea is that for accurate discretization, the relative class frequencies should be fairly consistent within an interval. If two adjacent intervals have a very similar distribution of classes, then the intervals can be merged. Otherwise, they should remain separate. It treats intervals as discrete categories. Initially, in the Chi Merge method each distinct value of a numerical attribute A is considered to be one interval.  $\chi^2$  test is performed for every pair of adjacent intervals. Adjacent intervals with the least  $\chi^2$  values are merged together, since low  $\chi^2$  values for a pair indicate similar class distributions. This merge process proceeds recursively until a predefined stopping criterion is met like significance level, max-interval, max inconsistency etc.
- **Cluster Analysis –** It is a popular data discretization technique. A clustering algorithm is applied to discretize a numerical attribute ‘A’ by partitioning the values of A into clusters. Clustering can produce high quality discretization result. By using the top-down splitting strategy or a bottom-up merging strategy, Clustering can be used to generate a concept hierarchy of A. Here, each cluster will form a node of the concept hierarchy. In the top-down approach, each cluster may be further decomposed into several subclusters forming a lower level of hierarchy. In the bottom-up approach, clusters are formed by repeatedly grouping neighbouring clusters in order to form higher-level concepts.
- **Intuitive Partitioning –** Intuitive partitioning or Natural partitioning is one of the easiest way of data discretization. **The 3-4-5 rule can be used to segment numerical data into relatively uniform intervals.** Here, partitions are created at 3, 4 or 5 relatively equal-width intervals, recursively and level by level, based on the value at the most significant digit (MSD).

The 3-4-5 rule is as follows –

- If an interval covers 3, 6, 7 or 9 distinct values at most significant digit, then create 3 intervals. Here, there can be 3 equal-width intervals for 3,6,9; and 3 intervals in the grouping of 2-3-2 each for 7.
- If it covers 2,4 or 8 distinct values at most significant digit, then create 4 sub-intervals of equal-width.
- If it covers 1,5 or 10 distinct values at the most significant digit, then partition the range into 5 equal-width intervals.

**For instance,** breaking up annual salaries in the range of into ranges like 50000 – 100000 are often more desirable than ranges like 51263 – 98765.

**Example** - Assume that we have records showing profits made in each sale throughout a financial year. Profit data range is -351976 to 47000896. (Negative profit value indicates loss.)

Now, we need to perform data smoothing so we will discard 5% data values from top and bottom of the dataset. This may avoid noisy data.

Suppose after discarding 5% of the data values, our new values for Low = -159876 and High = 183876. Observe that here MSD is at million position.

Now, we will round down the Low and High at MSD. So Low = -1000000 (rounding down -159876 to nearest million gives -1000000) and High = 2000000 (rounding 183876 to nearest million gives 2000000). So here Range is 2000000 – (-1000000) = 3000000. We will consider only MSD here, so range of this interval is 3.

Now, as the interval covers 3 distinct values at MSD, we will divide this interval into 3 equal-width size intervals.

Interval 1: (-1000000 to 0]

Interval 2: (0 to 1000000]

Interval 3: (1000000 to 2000000]

Observe the notation (a to b] in above intervals – it denotes the range that excludes ‘a’ but includes ‘b’.

Further, we can apply 3-4-5 rule recursively to each interval creating a concept hierarchy.

---

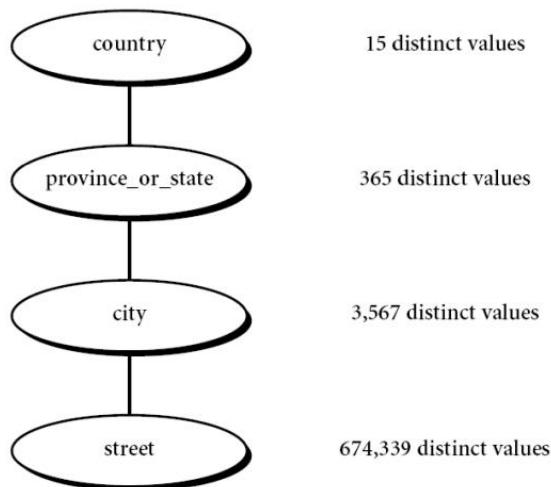
## 8.4 Concept Hierarchy Generation for Categorical Data

---

Generalization is the generation of concept hierarchies for categorical data. Categorical attributes have a finite number of distinct values, with no ordering among the values. Geographic location, product type, job category etc. can be its few examples.

Methods for generation of concept hierarchies for categorical data are as follows –

- **Specification of a partial ordering of attributes explicitly at the schema level by users or experts** – A user can easily define a concept hierarchy by specifying ordering of the attributes at schema level.  
**For instance**, dimension ‘location’ may contain a group of attributes like street, city, state and country. A hierarchy can be defined by specifying the total ordering among these attributes at schema level such as  
 Street < City < State < Country
- **Specification of a portion of a hierarchy by explicit data grouping** – We can easily specify explicit groupings for a small portion of intermediate-level data.  
**For instance**, after specifying that state and country form a hierarchy at schema level, a user can define some intermediate levels manually such as  
 {Jaisalmer, Jaipur, Udaipur} < Rajasthan
- **Specification of a set of attributes, but not of their partial ordering** – A user can specify a set of attributes forming a concept hierarchy, but may not explicitly state their partial ordering. The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept hierarchy. A concept hierarchy can be automatically generated based on the number of distinct values per attribute in the given attribute set. The attribute with the most distinct values is placed at the lowest level in the hierarchy. The lower the number of distinct values an attribute has, the higher it is placed in the hierarchy. You can observe in the below given diagram ‘street’ is placed at the lowest level as it has largest number of distinct values.



- **Specification of only a partial set of attributes** – Many a times user has only a vague idea about what should be included in a hierarchy. So, he/she may include only a subset of relevant attributes in the hierarchy specification. For instance, a user may include only street and city in the hierarchy specification of dimension ‘location’. To handle such partially defined hierarchies, it is suggested to embed data semantics in the database schema so that attributes with tight semantic connections can be pinned together.

## **8.5 Summary**

---

Data Reduction methods include data cube aggregation, attribute subset selection, dimensionality reduction, numerosity reduction and discretization can be used to obtain a reduced representation of the data while minimizing the loss of information and its quality.

Data discretization and generation of concept hierarchies for numerical data consists of methods like binning, histogram analysis, entropy-based discretization, chi-square analysis, cluster analysis and natural partitioning. Concept hierarchy for categorical data can be generated based on the number of distinct values of the attributes defining the hierarchy.

---

## **8.6 Exercises**

---

1. Which method according to you is the best method for data reduction?
  2. Write a short note on data cube aggregation.
  3. What is concept hierarchy in data mining?
  4. How concept hierarchy is generated for numeric data?
  5. How concept hierarchy is generated for categorical data?
- 

## **8.7 References**

---

- Dunham, Margaret H, Data mining: Introductory and advanced topics, Pearson Education India, 2006.
- Han, Jiawei, Jian Pei, and Micheline Kamber, Data mining: concepts and techniques, Second Edition, Elsevier, Morgan Kaufmann, 2011.

# **Unit 4**

## **Chapter 9: Association Rules**

- 9.0 Objectives
- 9.1 Introduction
- 9.2 Association Rule Mining
- 9.3 Support and Confidence
  - 9.3.1 Support
  - 9.3.2 Confidence
  - 9.3.3 Lift
- 9.4 Frequent Pattern Mining
  - 9.4.1 Market Basket Analysis
  - 9.4.2 Medical Diagnosis
  - 9.4.3 Census Data
  - 9.4.4 Protein Sequence
- 9.5 Market Basket Analysis
  - 9.5.1 Implementation of MBA
- 9.6 Apriori Algorithm
  - 9.6.1 Apriori Property
  - 9.6.2 Steps in Apriori
  - 9.6.3 Example of Apriori
  - 9.6.4 Apriori Pseudo Code
  - 9.6.5 Advantages and Disadvantages
  - 9.6.6 Method to Improve Apriori Efficiency
  - 9.6.7 Applications of Apriori
- 9.7 Associative Classification- Rule Mining
  - 9.7.1 Typical Associative Classification Methods
  - 9.7.2 Rules for Support and confidence in Associative Classification
- 9.8 Conclusion
- 9.9 Summary
- 9.10 References

---

## **9.0 Objectives**

---

In this chapter we will describe a class of unsupervised learning models that can be used when the dataset of interest does not include a target attribute. These are methods that derive association rules the aim of which is to identify regular patterns and recurrences within a large set of transactions. They are fairly simple and intuitive and are frequently used to investigate sales transactions in market basket analysis and navigation paths within websites.

Association rule mining represents a data mining technique and its goal is to find interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored in databases, many companies are becoming interested in mining association rules from their databases to increase their profits.

The Main objective of data mining is to find out the new, unknown and unpredictable information from the used database, which is useful and helps in decision making. There are a number of techniques used in data mining to identify the frequent pattern and mining rules includes clusters analysis, anomaly detection, association rule mining etc. In this Chapter we provide an overview of association rule research.

---

## **9.1 Introduction**

---

Data Mining is the discovery of hidden information found in databases and can be viewed as a step in the knowledge discovery process. Data mining functions include clustering, classification, prediction, and link analysis. One of the most important data mining applications is that of mining association rules.

Association rules, first introduced in 1993 and are used to identify relationships among a set of items in a database. These relationships are not based on inherent properties of the data themselves (as with functional dependencies), but rather based on co-occurrence of the data items. In Data mining significant data dig out from a huge database repository. It is the progression of selection of significant information from big amount of data by using convinced sophisticated algorithms. Data mining is becoming a gradually more vital tool to convert data into valuable information which facilitate in decision making.

It is the most important and deeply considered functions of mining of data. It not only provides a well-organized method of discovering the patterns and recognition of the model but also verifies the rule that exist, which ultimately helps in providing new rules. Association rules provide the effective scientific base for decision making. Rule of associations have been used in many applications to find frequent patterns in data. One of the key domains which use association rule is business field where it helps in a very effective and efficient decision making and marketing. Other field areas where association rule mining can be applied are market basket analysis, medical diagnosis, census data, fraud detection in web and DNA data analysis etc.

For example, in the field of electrical power engineering, the methods of data mining have been used for the condition monitoring of high voltage equipment. In medical field ARM is used to find frequently occur diseases in particular area and to diagnose different diseases. It is also used to attain information about the navigational activities of users in web Log data. Recently it is discovered that there are various algorithms for finding the association rules. For frequent pattern mining different frameworks have been defined.

---

## 9.2 Association rule mining

---

Association Rule Mining, as the name suggests, association rules are simple If/Then statements that help discover relationships between seemingly independent relational databases or other data repositories. Most algorithms work with numeric datasets and hence tend to be mathematical. However, association rule mining is suitable for non-numeric, categorical data and requires just a little bit more than simple counting.

Association rule mining is a procedure which aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases, and other forms of repositories. Main purpose of Association rule mining is,

- Finding frequent patterns, associations, correlations, or causal structures among sets of items in transaction databases.
- Understand customer buying habits by finding associations and correlations between the different items that customers place in their “shopping basket”.

**Association Rule Mining is one of the ways to find patterns in data. It finds:**

- Features (dimensions) which occur together.
- Features (dimensions) which are “correlated”.

We can use Association Rules in any dataset where features take only two values i.e., 0/1. Some examples are listed below:

- Market Basket Analysis is a popular application of Association Rules.
- People who visit webpage X are likely to visit webpage Y.
- People who have age-group [30,40] & income [>\$100k] are likely to own home.

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

- A minimum support threshold is applied to find all frequent itemsets in a database.
- A minimum confidence constraint is applied to these frequent itemsets in order to form rules.

While the second step is straightforward, the first step needs more attention.

---

## 9.3 Support and Confidence

---

A set of transactions process aims to find the rules that enable us to predict the occurrence of a specific item based on the occurrence of other items in the transaction.

**An association rule has 2 parts:**

- An Antecedent (if) and
- a Consequent (then)

An antecedent is something that's found in data, and a consequent is an item that is found in combination with the antecedent.

*Antecedent → Consequent [support, confidence]*

(**support** and **confidence** are user defined measures of interestingness)

Example:

*"If a customer buys bread, he's 70% likely of buying milk."*

In the above association rule, bread is the antecedent and milk is the consequent. These types of relationships where we can find out some association or relation between two items is known as *single cardinality*. It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly. So, to measure the associations between thousands of data items, there are several metrics. If the above rule is a result of a thorough analysis of some data sets, it cannot be only used to improve customer service but also improve the company's revenue.

Association rules are created by thoroughly analyzing data and looking for frequent if/then patterns. Then, depending on the following two parameters, the important relationships are observed:

**9.3.1 Support:** Support indicates how frequently the if/then relationship appears in the database. It is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the itemset X. If there are X datasets, then for transactions T, it can be written as:

$$\text{Supp}(X) = \frac{\text{Freq}(X)}{T}$$

**9.3.2 Confidence:** Confidence tells about the number of times these relationships have been found to be true. It indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.

$$\text{Confidence} = \frac{\text{Freq}(X,Y)}{\text{Freq}(X)}$$

**9.3.3 Lift:** Lift is the strength of any rule, which can be defined as below formula:

$$\text{Lift} = \frac{\text{Supp}(X,Y)}{\text{Supp}(X) \times \text{Supp}(Y)}$$

It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has three possible values:

If,

- **Lift=1:** The probability of occurrence of antecedent and consequent is independent of each other.
- **Lift>1:** It determines the degree to which the two itemsets are dependent to each other.
- **Lift<1:** It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

So, in a given transaction with multiple items, Association Rule Mining primarily tries to find the rules that govern how or why such products/items are often bought together.

## 9.4 Frequent Pattern Mining

Frequent Pattern Mining is also known as the Association Rule Mining. Finding frequent patterns, causal structures and associations in data sets and is an inquisitive process called pattern mining. When a series of transactions are given, pattern mining's main motive is to find the rules that enable us to speculate a certain item based on the happening of other items in the transaction.

Frequent Pattern Mining is an analytical process that finds frequent patterns, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other data repositories. Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set.

For instance, a set of items, such as pen and ink, often appears together in a set of data transactions, is called a recurrent item set. Purchasing a personal computer, later a digital camera, and then a hard disk, if all these events repeatedly occur in the history of shopping database, it is a (frequent) sequential pattern. If the occurrence of a substructure is regular in a graph database, it is called a (frequent) structural pattern.

Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Before we start defining the rule, let us first see the basic definitions.

**Support Count ( $\delta$ )**– Frequency of occurrence of a itemset.

Here  $\delta (\{\text{Milk, Bread, Diaper}\}) = 2$

**Frequent Itemset** – An itemset whose support is greater than or equal to minsup threshold.

**Association Rule** – An implication expression of the form  $X \rightarrow Y$ , where X and Y are any 2 itemsets.

**Example:**  $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

**From the above table,  $\{\text{Milk, Diaper}\} \Rightarrow \{\text{Beer}\}$**

$$\begin{aligned} \text{Support} &= (\{\text{Milk, Diaper, Beer}\}) \div |\text{T}| \\ &= 2/5 \\ &= 0.4 \end{aligned}$$

$$\begin{aligned} \text{Confidence} &= (\text{Milk, Diaper, Beer}) \div (\text{Milk, Diaper}) \\ &= 2/3 \\ &= 0.67 \end{aligned}$$

$$\begin{aligned} \text{Lift} &= \text{Supp}(\{\text{Milk, Diaper, Beer}\}) \div \text{Supp}(\{\text{Milk, Diaper}\}) * \text{Supp}(\{\text{Beer}\}) \\ &= 0.4 / (0.6 * 0.6) \\ &= .11 \end{aligned}$$

**Let's look at some areas where Association Rule Mining has helped quite a lot:**

#### 9.4.1 Market Basket Analysis:

This is the most typical example of association mining. Data is collected using barcode scanners in most supermarkets. This database, known as the “market basket” database, consists of a large number of records on past transactions. A single record lists all the items bought by a customer in one sale. Knowing which groups are inclined towards which set of

items gives these shops the freedom to adjust the store layout and the store catalog to place the optimally concerning one another.

#### **9.4.2 Medical Diagnosis:**

Association rules in medical diagnosis can be useful for assisting physicians for curing patients. Diagnosis is not an easy process and has a scope of errors which may result in unreliable end-results. Using relational association rule mining, we can identify the probability of the occurrence of illness concerning various factors and symptoms. Further, using learning techniques, this interface can be extended by adding new symptoms and defining relationships between the new signs and the corresponding diseases.

#### **9.4.3 Census Data:**

Every government has tonnes of census data. This data can be used to plan efficient public services (education, health, transport) as well as help public businesses. This application of association rule mining and data mining has immense potential in supporting sound public policy and bringing forth an efficient functioning of a democratic society.

#### **9.4.4 Protein Sequence:**

Proteins are sequences made up of twenty types of amino acids. Each protein bears a unique 3D structure which depends on the sequence of these amino acids. A slight change in the sequence can cause a change in structure which might change the functioning of the protein. This dependency of the protein functioning on its amino acid sequence has been a subject of great research.

**To understand the value of this applied technique, let's consider two business use cases.**

##### **Use Case One**

- **Business Problem:** A retail store manager wants to conduct Market Basket analysis to come up with a better strategy of product placement and product bundling.
- **Business Benefit:** Based on the rules generated, the store manager can strategically place the products together or in sequence leading to growth in sales and, in turn, revenue of the store. Offers such as “Buy this and get this free” or “Buy this and get % off on this” can be designed based on the rules generated.

##### **Use Case Two**

- **Business Problem:** A bank-marketing manager wishes to analyze which products are frequently and sequentially bought together. Each customer is represented as a transaction, containing the ordered set of products, and which products are likely to be purchased simultaneously/sequentially can then be predicted.
- **Business Benefit:** Based on the rules generated, banking products can be cross-sold to each existing or prospective customer to drive sales and bank revenue. For example, if savings, personal loan and credit cards are frequently/sequentially bought, then a new saving account customer can be cross-sold with a personal loan and credit card.

## 9.5 Market Basket Analysis

---

Association Rule Mining is sometimes referred to as “Market Basket Analysis”, as it was the first application area of association mining. It is a modeling technique that helps to identify which items should be purchased together. The aim is to discover associations of items occurring together more often than you’d expect from randomly sampling all the possibilities. The classic anecdote of Beer and Diaper will help in understanding this better.

Assume that, there are large number of items like Tea, Coffee, Milk, Sugar. Among these, the customer buys the subset of items as per the requirement and market gets the information of items which customer has purchased together. So, the market uses this information to put the items on different positions (or locations).

Market Basket Analysis method of determining customer obtained patterns by mining association from retailer transactional database. Now a day’s every product comes with the bar code. This data is rapidly documented by the business world as having the huge possible value in marketing. In detailed, commercial organizations are interested in “association rules” that identify the patterns of purchases, such that the occurrence of one item in a basket will indicate the presence of one or more additional items. This “market basket analysis” result can then be used to recommend the combinations of the products for special promotions or sales, devise a more actual store layout, and give vision into brand loyalty and co-branding. It will also lead the managers towards efficient and real strategic decision making.

### 9.5.1 Implementation of market based analysis

- The market basket analysis is used to decide the perfect location, where the items can be placed inside the store.  
**For example:** If the customer buys a coffee, it is possible that the customer may buy milk or sugar along with coffee.
- So keeping the coffee and sugar next to each other in store will be helpful to customers to buy the items together and improves sales of the company.
- The problem of large volume transactions can be minimized by using **differential market basket analysis**, which is capable of finding interesting results and eliminates the large volume.

Algorithm which is used in market basket analysis (MBA) is apriori algorithm because it is a candidate generation algorithm. It is founded on information that this algorithm uses the preceding knowledge of the regular item set possessions. Apriori procedure pays to an iterative tactic that is recognized as a level wise search in which k-item sets are used to discover (k+1) itemsets. Based on this possession, if a set cannot pass the minimum verge than all of its super sets will also fail the test as well.

Thus, if an item set is not a recurrent item set, then item set will not use to create large item set. Apriori procedure is the most recurrently used algorithm among the association rules algorithms that were used at the analysis phase. The problems occur in apriori algorithm are

that it scans the databases again and again to check the recurrent item sets and it also generate infrequent itemsets.

Strong associations have been observed among the purchased item sets group with regard to the purchase behaviour of the customers of the retail store. The customer's shopping information analyzed by using the association rules mining with the apriori algorithm. As a result of the analysis, strong and useful association rules were determined between the product groups with regard to understanding what kind of purchase behaviour customer's exhibit within a certain shopping visit from both in-category and from different product categories for the specialty store

#### **Example of Marketing and Sales Promotion:**

- Let the rule discovered be  
    {Bagels, ... } --> {Potato Chips}
- **Potato Chips as consequent =>** Can be used to determine what should be done to boost its sales.
- **Bagels in the antecedent =>** Can be used to see which products would be affected if the store discontinues selling bagels.
- **Bagels in antecedent and Potato chips in consequent =>** Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

---

## **9.6 Apriori algorithm**

---

With the quick growth in e-commerce applications, there is an accumulation vast quantity of data in months not in years. Data Mining, also known as Knowledge Discovery in Databases(KDD), to find anomalies, correlations, patterns, and trends to predict outcomes.

Apriori algorithm is a classical algorithm in data mining. It is used for mining frequent itemsets and relevant association rules. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store.

Apriori algorithm was the first algorithm that was proposed for frequent itemset mining. It uses prior(a-prior) knowledge of frequent itemset properties. A minimum threshold is set on the expert advice or user understanding.

**Apriori algorithm** is given by R. Agrawal and R. Srikant in 1994 for finding frequent itemsets in a dataset for boolean association rule. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets.

To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called *Apriori property* which helps by reducing the search space.

### **9.6.1 Apriori Property:**

All non-empty subset of frequent itemset must be frequent. The key concept of Apriori algorithm is its anti-monotonicity of support measure. Apriori assumes that All subsets of a

frequent itemset must be frequent (Apriori property). If an itemset is infrequent, all its supersets will be infrequent.

### 9.6.2 Steps in Apriori

Apriori algorithm is a sequence of steps to be followed to find the most frequent itemset in the given database. This data mining technique follows the join and the prune steps iteratively until the most frequent itemset is achieved. A minimum support threshold is given in the problem or it is assumed by the user.

**Step 1:** In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. The algorithm will count the occurrences of each item.

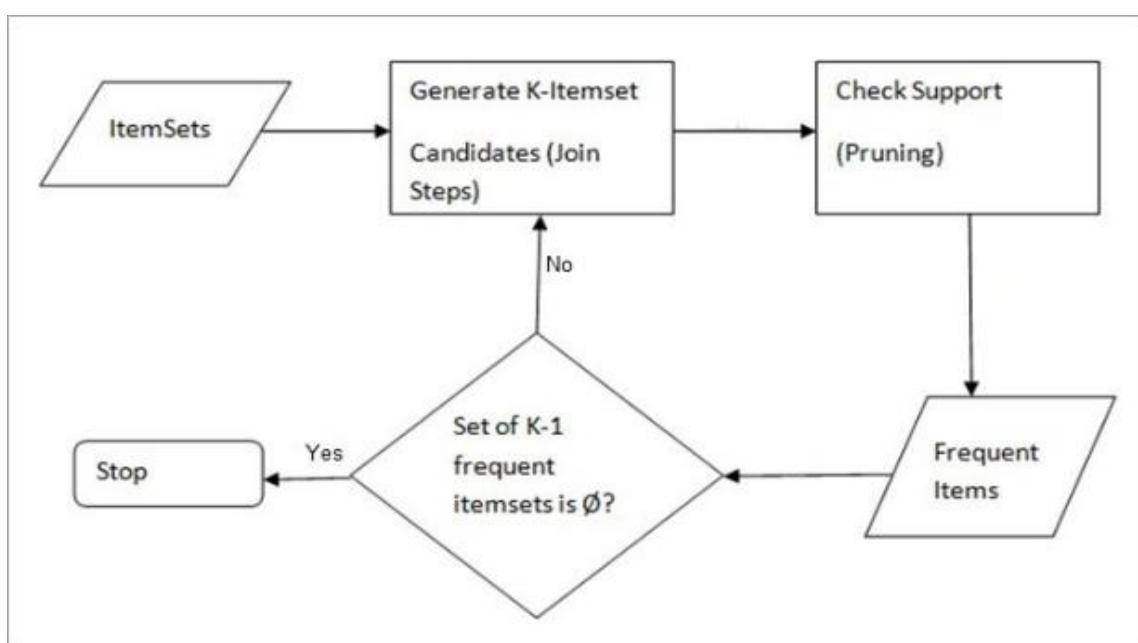
**Step 2:** Let there be some minimum support, min\_sup (eg 2). The set of 1 – itemsets whose occurrence is satisfying the min sup are determined. Only those candidates which count more than or equal to min\_sup, are taken ahead for the next iteration and the others are pruned.

**Step 3:** Next, 2-itemset frequent items with min\_sup are discovered. For this in the join step, the 2-itemset is generated by forming a group of 2 by combining items with itself.

**Step 4:** The 2-itemset candidates are pruned using min-sup threshold value. Now the table will have 2 –itemsets with min-sup only.

**Step 5:** The next iteration will form 3 –itemsets using join and prune step. This iteration will follow antimonotone property where the subsets of 3-itemsets, that is the 2 –itemset subsets of each group fall in min\_sup. If all 2-itemset subsets are frequent then the superset will be frequent otherwise it is pruned.

**Step 6:** Next step will follow making 4-itemset by joining 3-itemset with itself and pruning if its subset does not meet the min\_sup criteria. The algorithm is stopped when the most frequent itemset is achieved.



### 9.6.3 Example of Apriori:

**Support threshold=50%, Confidence= 60%**

**TABLE-1**

Transaction	List of items
T1	I1,I2,I3
T2	I2,I3,I4
T3	I4,I5
T4	I1,I2,I4
T5	I1,I2,I3,I5
T6	I1,I2,I3,I4

**Solution:**

Support threshold=50% =>  $0.5*6= 3 \Rightarrow \text{min\_sup}=3$

#### 1. Count of Each Item

**TABLE-2**

Item	Count
I1	4
I2	5
I3	4
I4	4
I5	2

**2. Prune Step:** TABLE -2 shows that I5 item does not meet min\_sup=3, thus it is deleted, only I1, I2, I3, I4 meet min\_sup count.

**TABLE-3**

Item	Count
I1	4
I2	5
I3	4
I4	4

**3. Join Step:** Form 2-itemset. From TABLE-1 find out the occurrences of 2-itemset.

**TABLE-4**

Item	Count
I1,I2	4
I1,I3	3
I1,I4	2

I2,I3	4
I2,I4	3
I3,I4	2

**4. Prune Step:** TABLE -4 shows that item set {I1, I4} and {I3, I4} does not meet min\_sup, thus it is deleted.

**TABLE-5**

Item	Count
I1,I2	4
I1,I3	3
I2,I3	4
I2,I4	3

**5. Join and Prune Step:** Form 3-itemset. From the TABLE- 1 find out occurrences of 3-itemset. From TABLE-5, find out the 2-itemset subsets which support min\_sup. We can see for itemset {I1, I2, I3} subsets, {I1, I2}, {I1, I3}, {I2, I3} are occurring in TABLE-5 thus {I1, I2, I3} is frequent.

We can see for itemset {I1, I2, I4} subsets, {I1, I2}, {I1, I4}, {I2, I4}, {I1, I4} is not frequent, as it is not occurring in TABLE-5 thus {I1, I2, I4} is not frequent, hence it is deleted.

**TABLE-6**

Item
I1,I2,I3
I1,I2,I4
I1,I3,I4
I2,I3,I4

Only {I1, I2, I3} is frequent.

**6. Generate Association Rules:** From the frequent itemset discovered above the association could be:

$$\{I1, I2\} \Rightarrow \{I3\}$$

$$\text{Confidence} = \text{support } \{I1, I2, I3\} / \text{support } \{I1, I2\} = (3/4) * 100 = 75\%$$

$$\{I1, I3\} \Rightarrow \{I2\}$$

$$\text{Confidence} = \text{support } \{I1, I2, I3\} / \text{support } \{I1, I3\} = (3/3) * 100 = 100\%$$

$$\{I2, I3\} \Rightarrow \{I1\}$$

$$\text{Confidence} = \text{support } \{I1, I2, I3\} / \text{support } \{I2, I3\} = (3/4) * 100 = 75\%$$

$$\{I1\} \Rightarrow \{I2, I3\}$$

Confidence = support {I1, I2, I3} / support {I1} =  $(3/4) * 100 = 75\%$

{I2} => {I1, I3}

Confidence = support {I1, I2, I3} / support {I2} =  $(3/5) * 100 = 60\%$

{I3} => {I1, I2}

Confidence = support {I1, I2, I3} / support {I3} =  $(3/4) * 100 = 75\%$

This shows that all the above association rules are strong if minimum confidence threshold is 60%.

#### 9.6.4 The Apriori Algorithm: Pseudo Code

C: Candidate item set of size k

L: Frequent itemset of size k

- Join Step:  $C_k$  is generated by joining  $L_{k-1}$  with itself
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
- Pseudo-code :  
 $C_k$ : Candidate itemset of size k  
 $L_k$ : frequent itemset of size k

```
 $L_1 = \{\text{frequent items}\};$ 
 $\text{for } (k = 1; L_k \neq \emptyset; k++) \text{ do begin}$ 
     $C_{k+1} = \text{candidates generated from } L_k;$ 
     $\text{for each transaction } t \text{ in database do}$ 
         $\text{increment the count of all candidates in } C_{k+1}$ 
         $\text{that are contained in } t$ 
     $L_{k+1} = \text{candidates in } C_{k+1} \text{ with min\_support}$ 
     $\text{end}$ 
 $\text{return } \cup_k L_k;$ 
```

#### 9.6.5 Advantages and Disadvantages

##### Advantages

1. Easy to understand algorithm
2. Join and Prune steps are easy to implement on large itemsets in large databases

##### Disadvantages

1. It requires high computation if the itemsets are very large and the minimum support is kept very low.
2. The entire database needs to be scanned.

### **9.6.6 Methods to Improve Apriori Efficiency**

**Many methods are available for improving the efficiency of the algorithm.**

1. **Hash-Based Technique:** This method uses a hash-based structure called a hash table for generating the k-itemsets and its corresponding count. It uses a hash function for generating the table.
2. **Transaction Reduction:** This method reduces the number of transactions scanning in iterations. The transactions which do not contain frequent items are marked or removed.
3. **Partitioning:** This method requires only two database scans to mine the frequent itemsets. It says that for any itemset to be potentially frequent in the database, it should be frequent in at least one of the partitions of the database.
4. **Sampling:** This method picks a random sample S from Database D and then searches for frequent itemset in S. It may be possible to lose a global frequent itemset. This can be reduced by lowering the min\_sup.
5. **Dynamic Itemset Counting:** This technique can add new candidate itemsets at any marked start point of the database during the scanning of the database.

### **9.6.7 Applications of Apriori Algorithm**

**Some fields where Apriori is used:**

1. **In Education Field:** Extracting association rules in data mining of admitted students through characteristics and specialties.
2. **In the Medical field:** For example Analysis of the patient's database.
3. **In Forestry:** Analysis of probability and intensity of forest fire with the forest fire data.
4. Apriori is used by many companies like Amazon in the **Recommender System** and by Google for the auto-complete feature.

---

## **9.7 Associative classification- Rule Mining**

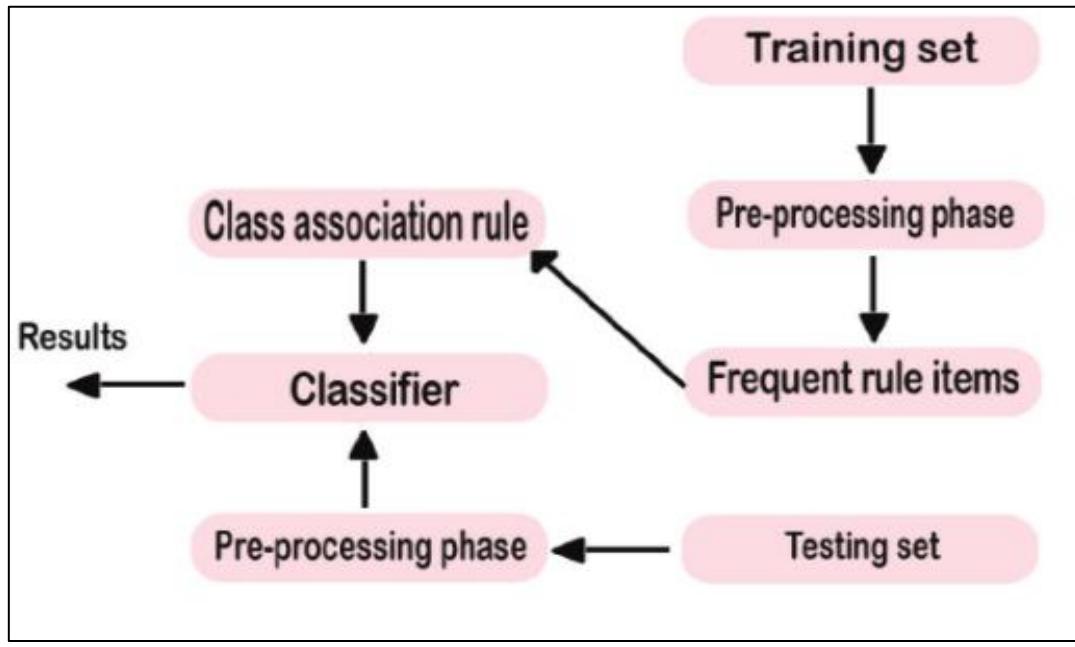
---

Associative classifiers are a classification system based on associative classification rules. Although associative classification is more accurate than a traditional classification approach, it cannot handle numerical data and its relationships.

Associative Classification, a combination of two important and different fields (classification and association rule mining), aims at building accurate and interpretable classifiers by means of association rules. A major problem in this field is that existing proposals do not scale well when Data Mining is considered.

Associative classification (AC) is a promising data mining approach that integrates classification and association rule discovery to build classification models (classifiers). Previously several AC algorithms have been proposed such as Classification based Association (CBA), Classification based on Predicted Association Rule (CPAR), Multi-class

Classification using Association Rule (MCAR), Live and Let Live ( $L^3$ ) and others. These algorithms use different procedures for rule learning, rule sorting, rule pruning, classifier building and class allocation for test cases.



**Data flow diagram of associative classification**

### Associative Classification

- Association rules are generated and analyzed for use in classification
- Search for strong associations between frequent patterns (conjunctions of attribute-value pairs) and class labels
- Classification: Based on evaluating a set of rules in the form of  $P_1 \wedge p_2 \dots \wedge p_l \rightarrow A_{\text{class}} = C \parallel (\text{conf}, \text{sup})$
- It explores highly confident associations among multiple attributes and may overcome some constraints introduced by decision-tree induction, which considers only one attribute at a time

#### **9.7.1 Typical Associative Classification Methods**

1. CBA (Classification By Association):
  - Mine association possible rules in the form of
  - Cond-set (a set of attribute-value pairs)  $\rightarrow$  class label
  - Build classifier: Organize rules according to decreasing precedence based on confidence and then support
2. CMAR (Classification based on Multiple Association Rules):
  - Classification: Statistical analysis on multiple rules
3. CPAR (Classification based on Predictive Association Rules):
  - Generation of predictive rules (FOIL-like analysis)
  - High efficiency, accuracy similar to CMAR

### **9.7.2 Rule support and confidence**

Given a training data set T, for a rule R: P → C

- The support of R, denoted as sup(R), is the number of objects in T matching R condition and having a class label c
- The confidence of R, denoted as conf(R), is the the number of objects matching R condition and having class label cover the number of objects matching R condition
- Any Item has a support larger than the user minimum support is called frequent itemset

---

## **9.8 Associative classification Conclusion**

---

- Associative classification is a promising approach in data mining.
- Since more than LLHs (Low-Level Heuristic) could improve the objective function in the hyperheuristic, we need a multi-label rules in the classifier.
- Associative classifiers produce more accurate classification models than traditional classification algorithms such as decision trees and rule induction approaches.
- One challenge in associative classification is the exponential growth of rules, therefore pruning becomes essential

---

## **9.9 Summary**

---

In this chapter we have presented some needed concepts for dealing with association rules, recalled previous efforts concerning association rule mining. Presented an efficient algorithm for identifying association rules of interest. Introduced detailed efforts on mining association rules; Association rule mining: support and confidence and frequent item sets, market basket analysis, Apriori algorithm and Associative classification.

---

## **9.10 References**

---

- P. N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining.
- Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques
- Tan, P., Steinbach, M., & Kumar, Introduction to data mining. Pearson Education
- Agrawal, R., Imielinski, T., Mining association rules between sets of items in large databases.
- [https://www.brainkart.com/article/Associative-Classification\\_8326/](https://www.brainkart.com/article/Associative-Classification_8326/)

- <https://www.softwaretestinghelp.com/apriori-algorithm/>

## **Unit 5**

### **Chapter 10: Classification – I**

- 10.0 Objectives
- 10.1 Introduction
- 10.2 Classification
  - 10.2.1 Training and Testing
  - 10.2.2 Categories of Classification
  - 10.2.3 Associated Tools and Languages
  - 10.2.4 Advantages and Disadvantages
- 10.3 Classification of Data Mining
  - 10.3.1 Two Important steps
    - a. Model Construction
    - b. Model Usage
  - 10.3.2 Classification Methods
- 10.4 Statistical-based Algorithms
  - 10.4.1 Regression
  - 10.4.2 Terminologies Related to the Regression Analysis
  - 10.4.3 Use of Regression Analysis
  - 10.4.4 Types of Regression
    - a. Linear Regression
    - b. Logistic Regression
    - c. Polynomial Regression
    - d. Support Vector Regression
    - e. Decision Tree Regression
    - f. Random Forest Regression
    - g. Ridge Regression
    - h. Lasso Regression
- 10.5 Naïve Bayesian Classification
  - 10.5.1 Working of Naïve Bayes
  - 10.5.2 Advantages and Disadvantages
  - 10.5.3 Types of Naïve Bayes Model
- 10.6 Distance-based algorithm
  - 10.6.1 K Nearest Neighbor
  - 10.6.2 Working of KNN Algorithm
  - 10.6.3 Advantages and Disadvantages
- 10.7 Conclusion

## 10.8 Summary

## 10.9 References

---

## **10.0 Objectives**

---

The main objective of classification is to create a courseware that focuses on creating materials to achieve the goal of helping the students get deeper understanding of the most used classification algorithms in data mining. The existing materials on the classification algorithms are completely textual and students find it difficult to grasp. With the help of this courseware, students will be able to learn the algorithms and then visualize the steps with the help of interactive examples that can be modified in many ways by the student to get a complete understanding of the algorithms. There is also information provided on how to make practical use of these algorithms using data mining tools.

---

### **10.1 Introduction**

---

Classification is used in data mining to classify data based on class labels. It involves building a model using training data set, and then using the built model to assign given items to specific classes/categories. In the model building process, also called training process, a classification algorithm finds relationships between the attributes of the data and the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a new data set in which the class assignments are unknown. According to Carlo Vercellis, Classification models are supervised learning methods for predicting the value of a categorical target attribute, unlike regression models which deal with numerical attributes. Starting from a set of past observations whose target class is known, classification models are used to generate a set of rules that allow the target class of future examples to be predicted.

Classification holds a prominent position in learning theory due to its theoretical implications and the countless applications it affords. From a theoretical viewpoint, the development of algorithms capable of learning from past experience represents a fundamental step in emulating the inductive capabilities of the human brain.

On the other hand, the opportunities afforded by classification extend into several different application domains: selection of the target customers for a marketing campaign, fraud detection, image recognition, early diagnosis of diseases, text cataloguing and spam email recognition are just a few examples of real problems that can be framed within the classification paradigm.

In this chapter, we will review the major classification methods: classification trees, Bayesian methods, neural networks, logistic regression and support vector machines. Statistical-based algorithms- Regression, Naïve Bayesian classification, Distance-based algorithm- K Nearest Neighbour, Decision Tree-based algorithms -ID3, C4.5, CART and so on.

---

## 10.2 Classification

---

Classification is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

**Example:** Before starting any project, we need to check its feasibility. In this case, a classifier is required to predict class labels such as ‘Safe’ and ‘Risky’ for adopting the Project and to further approve it. It is a two-step process such as:

1. **Learning Step (Training Phase):** Construction of Classification Model  
Different Algorithms are used to build a classifier by making the model learn using the training set available. The model has to be trained for the prediction of accurate results.
2. **Classification Step:** Model used to predict class labels and testing the constructed model on test data and hence estimate the accuracy of the classification rules.

### 10.2.1 Training and Testing:

Suppose there is a person who is sitting under a fan and the fan starts falling on him, he should get aside in order not to get hurt. So, this is his training part to move away. While Testing if the person sees any heavy object coming towards him or falling on him and moves aside then the system is tested positively and if the person does not move aside then the system is negatively tested. Same is the case with the data, it should be trained in order to get the accurate and best results.

There are certain data types associated with data mining that actually tells us the format of the file (whether it is in text format or in numerical format). Attributes – Represents different features of an object. Different types of attributes are:

1. **Binary:** Possesses only two values i.e. True or False  
Example: Suppose there is a survey evaluating some products. We need to check whether it's useful or not. So, the Customer has to answer it in Yes or No.  
Product usefulness: Yes / No
  - **Symmetric:** Both values are equally important in all aspects
  - **Asymmetric:** When both the values may not be important.
2. **Nominal:** When more than two outcomes are possible. It is in Alphabet form rather than being in Integer form.  
Example: One needs to choose some material but of different colors. So, the color might be Yellow, Green, Black, Red.  
Different Colors: Red, Green, Black, Yellow
  - **Ordinal:** Values that must have some meaningful order.  
Example: Suppose there are grade sheets of few students which might contain different grades as per their performance such as A, B, C, D  
Grades: A, B, C, D

- **Continuous:** May have an infinite number of values, it is in float type  
Example: Measuring the weight of few Students in a sequence or orderly manner i.e.  
50, 51, 52, 53  
Weight: 50, 51, 52, 53
- **Discrete:** Finite number of values.  
Example: Marks of a Student in a few subjects: 65, 70, 75, 80, 90  
Marks: 65, 70, 75, 80, 90

### 10.2.2 Categories of Classification

Classifiers can be categorized into two major types:

1. **Discriminative:** It is a very basic classifier and determines just one class for each row of data. It tries to model just by depending on the observed data, depends heavily on the quality of data rather than on distributions.

**Example:** Logistic Regression

Acceptance of a student at a University (Test and Grades need to be considered)

Suppose there are few students and the Result of them are as follows :

2. **Generative:** It models the distribution of individual classes and tries to learn the model that generates the data behind the scenes by estimating assumptions and distributions of the model. Used to predict the unseen data.

**Example:** Naive Bayes Classifier

Detecting Spam emails by looking at the previous data. Suppose 100 emails and that too divided in 1:4 i.e. Class A: 25%(Spam emails) and Class B: 75%(Non-Spam emails).

Now if a user wants to check that if an email contains the word cheap, then that may be termed as Spam.

It seems to be that in Class A(i.e. in 25% of data), 20 out of 25 emails are spam and rest not.

And in Class B(i.e. in 75% of data), 70 out of 75 emails are not spam and rest are spam.

So, if the email contains the word cheap, what is the probability of it being spam ?? (= 80%)

### 10.2.3 Associated Tools and Languages

Used for mining/ to extract useful information from raw data.

- **Main Languages used:** R, SAS, Python, SQL
- **Major Tools used:** RapidMiner, Orange, KNIME, Spark, Weka
- **Libraries used:** Jupyter, NumPy, Matplotlib, Pandas, ScikitLearn, NLTK, TensorFlow, Seaborn, Basemap, etc.

**Real-Life Examples :**

- **Market Basket Analysis:**

It is a modeling technique that has been associated with frequent transactions of buying

some combination of items.

**Example:** Amazon and many other Retailers use this technique. While viewing some products, certain suggestions for the commodities are shown that some people have bought in the past.

- **Weather Forecasting:**

Changing Patterns in weather conditions needs to be observed based on parameters such as temperature, humidity, wind direction. This keen observation also requires the use of previous records in order to predict it accurately.

#### **10.2.4 Advantages and Disadvantages**

##### **Advantages:**

- Mining Based Methods are cost-effective and efficient
- Helps in identifying criminal suspects
- Helps in predicting the risk of diseases
- Helps Banks and Financial Institutions to identify defaulters so that they may approve Cards, Loan, etc.

##### **Disadvantages:**

Privacy: When the data is either are chances that a company may give some information about their customers to other vendors or use this information for their profit.

Accuracy Problem: Selection of Accurate model must be there in order to get the best accuracy and result.

---

### **10.3 Classification of Data Mining Systems**

---

Data Mining is considered as an interdisciplinary field. It includes a set of various disciplines such as statistics, database systems, machine learning, visualization and information sciences. Classification of the data mining system helps users to understand the system and match their requirements with such systems.

Data mining systems can be categorized according to various criteria, as follows:

1. **Classification according to the application adapted:**

This involves domain-specific application. For example, the data mining systems can be tailored accordingly for telecommunications, finance, stock markets, e-mails and so on.

2. **Classification according to the type of techniques utilized:**

This technique involves the degree of user interaction or the technique of data analysis involved. For example, machine learning, visualization, pattern recognition, neural networks, database-oriented or data-warehouse oriented techniques.

### **3. Classification according to the types of knowledge mined:**

This is based on functionalities such as characterization, association, discrimination and correlation, prediction etc.

### **4. Classification according to types of databases mined:**

A database system can be classified as a ‘type of data’ or ‘use of data’ model or ‘application of data’.

#### **10.3.1 The two important steps of classification are:**

##### **a. Model construction**

- A predefine class label is assigned to every sample tuple or object. These tuples or subset data are known as training data set.
- The constructed model, which is based on training set is represented as classification rules, decision trees or mathematical formulae.

##### **b. Model usage**

- The constructed model is used to perform classification of unknown objects.
- A class label of test sample is compared with the resultant class label.
- Accuracy of model is compared by calculating the percentage of test set samples, that are correctly classified by the constructed model.
- Test sample data and training data sample are always different.

#### **10.3.2 Classification methods**

Classification is one of the most commonly used technique when it comes to classifying large sets of data. This method of data analysis includes algorithms for supervised learning adapted to the data quality. The objective is to learn the relation which links a variable of interest, of qualitative type, to the other observed variables, possibly for the purpose of prediction. The algorithm that performs the classification is the classifier while the observations are the instances. The classification method uses algorithms such as decision tree to obtain useful information. Companies use this approach to learn about the behavior and preferences of their customers. With classification, you can distinguish between data that is useful to your goal and data that is not relevant.

The study of classification in statistics is vast, and there are several types of classification algorithms you can use depending on the dataset you’re working with. Below are the most common algorithms in Data Mining.

- a) Statistical-based algorithms- Regression
- b) Naïve Bayesian classification
- c) Distance-based algorithm- K Nearest Neighbour
- d) Decision Tree-based algorithms -ID3
- e) C4.5

f) CART

---

## 10.4 Statistical-based algorithms

---

There are two main phases present to work on classification. The first can easily identify the statistical community. The second, “modern” phase concentrated on more flexible classes of models. In which many of which attempt has to take. That provides an estimate of the joint distribution of the feature within each class. That can, in turn, provide a classification rule. Generally, statistical procedures have to characterize by having a precise fundamental probability model. That used to provide a probability of being in each class instead of just a classification. Also, we can assume that techniques will use by statisticians.

### 10.4.1 Regression

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price**, etc.

Regression is an important and broadly used statistical tool. The key objective of regression-based tasks is to predict output labels or responses which are continuous numeric values, for the given input data. The output will be based on what the model has learned in training phase. Basically, regression models use the input data features (independent variables) and their corresponding continuous numeric output values (dependent or outcome variables) to learn specific association between inputs and corresponding outputs.

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, "Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum." The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Some examples of regression can be as:

- Prediction of rain using temperature and other factors
- Determining Market trends
- Prediction of road accidents due to rash driving.

### 10.4.2 Terminologies Related to the Regression Analysis:

- **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.
- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.
- **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.
- **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

#### **10.4.3        Use of Regression Analysis**

As mentioned above, Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future predictions such as weather condition, sales prediction, marketing trends, etc., for such case we need some technology which can make predictions more accurately. So for such case we need Regression analysis which is a statistical method and used in machine learning and data science.

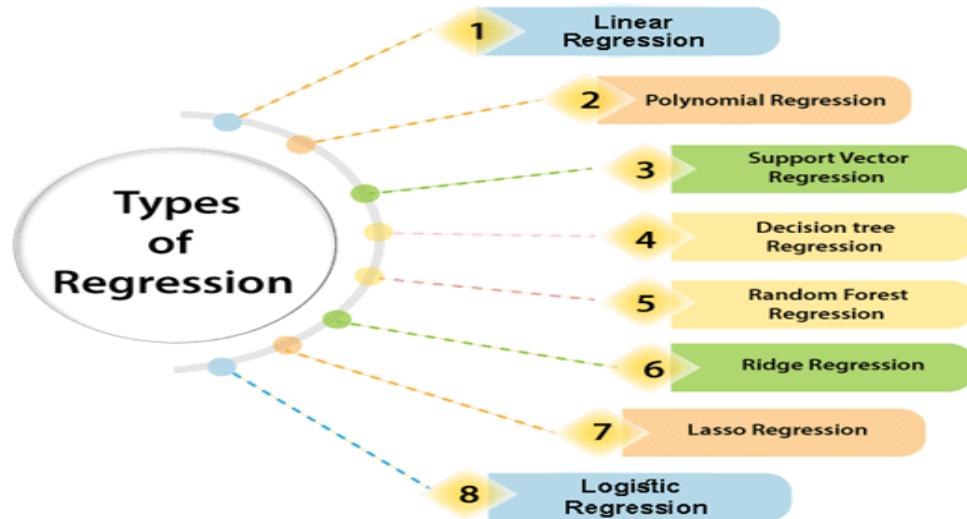
Below are some other reasons for using Regression analysis:

- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It helps to predict real/continuous values.
- By performing the regression, we can confidently determine the **most important factor, the least important factor, and how each factor is affecting the other factors**.

#### **10.4.4        Types of Regression**

- a. Linear Regression
- b. Logistic Regression

- c. Polynomial Regression
- d. Support Vector Regression
- e. Decision Tree Regression
- f. Random Forest Regression
- g. Ridge Regression
- h. Lasso Regression



#### a. Linear Regression:

It is one of the most widely known modeling techniques, as it is amongst the first elite regression analysis methods picked up by people at the time of learning predictive modeling. Here, the dependent variable is continuous and independent variable is more often continuous or discreet with a linear regression line.

Please note, in a multiple linear regression there is more than one independent variable and in a simple linear regression, there is only one independent variable. Thus, linear regression is best to be used only when there is a linear relationship between the independent and a dependent variable

Example: A business can use linear regression for measuring the effectiveness of the marketing campaigns, pricing, and promotions on sales of a product. Suppose a company selling sports equipment wants to understand if the funds they have invested in the marketing and branding of their products has given them substantial return or not. Linear regression is the best statistical method to interpret the results. The best thing about linear regression is it also helps in analyzing the obscure impact of each marketing and branding activity, yet controlling the constituent's potential to regulate the sales. If the company is running two or more advertising campaigns at the same time; as if one on television and two on radio, then linear regression can easily analyze the independent as well as the combined influence of running these advertisements together.

**b. Logistic Regression:**

Logistic regression is commonly used to determine the probability of event=Success and event=Failure. Whenever the dependent variable is binary like 0/1, True/False, Yes/No logistic regression is used. Thus, it can be said that logistic regression is used to analyze either the close-ended questions in a survey or the questions demanding numeric response in a survey.

Please note, logistic regression does not need a linear relationship between a dependent and an independent variable just like linear regression. The logistic regression applies a non-linear log transformation for predicting the odds' ratio; therefore, it easily handles various types of relationships between a dependent and an independent variable.

Example: Logistic regression is widely used to analyze categorical data, particularly for binary response data in business data modeling. More often logistic regression is used to when the dependent variable is categorical like to predict whether the health claim made by a person is real(1) or fraudulent, to understand if the tumor is malignant(1) or not. Businesses use logistic regression to predict whether the consumers in a particular demographic will purchase their product or will buy from the competitors based on age, income, gender, race, state of residence, previous purchase, etc.

**c. Polynomial Regression:**

Polynomial regression is commonly used to analyze the curvilinear data and this happens when the power of an independent variable is more than 1. In this regression analysis method, the best fit line is never a 'straight-line' but always a 'curve line' fitting into the data points.

Please note, polynomial regression is better to be used when few of the variables have exponents and few do not have any. Additionally, it can model non-linearly separable data offering the liberty to choose the exact exponent for each variable and that too with full control over the modeling features available.

Example: Polynomial regression when combined with response surface analysis is considered as a sophisticated statistical approach commonly used in multisource feedback research. Polynomial regression is used mostly in finance and insurance-related industries where the relationship between dependent and independent variable is curvilinear. Suppose a person wants to budget expense planning by determining how much time it would take to earn a definitive sum of money. Polynomial regression by taking into account his/her income and predicting expenses can easily determine the precise time he/she needs to work to earn that specific sum of amount.

**d. Support Vector Regression:**

Support Vector Machine is a supervised learning algorithm which can be used for regression as well as classification problems. So if we use it for regression problems, then it is termed as Support Vector Regression.

Support Vector Regression is a regression algorithm which works for continuous variables. Below are some keywords which are used in Support Vector Regression:

**Kernel:** It is a function used to map a lower-dimensional data into higher dimensional data.

**Hyperplane:** In general SVM, it is a separation line between two classes, but in SVR, it is a line which helps to predict the continuous variables and cover most of the datapoints.

**Boundary line:** Boundary lines are the two lines apart from hyperplane, which creates a margin for datapoints.

**Support vectors:** Support vectors are the datapoints which are nearest to the hyperplane and opposite class.

In SVR, we always try to determine a hyperplane with a maximum margin, so that maximum number of datapoints are covered in that margin. The main goal of SVR is to consider the maximum datapoints within the boundary lines and the hyperplane (best-fit line) must contain a maximum number of datapoints.

#### e. Decision Tree Regression:

Decision Tree is a supervised learning algorithm which can be used for solving both classification and regression problems. It can solve problems for both categorical and numerical data. Decision Tree regression builds a tree-like structure in which each internal node represents the "test" for an attribute, each branch represents the result of the test, and each leaf node represents the final decision or result. A decision tree is constructed starting from the root node/parent node (dataset), which splits into left and right child nodes (subsets of dataset). These child nodes are further divided into their children node, and themselves become the parent node of those nodes.

#### f. Random Forest Regression:

Random forest is one of the most powerful supervised learning algorithms which is capable of performing regression as well as classification tasks. The Random Forest regression is an ensemble learning method which combines multiple decision trees and predicts the final output based on the average of each tree output. Random forest uses Bagging or Bootstrap Aggregation technique of ensemble learning in which aggregated decision tree runs in parallel and do not interact with each other. With the help of Random Forest regression, we can prevent Overfitting in the model by creating random subsets of the dataset.

#### g. Ridge Regression:

Ridge regression is one of the most robust versions of linear regression in which a small

amount of bias is introduced so that we can get better long term predictions.

The amount of bias added to the model is known as Ridge Regression penalty. We can compute this penalty term by multiplying with the lambda to the squared weight of each individual features.

A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used.

Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as L2 regularization.

It helps to solve the problems if we have more parameters than samples.

#### **h. Lasso Regression:**

Lasso regression is another regularization technique to reduce the complexity of the model. It is similar to the Ridge Regression except that penalty term contains only the absolute weights instead of a square of weights. Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.

---

## **10.5 Naive Bayesian classification**

---

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. A Naive Bayesian method is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore, they are considered as naive.

Naive Bayes classification is one of the simplest and popular algorithms in data mining. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. It is mainly used in text classification that includes a high-dimensional training dataset.

For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness and diameter features. Although the assumption that the predictor (independent) variables are independent is not always accurate, it does simplify the classification task dramatically, since it allows the class

conditional densities to be calculated separately for each variable, i.e., it reduces a multidimensional task to a number of one-dimensional ones. In effect, Naive Bayes reduces a high-dimensional density estimation task to a one-dimensional kernel density estimation.

Furthermore, the assumption does not seem to greatly affect the posterior probabilities, especially in regions near decision boundaries, thus, leaving the classification task unaffected.

Now, before moving to the formula for Naive Bayes, it is important to know about Bayes' theorem.

### Bayes' Theorem

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

$P(A|B)$  is Posterior probability: Probability of hypothesis A on the observed event B.

$P(B|A)$  is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

$P(A)$  is Prior Probability: Probability of hypothesis before observing the evidence.

### 10.5.1 Working of Naive Bayes Classifier:

Working of Naïve Bayes' Classifier can be understood with the help of the below example:

Suppose we have a dataset of weather conditions and corresponding target variable "Play". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:

- Convert the given dataset into frequency tables.
- Generate Likelihood table by finding the probabilities of given features.
- Now, use Bayes theorem to calculate the posterior probability.

**Problem:** If the weather is sunny, then the Player should play or not?

**Solution:** To solve this, first consider the below dataset:

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No

<b>5</b>	Rainy	Yes
<b>6</b>	Sunny	Yes
<b>7</b>	Overcast	Yes
<b>8</b>	Rainy	No
<b>9</b>	Sunny	No
<b>10</b>	Sunny	Yes
<b>11</b>	Rainy	No
<b>12</b>	Overcast	Yes
<b>13</b>	Overcast	Yes

Frequency table for the Weather Conditions:

<b>Weather</b>	<b>Yes</b>	<b>No</b>
<b>Overcast</b>	5	0
<b>Rainy</b>	2	2
<b>Sunny</b>	3	2
<b>Total</b>	10	5

Likelihood table weather condition:

<b>Weather</b>	<b>No</b>	<b>Yes</b>	
<b>Overcast</b>	0	5	5/14= 0.35
<b>Rainy</b>	2	2	4/14=0.29
<b>Sunny</b>	2	3	5/14=0.35
<b>All</b>	4/14=0.29	10/14=0.71	

Applying Bayes' theorem:

$$P(\text{Yes}|\text{Sunny}) = P(\text{Sunny}|\text{Yes}) * P(\text{Yes}) / P(\text{Sunny})$$

$$P(\text{Sunny}|\text{Yes}) = 3/10 = 0.3$$

$$P(\text{Sunny}) = 0.35$$

$$P(\text{Yes}) = 0.71$$

$$\text{So } P(\text{Yes}|\text{Sunny}) = 0.3 * 0.71 / 0.35 = 0.60$$

$$P(\text{No}|\text{Sunny}) = P(\text{Sunny}|\text{No}) * P(\text{No}) / P(\text{Sunny})$$

$$P(\text{Sunny}|\text{NO}) = 2/4 = 0.5$$

$$P(\text{No}) = 0.29$$

$$P(\text{Sunny}) = 0.35$$

So  $P(\text{No}|\text{Sunny}) = 0.5 * 0.29 / 0.35 = 0.41$

So as we can see from the above calculation that  $P(\text{Yes}|\text{Sunny}) > P(\text{No}|\text{Sunny})$

Hence on a Sunny day, Player can play the game.

## 10.5.2 Advantages and Disadvantages

### Advantages of Naive Bayes Classifier:

- Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
- It can be used for Binary as well as Multi-class Classifications.
- It performs well in Multi-class predictions as compared to the other Algorithms.
- It is the most popular choice for text classification problems.

### Disadvantages of Naive Bayes Classifier:

- Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

### Applications of Naive Bayes Classifier:

- It is used for Credit Scoring.
- It is used in medical data classification.
- It can be used in real-time predictions because Naïve Bayes Classifier is an eager learner.
- It is used in Text classification such as Spam filtering and Sentiment analysis.

## 10.5.3 Types of Naive Bayes Model:

There are three types of Naive Bayes Model, which are given below:

- a. **Gaussian:** The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.
- b. **Multinomial:** The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc. The classifier uses the frequency of words for the predictors.
- c. **Bernoulli:** The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

---

## 10.6 Distance-based algorithm

---

Distance-based algorithms are nonparametric methods that can be used for classification. These algorithms classify objects by the dissimilarity between them as measured by distance functions.

It classifies queries by computing distances between these queries and a number of internally stored exemplars. Exemplars that are closest to the query have the largest influence on the classification assigned to the query. Here we will study one distance based algorithm, K Nearest Neighbor in detail.

### 10.6.1 K Nearest Neighbor

k-Nearest Neighbours is a non-parametric lazy classification algorithm. It is one of the simplest data-mining algorithms where if there are N given training samples and a sample point S is given, the algorithm identifies k closest neighbors to S. The algorithm is lazy because it doesn't have any training phase before making decisions and it is non-parametric because it does not make any assumptions of the input data. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data

### 10.6.2 Working of KNN Algorithm

K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of following steps –

**Step 1** – For implementing any algorithm, we need dataset. So during the first step of KNN, we must load the training as well as test data.

**Step 2** – Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.

**Step 3** – For each point in the test data do the following –

- **3.1** – Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
- **3.2** – Now, based on the distance value, sort them in ascending order.
- **3.3** – Next, it will choose the top K rows from the sorted array.

- **3.4** – Now, it will assign a class to the test point based on most frequent class of these rows.

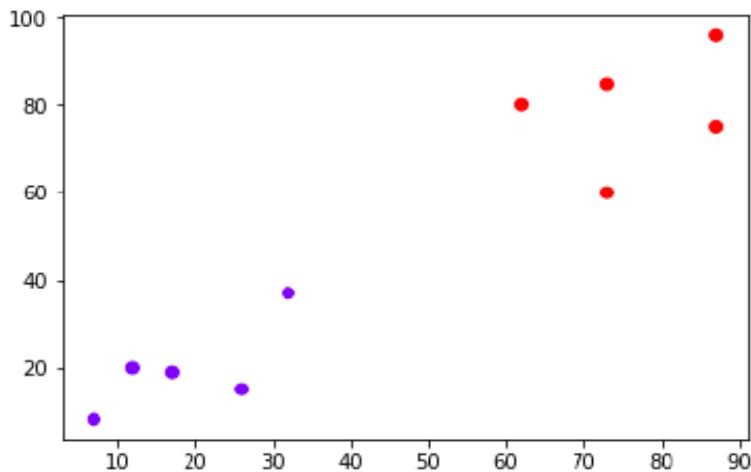
#### **Step 4** – End

Example

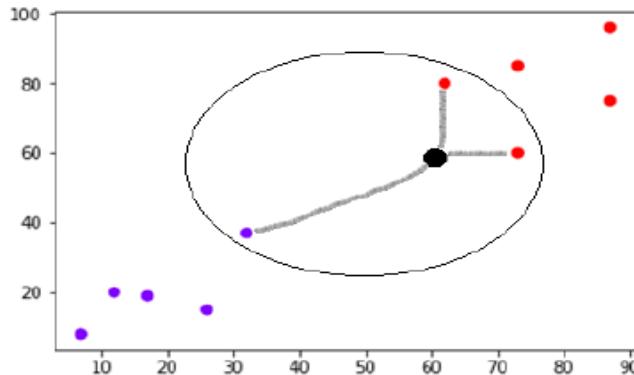
The following is an example to understand the concept of K and working of KNN algorithm

---

Suppose we have a dataset which can be plotted as follows,



Now, we need to classify new data point with black dot (at point 60,60) into blue or red class. We are assuming K = 3 i.e. it would find three nearest data points. It is shown in the next diagram.



We can see in the above diagram the three nearest neighbors of the data point with black dot. Among those three, two of them lies in Red class hence the black dot will also be assigned in red class.

In the case of KNN, indeed it doesn't "compare" the new unclassified data with all other, actually he performs a mathematical calculation to measure the distance between the data to makes the classification.

**The KNN's steps are:**

- 1 — Receive an unclassified data;
- 2 — Measure the distance (Euclidian, Manhattan, Minkowski or Weighted) from the new data to all others data that is already classified;

- 3 — Gets the K(K is a parameter that you define) smaller distances;
  - 4 — Check the list of classes had the shortest distance and count the amount of each class that appears;
  - 5 — Takes as correct class the class that appeared the most times;
  - 6 — Classifies the new data with the class that you took in step 5
- Calculating distance:**

To calculate the distance between two points (your new sample and all the data you have in your dataset) is very simple, as said before, there are several ways to get this value, in this article we will use the Euclidean distance.

The Euclidean distance's formula is like the image below:

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}.$$

Using this formula that you will check the distance between 1 point and 1 other point in your dataset, one by one in all your dataset, the smaller the result of this calculation is the most similar between these two data.

To make it simple,

Let's use the example from the previous worksheet but now with 1 unclassified data, that's the information we want to discover.

We have 5 data (lines) in this example, each sample (data/line) have your attributes (characteristics), let's imagine that all these are images, each line would be an image and each column would be an image's pixel.

	A	B	C	D	E	F	G	H	I
1	1	1	1	1	1	1	1	1	
2	2	2	2	2	2	2	2	2A	
3	2	2	2	2	2	2	2	2A	
4	1	1	1	1	1	1	1	1R	
5	1	1	1	1	1	1	1	1R	

So let's start by that confuse explain above.,

Let's take the first line, which is the data that we want to classify and let's measure the Euclidean distance to line 2

### 1 — Subtraction

Let's subtract each attribute (column) from row 1 with the attributes from row 2, example:

$$(1-2) = -1$$

### 2 — Exponentiation:

After you had subtract column 1 from row 1, with column 1 from row 2, we will get the squared root, so the result numbers are always positive, example:

$$(1-2)^2 = (-1)^2 = 1$$

### 3 — Sum

After you have done step 2, for all the row 1's columns and row 2's columns, we will sum all these results, let's make an example using the spreadsheet's columns's image and we will have the following result:

	A	B	C	D	E	F	G	H	I
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2A
3	2	2	2	2	2	2	2	2	2A
4	1	1	1	1	1	1	1	1	1R
5	1	1	1	1	1	1	1	1	1R

$$(1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2 = 8$$

Note that we have 8 attribute columns, both in row 1 and row 2, and we performed step 2 for each dataset's attribute, so our final result was 8, but why 8? Because each time we run step 2, the result gave 1, by a “coincidence” we have the same data in all columns and the result of  $(1-2)^2$  is equal to 1, I used those values to facilitate the math here, but no, this attributes doesn't need always be the same number, later we will see this better in implementing a code with this algorithm.

### 4 — Square root:

After performed step 3, we will get our subtractions's sum's square root. In step 3, the result was 8, so let's take the square root of number 8:

$$\sqrt{8} = 2,83 \text{ ou } 8^{(1/2)} = 2,83$$

(Note: ou – Euclidean distance.)

Now you have the Euclidean distance from line 1 to line 2, look, it was not so difficult, you could do it in a simple paper!

Now, you only need to make these for all dataset's lines, from line 1 to all other lines, when you do this, you will have the Euclidean distance from line 1 to all other lines, then you will sort it to get the “k”(e.g. k = 3) smallest distances, so you will check which is the class that most appears, the class that appears the most times will be the class that you will use to classify the line 1 (which was not classified before).

### 10.6.3 Advantages and Disadvantages of KNN

Advantages

- It is very simple algorithm to understand and interpret.
- It is very useful for nonlinear data because there is no assumption about data in this algorithm.
- It is a versatile algorithm as we can use it for classification as well as regression.

- It has relatively high accuracy but there are much better supervised learning models than KNN.

#### Disadvantages

- It is computationally a bit expensive algorithm because it stores all the training data.
- High memory storage required as compared to other supervised learning algorithms.
- Prediction is slow in case of big N.
- It is very sensitive to the scale of data as well as irrelevant features.

---

## **10.7 Conclusion**

---

Data mining offers promising ways to uncover hidden patterns within large amounts of data. These hidden patterns can potentially be used to predict future behavior. The availability of new data mining algorithms, however, should be met with caution. First of all, these techniques are only as good as the data that has been collected. Good data is the first requirement for good data exploration. Assuming good data is available, the next step is to choose the most appropriate technique to mine the data. However, there are tradeoffs to consider when choosing the appropriate data mining technique to be used in a certain application. There are definite differences in the types of problems that are conductive to each technique

---

## **10.8 Summary**

---

The presented discussion on knowledge extraction from medical databases is merely a short summary of the ongoing efforts in this area. It does, however, point to interesting directions of our research, where the aim is to apply hybrid classification schemes and create data mining tools well suited to the crucial demands with using statistical based algorithms and Distance based algorithms. Different techniques with Regression Analysis, Naïve Bayesian Classification, K Nearest Neighbor are explained with examples.

The objectives listed above would have been achieved if readers can gain a good understanding of data mining and be able to develop data mining applications. There is no doubt that data mining can be a very powerful technology and methodology for generating information from raw data to address business and other problems. This usefulness, however, will not be realised unless knowledge of data mining is put to good use.

---

## **10.9 References**

---

- Anurag Upadhyay, Suneet Shukla, Sudsanshu Kumar, Empirical Comparison by data mining Classification algorithms
- Jaiwei Han and Micheline Kamber, Data Mining Concepts and Techniques.Morgan Kaufmann Publishers.
- <https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm>

- <https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/>

## Unit 5

### Chapter 11: Classification – II

- 11.0 Objectives
- 11.1 Introduction
- 11.2 Decision Tree
  - 11.2.1 Decision Tree Terminologies
  - 11.2.2 Decision Tree Algorithm
  - 11.2.3 Decision Tree Example
  - 11.2.4 Attribute Selection Measures
    - a. Information Gain
    - b. Gini Index
    - c. Gain Ratio
  - 11.2.5 Overfitting in Decision Trees
    - a. Pruning Decision Trees
    - b. Random Forest
  - 11.2.6 Better Linear of Tree-based Model
  - 11.2.7 Advantages and Disadvantages
- 11.3 Iterative Dichotomiser 3 (ID3)
  - 11.3.1 History of ID3
  - 11.3.2 Algorithm of ID3
  - 11.3.3 Advantages and Disadvantages
- 11.4 C4.5
  - 11.4.1 Algorithm of C4.5
  - 11.4.2 Pseudocode of C4.5
  - 11.4.3 Advantages
- 11.5 CART (Classification and Regression Tree)
  - 11.5.1 Classification Tree
  - 11.5.2 Regression Tree
  - 11.5.3 Difference between Classification and Regression Trees
  - 11.5.4 Advantages of Classification and Regression Trees
  - 11.5.5 Limitations of Classification and Regression Trees
- 11.6 Conclusion
- 11.7 Summary
- 11.8 References

---

## **11.0 Objectives**

---

After going through this lesson you will be able to learn following things.

1. Learn about the decision tree algorithm for classification problems.
  2. Decision tree-based classification algorithms serve as the fundamental step in application of the decision tree method, which is a predictive modeling technique for classification of data.
  3. This chapter provides a broad overview of decision tree-based algorithms that are among the most commonly used methods for constructing classifiers.
  4. You will also learn the various decision tree methods like ID3, C4.5, CART etc. in detail.
- 

## **11.1 Introduction**

---

As you studied in previous chapter, that Classification is a two-step process, learning step and prediction step, in Data mining. In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response for given data. Decision Tree is one of the easiest and popular classification algorithms to understand and interpret.

In this chapter, you will learn Tree based algorithms which are considered to be one of the best and mostly used supervised classification methods. Tree based algorithms empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand.

---

## **11.2 Decision Tree**

---

A decision tree is a plan that includes a root node, branches, and leaf nodes. Every internal node characterizes an examination on an attribute, each division characterizes the consequence of an examination, and each leaf node grasps a class tag. The primary node in the tree is the root node. Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand. The logic behind the decision tree can be easily understood because it shows a tree-like structure.

Using a decision tree, we can visualize the decisions that make it easy to understand and thus it is a popular data mining classification technique. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data). In Decision Trees, for

predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

### **11.2.1 Decision Tree Terminologies**

1. **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
2. **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
3. **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
4. **Branch/Sub Tree:** A tree formed by splitting the tree.
5. **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
6. **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

### **11.2.2 Decision tree algorithm**

- Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.
- Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.
- We can represent any boolean function on discrete attributes using the decision tree.

#### **Working of Decision Tree algorithm**

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

**Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.

**Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM)**.

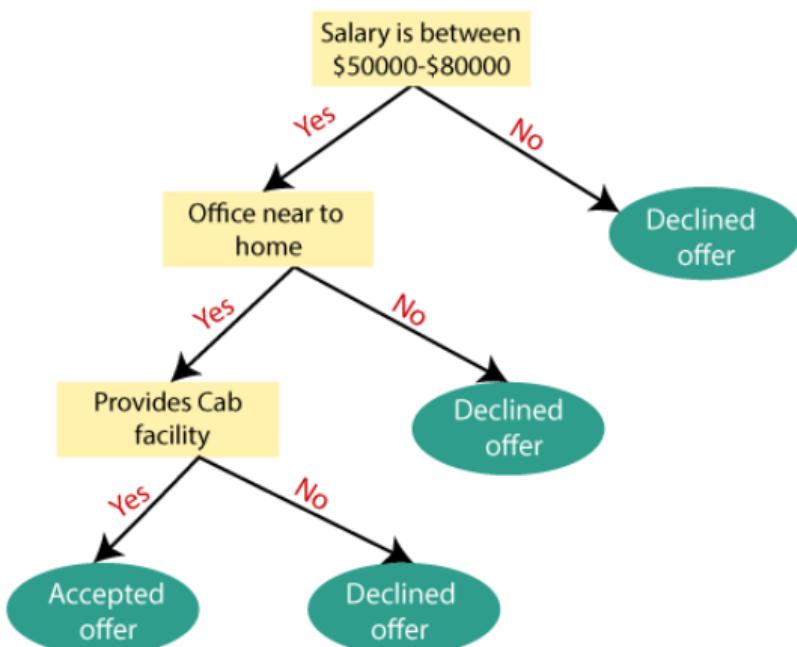
**Step-3:** Divide the S into subsets that contains possible values for the best attributes.

**Step-4:** Generate the decision tree node, which contains the best attribute.

**Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

### 11.2.3 Example

Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:



### 11.2.4 Attribute Selection Measures

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as Attribute selection measure or ASM. By this measurement, we can easily select the best attribute for the nodes of the tree. There are following popular techniques for ASM, which are:

- a. Information Gain
- b. Gini Index
- c. Gain ratio

#### a. Information Gain

This method is the main method that is used to build decision trees. It reduces the information that is required to classify the tuples. It reduces the number of tests that are needed to classify the given tuple. The attribute with the highest information gain is selected.

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, we split the node and build the decision tree.
- A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

$$\text{Information Gain} = \text{Entropy}(S) - [(\text{Weighted Avg}) * \text{Entropy (each feature)}]$$

**Entropy:** Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

$$\text{Entropy}(s) = -P(\text{yes})\log_2 P(\text{yes}) - P(\text{no})\log_2 P(\text{no})$$

**Where,**

- **S**= Total number of samples
- **P(yes)**= probability of yes
- **P(no)**= probability of no

### b. Gini Index

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART (Classification and Regression Tree) algorithm.
- An attribute with the low Gini index should be preferred as compared to the high Gini index.
- It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
- Gini index can be calculated using the below formula:

$$\text{Gini Index} = 1 - \sum_j P_j^2$$

### c. Gain Ratio

- Information gain is biased towards choosing attributes with a large number of values as root nodes. It means it prefers the attribute with a large number of distinct values.

- C4.5, an improvement of ID3, uses Gain ratio which is a modification of Information gain that reduces its bias and is usually the best option.
- Gain ratio overcomes the problem with information gain by taking into account the number of branches that would result before making the split.
- It corrects information gain by taking the intrinsic information of a split into account.

$$\text{Gain Ratio (A)} = \text{Gain (A)} / \text{Splitinfo (D)}$$

### 11.2.5 Overfitting in Decision Trees

The common problem with Decision trees, especially having a table full of columns, they fit a lot. Sometimes it looks like the tree memorized the training data set. If there is no limit set on a decision tree, it will give you 100% accuracy on the training data set because in the worst case it will end up making 1 leaf for each observation. Thus this affects the accuracy when predicting samples that are not part of the training set. Overfitting reduces the predictive nature of the decision tree.

Here are two ways to remove overfitting:

- a. Pruning Decision Trees.
- b. Random Forest
- a. **Pruning Decision Trees:** The splitting process results in fully grown trees until the stopping criteria are reached. But, the fully grown tree is likely to overfit the data, leading to poor accuracy on unseen data. Pruning is the method of removing the unused branches from the decision tree. Some branches of the decision tree might represent outliers or noisy data. Tree pruning is the method to reduce the unwanted branches of the tree. This will reduce the complexity of the tree and help in effective predictive analysis. It reduces the overfitting as it removes the unimportant branches from the trees.
- b. **Random Forest:** Many decision trees can produce more accurate predictions than just one single decision tree by itself. Indeed, the random forest algorithm is a supervised classification algorithm that builds N slightly differently trained decision trees and merges them together to get more accurate and stable predictions. Random Forest is an example of ensemble learning; in which we combine multiple machine learning algorithms to obtain better predictive performance.

## **11.2.6 Better Linear or tree-based models**

It depends on the kind of problem you are solving.

1. If the relationship between dependent & independent variables is well approximated by a linear model, linear regression will outperform the tree-based model.
2. If there is a high non-linearity & complex relationship between dependent & independent variables, a tree model will outperform a classical regression method.
3. If you need to build a model that is easy to explain to people, a decision tree model will always do better than a linear model. Decision tree models are even simpler to interpret than linear regression!

## **11.2.7 Advantages and Disadvantages**

### **Advantages of the Decision Tree**

- It is simple to understand as it follows the same process which a human follows while making any decision in real-life.
- It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.
- There is less requirement of data cleaning compared to other algorithms.

### **Disadvantages of the Decision Tree**

- The decision tree contains lots of layers, which makes it complex.
- It may have an overfitting issue, which can be resolved using the **Random Forest algorithm**.
- For more class labels, the computational complexity of the decision tree may increase.

---

## **11.3 Iterative Dichotomiser 3 (ID3)**

---

ID3 stands for Iterative Dichotomiser 3 and is named such because the algorithm iteratively (repeatedly) dichotomizes(divides) features into two or more groups at each step. ID3 algorithm uses Information Gain to decide which attribute is to be used classify the current subset of the data. For each level of the tree, information gain is calculated for the remaining data recursively.

ID3 is a precursor to the C4.5 Algorithm. The ID3 follows the Occam's razor principle. Attempts to create the smallest possible decision tree. It uses a top-down greedy approach

to build a decision tree. In simple words, the top-down approach means that we start building the tree from the top and the greedy approach means that at each iteration we select the best feature at the present moment to create a node.

### 11.3.1 History of ID3

The ID3 algorithm was invented by Ross Quinlan. Quinlan was a computer science researcher in data mining, and decision theory. Received doctorate in computer science at the University of Washington in 1968.

### 11.3.2 ID 3 Algorithm

- Calculate the entropy of every attribute using the data set
- Split the set into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum)
- Make a decision tree node containing that attribute
- Recurse on subsets using remaining attributes

#### Entropy

- In order to define information gain precisely, we need to discuss entropy first.
- A formula to calculate the homogeneity of a sample.
- A completely homogeneous sample has entropy of 0 (leaf node).
- An equally divided sample has entropy of 1.

The formula for entropy is:

$$\text{Entropy}(S) = -p(I) \log_2 p(I)$$

where  $p(I)$  is the proportion of  $S$  belonging to class  $I$ .

$\sum$  is over total outcomes.

Log2 is log base 2.

Example

If  $S$  is a collection of 14 examples with 9 YES and 5 NO examples

Then,

$$\text{Entropy}(S) = - (9/14) \log_2 (9/14) - (5/14) \log_2 (5/14) = 0.940$$

### 11.3.3 Advantages and Disadvantages of ID3

#### Advantage of ID3

- Understandable prediction rules are created from the training data.
- Builds the fastest tree.
- Builds a short tree.
- Only need to test enough attributes until all data is classified.

- Finding leaf nodes enables test data to be pruned, reducing number of tests.

### **Disadvantage of ID3**

- Data may be over-fitted or overclassified, if a small sample is tested.
- Only one attribute at a time is tested for making a decision.
- Classifying continuous data may be computationally expensive, as many trees must be generated to see where to break the continuum.

## 11.4 C4.5

The C4.5 algorithm was proposed in 1993, again by Ross Quinlan, to overcome the limitations of ID3 algorithm. This algorithm is a famous algorithm in Data Mining. The C4.5 algorithm acts as a Decision Tree Classifier. The C4.5 algorithm is very helpful to generate a useful decision, that is based on a sample of data.

This algorithm uses either Information gain or Gain ratio to decide upon the classifying attribute. It is a direct improvement from the ID3 algorithm as it can handle both continuous and missing attribute values. C4.5 is given a set of data representing things that are already classified. When we generate the decision trees with the help of C4.5 algorithm, then it can be used for classification of the dataset, and that is the main reason due to which C4.5 is also known as a statistical classifier.

C4.5 algorithm is an algorithm to form a decision tree by counting the value of the gain, where the biggest gains are to be used as an initial node or the root node. C4.5 algorithms step in building a decision tree as follows:

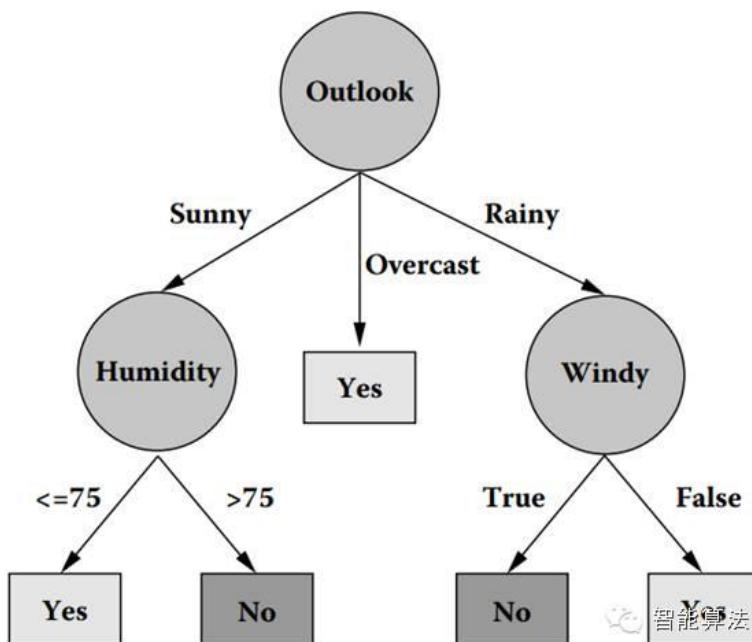
- Select the attribute with the largest gain value as the root.
- Create a branch for each value.
- For the case of the branches.
- Repeat the process for each branch until all cases the branches have the same class.

\*Note: Entropy and Gain formula given in\_\_\_\_\_.

The following figure is a decision tree generated by a typical C4.5 algorithm on a data set. The data set is shown in Figure 1, which represents the relationship between weather conditions and whether to go golfing.

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

Data set 1



Decision tree generated by C4.5 on the data set

#### 11.4.1 Algorithm description

C4.5 is not an algorithm, but a set of algorithms—C4.5, non-prune C4.5 and C4.5 rules. The pseudo code in the figure below will give the basic workflow of C4.5:

##### Algorithm 1.1 C4.5(D)

**Input:** an attribute-valued dataset  $D$

- 1: Tree = {}
- 2: **if**  $D$  is “pure” OR other stopping criteria met **then**
- 3:     **terminate**
- 4: **end if**
- 5: **for all** attribute  $a \in D$  **do**
- 6:     Compute information-theoretic criteria if we split on  $a$
- 7: **end for**
- 8:  $a_{best}$  = Best attribute according to above computed criteria
- 9: Tree = Create a decision node that tests  $a_{best}$  in the root
- 10:  $D_v$  = Induced sub-datasets from  $D$  based on  $a_{best}$
- 11: **for all**  $D_v$  **do**
- 12:     Tree <sub>$v$</sub>  = C4.5( $D_v$ )
- 13:     Attach Tree <sub>$v$</sub>  to the corresponding branch of Tree
- 14: **end for**
- 15: **return** Tree



C4.5 algorithm flow

### **11.4.2 Pseudocode**

- Check for the above base cases.
- For each attribute  $a$ , find the normalised information gain ratio from splitting on  $a$ .
- Let  $a_{best}$  be the attribute with the highest normalized information gain.
- Create a decision node that splits on  $a_{best}$ .
- Recur on the sublists obtained by splitting on  $a_{best}$ , and add those nodes as children of node.

### **11.4.3 Advantages of C4.5 over other Decision Tree systems:**

- The algorithm inherently employs Single Pass Pruning Process to Mitigate overfitting.
- It can work with both **Discrete** and **Continuous Data**
- C4.5 can handle the issue of incomplete data very well

Further, it is important to know that C4.5 is not the best algorithm in all cases, but it is very useful in some situations.

---

## **11.5 CART (Classification and Regression Tree)**

---

As name include the CART or Classification & Regression Trees methodology refers to these two types of decision trees. It is a dynamic learning algorithm which can produce a classification tree as well as regression tree depending upon the dependent variable. Classification and regression trees (CART) are a set of techniques for classification and prediction. The technique is aimed at producing rules that predict the value of an outcome (target) variable from known values of predictor variables. The predictor variables may be a mixture of categorical and continuous variables. The Classification and regression tree(CART) methodology are one of the oldest and most fundamental algorithms. It is used to predict outcomes based on certain predictor variables.

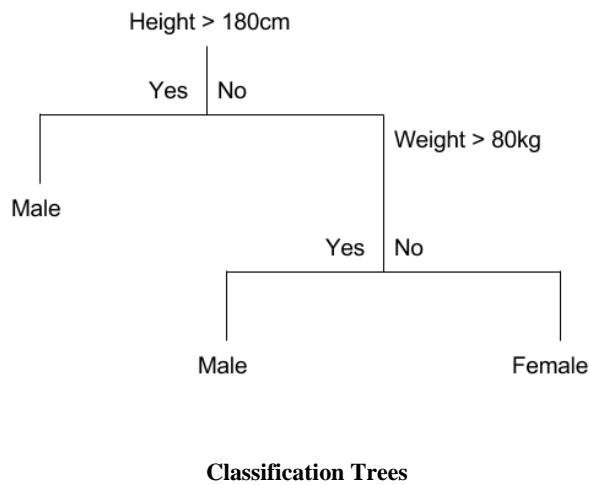
### **11.5.1 Classification Trees:**

A classification tree is an algorithm where the target variable is fixed or categorical. The algorithm is then used to identify the “class” within which a target variable would most likely fall.

An example of a classification-type problem would be determining who will or will not subscribe to a digital platform; or who will or will not graduate from high school.

These are examples of simple binary classifications where the categorical dependent variable can assume only one of two, mutually exclusive values. In other cases, you might have to predict among a number of different variables. For instance, you may have to predict which type of smartphone a consumer may decide to purchase.

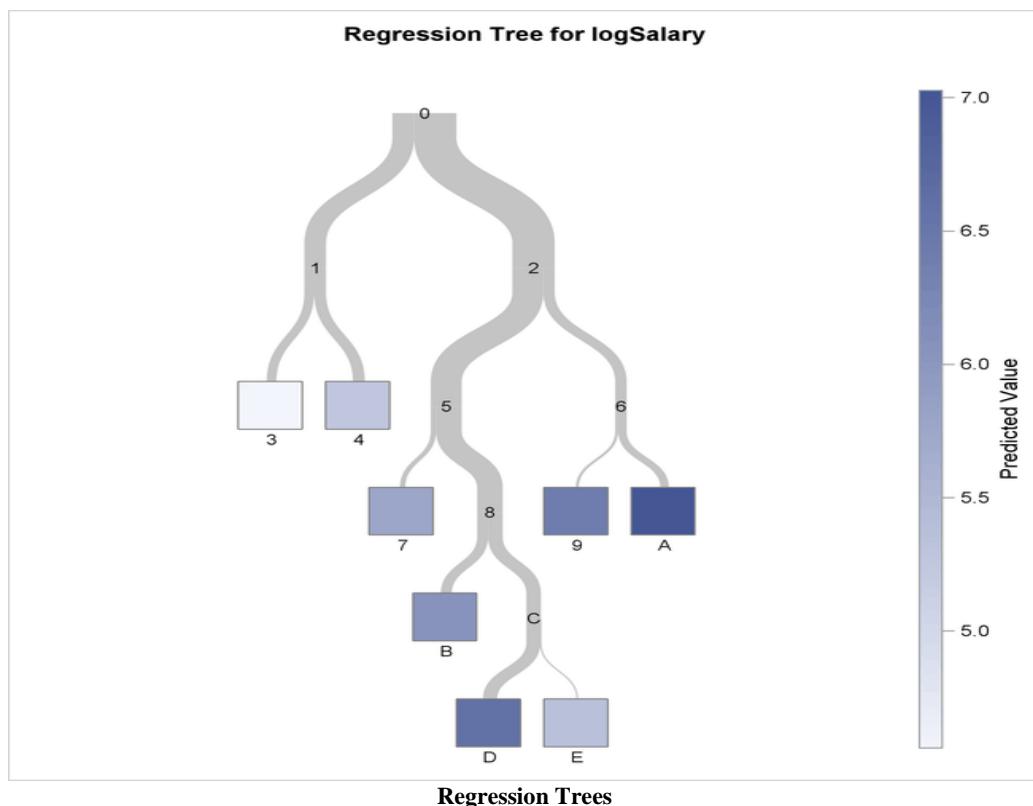
In such cases, there are multiple values for the categorical dependent variable. Here's what a classic classification tree looks like.



### 11.5.2 Regression Trees:

A regression tree refers to an algorithm where the target variable is and the algorithm is used to predict its value. As an example of a regression type problem, you may want to predict the selling prices of a residential house, which is a continuous dependent variable.

This will depend on both continuous factors like square footage as well as categorical factors like the style of home, area in which the property is located, and so on.



### **11.5.3            Difference Between Classification and Regression Trees**

Decision trees are easily understood and there are several classification and regression trees ppts to make things even simpler. However, it's important to understand that there are some fundamental differences between classification and regression trees.

#### **When to use Classification and Regression Trees**

Classification trees are used when the dataset needs to be split into classes that belong to the response variable. In many cases, the classes Yes or No.

In other words, they are just two and mutually exclusive. In some cases, there may be more than two classes in which case a variant of the classification tree algorithm is used.

Regression trees, on the other hand, are used when the response variable is continuous. For instance, if the response variable is something like the price of a property or the temperature of the day, a regression tree is used.

In other words, regression trees are used for prediction-type problems while classification trees are used for classification-type problems.

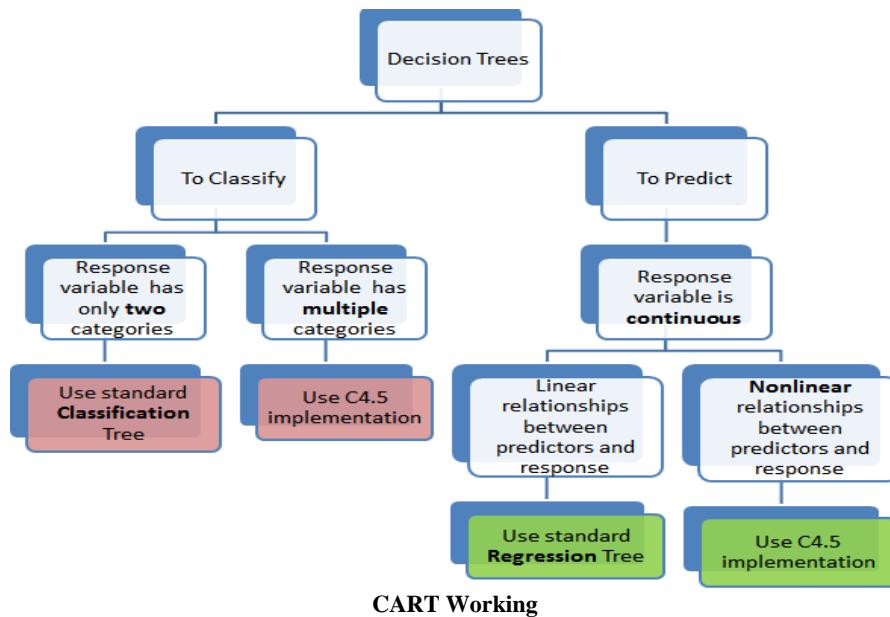
#### **How Classification and Regression Trees Work**

A classification tree splits the dataset based on the homogeneity of data. Say, for instance, there are two variables; income and age; which determine whether or not a consumer will buy a particular kind of phone.

If the training data shows that 95% of people who are older than 30 bought the phone, the data gets split there and age becomes a top node in the tree. This split makes the data “95% pure”. Measures of impurity like entropy or Gini index are used to quantify the homogeneity of the data when it comes to classification trees.

In a regression tree, a regression model is fit to the target variable using each of the independent variables. After this, the data is split at several points for each independent variable.

At each such point, the error between the predicted values and actual values is squared to get “A Sum of Squared Errors” (SSE). The SSE is compared across the variables and the variable or point which has the lowest SSE is chosen as the split point. This process is continued recursively.



#### 11.5.4 Advantages of Classification and Regression Trees

The purpose of the analysis conducted by any classification or regression tree is to create a set of if-else conditions that allow for the accurate prediction or classification of a case.

Classification and regression trees work to produce accurate predictions or predicted classifications, based on the set of if-else conditions. They usually have several advantages over regular decision trees.

##### (i) The Results are Simplistic

The interpretation of results summarized in classification or regression trees is usually fairly simple. The simplicity of results helps in the following ways.

1. It allows for the rapid classification of new observations. That's because it is much simpler to evaluate just one or two logical conditions than to compute scores using complex nonlinear equations for each group.
2. It can often result in a simpler model which explains why the observations are either classified or predicted in a certain way. For instance, business problems are much easier to explain with if-then statements than with complex nonlinear equations.

##### (ii) Classification and Regression Trees are Nonparametric & Nonlinear

The results from classification and regression trees can be summarized in simplistic if-then conditions. This negates the need for the following implicit assumptions.

1. The predictor variables and the dependent variable are linear.

2. The predictor variables and the dependent variable follow some specific nonlinear link functions.
3. The predictor variables and the dependent variable are monotonic.

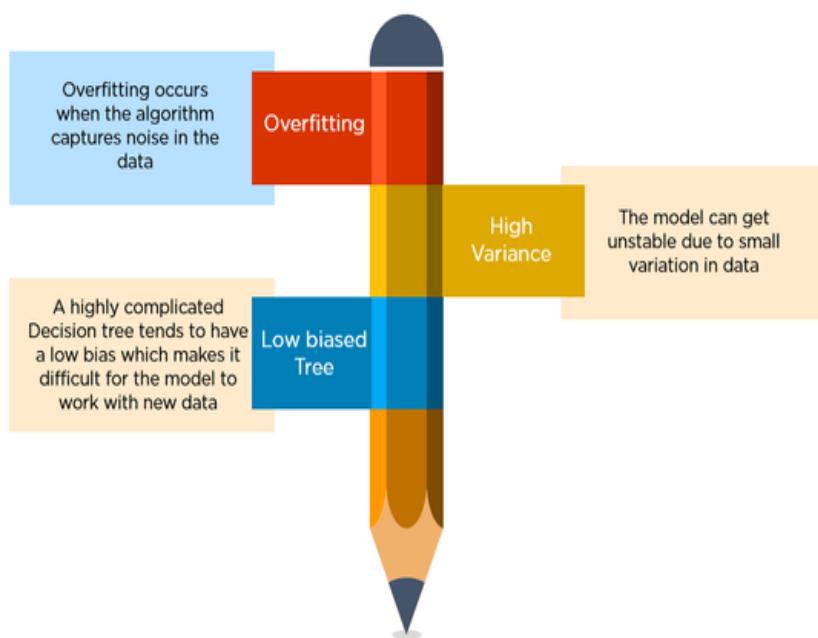
Since there is no need for such implicit assumptions, classification and regression tree methods are well suited to data mining. This is because there is very little knowledge or assumptions that can be made beforehand about how the different variables are related.

As a result, classification and regression trees can actually reveal relationships between these variables that would not have been possible using other techniques.

### **(iii) Classification and Regression Trees Implicitly Perform Feature Selection**

Feature selection or variable screening is an important part of analytics. When we use decision trees, the top few nodes on which the tree is split are the most important variables within the set. As a result, feature selection gets performed automatically and we don't need to do it again.

#### **11.5.5        Limitations of Classification and Regression Trees**



Classification and regression tree tutorials, as well as classification and regression tree ppts, exist in abundance. This is a testament to the popularity of these decision trees and how frequently they are used. However, these decision trees are not without their disadvantages. There are many classification and regression tree examples where the use of a decision tree has not led to the optimal result. Here are some of the limitations of classification and regression trees.

##### **i. Overfitting**

Overfitting occurs when the tree takes into account a lot of noise that exists in the data and comes up with an inaccurate result.

### **ii. High variance**

In this case, a small variance in the data can lead to a very high variance in the prediction, thereby affecting the stability of the outcome.

### **iii. Low bias**

A decision tree that is very complex usually has a low bias. This makes it very difficult for the model to incorporate any new data.

---

## **11.6 Conclusion**

---

Decision Trees are data mining techniques for classification and regression analysis. This technique is now spanning over many areas like medical diagnosis, target marketing, etc. These trees are constructed by following an algorithm such as ID3, CART. These algorithms find different ways to split the data into partitions.

Decision tree algorithms transform the raw data into rule based mechanism. Even though decision tree algorithms are powerful, they have long training time. On the other hand, they tend to fall over-fitting. It is the most widely known supervised learning technique that is used in machine learning and pattern analysis. The decision trees predict the values of the target variable by building models through learning from the training set provided to the system.

---

## **11.7 Summary**

---

In this Chapter, we have mentioned one of the most common decision tree algorithm named as ID3. They can use nominal attributes whereas most of common machine learning algorithms cannot. However, it is required to transform numeric attributes to nominal in ID3. Besides, its evolved version C4.5 exists which can handle nominal data. CART methodology are one of the oldest and most fundamental algorithms. All these are excellent for data mining tasks because they require very little data pre-processing. Decision tree models are easy to understand and implement which gives them a strong advantage when compared to other analytical models.

---

## **11.9 References**

---

- Anurag Upadhyay, Suneet Shukla, Sudsanshu Kumar, Empirical Comparison by data mining Classification algorithms
- Jaiwei Han and Micheline Kamber, Data Mining Concepts and Techniques.Morgan Kaufmann Publishers.
- <https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm>
- <https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/>

## UNIT STRUCTURE

### MODULE 6 – CHAPTER 1

#### ADVANCED DATABASE MANAGEMENT SYSTEM

- 6.1 What is Clustering?
- 6.2 Requirements of Clustering
- 6.3 Clustering Vs Classification
- 6.4 Types of Clusters
- 6.5 Distinctions between Sets of Clusters
- 6.6 What is Cluster Analysis?
- 6.7 Applications of Cluster Analysis
- 6.8 What kind of classification is not considered a cluster analysis?
- 6.9 General Algorithmic Issues

#### **6.1 What is Clustering?**

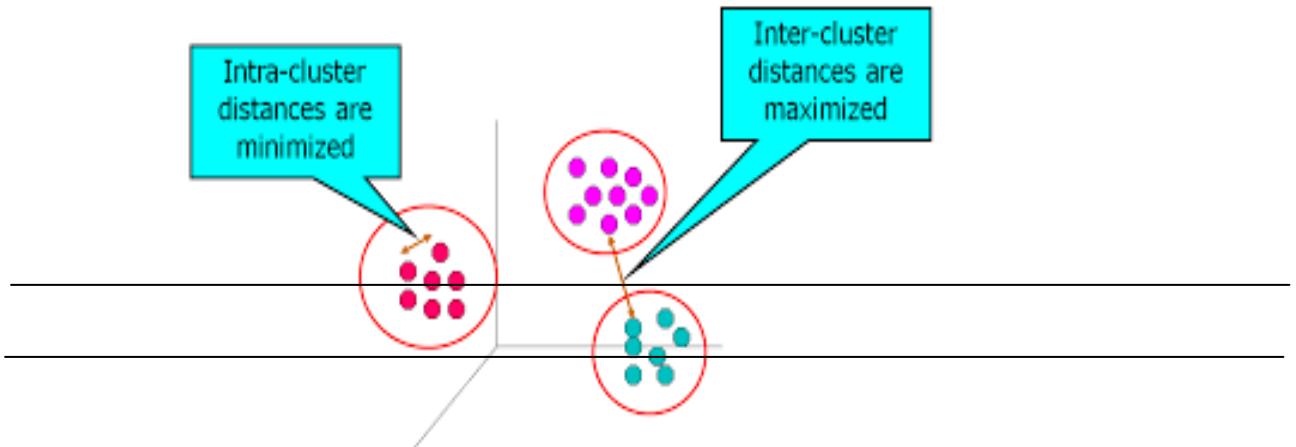
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Let's understand this with an example.

Suppose, you are the head of a rental store and wish to understand preferences of your customers to scale up your business. Is it possible for you to look at details of each customer and devise a unique business strategy for each one of them? What you can do is to cluster all of your customers into say 10 groups based on their purchasing habits and use a separate strategy for customers in each of these 10 groups which is called clustering

Cluster is a collection of data objects

- Similar (or related) to one another within the same group
- Dissimilar (or unrelated) to one another in other groups



### Points to Remember

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, first the set of data is partitioned into groups based on data similarity and then labels are assigned to the groups.
- The main advantage of clustering is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

## 6.2 Requirements of Clustering

This section is to make you learn about the requirements for clustering as a data mining tool, as well as aspects that can be used for comparing clustering methods. The following are typical requirements of clustering in data mining.

- **Scalability:**

Many clustering algorithms work well on small data sets containing fewer than several hundred data objects; however, a large database may contain millions or even billions of objects, particularly in Web search scenarios. Clustering on only a sample of a given large data set may lead to biased results. Therefore, highly scalable clustering algorithms are needed.

- **Ability to deal with different types of attributes:**

Many algorithms are designed to cluster numeric (interval-based) data. However, applications may require clustering other data types, such as binary, nominal (categorical), and ordinal data, or mixtures of these data types. Recently, more and more applications need clustering techniques for complex data types such as graphs, sequences, images, and documents.

- **Discovery of clusters with arbitrary shape:**

Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures (Chapter 2). Algorithms based on such distance measures tend to find spherical clusters with similar size and density. However, a cluster could be of any shape. Consider sensors, for example, which are often deployed for environment surveillance. Cluster analysis on sensor readings can detect interesting phenomena. We may want to use clustering to find the frontier of a running forest fire, which is often not spherical. It is important to develop algorithms that can detect clusters of arbitrary shape.

- **Requirements for domain knowledge to determine input parameters:**

Many clustering algorithms require users to provide domain knowledge in the form of input parameters such as the desired number of clusters. Consequently, the clustering results may be sensitive to such parameters. Parameters are often hard to determine, especially for high-dimensionality data sets and where users have yet to grasp a deep understanding of their data. Requiring the specification of domain knowledge not only burdens users, but also makes the quality of clustering difficult to control.

- **Ability to deal with noisy data:**

Most real-world data sets contain outliers and/or missing, unknown, or erroneous data. Sensor readings, for example, are often noisy—some readings may be inaccurate due to the sensing mechanisms, and some readings may be erroneous due to interferences from surrounding transient objects. Clustering algorithms can be sensitive to such noise and may produce poor-quality clusters. Therefore, we need clustering methods that are robust to noise.

- **Incremental clustering and insensitivity to input order:**

In many applications, incremental updates (representing newer data) may arrive at any time. Some clustering algorithms cannot incorporate incremental updates into existing clustering structures and, instead, have to recompute a new clustering from scratch. Clustering algorithms may also be sensitive to the input data order. That is, given a set of data objects, clustering algorithms may

return dramatically different clusterings depending on the order in which the objects are presented. Incremental clustering algorithms and algorithms that are insensitive to the input order are needed.

- **Capability of clustering high-dimensionality data:**

A data set can contain numerous dimensions or attributes. When clustering documents, for example, each keyword can be regarded as a dimension, and there are often thousands of keywords. Most clustering algorithms are good at handling low-dimensional data such as data sets involving only two or three dimensions. Finding clusters of data objects in a high-dimensional space is challenging, especially considering that such data can be very sparse and highly skewed.

- **Constraint-based clustering:**

Real-world applications may need to perform clustering under various kinds of constraints. Suppose that your job is to choose the locations for a given number of new automatic teller machines (ATMs) in a city. To decide upon this, you may cluster households while considering constraints such as the city's rivers and highway networks and the types and number of customers per cluster. A challenging task is to find data groups with good clustering behavior that satisfy specified constraints.

- **Interpretability and usability:**

Users want clustering results to be interpretable, comprehensible, and usable. That is, clustering may need to be tied in with specific semantic interpretations and applications. It is important to study how an application goal may influence the selection of clustering features and clustering methods.

---

### 6.3 Clustering Vs Classification

---

#### Difference between Clustering and Classification

Clustering and classification techniques are used in machine-learning, information retrieval, image investigation, and related tasks. These two strategies are the two main divisions of data mining processes. In the data analysis world, these are essential in managing algorithms. Specifically, both of these processes divide data into sets. This task is highly relevant in today's information age as the immense increase of data coupled with development

needs to be aptly facilitated. Notably, clustering and classification help solve global issues such as crime, poverty, and diseases through data science.

## **What is Clustering?**

Basically, clustering involves grouping data with respect to their similarities. It is primarily concerned with distance measures and clustering algorithms which calculate the difference between data and divide them systematically.

For instance, students with similar learning styles are grouped together and are taught separately from those with differing learning approaches. In data mining, clustering is most commonly referred to as “unsupervised learning technic” as the grouping is based on a natural or inherent characteristic. It is applied in several scientific fields such as information technology, biology, criminology, and medicine.

## **Characteristics of Clustering**

- **No Exact Definition**

Clustering has no precise definition that is why there are various clustering algorithms or cluster models. Roughly speaking, the two kinds of clustering are hard and soft. Hard clustering is concerned with labeling an object as simply belonging to a cluster or not. In contrast, soft clustering or fuzzy clustering specifies the degree as to how something belongs to a certain group.

- **Difficult to be Evaluated**

The validation or assessment of results from clustering analysis is often difficult to ascertain due to its inherent inexactness.

- **Unsupervised**

As it is an unsupervised learning strategy, the analysis is merely based on current features; thus, no stringent regulation is needed.

## **What is Classification?**

Classification entails assigning labels to existing situations or classes; hence, the term “classification”. For example, students exhibiting certain learning characteristics are classified as visual learners. Classification is also known as “supervised learning technic” wherein machines learn from already labeled or classified data. It is highly applicable in pattern recognition, statistics, and biometrics.

## **Characteristics of Classification**

- **Utilizes a “Classifier”**

To analyze data, a classifier is a defined algorithm that concretely maps information to a specific class. For example, a classification algorithm would train a model to identify whether a certain cell is malignant or benign.

- **Evaluated Through Common Metrics**

The quality of a classification analysis is often assessed via precision and recall which are popular metric procedures. A classifier is evaluated regarding its accuracy and sensitivity in identifying the output.

- **Supervised**

Classification is a supervised learning technic as it assigns previously determined identities based on comparable features. It deduces a function from a labeled training set.

## **Differences between Clustering and Classification**

### **Supervision**

The main difference is that clustering is unsupervised and is considered as “self-learning” whereas classification is supervised as it depends on predefined labels.

### **Use of Training Set**

Clustering does not poignantly employ training sets, which are groups of instances employed to generate the groupings, while classification imperatively needs training sets to identify similar features.

### **Labeling**

Clustering works with unlabeled data as it does not need training. On the other hand, classification deals with both unlabeled and labeled data in its processes.

### **Goal**

Clustering groups objects with the aim to narrow down relations as well as learn novel information from hidden patterns while classification seeks to determine which explicit group a certain object belongs to.

### **Specifics**

While classification does not specify what needs to be learned, clustering specifies the required improvement as it points out the differences by considering the similarities between data.

## **Phases**

Generally, clustering only consists of a single phase (grouping) while classification has two stages, training (model learns from training data set) and testing (target class is predicted).

## **Boundary Conditions**

Determining the boundary conditions is highly important in the classification process as compared to clustering. For instance, knowing the percentage range of “low” as compared to “moderate” and “high” is needed in establishing the classification.

## **Prediction**

As compared to clustering, classification is more involved with prediction as it particularly aims to identify target classes. For instance, this may be applied in “facial key points detection” as it can be used in predicting whether a certain witness is lying or not.

## **Complexity**

Since classification consists of more stages, deals with prediction, and involves degrees or levels, its’ nature is more complicated as compared to clustering which is mainly concerned with grouping similar attributes.

## **Number of Probable Algorithms**

Clustering algorithms are mainly linear and nonlinear while classification consists of more algorithmic tools such as linear classifiers, neural networks, Kernel estimation, decision trees, and support vector machines.

## **Clustering vs Classification: Table comparing the difference between Clustering and Classification**

<b>Clustering</b>	<b>Classification</b>
Unsupervised data	Supervised data
Does not highly value training sets	Does highly value training sets
Works solely with unlabeled data	Involves both unlabeled and labeled data
Aims to identify similarities among data	Aims to verify where a datum belongs to

Specifies required change	Does not specify required improvement
Has a single phase	Has two phases
Determining boundary conditions is not paramount	Identifying the boundary conditions is essential in executing the phases
Does not generally deal with prediction	Deals with prediction
Mainly employs two algorithms	Has a number of probable algorithms to use
Process is less complex	Process is more complex

---

## 6.4 Types of clusters

---

Broadly speaking, clustering can be divided into two subgroups:

- **Hard Clustering:**

In hard clustering, each data point either belongs to a cluster completely or not. For example, in the above example each customer is put into one group out of the 10 groups.

- **Soft Clustering:**

In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, from the above scenario each costumer is assigned a probability to be in either of 10 clusters of the retail store.

Apart from which clusters can also be divided into

- Well separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or conceptual
- Described by an objective function

## Well separated clusters

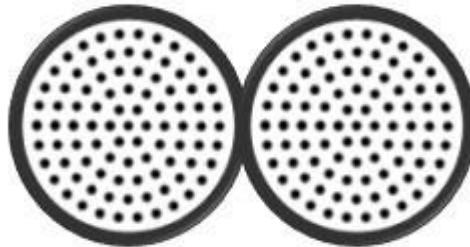
A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster



- These clusters need not be globular but, can have any shape.
- Sometimes a threshold is used to specify that all the objects in a cluster must sufficiently close to one another. Definition of a cluster is satisfied only when the data contains natural clusters.

## Centre based cluster

- A cluster is a set of points such that a point in a cluster is closer (more similar) to the “center” of that cluster than to the center of any other cluster.



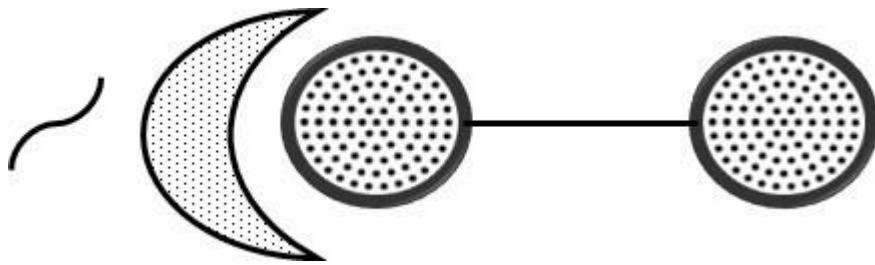
- The center of a cluster can be either centroid or medoid

If the data is numerical, the prototype of the cluster is often a **centroid** i.e., the average of all the points in the cluster.

If the data has categorical attributes, the prototype of the cluster is often a **medoid** i.e., the most representative point of the cluster.

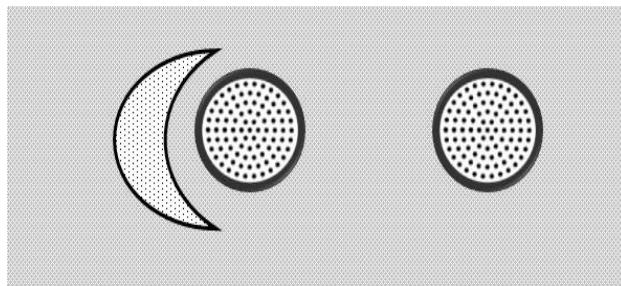
- “Center-Based” Clusters can also be referred as Prototype based clusters.
- These clusters tend to be globular.
- **K-Means and K-Medoids** are the examples of Prototype Based Clustering algorithms

## Contiguous cluster (Nearest neighbour or transitive clustering)



- Two objects are connected only if they are within a specified distance of each other.
- Each point in a cluster is closer to at least one point in the same cluster than to any point in a different cluster.
- Useful when clusters are irregular and intertwined.
- **Clique** is another type of Graph Based
- **Agglomerative hierarchical clustering** has close relation with Graph based clustering technique.

### **Density based cluster definition**



- A cluster is dense region of points, which is separated by low-density regions, from other regions of high density.
- Density based clusters are employed when the clusters are irregular, intertwined and when noise and outliers are present.
- Points in low density region are classified as noise and omitted.
- **DBSCAN** is an example of Density based clustering algorithm.

### **Shared property or conceptual clusters**

- A cluster is set of objects that share some common property or represent a particular concept.
- The most general notion of a cluster; in some ways includes all other types.

### **Clusters defined by an objective function**

- Set of clusters minimizes or maximizes some objective function.
- Enumerate all possible ways of dividing the points into clusters and evaluate the 'goodness' of each potential set of clusters by using the given objective function (NP-hard).
- Can have global or local objective function.

- Hierarchical clustering algorithms typically have local objective function.
  - Partitional algorithms typically have global objective function.
- 

## 6.5 Distinctions between Sets of Clusters

---

- Exclusive versus non-exclusive
- Fuzzy versus non-fuzzy
- Partial versus complete
- Heterogeneous versus homogeneous

### Exclusive versus non-exclusive

In non-exclusive clusterings, points may belong to multiple clusters. – Can represent multiple classes or ‘border’ points

### Fuzzy versus non-fuzzy

In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1 – Weights must sum to 1 – Probabilistic clustering has similar characteristics

### Partial versus complete

In some cases, we only want to cluster some of the data

### Heterogeneous versus homogeneous

Cluster of widely different sizes, shapes, and densities

---

## 6.6 What is cluster analysis?

---

Cluster analysis or clustering or data segmentation is a type of strategy that is used to categorize objects or cases into proximate groups called clusters. For instance, in the insurance providers, these steps in cluster analysis help segregate fraudulent access of the customer data. Cluster analysis is applied in

- Data reduction
  - Summarization: Preprocessing for regression, PCA, Classification and association analysis
- Hypothesis generation and testing
- Prediction based on groups:
  - Cluster and find characteristics/patterns for each group
- Find K-nearest neighbors
- Localizing search to one or a small number of clusters
- Outlier detection:
  - Outliers are often viewed as those “far away” from any cluster

## Types Of Data Used In Cluster Analysis Are:

- Interval-Scaled variables
- Binary variables
- Nominal, Ordinal, and Ratio variables
- Variables of mixed types

### Interval-Scaled Variables

Interval-scaled variables are continuous measurements of a roughly linear scale. Typical examples include weight and height, latitude and longitude coordinates (e.g., when clustering houses), and weather temperature. The measurement unit used can affect the clustering analysis. For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to a very different clustering structure.

In general, expressing a variable in smaller units will lead to a larger range for that variable, and thus a larger effect on the resulting clustering structure. To help avoid dependence on the choice of measurement units, the data should be standardized. Standardizing measurements attempts to give all variables an equal weight.

This is especially useful when given no prior knowledge of the data. However, in some applications, users may intentionally want to give more weight to a certain set of variables than to others. For example, when clustering basketball player candidates, we may prefer to give more weight to the variable height.

### Binary Variables

A binary variable is a variable that can take only 2 values. For example, generally, gender variables can take 2 variables male and female.

#### Contingency Table For Binary Data

Let us consider binary values 0 and 1

Let  $p=a+b+c+d$

**Simple matching coefficient** (invariant, if the binary variable is symmetric):  
**Jaccard coefficient** (noninvariant if the binary variable is asymmetric):

### Nominal or Categorical Variables

A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green.

### Method 1: Simple matching

The dissimilarity between two objects  $i$  and  $j$  can be computed based on the simple matching.

**m:** Let  $m$  be no of matches (i.e., the number of variables for which  $i$  and  $j$  are in the same state).

**p:** Let  $p$  be total no of variables.

### Method 2: use a large number of binary variables

Creating a new binary variable for each of the  $M$  nominal states.

## Ordinal Variables

An ordinal variable can be discrete or continuous. In this order is important, e.g., rank. It can be treated like interval-scaled. By replacing  $x_{if}$  by their rank, By mapping the range of each variable onto  $[0, 1]$  by replacing the  $i$ -th object in the  $f$ -th variable by, Then compute the dissimilarity using methods for interval-scaled variables.

## Ratio-Scaled Intervals

**Ratio-scaled variable:** It is a positive measurement on a nonlinear scale, approximately at an exponential scale, such as  $Ae^{Bt}$  or  $A^e \cdot Bt$ .

### Methods:

- First, treat them like interval-scaled variables — not a good choice! (why?)
- Then apply logarithmic transformation i.e.  $y = \log(x)$
- Finally, treat them as continuous ordinal data treat their rank as interval-scaled.

## Variables of Mixed Type

A database may contain all the six types of variables symmetric binary, asymmetric binary, nominal, ordinal, interval, and ratio. And those combinedly called as mixed-type variables.

## Types of Data Structures

Suppose that a data set to be clustered contains  $n$  objects, which may represent persons, houses, documents, countries, and so on. Main memory-based clustering algorithms typically operate on either of the following two data structures.

Types of data structures in cluster analysis are

- **Data Matrix** (or object by variable structure)
- **Dissimilarity Matrix** (or object by object structure)

## **Data Matrix**

This represents n objects, such as persons, with p variables (also called measurements or attributes), such as age, height, weight, gender, race and so on. The structure is in the form of a relational table, or n-by-p matrix (n objects x p variables). The Data Matrix is often called a two-mode matrix since the rows and columns of this represent the different entities.

## **Dissimilarity Matrix**

This stores a collection of proximities that are available for all pairs of n objects. It is often represented by a n – by – n table, where  $d(i,j)$  is the measured difference or dissimilarity 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 higher similar or “near” each other and becomes larger the more they differ. Since  $d(i,j) = d(j,i)$  and  $d(i,i) = 0$ . This is also called as one mode matrix since the rows and columns of this represent the same entity.

---

## **6.7 Applications of cluster Analysis**

---

**Marketing:** Finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;

**Biology:** Classification of plants and animals given their features;

**Libraries:** Book ordering;

**Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;

**City-planning:** Identifying groups of houses according to their house type, value and geographical location;

**Earthquake studies:** Clustering observed earthquake epicenters to identify dangerous zones;

**WWW document classification:** Clustering weblog data to discover groups of similar access patterns.

---

## **6.8 What kind of classification is not considered a cluster analysis?**

---

- **Graph Partitioning**

The type of classification where areas are not the same and are only classified based on mutual synergy and relevance is not cluster analysis.

- **Results of a query**

In this type of classification, the groups are created based on the specification given from external sources. It is not counted as a Cluster Analysis.

- **Simple Segmentation**

Division of names into separate groups of registration based on the last name does not qualify as Cluster Analysis.

- **Supervised Classification**

These types of classifications which is classified using label information cannot be said as Cluster Analysis because cluster analysis involves group based on the pattern.

---

## 6.9 General Algorithmic Issues

---

- Assessment of Results
- How Many Clusters?
- Data Preparation
- Proximity Measures
- Handling Outliers

### Assessment of Results

The data mining clustering process starts with the assessment of whether any cluster tendency has a place at all, and correspondingly includes, appropriate attribute selection, and in many cases feature construction. It finishes with the validation and evaluation of the resulting clustering system. The clustering system can be assessed by an expert, or by a particular automated procedure. Traditionally, the first type of assessment relates to two issues:

- Cluster interpretability,
- Cluster visualization.

Interpretability depends on the technique used.

### How Many Clusters?

In many methods number  $k$  of clusters to construct is an input user parameter. Running an algorithm several times leads to a sequence of clustering systems. Each system consists of more granular and less-separated clusters. In the case of k-means, the objective function is monotone decreasing. Therefore, the answer to the question

of which system is preferable is not trivial. Many criteria have been introduced to find an optimal k.

## Data Preparation

Irrelevant attributes make chances of a successful clustering futile, because they negatively affect proximity measures and eliminate clustering tendency. Therefore, sound exploratory data analysis (EDA) is essential.

## Proximity Measures

Both hierarchical and partitioning methods use different distances and similarity measures

**Euclidean (p=2) distance** is by far the most popular choice used in k-means objective function (sum of squares of distances between points and centroids) that has a clear statistical meaning of total inter-clusters variance.

**Manhattan distance** corresponds to p=1. The distance that returns the maximum of absolute difference in coordinates is also used and corresponds to  $\leq p < \infty$   $p = \infty$ . In many applications (profile analyses) points are scaled to have a unit norm

## Handling Outliers

Applications that derive their data from measurements have an associated amount of noise, which can be viewed as outliers. Alternately, outliers can be viewed as legitimate records having abnormal behavior. In general, clustering techniques do not distinguish between the two: neither noise nor abnormalities fit into clusters. Correspondingly, the preferable way to deal with outliers in partitioning the data is to keep one extra set of outliers, so as not to pollute factual clusters. There are multiple ways of how descriptive learning handles outliers. If a summarization or data preprocessing phase is present, it usually takes care of outliers.

## UNIT STRUCTURE

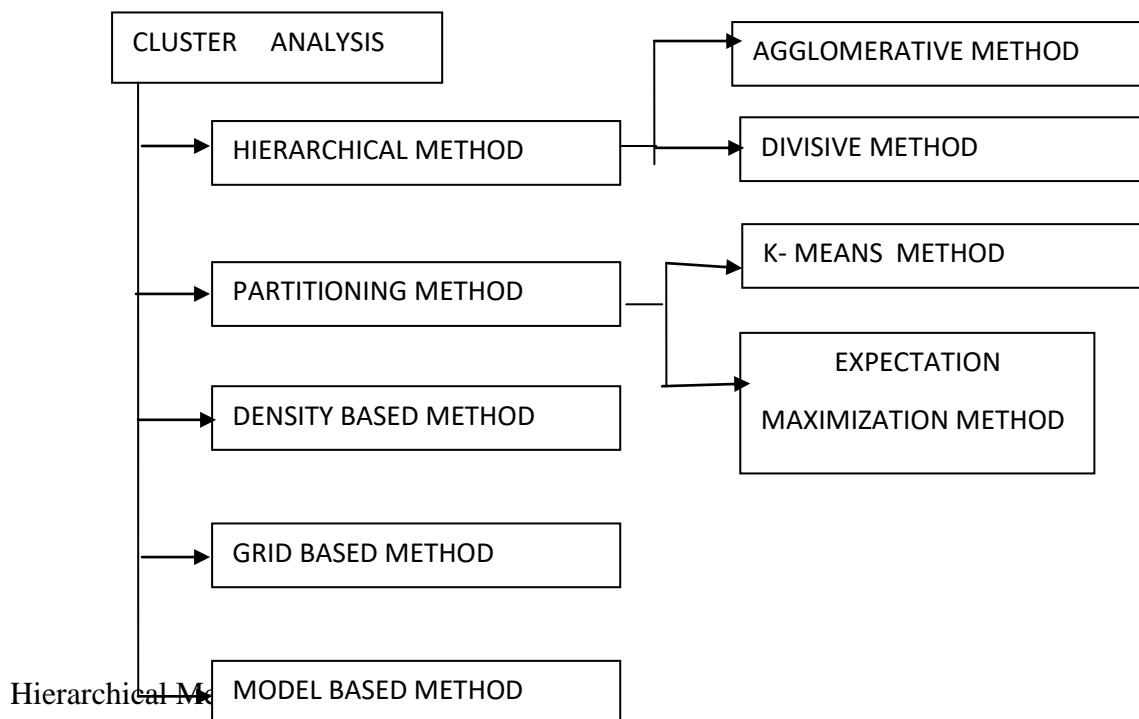
### MODULE 6 – CHAPTER 2

#### ADVANCED DATABASE MANAGEMENT SYSTEM

- 6.10 Clustering Methods
- 6.11 Clustering Algorithm Applications
- 6.12 Summary
- 6.13 Reference for further reading
- 6.14 Model Questions

## 6.10 Clustering Methods

- Hierarchical Method
- Partitioning Method
- Density-based Method
- Grid-Based Method
- Model-Based Method



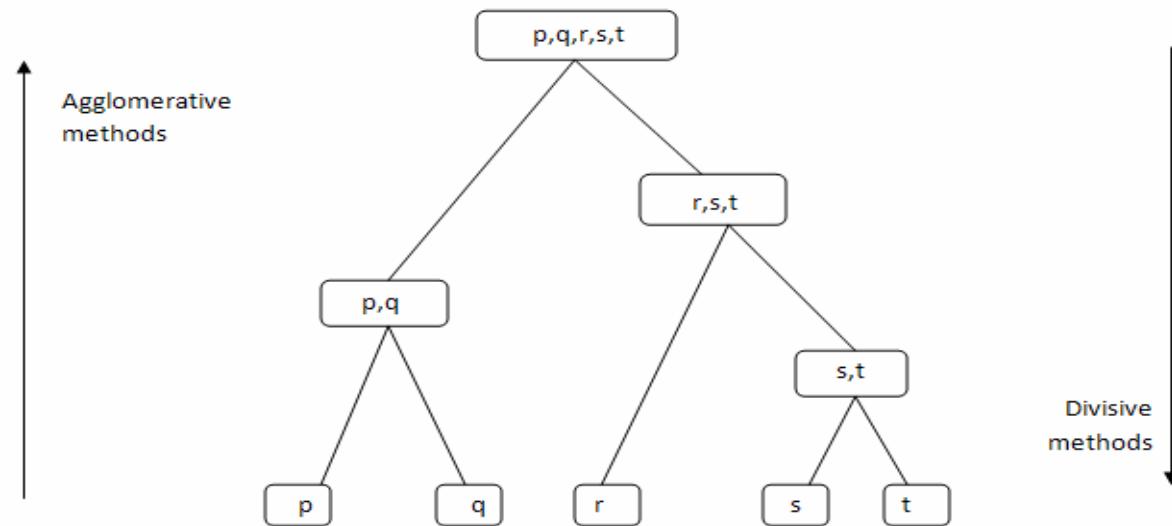
A **Hierarchical clustering** method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data points as a separate cluster. Then, it repeatedly executes the subsequent steps:

- Identify the 2 clusters which can be closest together, and
- Merge the 2 maximum comparable clusters.
- Continue these steps until all the clusters are merged together.

In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called **Dendrogram** (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or cluster are break up (top-down view).

There are two types of hierarchical clustering methods:

- Agglomerative Clustering
  - Divisive Clustering
- 
- Agglomerative — Bottom up approach. Start with many small clusters and merge them together to create bigger clusters.
  - Divisive — Top down approach. Start with a single cluster than break it up into smaller clusters.



Some pros and cons of Hierarchical Clustering

### Pros

- No apriori information about the number of clusters required.
- Easy to implement and gives best result in some cases.

### Cons

- Algorithm can never undo what was done previously.
- Time complexity of at least  $O(n^2 \log n)$  is required, where ' $n$ ' is the number of data points.
- Based on the type of distance matrix chosen for merging different algorithms can suffer with one or more of the following:
  - Sensitivity to noise and outliers
  - Breaking large clusters

- Difficulty handling different sized clusters and convex shapes
- No objective function is directly minimized
- Sometimes it is difficult to identify the correct number of clusters by the dendrogram.

The basic method to generate Agglomerative Hierarchical Clustering:

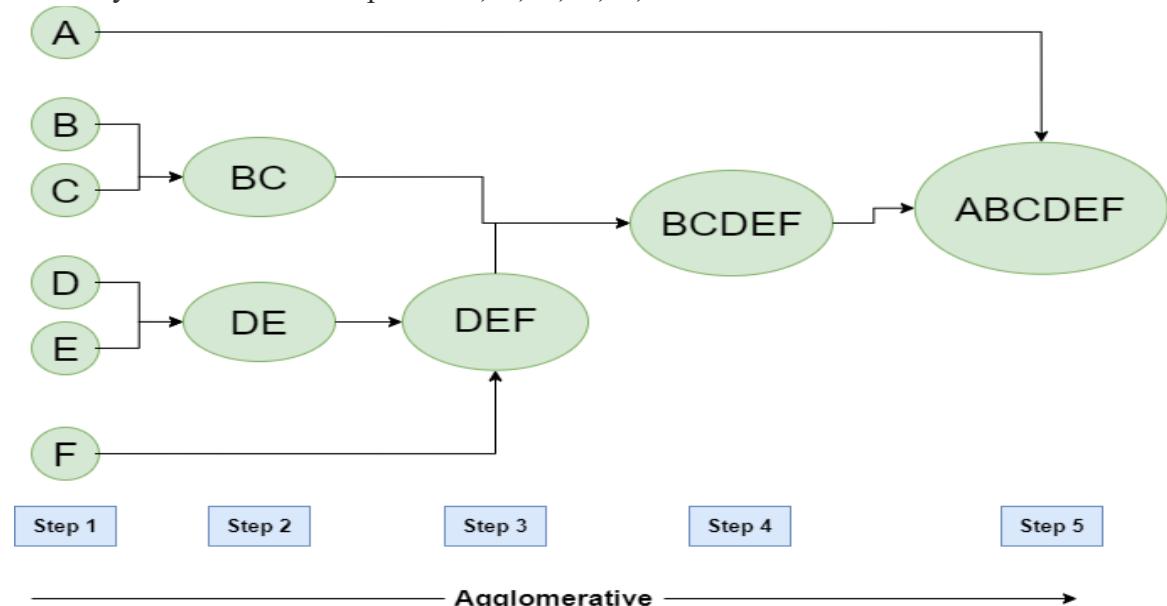
Initially consider every data point as an **individual** Cluster and at every step, **merge** the nearest pairs of the cluster. (It is a bottom-up method). At first every data set is considered as individual entity or cluster and at every iteration, the clusters merge with different clusters until one cluster is formed.

Algorithm for Agglomerative Hierarchical Clustering:

- Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix)
- Consider every data point as a individual cluster
- Merge the clusters which are highly similar or close to each other.
- Recalculate the proximity matrix for each cluster
- Repeat Step 3 and 4 until only a single cluster remains.

Graphical representation of the algorithm using a dendrogram is given below.

Let's say there are six data points **A, B, C, D, E, F**.



**Figure – Agglomerative Hierarchical clustering**

- **Step-1:**  
Consider each alphabet as a single cluster and calculate the distance of one cluster from all the other clusters.
- **Step-2:**  
In the second step comparable clusters are merged together to form a single cluster. Let's say cluster (B) and cluster (C) are very similar to each other therefore merge them in the second step similarly with cluster (D) and (E) and at last, get the clusters [(A), (BC), (DE), (F)]
- **Step-3:**  
Recalculate the proximity according to the algorithm and merge the two nearest clusters([(DE), (F)]) together to form new clusters as [(A), (BC), (DEF)]
- **Step-4:**  
Repeating the same process; The clusters DEF and BC are comparable and merged together to form a new cluster. Finally left with clusters [(A), (BCDEF)].
- **Step-5:**  
At last the two remaining clusters are merged together to form a single cluster [(ABCDEF)].

### How to join two clusters to form one cluster?

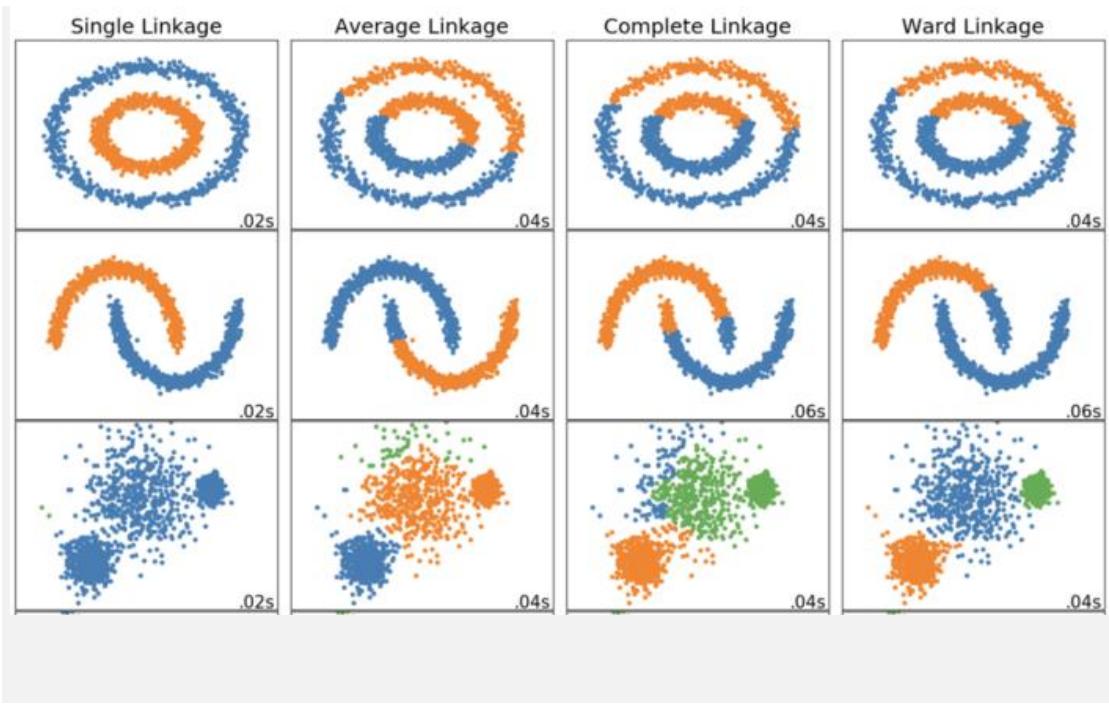
To obtain the desired number of clusters, the number of clusters needs to be reduced from initially being n cluster (n equals the total number of data-points). Two clusters are combined by computing the similarity between them.

There are some methods which are used to calculate the similarity between two clusters:

- Distance between two closest points in two clusters.
- Distance between two farthest points in two clusters.
- The average distance between all points in the two clusters.
- Distance between centroids of two clusters.

### Linkage Criteria

Similar to gradient descent, certain parameters can be tweaked to get drastically different results.



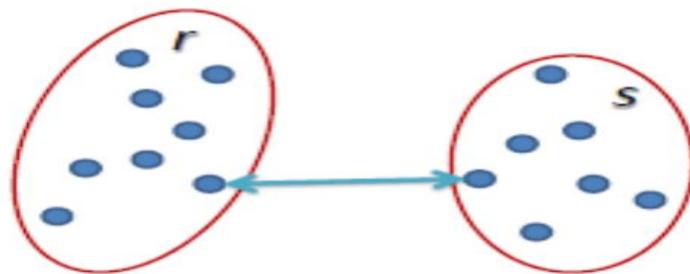
The linkage criteria refers to how the distance between clusters is calculated

Take the two closest clusters and make them one cluster



### Single Linkage

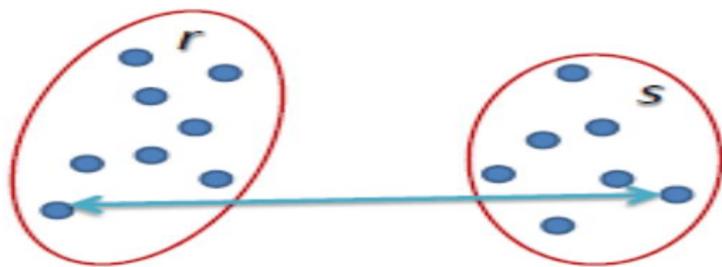
The distance between two clusters is the shortest distance between two points in each cluster



$$L(r, s) = \min(D(x_{ri}, x_{sj}))$$

### Complete Linkage

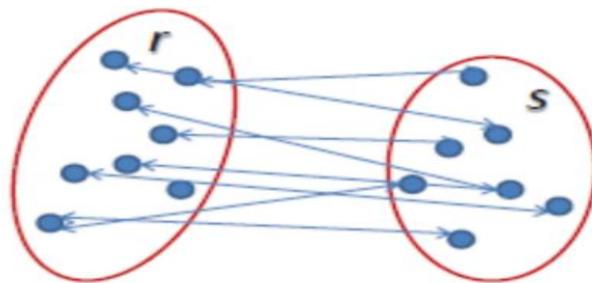
The distance between two clusters is the longest distance between two points in each cluster



$$L(r, s) = \max(D(x_{ri}, x_{sj}))$$

### Average Linkage

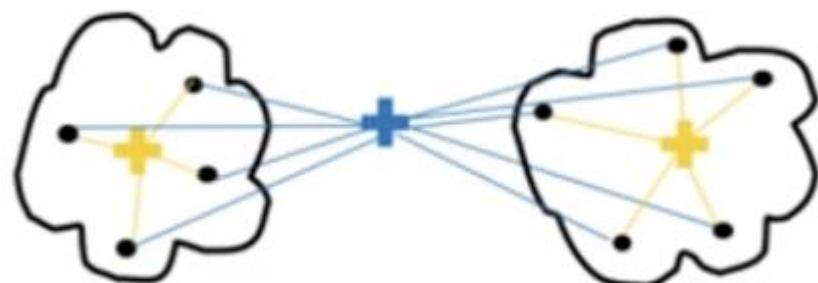
The distance between clusters is the average distance between each point in one cluster to every point in other cluster



$$L(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

### Ward Linkage

The distance between clusters is the sum of squared differences within all clusters

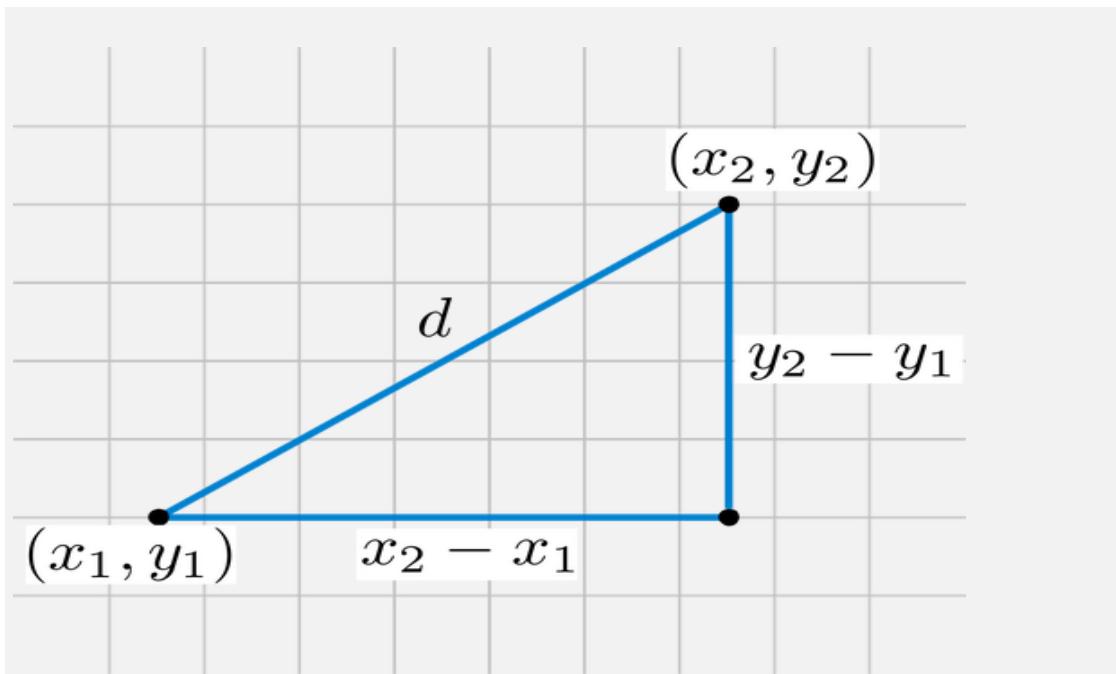


## Distance Metric

The method used to calculate the distance between data points will affect the end result.

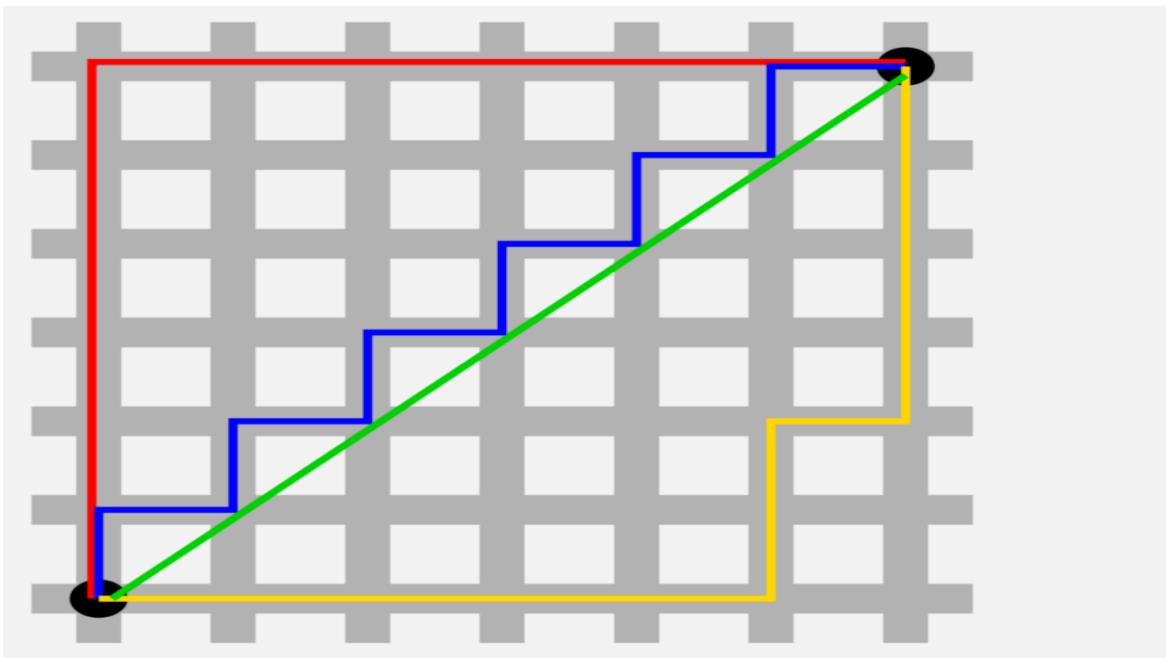
### Euclidean Distance

The shortest distance between two points. For example, if  $x=(a,b)$  and  $y=(c,d)$ , the Euclidean distance between  $x$  and  $y$  is  $\sqrt{(a-c)^2+(b-d)^2}$



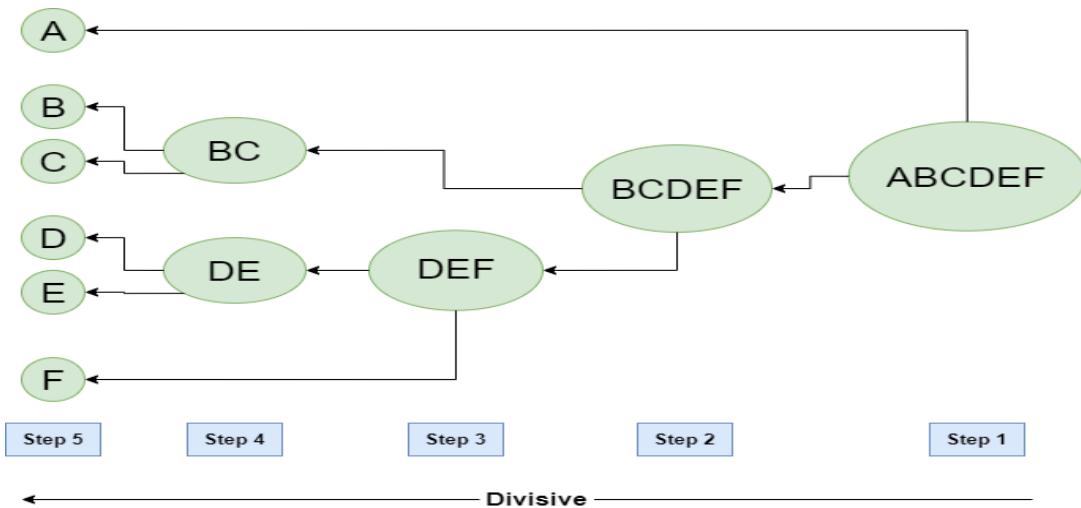
### Manhattan Distance

Imagine you were in the downtown center of a big city and you wanted to get from point A to point B. You wouldn't be able to cut across buildings, rather you'd have to make your way by walking along the various streets. For example, if  $x=(a,b)$  and  $y=(c,d)$ , the Manhattan distance between  $x$  and  $y$  is  $|a-c|+|b-d|$



### Divisive:

The Divisive Hierarchical clustering is precisely **opposite** of the Agglomerative Hierarchical clustering. The divisive clustering algorithm is a top-down clustering approach in which all the data points are taken as a single cluster and in every iteration, the data points are separated from the clusters which aren't comparable. In the end, it is left with N clusters.

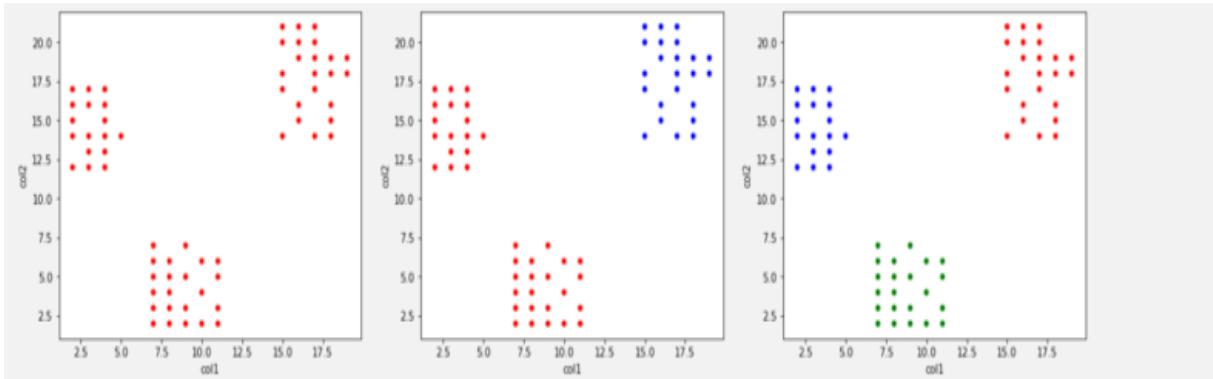


**Figure –** Divisive Hierarchical clustering

### Steps of Divisive Clustering

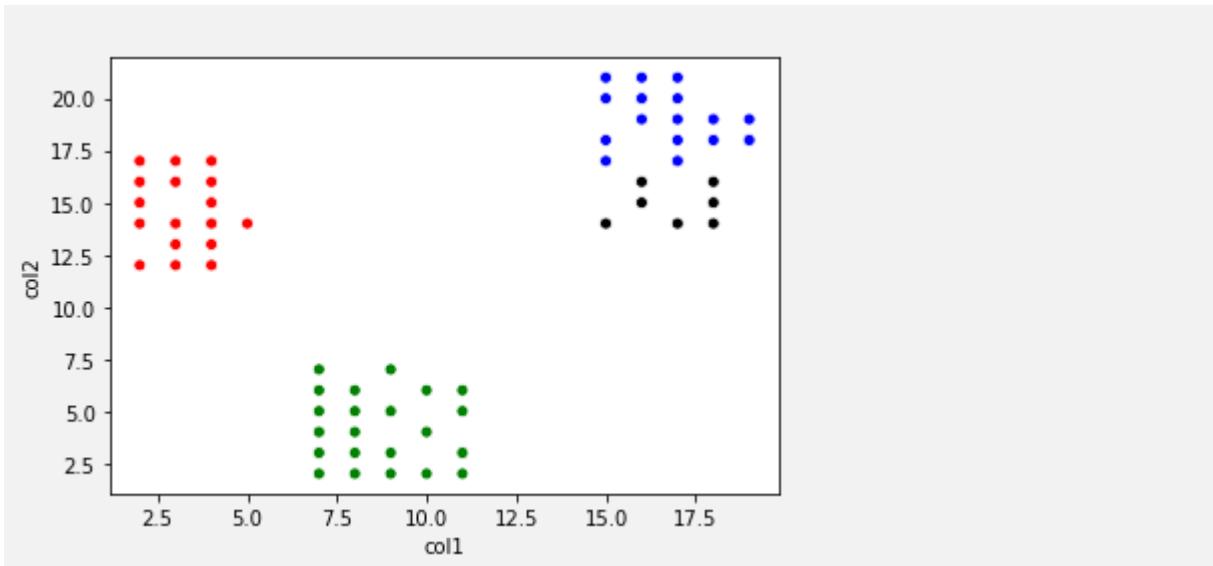
- Initially, all points in the dataset belong to one single cluster.
- Partition the cluster into two least similar cluster

- Proceed recursively to form new clusters until the desired number of clusters is obtained.



1st Image: All the data points belong to one cluster, 2nd Image: 1 cluster is separated from the previous single cluster, 3rd Image: Further 1 cluster is separated from the previous set of clusters.

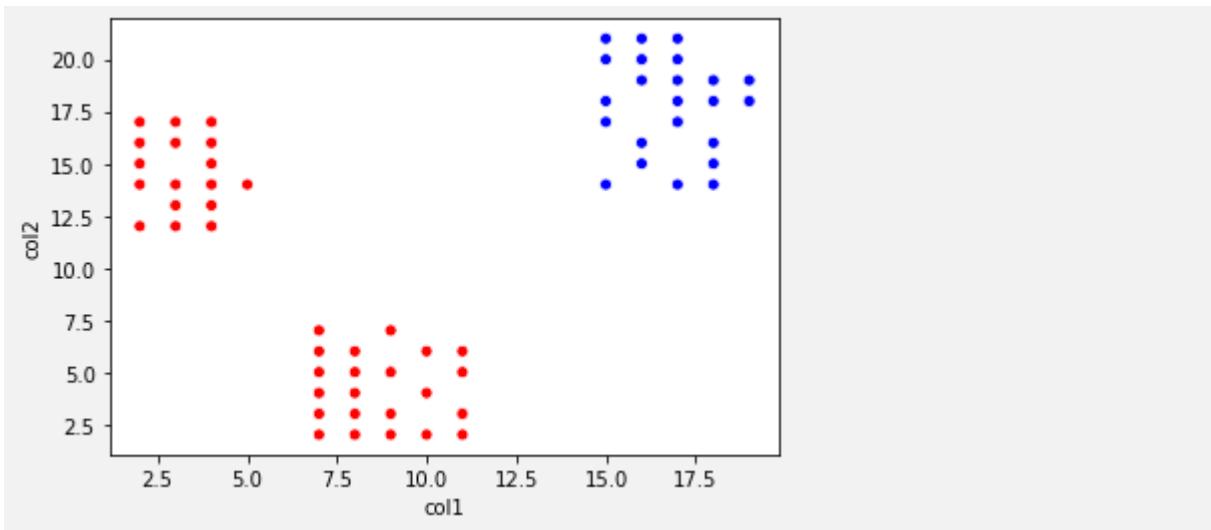
In the above sample dataset, it is observed that there is 3 cluster that is far separated from each other so stopped after getting 3 clusters. Even separating for further more clusters are done , below is the obtained result.



Sample dataset separated into 4 clusters

### How to choose which cluster to split?

Check the sum of squared errors of each cluster and choose the one with the largest value. In the below 2-dimension dataset, currently, the data points are separated into 2 clusters, for further separating it to form the 3rd cluster find the sum of squared errors (SSE) for each of the points in a red cluster and blue cluster.



Sample dataset separated into 2clusters

The cluster with the largest SSE value is separated into 2 clusters, hence forming a new cluster. In the above image, it is observed red cluster has larger SSE so it is separated into 2 clusters forming 3 total clusters.

### **How to split the above-chosen cluster?**

Once it is decided to split which cluster, then the question arises on how to split the chosen cluster into 2 clusters. One way is to use Ward's criterion to chase for the largest reduction in the difference in the SSE criterion as a result of the split.

### **How to handle the noise or outlier?**

Due to the presence of outlier or noise, it can result in forming a new cluster of its own. To handle the noise in the dataset using a threshold to determine the termination criterion that means do not generate clusters that are too small.

### **Approaches to Improve Quality of Hierarchical Clustering**

Here are the two approaches that are used to improve the quality of hierarchical clustering –

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

### **Difference between Hierarchical Clustering and Non Hierarchical Clustering**

#### **Hierarchical Clustering:**

Hierarchical clustering is basically an unsupervised clustering technique which

involves creating clusters in a predefined order. The clusters are ordered in a top to bottom manner. In this type of clustering, similar clusters are grouped together and are arranged in a hierarchical manner. It can be further divided into two types namely agglomerative hierarchical clustering and Divisive hierarchical clustering. In this clustering, we link the pairs of clusters all the data objects are there in the hierarchy.

### **Non Hierarchical Clustering:**

Non Hierarchical Clustering involves formation of new clusters by merging or splitting the clusters. It does not follow a tree like structure like hierarchical clustering. This technique groups the data in order to maximize or minimize some evaluation criteria. K means clustering is an effective way of non hierarchical clustering. In this method the partitions are made such that non-overlapping groups having no hierarchical relationships between themselves.

### **Difference between Hierarchical Clustering and Non Hierarchical Clustering**

S.NO.	Hierarchical Clustering:	Non Hierarchical Clustering:
1.	Hierarchical Clustering involves creating clusters in a predefined order from top to bottom .	Non Hierarchical Clustering involves formation of new clusters by merging or splitting the clusters instead of following a hierarchical order.
2.	It is considered less reliable than Non Hierarchical Clustering.	It is comparatively more reliable than Hierarchical Clustering.
3.	It is considered slower than Non Hierarchical Clustering.	It is comparatively more faster than Hierarchical Clustering.
4.	It is very problematic to apply this technique when we have data with high level of error.	It can work better than Hierarchical clustering even when error is there.
5.	It is comparatively easier to read and understand.	The clusters are difficult to read and understand as compared to Hierarchical clustering.
6.	It is relatively unstable than Non Hierarchical clustering.	It is a relatively stable technique.

## Partitioning Method

Suppose we are given a database of ‘n’ objects and the partitioning method constructs ‘k’ partition of data. Each partition will represent a cluster and  $k \leq n$ . It means that it will classify the data into k groups, which satisfy the following requirements –

- Each group contains at least one object.
- Each object must belong to exactly one group.

## Points to remember

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

## What is K-Means Algorithm?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if  $K=2$ , there will be two clusters, and for  $K=3$ , there will be three clusters, and so on. It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs to only one group that has similar Properties.

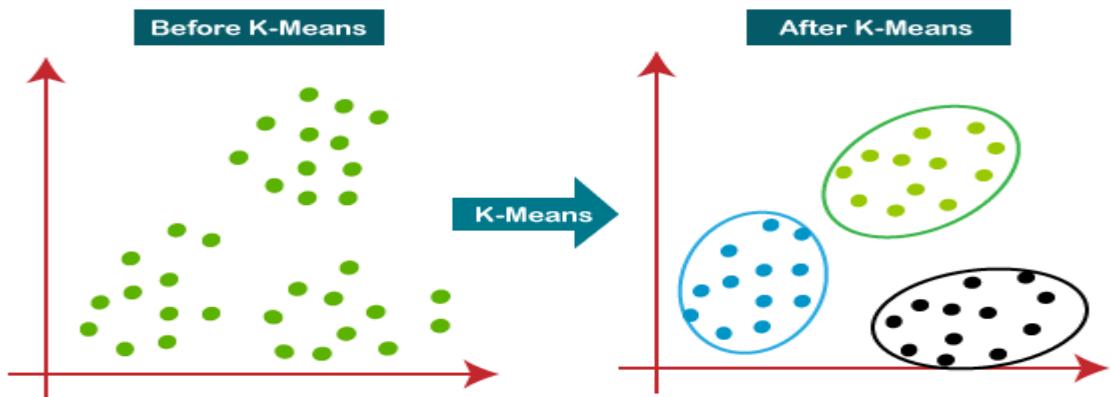
It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.



Working of K-Means algorithm

How does the K-Means Algorithm Work?

**Step-1:** Select the number K to decide the number of clusters.

**Step-2:** Select random K points or centroids. (It can be other from the input dataset).

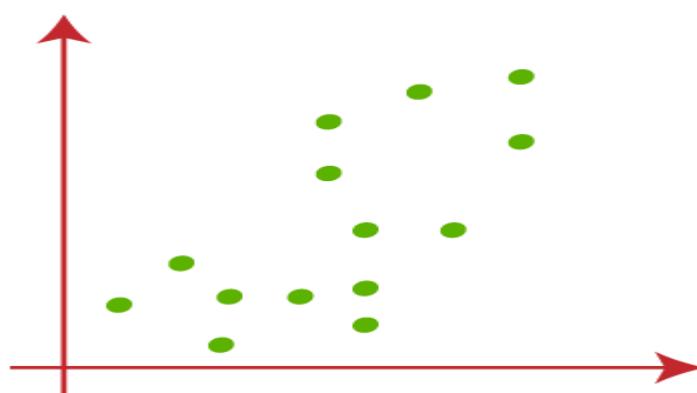
**Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.

**Step-4:** Calculate the variance and place a new centroid of each cluster.

**Step-5:** Repeat the third steps, which mean reassign each datapoint to the new closest centroid of each cluster.

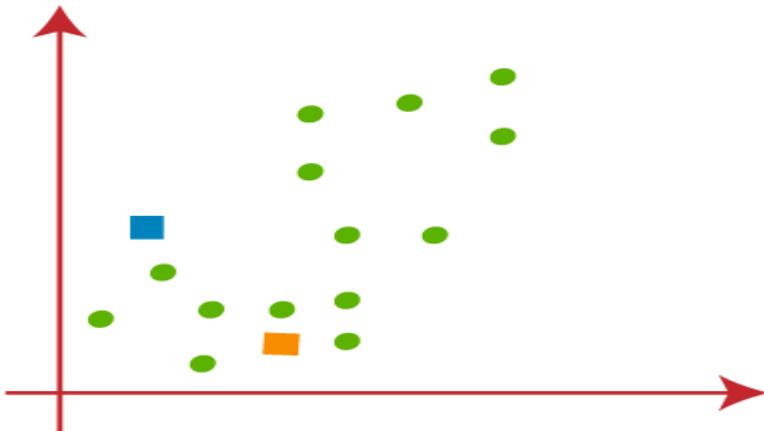
**Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.

**Step-7:** The model is ready.

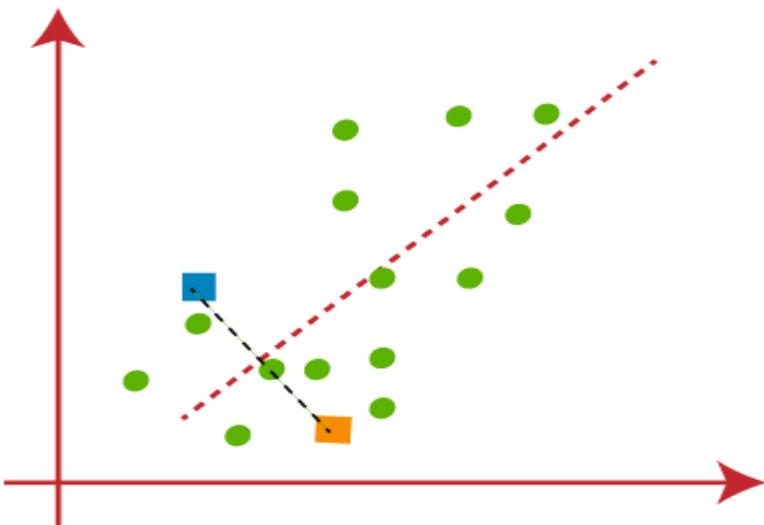


x-y axis scatter plot of two variables M1 and M2

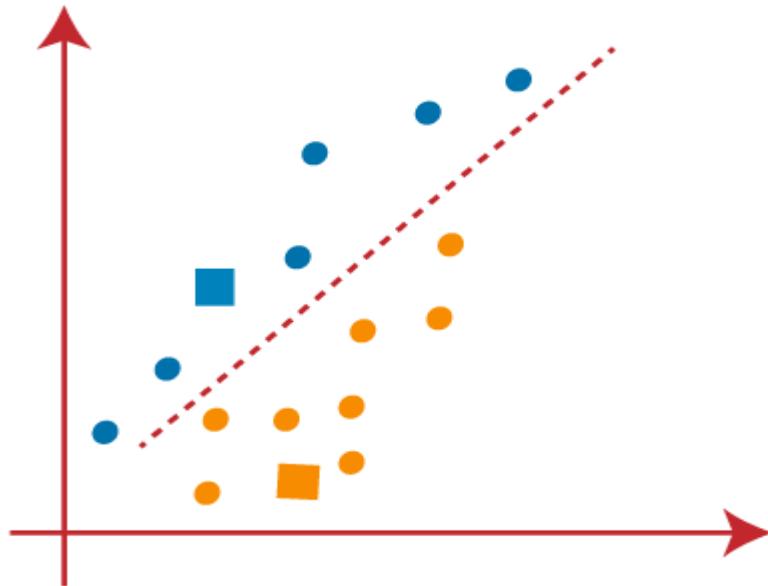
- Let's take number  $k$  of clusters, i.e.,  $K=2$ , to identify the dataset and to put them into different clusters. It means here we will try to group these datasets into two different clusters.
- Need to choose some random  $k$  points or centroid to form the cluster. These points can be either the points from the dataset or any other point. So, here select the below two points as  $k$  points, which are not the part of our dataset. Consider the below image:



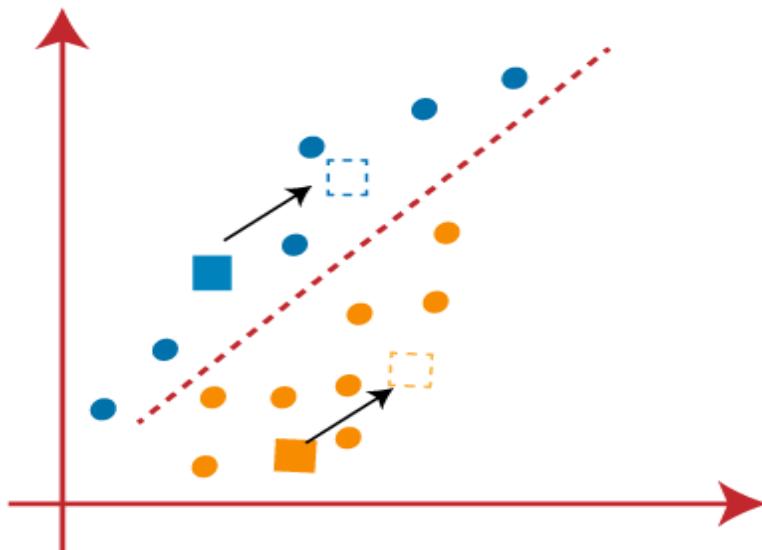
- Now assign each data point of the scatter plot to its closest K-point or centroid. Compute it by applying some mathematics that calculate the distance between two points and draw a median between both the centroids.



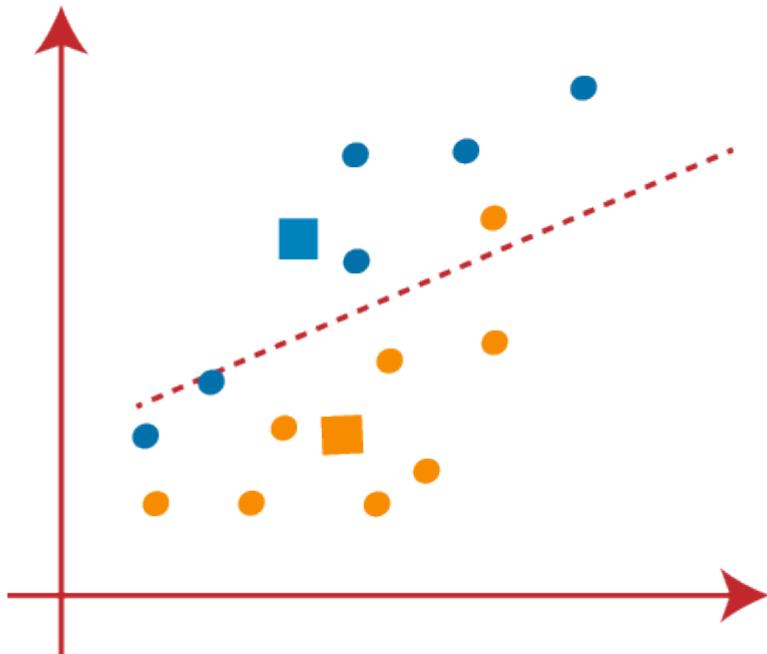
From the above image, it is clear that points left side of the line is near to the K1 or blue centroid, and points to the right of the line are close to the yellow centroid. Color them as blue and yellow for clear visualization.



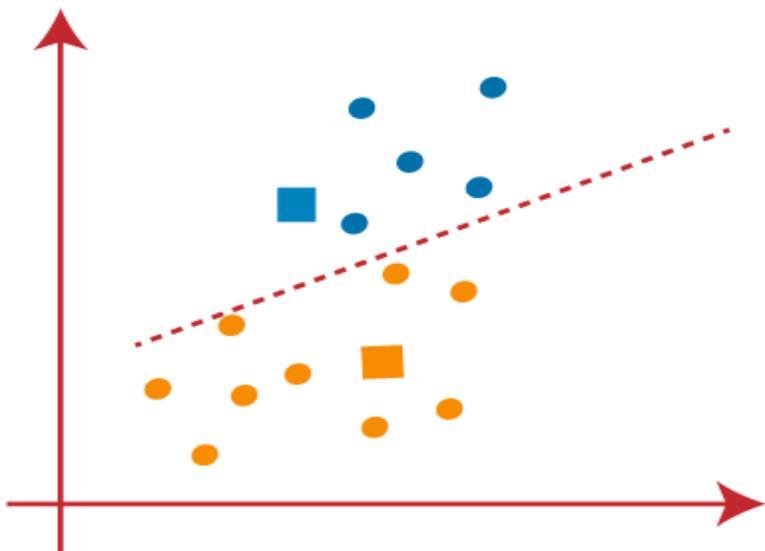
- Find the closest cluster and repeat the process by choosing **a new centroid**. To choose the new centroids, compute the center of gravity of these centroids, and will find new centroids as below:



- Next, reassign each datapoint to the new centroid. For this repeat the same process of finding a median line. The median will be like below image:

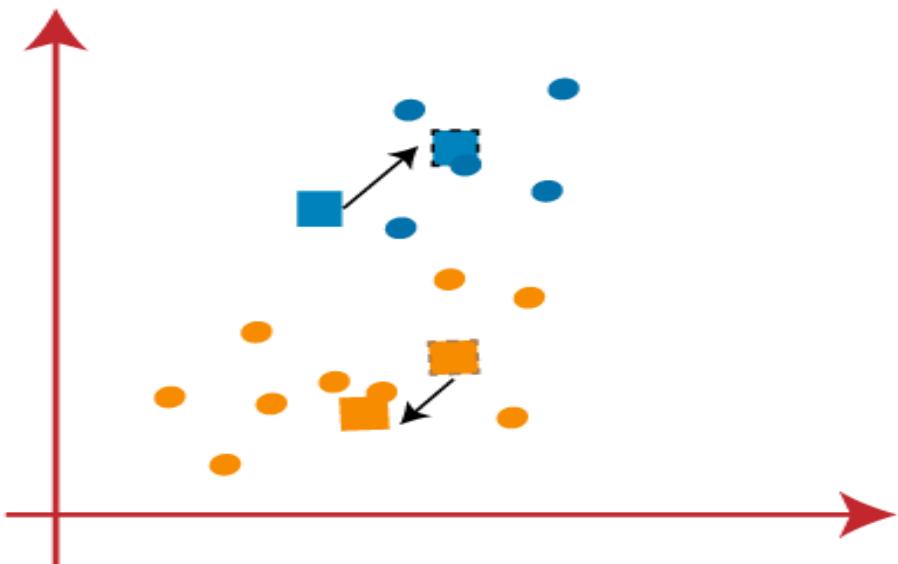


From the above image, one yellow point is on the left side of the line, and two blue points are right to the line. So, these three points will be assigned to new centroids.

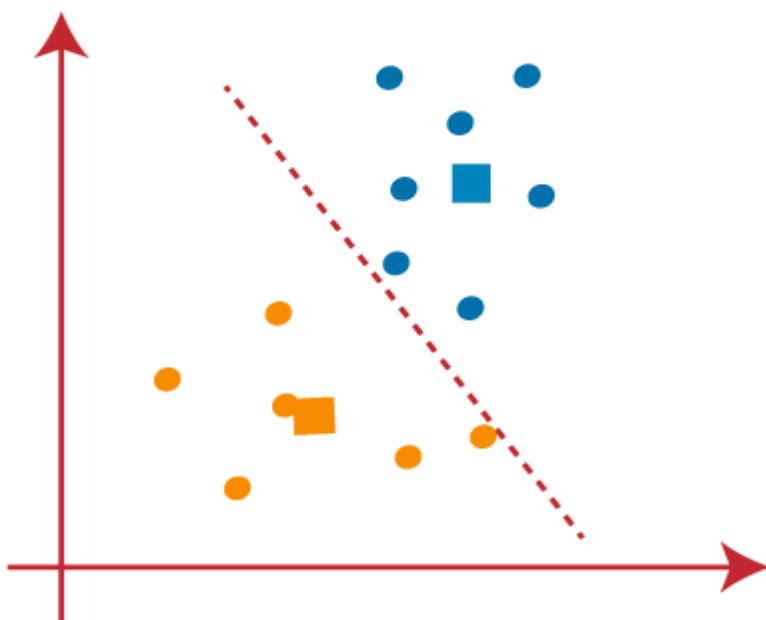


As reassignment has taken place, again go to the step-4, which is finding new centroids or K-points.

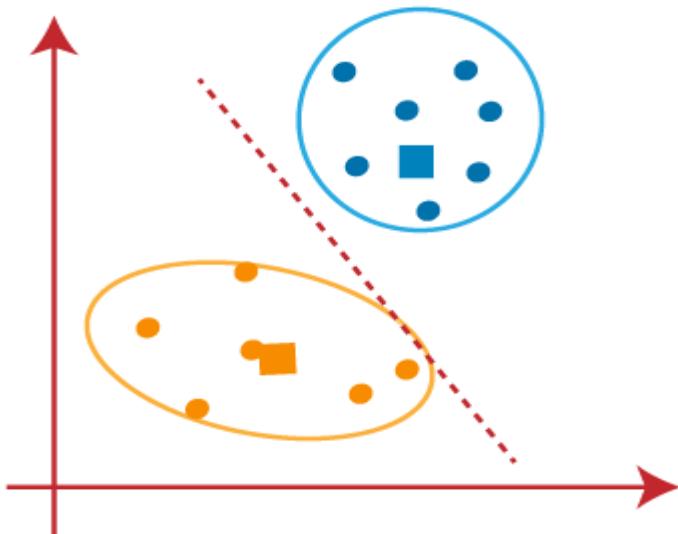
- Repeat the process by finding the center of gravity of centroids, so the new centroids will be as shown in the below image:



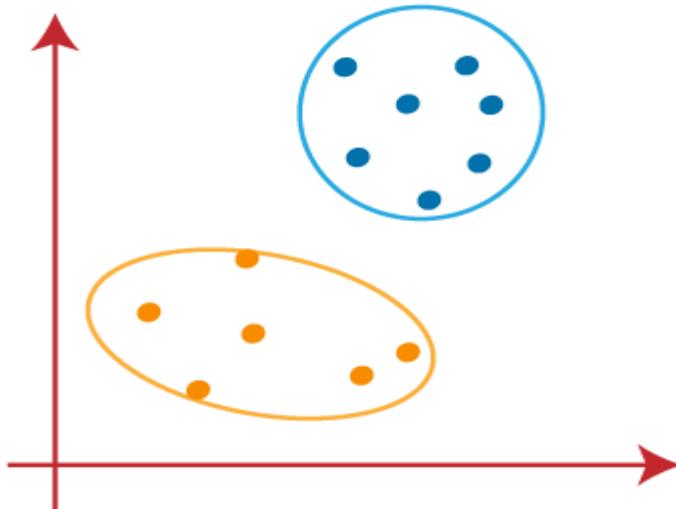
- As new centroids are formed again draw the median line and reassign the data points. So, the image will be:



- As per the above image; there are no dissimilar data points on either side of the line, which means the model is formed. Consider the below image:



As the model is ready, now remove the assumed centroids, and the two final clusters will be as shown in the below image:



### **How to choose the value of "K number of clusters" in K-means Clustering?**

The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms. But choosing the optimal number of clusters is a big task. There are some different ways to find the optimal number of clusters, but here the discussion is on the most appropriate method to find the number of clusters or value of K.

## Elbow Method

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. **WCSS** stands for **Within Cluster Sum of Squares**, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:

$$\text{WCSS} = \sum_{P_i \text{ in Cluster1}} \text{distance}(P_i C_1)^2 + \sum_{P_i \text{ in Cluster2}} \text{distance}(P_i C_2)^2 + \sum_{P_i \text{ in Cluster3}} \text{distance}(P_i C_3)^2$$

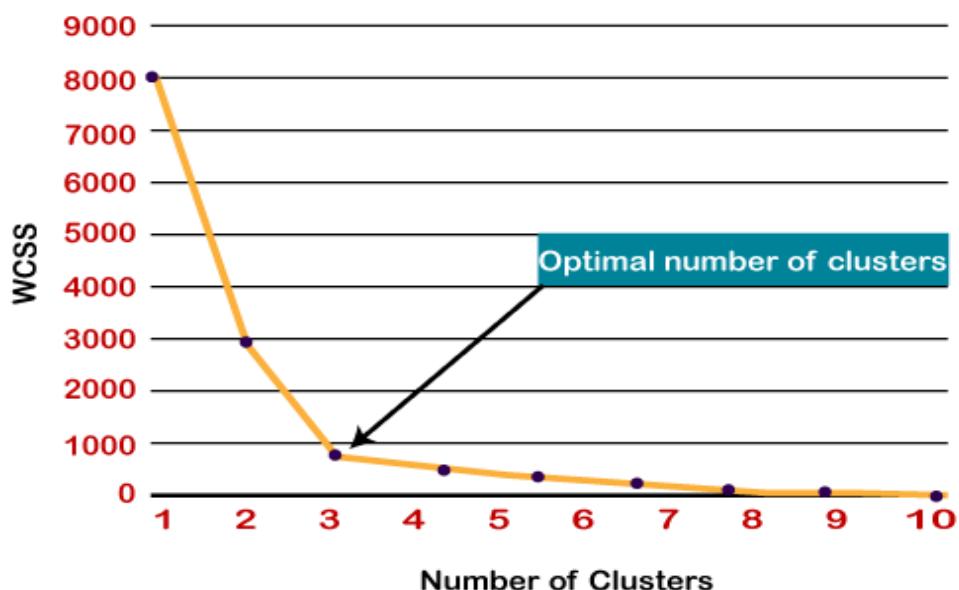
In the above formula of WCSS,

$\sum_{P_i \text{ in Cluster1}} \text{distance}(P_i C_1)^2$ : It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms. To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

To find the optimal value of clusters, the elbow method follows the below steps:

- It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
- For each value of K, calculates the WCSS value.
- Plots a curve between calculated WCSS values and the number of clusters K.
- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

Since the graph shows the sharp bend, which looks like an elbow, hence it is known as the elbow method. The graph for the elbow method looks like the below image:



## Applications

K-means algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression, etc. The goal usually when we undergo a cluster analysis is either:

- Get a meaningful intuition of the structure of the data we're dealing with.
- Cluster-then-predict where different models will be built for different subgroups if there is a wide variation in the behaviors of different subgroups. An example is clustering patients into different subgroups and building a model for each subgroup to predict the probability of the risk of having heart attack.

## K-means algorithm Disadvantages

- The learning algorithm requires apriori specification of the number of cluster centers.
- The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.
- The learning algorithm is not invariant to non-linear transformations i.e. with different representation of data different results are obtained (data represented in form of cartesian co-ordinates and polar co-ordinates will give different results).
- Euclidean distance measures can unequally weight underlying factors.
- The learning algorithm provides the local optima of the squared error function.
- Randomly choosing of the cluster center cannot lead us to the fruitful result.
- Applicable only when mean is defined i.e. fails for categorical data.
- Unable to handle noisy data and outliers.
- Algorithm fails for non-linear data set.

## K-Medoids (also called as Partitioning Around Medoid) algorithm

K-Medoids (also called as Partitioning Around Medoid) algorithm was proposed in 1987 by Kaufman and Rousseeuw. A medoid can be defined as the point in the cluster, whose dissimilarities with all the other points in the cluster is minimum.

The dissimilarity of the medoid( $C_i$ ) and object( $P_i$ ) is calculated by using  $E = |P_i - C_i|$  *The cost in K-Medoids algorithm is given as*

### Algorithm:

1. Initialize: select k random points out of the n data points as the medoids.
2. Associate each data point to the closest medoid by using any common distance metric methods.
3. While the cost decreases:

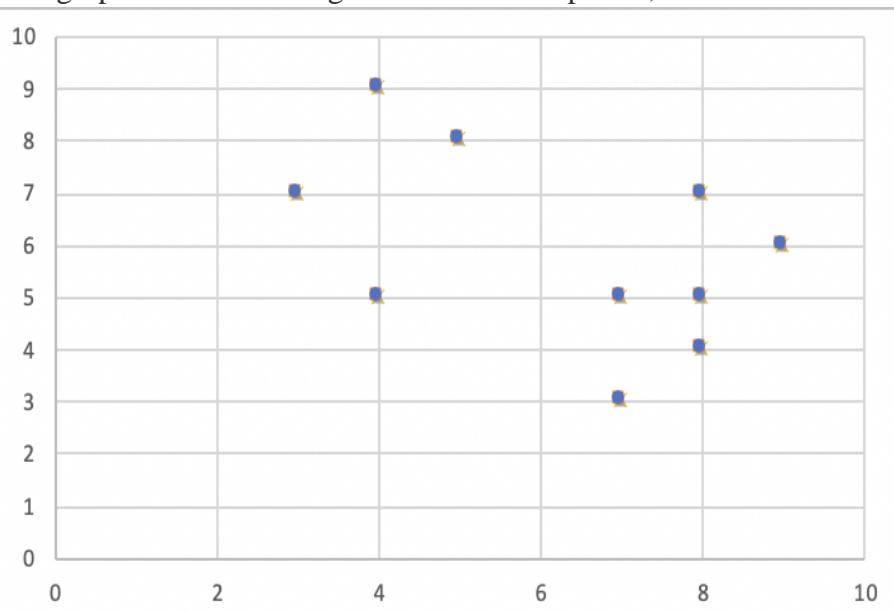
For each medoid m, for each data o point which is not a medoid:

1. Swap m and o, associate each data point to the closest medoid, recompute the cost.
2. If the total cost is more than that in the previous step, undo the swap.

Let's consider the following example:

	<b>X</b>	<b>Y</b>
0	8	7
1	3	7
2	4	9
3	9	6
4	8	5
5	5	8
6	7	3
7	8	4
8	7	5
9	4	5

If a graph is drawn using the above data points, we obtain the following:



### **Step 1:**

Let the randomly selected 2 medoids, so select  $k = 2$  and let **C1 -(4, 5)** and **C2 -(8, 5)** are the two medoids.

### **Step 2: Calculating cost.**

The dissimilarity of each non-medoid point with the medoids is calculated and tabulated:

	X	Y	Dissimilarity from C1	Dissimilarity from C2
0	8	7	6	2
1	3	7	3	7
2	4	9	4	8
3	9	6	6	2
4	8	5	-	-
5	5	8	4	6
6	7	3	5	3
7	8	4	5	1
8	7	5	3	1
9	4	5	-	-

Each point is assigned to the cluster of that medoid whose dissimilarity is less. The points 1, 2, 5 go to cluster C1 and 0, 3, 6, 7, 8 go to cluster C2. The Cost =  $(3 + 4 + 4) + (3 + 1 + 1 + 2 + 2) = 20$

**Step 3: randomly select one non-medoid point and recalculate the cost.**  
 Let the randomly selected point be (8, 4). The dissimilarity of each non-medoid point with the medoids – C1 (4, 5) and C2 (8, 4) is calculated and tabulated.

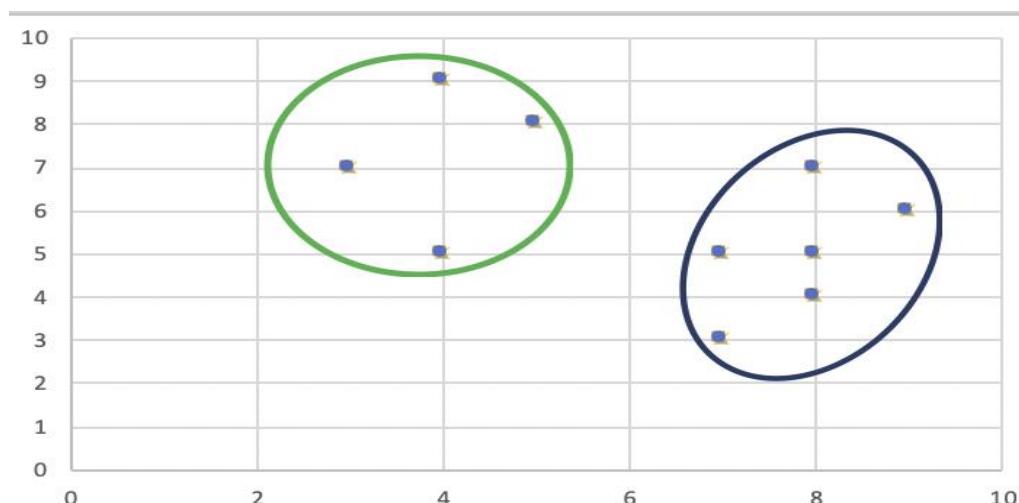
	X	Y	Dissimilarity from C1	Dissimilarity from C2
0	8	7	6	3
1	3	7	3	8
2	4	9	4	9
3	9	6	6	3
4	8	5	4	1
5	5	8	4	7
6	7	3	5	2
7	8	4	-	-
8	7	5	3	2
9	4	5	-	-

Each point is assigned to that cluster whose dissimilarity is less. So, the points 1, 2, 5 go to cluster C1 and 0, 3, 6, 7, 8 go to cluster C2.

The New cost =  $(3 + 4 + 4) + (2 + 2 + 1 + 3 + 3) = 22$

Swap Cost = New Cost – Previous Cost =  $22 - 20$  and  $2 > 0$

As the swap cost is not less than zero, we undo the swap. Hence (3, 4) and (7, 4) are the final medoids. The clustering would be in the following way



#### Advantages:

1. It is simple to understand and easy to implement.
2. K-Medoid Algorithm is fast and converges in a fixed number of steps.
3. PAM is less sensitive to outliers than other partitioning algorithms.

#### Disadvantages:

1. The main disadvantage of K-Medoid algorithms is that it is not suitable for clustering non-spherical (arbitrary shaped) groups of objects. This is because it relies on minimizing the distances between the non-medoid objects and the medoid (the cluster centre) – briefly, it uses compactness as clustering criteria instead of connectivity.
2. It may obtain different results for different runs on the same dataset because the first k medoids are chosen randomly.

### **Important distinction between hierarchical and partitional clustering**

#### **Partitional :**

data points are divided into finite number of partitions (non-overlapping subsets) i.e., each data point is assigned to exactly one subset

#### **Hierarchical :**

data points placed into a set of nested clusters are organized into a hierarchical tree i.e., tree expresses a continuum of similarities and clustering

### Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

### Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

### **Advantages**

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

### Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

## Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

---

### 6.11 Clustering Algorithm Applications

---

Clustering algorithms is used in various areas or fields of real-life examples such as data mining, web cluster engines, academics, bioinformatics, image processing & transformation.

- Recommendation engines

The recommendation engines is a widely used method for providing automated personalized suggestions about products, services and information where collaborative filtering is one of the famous recommendation system and techniques. In this method, the clustering algorithm provide an idea of like-minded users. The computation/estimation as data provided by several users is leveraged for improving the performance of collaborative filtering methods. And this can be implemented for rendering recommendations in diverse applications.

For example, the recommendation engine is broadly used in Amazon, Flipkart to recommend product and Youtube to suggest songs of the same genre.

Even though dealing with extensive data clustering is suitable as the first step for narrowing the choice of underlying relevant neighbours in collaborative filtering algorithms that also enhances the performance of complex recommendation engines. Essentially, each cluster will be assigned to specific preferences on the basis of customers' choices that belong to the cluster. And then, within each cluster, customers would receive recommendations estimated at the cluster level.

- Market and Customer segmentation

A process of splitting the target market into smaller and more defined categories is known as market segmentation. This segments customers/audiences into groups of similar characteristics (needs, location, interests or demographics) where target and personalization, under it, is an immense business.

For instance, a business is looking to get the best return on investment, it is necessary to target customers in a proper way. If wrong decisions are made then there is a high risk of not making any sales and ruining customers trust. So, the right approach is looking at specific characteristics of people and sharing campaigns with them that are

also helpful in engaging with more people of similar behaviour. Clustering algorithms are capable of grouping people with identical traits and prospects to purchase.

For example, once the groups are created, you can conduct a test campaign on each group by sending marketing copy and according to response, you can send more target messages (consisting information about products and services) to them in future.

Under the customer segmentation application, various clusters of customers are made with respect to their particular attributes. On the basis of user-based analysis, a company can identify potential customers for their products or services.

As groups of identical customers are made by clustering method in this area, it is very similar to collaborative filtering while embracing the very fine difference, here, irregular characteristics of objects are deployed for clustering purposes rather than rating/review information. Clustering methods enable us to segment customers into diverse clusters, depending on which companies can consider novel strategies to apply to their customer base.

For example, K-means clustering is helpful for marketers to improve customer base, work on targeted areas, and divide customers on the basis of purchase history, interests or activities.

Another example, a telecom company makes a cluster of prepaid users to understand the pattern/behaviour in the form of recharging amount, sending SMS, and using the internet, this also helps a company to make segments and plan any campaigns for targeted users (specific cluster of users).

- Social Network Analysis (SNA)

It is the process of examining qualitative and quantitative social structures by utilizing Graph Theory (a major branch of discrete mathematics) and networks. Here the mapping of social networks structure is arranged in terms of nodes (individual personality, people, or other entity inside the network) and the edges or links (relationships, interaction, or communication) that connect them.

Clustering methods are required in such analysis in order to map and measure the relationship and conflicts amid people, groups, companies, computer networks, and other similar connected information/knowledge entities. Clustering analysis can provide a visual and mathematical analysis/presentation of such relationships and give social network summarization.

For example, for understanding a network and its participants, there is a need to evaluate the location and grouping of actors in the network, where the actors can be

individual, professional groups, departments, organizations or any huge system-level unit.

Now, through a clustering approach, SNA can visualize the interaction among participants and obtain insights about several roles and groupings in the network, such as who are connectors, bridges, and experts, who are isolated actors and much similar information. It also tells where there are clusters, who are into them, who are at the gist in the network or on the outer edge.

- **Search Result Clustering**

You must have encountered similar results obtained while searching something particular at Google, these results are a mixture of the similar matches of your original query. Basically, this is the result of clustering, it makes groups of similar objects in a single cluster and renders to you, i.e provides results of searched data in terms with most closely related objects that are clustered across the data to be searched.

Therefore, the concept of similar objects serves as a backbone in getting searched results. Even though, most of the parameters are taken into consideration for defining the portrait of similar objects.

Depending on the closest similar objects/properties, the data is assigned to a single cluster, giving the plethora sets of similar results of the users. In simple terms, the search engine attempts to group identical objects in one cluster and non-identical objects in another cluster.

- **Biological Data Analysis, Medical Imaging Analysis and Identification of Cancer Cells**

One of the means to connect analytical tools with biological content is Biological data analysis for a heavy and extended understanding of the relationships identified as to be linked with experimental observations. On the other side, from the past few years, the exploitation of research done on the human genome and the expanding facility of accumulating diverse types of gene expression data lead to evolving biological data analysis exponentially.

Clustering helps in extracting useful knowledge from huge datasets collected in biology, and other life sciences realm as medicine or neuroscience with the fundamental aim of providing prediction and description of data structure.

Using clustering algorithms, cancerous datasets can be identified; a mix datasets involving both cancerous and non-cancerous data can be analyzed using clustering algorithms to understand the different traits present in the dataset, depending upon algorithms produces resulting clusters. On feeding to unsupervised clustering algorithms, we obtain accurate results from cancerous datasets.

- Identifying Fake News

Fake news is not a new phenomenon, but it is one that is becoming prolific.

What the problem is: Fake news is being created and spread at a rapid rate due to technology innovations such as social media. The issue gained attention recently during the 2016 US presidential campaign. During this campaign, the term Fake News was referenced an unprecedented number of times.

How clustering works: In a paper recently published by two computer science students at the University of California, Riverside, they are using clustering algorithms to identify fake news based on the content. The way that the algorithm works is by taking in the content of the fake news article, the corpus, examining the words used and then clustering them. These clusters are what helps the algorithm determine which pieces are genuine and which are fake news. Certain words are found more commonly in sensationalized, click-bait articles. When you see a high percentage of specific terms in an article, it gives a higher probability of the material being fake news.

- Spam filter

You know the junk folder in your email inbox? It is the place where emails that have been identified as spam by the algorithm. Many machine learning courses, such as Andrew Ng's famed Coursera course, use the spam filter as an example of unsupervised learning and clustering.

**What the problem is:** Spam emails are at best an annoying part of modern day marketing techniques, and at worst, an example of people phishing for your personal data. To avoid getting these emails in your main inbox, email companies use algorithms. The purpose of these algorithms is to flag an email as spam correctly or not.

**How clustering works:** K-Means clustering techniques have proven to be an effective way of identifying spam. The way that it works is by looking at the different sections of the email (header, sender, and content). The data is then grouped together. These groups can then be classified to identify which are spam. Including clustering in the classification process improves the accuracy of the filter to 97%. This is excellent news for people who want to be sure they're not missing out on your favorite newsletters and offers.

- Classifying network traffic

Imagine you want to understand the different types of traffic coming to your website. You are particularly interested in understanding which traffic is spam or coming from bots.

**What the problem is:** As more and more services begin to use APIs on your application, or as your website grows, it is important you know where the traffic is coming from. For example, you want to be able to block harmful traffic and double

down on areas driving growth. However, it is hard to know which is which when it comes to classifying the traffic.

**How clustering works:** K-means clustering is used to group together characteristics of the traffic sources. When the clusters are created, you can then classify the traffic types. The process is faster and more accurate than the previous Autoclass method. By having precise information on traffic sources, you are able to grow your site and plan capacity effectively.

- Identifying fraudulent or criminal activity

In this scenario, we are going to focus on fraudulent taxi driver behavior. However, the technique has been used in multiple scenarios.

**What is the problem:** You need to look into fraudulent driving activity. The challenge is how do you identify what is true and which is false?

**How clustering works:** By analysing the GPS logs, the algorithm is able to group similar behaviors. Based on the characteristics of the groups you are then able to classify them into those that are real and which are fraudulent.

- Document analysis

There are many different reasons why you would want to run an analysis on a document. In this scenario, you want to be able to organize the documents quickly and efficiently.

**What the problem is:** Imagine you are limited in time and need to organize information held in documents quickly. To be able to complete this ask you need to: understand the theme of the text, compare it with other documents and classify it.

**How clustering works:** Hierarchical clustering has been used to solve this problem. The algorithm is able to look at the text and group it into different themes. Using this technique, you can cluster and organize similar documents quickly using the characteristics identified in the paragraph.

- Fantasy Football and Sports

Ok so up until this point we have looked into different business problems and how clustering algorithms have been applied to solve them. But now for the critical issues - fantasy football!

**What is the problem:** Who should you have in your team? Which players are going to perform best for your team and allow you to beat the competition? The challenge at the start of the season is that there is very little if any data available to help you identify the winning players.

**How clustering works:** When there is little performance data available to train your model on, you have an advantage for unsupervised learning. In this type of machine learning problem, you can find similar players using some of their characteristics. This

has been done using K-Means clustering. Ultimately this means you can get a better team more quickly at the start of the year, giving you an advantage.

---

## 6.12 Summary

---

The objective of clustering is to assign observations to groups ("clusters") so that observations within each group are similar to one another with respect to variables or attributes of interest, and the groups themselves stand apart from one another. In other words, the objective is to divide the observations into homogeneous and distinct groups.

There are a number of clustering methods. One method, for example, begins with as many groups as there are observations, and then systematically merges observations to reduce the number of groups by one, two, :: :, until a single group containing all observations is formed. Another method begins with a given number of groups and an arbitrary assignment of the observations to the groups, and then reassigns the observations one by one so that ultimately each observation belongs to the nearest group.

Cluster analysis is also used to group variables into homogeneous and distinct groups. This approach is used, for example, in revising a questionnaire on the basis of responses received to a draft of the questionnaire. The grouping of the questions by means of cluster analysis helps to identify redundant questions and reduce their number, thus improving the chances of a good response rate to the final version of the questionnaire.

---

## 6.13 Reference for further reading

---

<https://www.Javatpoint.com>

<https://www.Geeksforgeeks.org>

<https://tutorialspoint.com>

<https://datafloq.com/read/7-innovative-uses-of-clustering-algorithms/6224>

[https://en.wikipedia.org/wiki/Cluster\\_analysis](https://en.wikipedia.org/wiki/Cluster_analysis)

---

## 6.14 Model Questions

---

1. What is Cluster Analysis? List and explain requirements of clustering in data mining
2. Discuss on types of clusters?
3. What are the applications of clustering?

4. Describe the features of partition based clustering algorithms?
5. List out some clustering methods?
6. Explain k-means partitioning algorithm in Cluster Analysis?
7. What is Hierarchical method?
8. Explain the General Steps of Hierarchical Clustering?
9. Explain the Methods of Hierarchical Clustering and give example for each one
10. Differentiate between Agglomerative and Divisive Hierarchical Clustering Algorithm?
11. Discuss the following clustering algorithm using examples :
  - a. K-means.
  - b. K-medoid.
12. What is the difference between hierarchical clustering and non hierarchical clustering?
13. Discuss in detail various types of data that are considered in the cluster analysis.?
14. Given two objects represented by the tuples (22,1,42,10) and (20,0,36,8) Compute the Manhattan distance between the two objects.

## **Chapter 1 Web Mining**

### **1.0 Objectives**

#### **1.1 Introduction**

#### **1.2 An Overview**

##### **1.2.1 What is Web mining?**

##### **1.2.2 Applications of web mining**

##### **1.2.3 Types of techniques of mining**

##### **1.2.4 Difference Between Web Content, Web Structure, and Web Usage Mining**

##### **1.2.5 Comparison between data mining and web mining**

#### **1.3 Future Trends**

##### **1.3.1 To Adopt a Framework**

##### **1.3.2 A Systematic Approach**

##### **1.3.3 Emerging Standards to Address the Redundancy and Quality of Web-Based Content**

##### **1.3.4 Development of Tools and Standards**

##### **1.3.5 Use of Intelligent Software**

#### **1.4 Web Personalization**

##### **1.4.1 Personalization Process**

- 1.4.2 Data Acquisition
- 1.4.3 Data Analysis
- 1.5 Tools and Standards
- 1.6 Trends and challenges in personalization
- 1.7 Let us Sum Up
- 1.8 List of References
- 1.9 Bibliography
- 1.10 Chapter End Exercises

## Chapter 1

### 1.0 Objectives

After going through this chapter, you will be able to understand:

- Web Mining, Applications of Web mining
- Techniques of Mining
- **Difference Between Web Content, Web Structure, and Web Usage Mining**
- Comparison between data mining and web mining
- Need for Web Warehousing and Knowledge Management
- Future trends
- Web personalization
- Tools and standards
- Trends and challenges

### 1.1 Introduction

Web mining is pushing the World Wide Web toward a more valuable environments in which clients/users can rapidly and effectively discover the data they need. It incorporates the disclosure and examination of information, reports, and interactive media from the World

Wide Web. Web mining utilizes archive content, hyperlink design, and utilization insights to help clients in gathering their data needs. The actual Web and web search tools contain relationship data about records. In this, Content mining is coming first. Discovering keywords and discovering the connection between a Web page content and a query content will be content mining. Hyperlinks give data about different records on the Web thought to be critical to another report. These connections add profundity to the report, giving the multi-dimensionality that describes the Web. Mining this connection structure is the second space of Web mining. At long last, there is a relationship to different records on the Web that are recognized by past look. These relationships are recorded in logs of searches and gets to.

Mining these logs is the third space of Web mining. Understanding the client is likewise a significant piece of Web mining. Examination of the client's past meetings favoured showcase of data, and communicated inclinations might impact the Web pages returned in response to a query. Web mining is interdisciplinary in nature, spreading over across such fields as data recovery, natural language preparing, data extraction, AI, data set, data mining, data warehousing, UI plan, and visual representation. Strategies for mining the Web have down to earth application in m-commerce, online business, e-government, e-learning, distance learning, virtual associations, knowledge management and digital libraries.

## **1.2 An Overview**

Information, data, and information are so basic to an association's generally speaking functional achievement, Web mining, warehousing and KM are intelligent augmentations of existing functional exercises. In the journey for opportune, exact choices, a fundamental component is to get the most ideal DIKs to create fitting and compelling approaches. The idea of Web warehousing started with the improvement of information warehousing. W.H. Inmon (1992) characterized information warehousing as a "subject-arranged, coordinated, non-unstable, time variation assortment of information on the side of the board's choices". The solitary contrast among information and Web warehousing is that in the last mentioned, the hidden data set is the whole World Wide Web. As a promptly available asset, the Web is a gigantic information stockroom that contains unpredictable data that is assembled and extricated into something important for use in the association circumstance. Utilizing customary information mining philosophies and strategies (Tech Reference, 2003),

the Web mining is the way toward extricating information from the Web and arranging them into recognizable examples and connections.

Data is backbone of any organization and so critical to any organization's overall operational success, Web mining, warehousing and KM are logical extensions of existing operational activities. Timely, accurate decisions, an important element is to get the simplest DIKs possible to supply appropriate and effective courses of action. The concept of Web warehousing originated with the event of knowledge warehousing.

The only difference between data and Web warehousing is that within the latter, the underlying database is that the entire World Wide Web. As a readily accessible resource, the online may be a huge data warehouse that contains volatile information that's gathered and extracted into something valuable to be used within the organization situation. Using traditional data processing methodologies and techniques (Tech Reference, 2003), the Web mining is that the process of extracting data from the web and sorting them into identifiable patterns and relationships.

### **1.2.1 What is Web Mining?**

Web Mining is the cycle of Data Mining strategies to consequently find and concentrate information from Web and various services. The principal motivation behind web mining is finding valuable data from the World-Wide Web and its usage.

### **1.2.2 Applications of Web Mining**

1. Web mining assists to improve the power of web search engine by classifying and identifying the web pages.
2. It is utilized for Web Searching e.g., Google, Yahoo etc and Vertical Searching.
3. Web mining is utilized to predict behaviour.
4. Web mining is extremely helpful of a particular website and e-service

### **1.2.3 Types of techniques of mining**

**Various techniques described below:**

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

### **Web Content Mining**

This technique extracts useful information from the content of the online documents. It contains several sorts of data – text, image, audio, video etc. Content data is that the group of facts and gives pattern about user needs. Text documents are associated with text mining, machine learning and NLP. This mining is additionally referred to as text mining

Two methodologies are discussed below:

**(i) Agent-based approach:**

In this methodology applicable sites are recognized utilizing shrewd frameworks.

**(ii) Data-based approach:**

This methodology is used to sort out semi-organized information to organized information.

### **Web Structure Mining**

Web structure mining is the use of finding structure data from the web. The design of the web chart comprises of site pages as nodes, and hyperlinks as edges connecting related pages. Structure mining fundamentally shows the organized summary of a specific site. It recognizes connection between site pages connected by data or direct connection association. To decide the association between two business sites, Web structure mining can be exceptionally helpful.

### **Web Usage Mining**

Web use mining is the use of distinguishing or finding fascinating use designs from enormous informational indexes. Also, these examples empower you to comprehend the client practices or something to that effect. In web use mining, client access information on the web and gather information in type of logs. Thus, web usage mining is likewise called log mining.

#### **1.2.4 Difference Between Web Content, Web Structure, and Web Usage Mining**

Criterion	Web Content		Web Structure	Web Usage
	IR VIEW	DB VIEW		

<b>View of data</b>	Unstructured Structured	Semi-structured Website as DB	Link structure	Interactivity
<b>Main Data</b>	Text Document Hyper Documents	Hyper Documents	Link structure	Server logs Browser logs
<b>Method</b>	Machine Learning Statistical (Including NLP)	Proprietary algorithm Association rules	Proprietary algorithm	Machine learning Statistical Association Rules
<b>Representation</b>	Bag of words, Phrases, concepts, or ontology	Edged Labeled graph Relational	Graph	Relational Table Graph
<b>Application Categories</b>	Categorization <ul style="list-style-type: none"> <li>• Clustering</li> <li>• Finding Extract rules</li> <li>• Finding Patterns in text</li> </ul>	Finding frequent sub structures <ul style="list-style-type: none"> <li>• Web site schema discovery</li> </ul>	1.Categorization 2 Clustering	1. Site construction 2. Adaptation and management

### 1.2.5 Comparison between data mining and web mining

Here, differences between Data Mining and Web Mining given below,

Data Mining	Web Mining
-------------	------------

<p>“Data Mining” works with the plan which is distinguished from the data that’s already accessible within the framework,</p>	<p>The technique of “Web Mining” works with the plan which is perceived with the help of various “Web Data” exclusively.</p>
<p>It is generally used in various kinds of organizations which works on “Artificial Intelligence”, where different trade decisions are upgraded with the assistance of different choice-making activities with the assistance of innovation,</p>	<p>It is used in the field of “Data Analytics”, where the available unrefined data is changed over/and additionally changed into a huge mastermind separately.</p>
<p>It incorporates the processes like Data Extraction, Design Disclosure, Algorithm Fathoming, etc.</p> <p>,</p>	<p>It also incorporates the processes like Information Extraction, Design Revelation, Algorithm Understanding, and so on however these all process occurs with the help of "Web" which excessively on various "Web Servers" and "Web Documents" independently.</p>
<p>It is, generally, done by the specialists like unique different “Data Engineers”, ‘Data Scientists’.</p>	<p>It is carried out by the specialists like “Data Analysts” and different “Data Scientists”, “Data Engineers”</p>
<p>The various tools which are used to handle of “Data Mining” are “Machine Learning Algorithms” etc.</p>	<p>The various devices which are used of “Web Mining” are “Apache Logs”, “Scrapy”, “PageRank” etc., with the help of these processes are</p>

	completed independently.
Various organizations are relying upon data mining for decision making.	Web mining is vital to drag the current data/information mining measure.
Skills needed for this and data cleaning and AI, machine learning algorithms,	Skills needed for Web mining are Application-level knowledge, "Data engineering", "statistics".
Applications are "Financial Data Analysis", "Retail Industry", "Telecommunication Industry", "Biological Data Analysis".	Web mining is the use of data mining techniques to remove data from web data, checking web reports, hyperlinks between documents

### 1.3 To Adopt a Framework for Managing Information and Data

Information and metadata will turn into an "resource" as a feature of the KM work, hierarchical pioneers will be confronted with serious circumstances that will direct their requirement for it. With circulated Web warehousing exercises, the vital job of DIKs is to turn into another asset in an association's current circumstance. A brought together coherent way to deal with create and oversee metadata content information bases might be important. With unified intelligent metadata that is listed, looked, and handled by tools, it turns into an empowering agent for associations to survey and utilize its information. This IRM metadata archive and information-based environment upholds the general mission and objectives of the association. The ramification of this change is that more powerful insightful tools will be required. They should uphold programmed ordering, approval, looking, planning of relaxed connections, versatile displaying, and fluffy set rationale advances to quantify the metadata's utility with regards to authoritative conveyance and variation to their current circumstance.

Numerous structures will foster glossaries, word references, subject scientific categorizations (nonexclusive to-explicit progressions), semantic guides (term/definition-to-homonym(s)- synonym(s)), representation methods to assemble relationship maps and new dialects to help relationship structures. This infers that overseeing metadata, information and data is plainly and unequivocally connected to the association's procedure with a genuine comprehension of its information advantage.

### **1.3.1 A Systematic Approach for Adopting a Knowledge Management Practice**

Fundamental standards are the establishment for accomplishment in giving an orderly methodology.

They are:

- (1) to catch, classify, and share both data and information.
- (2) to focus on the synergistic endeavours among individuals and networks with an accentuation on learning and preparing; and
- (3) to focus on the information and expertise utilized in the everyday work environment.

Keep in mind, in the underlying evaluation of an association's necessities, one KM best practice may not work in another circumstance because of hierarchical culture and setting (Glick, 2002). An arrangement can be set up utilizing this precise methodology and investigation, explicit venture objectives, necessities, assets, and measurements. The development of an association's knowledge management practice can move from present moment to longer-term activities to full acknowledgment dependent on the achievement of every information drive inside its legitimate setting. This infers a convincing vision and design for your association's KM mission and objective.

### **1.3.2 Emerging Standards to Address the Redundancy and Quality of Web-Based Content**

The planning of connections between various articles in the meta-model of models is beginning to assume a critical part in both hierarchical and industry guidelines. The advancement of different industry-based scientific categorizations (a class of articles and connections that exist between them in a nonexclusive to-explicit design) has begun. The planning of equivalent words, homonyms, and information with definitional clearness turns into a fundamental fixing to the exchange of DIKs. Ontologies (information bases that

contains the planned connections of one item to at least one articles) will be every now and again used to give the ability to gather new realities and perceptions. Ideas examined in the InfoMap (Burk, 1988), particularly those that attention on estimating cost and the maintenance rehearses for DIKs, will be of basic significance to the utility of the metadata information base store.

Data frameworks vendors and expert associations will uphold endeavours to construct these scientific classifications and ontologies to give equipment, programming and consulting administrations that help their utilization. The ramifications are that if programming items are created to help these exercises, the items might be restrictive in nature. The American National Standards Institute (ANSI) is one of the bodies that propose U.S. guidelines be acknowledged by the International Standards Organization (ISO). Every guideline association is comprised of volunteers who create, evaluate, and present a norm for formal endorsement. There can be contending principles that are created, and some must be accommodated. This could affect the turn of events and temporary endeavours expected to construct powerful scientific classifications and ontologies. This infers an efficient methodology for capturing explicit and verifiable information into a well created infrastructure.

### **1.3.3 Development of Tools and Standards to Assess Semantic Integrity of Web Delivered Pages**

The efforts to foster a semantic Web by the W3C gathering will be key being developed of guidelines. This work from different members will address an enormous chance to build the utility of the Web. The most troublesome service will be the semantic interpretation between a Web client's solicitation and utility of the outcomes. Items being produced for natural language, conversational, and parametric looking (Abrams, 2003) will be essential to their successful executions.

### **1.3.4 Use of Intelligent Software**

The efforts supporting the improvement of DARPA's Agent Markup Language (DAML) and Resource Definition Framework (RDF) are centred around the semantic combination, recovery, and utility of the Web's assets. A key exertion is to have a binding together meta-model to store the metadata adequately in its information-based repository for viable and

effective use in the association. The achievement of the information put together environment is predicated with respect to connecting different advancements and practices with basic spotlight on the nature of the information, data, and keeping up with information. This will be one of the greatest authoritative difficulties!

## **1.4 Web Personalization**

The Web has become a gigantic store of data and continues to develop dramatically under no control, while the human capacity to discover, peruse and comprehend content remaining parts steady. Furnishing individuals with admittance to data isn't the issue; the issue is that individuals with fluctuating requirements and inclinations explore through enormous Web structures, missing the objective of their request. Web personalization is quite possibly the most encouraging methodologies for easing this data over-burden, giving custom-made Web encounters. This part investigates the various essences of personalization, follows back its underlying foundations and follows its encouraging. It portrays the modules ordinarily including a personalization interaction, shows its nearby connection to Web mining, portrays the specialized issues that emerge, suggests arrangements whenever the situation allows, and talks about the adequacy of personalization and the connected concerns. Besides, the part represents latest things in the field proposing headings that might prompt new logical outcomes.

### **1.4.1 Personalization Process**

In this part we examine the general personalization measure as far as the discrete modules including it: data acquisition, data analysis and personalized output. We portray exhaustively the destinations of every module and survey the methodologies taken so far by researchers working in the field, the deterrents met, and whenever the situation allows, the solutions suggested.

### **1.4.2 Data Acquisition**

In the larger part of cases, Web personalization is an information concentrated assignment that depends on three general sorts of information: information about the client, information about the Website use and information about the product and equipment accessible on the client's side.

- User information. This class signifies data about personal attributes of the client. A few such kinds of information have been utilized in personalization applications, for example,
- Demographics (name, telephone number, geographic data, age, sex, instruction, pay, and so on)

Client's information on ideas and connections between ideas in an application space (input that has been of broad use in natural language processing system) or area explicit aptitude; Skills and abilities (as in separated from "what" the client knows, as a rule it is of equivalent significance to know what the client knows "how" to do, or significantly further, to recognize what the client knows about and what she can really achieve);

- Interests and inclinations

**Goals and plans** (plan acknowledgment methods and recognized objectives permit the Website to anticipate client interests and needs and change its substance for simpler and quicker objective accomplishment). There are two general methodologies for gaining client information of the kinds depicted above: either the client is asked expressly to give the information (utilizing surveys, fill-in inclination discourses, or even through machine readable data-carriers, like smart cards), or the framework verifiably infers such data without starting any association with the client (utilizing procurement rules, plan acknowledgment, and generalization thinking).

- **Usage Data.** Utilization information might be straightforwardly noticed and recorded or obtained by examining detectable information (whose sum and detail differ contingent upon the advances utilized during Website execution, i.e., java applets, and so on), an interaction previously referred to in this part as Web use mining. Use information may either be:

**Observable information** involving specific activities like tapping on a connection, information with respect to the fleeting review conduct, ratings (utilizing binary or a restricted, discrete scale) and other corroborative or disconformity activities (purchases, , messaging/saving/printing a record, bookmarking a Web page and that's just the beginning), or

Data that get from additional handling the noticed and respect use normalities (estimations of recurrence of choosing a choice/connect/administration, creation of ideas/proposals dependent on circumstance activity relationships, or varieties of this methodology, for example recording activity arrangements).

### **1.4.3 Data Analysis**

Client profiling drastically influences the sorts of examination that can be applied get-togethers period of information obtaining to achieve more complex personalization. The strategies that might be applied for additional breaking down and growing client profiles to determine deductions differ and come from various logical regions that include AI, insights, and data recovery. In this section, we follow the methodology of data recovery and set our emphasis on conveying Web digging for breaking down client conduct and inducing "intriguing" designs, similitudes, groups, and connections among clients or potentially page demands. In the previous years, a few analysts have applied Web usage mining for developing client profiles and settling on personalization choices. Web usage mining utilizes worker logs as its wellspring of data and the way toward getting important data from them advances as indicated by the accompanying stages (Srivastava et al., 2000): information readiness and pre-preparing, design revelation and example examination.

#### **Data Preparation and Pre processing**

The target of this stage is to determine a bunch of worker meetings from crude utilization information, as recorded as Web worker logs. Prior to continuing with a more point by point depiction of information readiness, it is important to give a bunch of information reflections as presented by the W3C1 (World Wide Web Consortium) for portraying Web utilization. A worker meeting is characterized as a bunch of site hits served because of a progression of HTTP demands from a solitary client to a solitary Web worker.

A site visit is a bunch of page documents that add to a solitary presentation in a Web program window (the meaning of the site hit is fundamental on the grounds that for investigating client conduct what is of worth is the total site hit and not every last one of the continuous separate demands that are produced consequently for obtaining portions of the site hit like contents, designs, and so on) Figuring out which log passages allude to a solitary

site hit (an issue known as online visit distinguishing proof) requires data about the webpage organizing and substance.

A consecutive series of online visit demands is named click stream and it is its full substance that we preferably need to know for solid ends. A client meeting is the clickstream of site hits for a solitary client across the whole Web, while a worker meeting is the arrangement of site hits in a client meeting for a specific site. During information planning the undertaking is to recognize the log information sections that allude to designs or traffic naturally produced by arachnids and specialists.

These sections in the majority of the cases are taken out from the log information, as they don't uncover genuine utilization data. In any case, an official conclusion on the most ideal approach to deal with them relies upon the particular application. In the wake of cleaning, log sections are typically parsed into information fields for simpler control. Aside from eliminating passages from the log information, by and large information planning likewise incorporates improving the use data by adding the missing snaps to the client clickstream. The explanation directing this assignment is customer and intermediary storing, which cause numerous solicitations not to be recorded in the worker logs and to be served by the reserved site visits. The way toward re-establishing the total snap stream is called way fruition and it is the last advance for pre-handling utilization information.

Missing site visit solicitations can be identified when the referrer page document for a site hit isn't essential for the past site hit. The lone sound approach to have the total client way is by utilizing either a product specialist or an altered program on the customer side. In any remaining cases the accessible arrangements (utilizing for example, aside from the referrer field, information about the connection design of the site) are heuristic in nature and can't ensure exactness. With the exception of the way fulfilment issue, there stays a bunch of other specialized impediments that should be defeated during information planning and pre-preparing.

All the more explicitly, a significant such issue is client ID. Various strategies are sent for client distinguishing proof and the general appraisal is that the more precise a strategy is, the higher the protection intrusion issue it faces. Expecting that every IP address/specialist pair distinguishes a novel client isn't generally the situation, as numerous clients might

utilize a similar PC to get to the Web and a similar client might get to the Web from different PCs. An implanted meeting ID requires dynamic locales and keeping in mind that it recognizes the different clients from a similar IP/Agent, it neglects to distinguish similar client from various IPs. Treats and programming specialists achieve the two destinations yet are normally not very much acknowledged (or even dismissed and incapacitated) by most clients.

Enlistment additionally gives dependable ID yet not all clients will go through such a technique or review logins and passwords. On the other hand, adjusted programs might give exact records of client conduct even across Websites, however they are not a practical arrangement in most of cases as they require establishment and just a set number of clients will introduce and utilize them. To wrap things up, there emerges the issue of meeting ID. Unimportant arrangements tackle this by setting a base time limit and accepting that resulting demands from a similar client surpassing it have a place with various meetings (or utilize a greatest edge for finishing up individually). Example Discovery Pattern revelation expects to identify fascinating examples with regards to the pre-handled Web utilization information by sending measurable and information mining strategies. These strategies typically involve (Eirinaki and Vazirgiannis, 2003):

- **Association rule mining:**

A procedure utilized for discovering continuous examples, affiliations and relationships among sets of things. In the Web personalization space, this strategy might demonstrate relationships between pages not straightforwardly associated and uncover beforehand obscure relationship between gatherings of clients with explicit interests. Such data might demonstrate important for online business stores to further develop Customer Relationship Management (CRM).

- **Clustering:**

A technique utilized for gathering things that have comparative attributes. For our situation things may either be clients (that exhibit comparative online conduct) or pages (that are comparability used by clients).

- **Classification:**

A process that figures out how to allot information things to one of a few predefined classes. Classes normally address distinctive user profiles, and classification is performed utilizing chosen highlights with high discriminative capacity as alludes to the arrangement of classes portraying each profile.

- **Sequential pattern discovery:**

An expansion to the affiliation rule mining method, utilized for uncovering examples of co-event, consequently consolidating the thought of time sequence. An example for this situation might be a Web page or a bunch of pages got to following another arrangement of pages. In this last stage the goal is to change over found standards, examples and measurements into information or knowledge including the Website being dissected. Information here is a theoretical thought that generally depicts the change from data to comprehension; it is along these lines profoundly subject to the human playing out the examination and arriving at resolutions. In the vast majority of the cases, perception procedures are utilized for "imparting" better the information to the investigator.

This methodology is for sure better than other more conventional strategies, (for example, collective or content-based sifting) as far as both adaptability and dependence on target input information (and not, for example, on abstract client appraisals). In any case, utilization-based personalization can likewise be dangerous when little use information are free relating to certain articles, or when the site content changes consistently. Mobasher et al. (2000a) claims that for more viable personalization, both utilization and content credits of a site should be coordinated into the information examination stage and be utilized consistently as the premise of all personalization choices.

This way semantic information is consolidated into the cycle by addressing area ontologies in the pre-handling and example disclosure stages and utilizing powerful procedures to get uniform profiles portrayal and tell the best way to utilize such profiles for performing continuous personalization (Mobasher&Dai, 2001).

## **1.5 Tools and Standards**

From the past, clearly customizing the Web insight for clients by tending to address necessities and inclinations is a difficult task for the Web industry. Electronic applications (e.g., portals, internet business destinations, e-learning conditions, and so forth) can work

on their presentation by utilizing appealing new tools, for example, dynamic suggestions dependent on singular qualities and recorded navigational history. Nonetheless, the inquiry that emerges is the means by which this can be really refined. Both the web industry and specialists from assorted logical regions have zeroed in on different parts of the point. The exploration draws near, and the business apparatuses that convey customized Web encounters dependent on business rules, Website content and construction, just as the client conduct recorded in Web log documents are various.

WebWatcher incorporates three learning approaches:

- (a) learning from previous tours,
- (b) learning from the hypertext structure and
- (c) combination of the first two approaches.

A proposal framework that helps Web search and customizes the aftereffects of an inquiry dependent on close to home history and inclinations (substance and evaluations of visited pages) is Fab (Balabanovic and Shoham, 1997). By joining both community oriented and content-based strategies, it prevails to dispose of a considerable lot of the shortcomings found in each approach. Humos/Wifs (Ambrosini et al., 1997) has two segments, the Hydrid User Modeling Subsystem and the Web-situated Information Filtering Subsystem, helping Web search and customizing the consequences of a question dependent on an inside portrayal of client interests (construed by the framework through an exchange). It utilizes a mixture way to deal with client demonstrating (combination of case-based segments and counterfeit neural organization) and exploits semantic organizations, just as an all-around organized information base, to perform exact sifting. Another specialist that learns clients' inclinations by taking a gander at their visit records and afterward gives them refreshed data about the Website is SiteHelper (Ngu and Wu, 1997).

The specialist does two kinds of gradual learning: intuitive learning, by asking client for criticism, and quiet learning, by utilizing the log records. Individual WebWatcher (Mladenic, 1999) is a "individual" specialist, propelled essentially by WebWatcher, that helps Web perusing and features valuable connections from the current page utilizing individual history (content of visited pages), while Let's Browse (Lieberman et al., 1999) executed as an augmentation to Letizia, upholds programmed discovery of the presence of clients,

robotized "station surfing" perusing, dynamic presentation of the client profiles and clarification of proposals.

The utilization of affiliation rules was first proposed in Agrawal et al. (1993) and Agrawal and Srikant (1994). Chen et al. (1998) use affiliation rules calculations to find "fascinating" relationships among client meetings, while the meaning of a meeting as a bunch of maximal forward references (which means an arrangement of Web pages got to by a client) was presented in Chen et al. (1996). This work is likewise the premise of SpeedTracer (Wu et al., 1998), which utilizes referrer and specialist data in the pre-preparing schedules to distinguish clients and worker meetings without extra customer side data, and afterward recognizes the most much of the time visited gatherings of Web pages.

Krishnan et al. (1998) depict way profiling procedures to anticipate future solicitation practices. Along these lines, content can be progressively created before the client demands it. Manber et al. (2000) presents Yahoo! personalization experience. Yippee! was one of the main websites to utilize personalization for an enormous scope. This work examines three instances of personalization: Yahoo! Friend, Inside Yahoo! Search and My Yahoo! application, which were presented in July 1996.

Cingil et al. (2000) depict the requirement for interoperability when mining the Web and how the different norms can be utilized for accomplishing personalization. Besides, they set up a design for giving Web workers consequently produced, machine processable, unique client profiles, while adjusting to clients' protection inclinations.

Mobasher et al. (2000b) depict an overall design for programmed Web personalization utilizing Web use mining strategies. WebPersonalizer is a high-level framework targeting mining Web log records to find information for the creation of customized proposals for the current client dependent on her similitudes with past clients. These client inclinations are consequently gained from Web utilization information, dispensing with in this way the subjectivity from profile information, just as keeping them refreshed. The pre-handling steps illustrated in Cooley et al. (1999a) are utilized to change over the worker signs into worker meetings.

The framework suggests pages from groups that intently match the current meeting. For customizing a site as per the necessities of every client, Spiliopoulou (2000) portrays an

interaction dependent on finding and breaking down client navigational examples. Mining these examples, we can acquire knowledge into a website's use and optimality regarding its present client populace. Utilization designs removed from Web information have been applied to a wide scope of uses. WebSIFT (Cooley et al., 1997, 1999b, 2000) is a site data channel framework that joins use, content, and design data about a website. The data channel consequently recognizes the found examples that have a serious level of emotional intriguing quality.

As referenced previously, the strategies applied for Web personalization ought to be founded on principles and dialects guaranteeing interoperability, better use of the put away data, just as close to home honesty and protection (Cingil et al., 2000). Extensible Markup Language (XML)<sup>2</sup> is a straightforward, entirely adaptable book design initially intended to address the difficulties of enormous scope electronic distributing. XML assumes an inexorably significant part in the trading of a wide assortment of information on the Web and the XML Query Language<sup>3</sup> can be utilized for removing information from XML archives. Asset Description Framework (RDF)<sup>4</sup> is an establishment for handling metadata and comprises a proposal of W3C. It gives interoperability between applications that trade machine-reasonable data on the Web and its grammar can utilize XML.

RDF applications incorporate asset disclosure, content portrayal/connections, information sharing and trade, Web pages' licensed innovation rights, clients' protection inclinations, Websites' security approaches, etc. Stage for Privacy Preferences (P3P)<sup>5</sup> was created by the W3C in 1999 and involves a standard that gives a straightforward and mechanized way for clients to deal with their own data when visiting Websites. Individual profiling is a type of Website guest reconnaissance and prompts various moral contemplations.

Site guests should be persuaded that any gathered data will stay private and secure. P3P empowers Websites to communicate their protection rehearses in a standard arrangement that can be recovered naturally and deciphered effectively by client specialists. P3P client specialists permit clients to be educated regarding site rehearses (in both machine and intelligible arrangements) and to mechanize dynamic dependent on these practices when proper. Along these lines, clients need not read the security approaches at each site they visit. Nonetheless, while P3P gives a standard system to portraying security rehearses, it doesn't guarantee that Websites really follow them.

Open Profiling Standard (OPS)<sup>6</sup> is a proposed standard by Netscape that empowers Web personalization. It permits clients to keep profile records on their hard drives, which can be gotten to by approved Web workers. The clients approach these records and can handle the introduced data. These records can supplant treats and manual online enlistment.

The OPS has been analyzed by the W3C, and its key thoughts have been consolidated into P3P. Client Profile Exchange (CPEX)<sup>7</sup> is an open norm for working with the security empowered trade of client data across dissimilar endeavor applications and frameworks. It coordinates on the web/disconnected client information in a XML-based information model for use inside different venture applications both on and off the Web, coming about in an arranged, client centered climate. The CPEX working gathering plans to foster open-source reference execution and engineer rules to speed selection of the norm among sellers<sup>1.6</sup>.

## TRENDS AND CHALLENGES IN PERSONALIZATION

While personalization looks significant and engaging for the Web insight, a few issues actually stay hazy. One such issue is privacy preserving and comes from the way that personalization requires gathering and putting away undeniably more close to personal information than common non-customized Websites. As per Earp and Baumer (2003), there is minimal legitimate insurance of buyer data gained on the web — either intentionally or automatically — while frameworks attempt to gather however much information as could reasonably be expected from clients, generally without clients' drive and at times without their mindfulness, to stay away from client interruption. Various overviews effectively accessible delineate client inclinations concerning on the web security (Kobsa and Schreck, 2003), with the necessity for protection of namelessness while interfacing with an online framework winning.

### 1.7 Let us Sum Up

- World Wide Web toward a more significant conditions in which customers/clients can quickly and adequately find the information they need.

- "Web Mining" is used inside the field of "Information Analytics"
- Web Content Mining method extricates helpful data from the substance of the online records.
- Web structure mining is the use of finding structure data from the web.

### **1.8 List of References**

1. Web Mining: Applications and Techniques by Anthony Scime State University of New York College at Brockport, USA
2. Web Mining and Social Networking, Techniques and Applications by Authors: **Xu, Guandong, Zhang, Yanchun, Li, Lin**
3. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications) by Bing Lu
4. Mining the Web by Soumen Chakrabarti

### **1.9 Bibliography**

1. Wang Bin and Liu Zhijing, "Web mining research," *Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003*, 2003, pp. 84-89, doi: 10.1109/ICCIMA.2003.1238105.
2. G. Dileep Kumar, Manohar Gosul, Web Mining Research and Future Directions, Advances in Network Security and, Applications, 2011, Volume 196, ISBN : 978-3-642-22539-0
3. Lorentzen, D.G. Webometrics benefitting from web mining? An investigation of methods and applications of two research fields. *Scientometrics* **99**, 409–445 (2014).
4. Jeong, D.H, Hwang, M., Kim, J., Song, S.K., Jung, H., Peters, C., Pietras, N., Kim, D.W.: Information Service Quality Evaluation Model from the Users Perspective, The 2nd International Semantic Technology (JIST) Conference 2012, Nara, Japan, 2012.
5. Helena Ahonen, Oskari Heinonen, Mika Klemettinen, A. Inkeri Verkamo, (1997), Applying Data Mining Techniques in Text Analysis, Report C-1997-23, Department of Computer Science, University of Helsinki, 1997

6. Web Mining: Applications and Techniques by Anthony Scime State University of New York College at Brockport, USA
7. Web Mining and Social Networking, Techniques and Applications by Authors: **Xu**, Guandong, **Zhang**, Yanchun, **Li**, Lin
8. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications) by Bing Lu
9. Mining the Web by Soumen Chakrabarti

### **1.10 Chapter End Exercises**

1. What is Web Mining?
2. How do you suggest we could estimate the size of the estimate the size of the web?
3. Why is Web Information Retrieval Important?
4. Why is Web Information Retrieval Difficult?
5. What is web Content Mining?
6. Explain Web structure mining.
7. What is Web Usage Mining?
8. Compare Data mining and Web Mining?
9. List out difference Between Web Content, Web Structure, and Web Usage Mining.

### **10. Explain Web Personalization.**

## 2.0 Objectives

### 2.1 Introduction

### 2.2 An Overview

#### 2.2.1 What is Text mining?

#### 2.2.2 Text Mining Techniques

#### 2.2.3 Information Retrieval Basics

#### 2.2.4 Text Databases and Information Retrieval

### 2.3 Basic methods of Text Retrieval

#### 2.3.1 Text Retrieval Methods

#### 2.3.2 Boolean Retrieval Model

#### 2.3.3 Vector Space Model

### 2.4 Usage of Text Mining

### 2.5 Areas of text mining in data mining

#### 2.5.1 Information Extraction

#### 2.5.2 Natural Language Processing

#### 2.5.3 Data Mining and IR

### 2.6 Text Mining Process

### 2.7 Text Mining Approaches in Data Mining

#### **2.7.1 Keyword-based Association Analysis**

#### **2.7.2 Document Classification Analysis**

### 2.8 Numericizing text

### **2.8.1 Stemming algorithms**

### **2.8.2 Support for different languages**

### **2.8.3 Exclude certain character**

### **2.8.4 Include lists, exclude lists (stop-words)**

2.9 What is Natural Language Processing (NLP)?

2.9.1 Machine Learning and Natural Language Processing

2.10 Big Data and the Limitations of Keyword Search

2.11 Ontologies, Vocabularies and Custom Dictionaries

2.12 Enterprise-Level Natural Language Processing

2.13 Analytical Tools

2.14 Scalability

2.15 Issues in Text Mining Field

2.16 Let us Sum Up

2.17 List of References

2.18 Bibliography

2.19 Chapter End Exercises

## **2.0 Objectives**

After going through this chapter, you will be able to understand:

- Can be able to define Text mining
- Knowing various Text mining Techniques
- Knowledge about text databases and information retrieval
- Usage of text mining
- Areas of text mining
- Text Mining Approaches
- Natural Language Processing
- Looking into issues in Text Mining

## **2.1 Introduction**

Text mining is a minor departure from a field called data mining that attempts to discover intriguing examples from huge data sets. Text data sets are quickly becoming because of the expanding measure of data accessible in electronic structure, like electronic publications, different sorts of electronic records, email, and the World Wide Web. These days the vast majority of the data in government, industry, business, and different organizations are put away electronically, as text information bases.

## **2.2 An Overview**

Information put away in most content data sets are semi organized information in that they are neither totally unstructured nor totally organized. For instance, a report might contain a couple of organized fields, like title, creators, distribution date, and class, etc, yet in addition contain some generally unstructured content parts, like

theoretical and substance. There has been a lot of studies on the demonstrating and execution of semi organized information in late data set examination. In addition, data recovery strategies, for example, text ordering techniques, have been created to deal with unstructured reports. Customary data recovery strategies become insufficient for the undeniably tremendous measures of text information. Ordinarily, just a little part of the numerous accessible archives will be applicable to a given individual client.

### **2.2.1 Text Mining**

Text information mining can be portrayed as the way toward extricating fundamental information from standard language text. Every one of the information that we produce through instant messages, records, messages, documents are written in like manner language text. Text mining is basically used to draw valuable experiences or patterns from such information.

### **2.2.2 Text Mining Techniques**

Text mining frequently incorporates the accompanying methods:

- Information extraction is a strategy for extricating space explicit data from messages. Text pieces are planned to field or layout templates that have a semantic method:
- Text summarization synopsis includes recognizing, summing up and arranging related content so clients can effectively manage data in huge records.
- Text categorization includes sorts out archives into a scientific categorization, in this manner considering more effective ventures. It includes the task of subject descriptors or arrangement codes or theoretical ideas to finish messages.
- Text clustering includes naturally grouping archives into groups where records inside each gathering share normal highlights.

All content mining approaches use data recovery components. Surely, the qualification between data recovery techniques and text mining is obscured. In the following area data recovery essentials are examined. Various complex expansions

to fundamental data recovery progressed in the lawful field are portrayed. We then, at that point talk about instances of data extraction, text outline, text arrangement and text grouping in law.

### **2.2.3 Information Retrieval Basics**

The point of proficient data recovery ought to be to recover that data, and just that data which is considered pertinent to a given question. Salton states that a regular data recovery framework chooses archives from an assortment because of a client's question and positions these reports as indicated by their importance to the inquiry. This is essentially refined by coordinating with a book portrayal with a portrayal of the inquiry.

Data recovery and information base frameworks have a few similitudes. While data set frameworks have zeroed in on question preparing and exchanges identifying with organized information, data recovery is worried about the association and data from an enormous number of text-based records. The undertaking of questioning information bases and text recovery frameworks is totally different. For text recovery frameworks, the coordinating isn't deterministic and regularly joins a component of vulnerability. Recovery models commonly rank the recovered record as per their possible importance to the inquiry.

Legitimate data recovery thinks about looking through both organized and unstructured substance. For organized data, the semantics can be obviously not set in stone and can be portrayed with straightforward and clear ideas. This data classification contains, for example, distinguishing proof information of the writings, information for rendition the board and the capacity and job of specific parts. This information are regularly included the type of metadata (i.e information that depict different information) to the reports. Unstructured data regularly happens in normal language messages or in different configurations like sound and video and for the most part has an intricate semantics.

The listed terms chosen concern single words and multi word stages and are expected to mirror the substance of the content. She asserts that a pervasive cycle

of choosing regular language file terms from messages that mirror its substance is made out of the accompanying advances:

1. Lexical investigation the content is parsed, and singular words are perceived.
2. The expulsion of stopwords – a book recovery framework regularly relates a stop list with a bunch of reports. A stop list is a bunch of words that are considered unessential, (for example, a , the, for) or if nothing else unimportant for the given question.
3. The discretionary decrease of the leftover words to their stem structure A gathering of various words might have a similar word stem.

The content recovery framework needs to recognize gatherings of words that have a little syntactic variety from one another and just utilize single word from each gathering of. just use penetrate rather than breaks, penetrate, penetrated. There are various strategies for stemming, a considerable lot of which depend upon phonetic information on the assortment's language.

4. The discretionary definition of expressions as list terms. Strategies of expression acknowledgment utilize the measurements of co-events of words or depend upon semantic information on the assortment's language.
5. The alternative substitution of words, word stems or expressions by their thesaurus class terms – A thesaurus replaces the individual words or expressions of a book by more uniform ideas.
6. The calculation of the significance marker or term weight of each leftover word stem or word, thesaurus class term or expressions term.

#### **2.2.4 Text Databases and Information Retrieval:**

Text databases (record information bases) are huge assortments of reports from different sources: news, stories, research papers, books, computerized libraries, E-mail messages, and Web pages, library data set, and so on. Information put away is

typically semi-organized and Traditional inquiry strategies become insufficient for the inexorably tremendous measures of text information

### **Information retrieval (IR)**

A field created in corresponding with data set frameworks Information is coordinated into (an enormous number of) records IR manages the issue of finding significant archives as for the client information or inclination

IR Systems and DBMS manage various issues Typical DBMS issues are update, exchange the executives, complex items

IR issues are management of unstructured reports, rough hunt utilizing catchphrases and importance

1. Typical IR systems Online library catalogues
2. Online document management systems
3. Main IR approaches
4. pull for short-term information need
5. push for long-term information need (e.g., recommender systems)

### **2.3 Basic methods of Text Retrieval:**

#### **2.3.1 Text Retrieval Methods**

- Document Selection Query characterizes a set of imperatives
- Only the reports that fulfill the inquiry are returned A regular methodology is the Boolean Retrieval Model
- Document Ranking
- Documents are positioned based on their significance concerning the client question
- For each report a level of importance to the question is estimated
- A regular methodology is the Vector Space Model.

Generally, not suitable to satisfy information need Useful only in very specific domain where users have a big expertise

How to select keywords to capture Basic concepts? How to assign weights to each term?

### **2.3.2 Boolean Retrieval Model**

An inquiry is made out of keyword connected by the three consistent connectives: not, and, or

E.g.: vehicle and fix, plane, or plane

In the Boolean model each record is either pertinent or non-important, contingent upon it matches or not the question

Limits

For the most part, not reasonable to fulfill data need Useful just in quite certain area where clients have a major skill

### **2.3.3 Vector Space Model:**

A record and a question are addressed as vectors in high dimensional space comparing to every one of the keywords. Pertinence is estimated with a suitable closeness measure characterized over the vector space.

Issues:

How to choose Basic ideas? How to allot loads to each term?

## **2.4 Usage of Text Mining:**

Text Mining intends to separate helpful information from text reports

- Approaches Keyword-based
- Relies on IR methods Tagging
- Manual labeling
- Automatic order Information-extraction
- Natural Language Processing (NLP) Keyword-Based Association Analysis Document Classification

Natural Language Processing is a field of software engineering, AI, and semantics concerned about the associations among PCs and human (normal) dialects. In that capacity, NLP is identified with the space of human computer interaction. Numerous difficulties in NLP include regular language understanding, that is, empowering PCs to get importance from human or normal language info, and others include normal language age

Modern NLP calculations depend on AI, particularly statistical machine learning. The worldview of AI is not quite the same as that of earlier endeavours at language preparing. Earlier executions of language-handling undertakings normally elaborate the immediate hand coding of huge arrangements of rules. The AI worldview calls rather for utilizing general learning algorithms regularly, although not generally, grounded in statistical inference to naturally learn such principles through the investigation of enormous corpora of ordinary certifiable models. A corpus (plural, "corpora") is a bunch of records (or now and again, singular sentences) that have been hand-clarified with the right qualities to be learned.

Various classes of AI calculations have been applied to NLP undertakings. These algorithms take as information an enormous arrangement of "highlights" that are produced from the information. Some of the earliest-used algorithms, such as decision trees, produced systems of hard if-then rules similar to the systems of hand-written rules that were then common. Progressively, nonetheless, research has focused in on measurable models, which make delicate, probabilistic choices dependent on connecting genuine esteemed loads to each input highlight. Such models enjoy the benefit that they can communicate the general sureness of a wide

range of potential answers instead of just one, creating more dependable outcomes when a particular model is incorporated as a segment of a bigger framework.

Systems based on machine-learning algorithms have many advantages over hand-produced rules:

- The learning systems utilized during AI naturally centre around the most widely recognized cases, though when composing rules by hand it is normal not clear at all where the work ought to be coordinated.
- Automatic learning systems can utilize measurable deduction calculations to create models that are vigorous to new information (for example containing words or constructions that have not been seen previously) and to mistake information (for example with incorrectly spelled words or words inadvertently precluded). By and large, taking care of such info smoothly with transcribed guidelines or all the more for the most part, making frameworks of written by hand decides that settle on delicate choices is very troublesome, mistake inclined and tedious.
- Systems dependent on consequently learning the standards can be made more exact basically by providing more info information. Notwithstanding, frameworks dependent on transcribed principles must be made more precise by expanding the intricacy of the guidelines, which is a significantly more troublesome undertaking. Specifically, there is a breaking point to the intricacy of frameworks dependent available made guidelines, past which the frameworks become increasingly unmanageable. Notwithstanding, making more information to contribution to AI frameworks essentially requires a relating expansion in the quantity of worker hours worked, by and large without huge expansions in the intricacy of the explanation interaction.

## **2.5 Areas of text mining in data mining:**

### **2.5.1 Information Extraction:**

The automatic extraction of structured data such as entities, entities relationships, and attributes describing entities from an unstructured source is called information extraction.

### **2.5.2 Natural Language Processing:**

NLP represents Natural language preparing. PC programming can comprehend human language however same as it very well might be spoken. NLP is essentially a part of counterfeit intelligence (AI). The improvement of the NLP application is troublesome on the grounds that PCs by and large anticipate that humans should "Talk" to them in a programming language that is precise, clear, and particularly organized. Human discourse is generally not legitimate with the goal that it can rely upon numerous perplexing factors, including slang, social setting, and local lingos

### **2.5.3 Data Mining and IR**

Information mining alludes to the extraction of valuable information, concealed examples from huge informational indexes. Information mining instruments can foresee practices and future patterns that permit organizations to settle on a superior information driven choice. Information mining instruments can be utilized to determine numerous business issues that have generally been too tedious.

Data recovery manages recovering valuable information from information that is put away in our frameworks. Then again, as a similarity, we can see web indexes that occur on sites, for example, online business destinations or some other locales as a feature of data recovery.

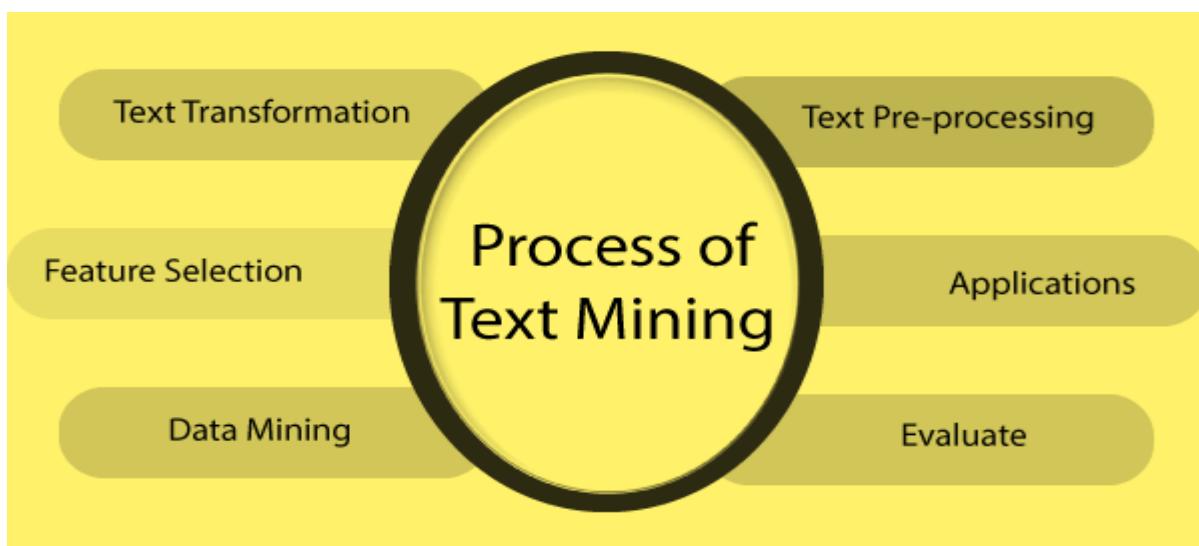
The content mining market has encountered remarkable development and selection throughout the most recent couple of years and furthermore expected to acquire critical development and reception in the coming future. One of the essential purposes for the reception of text mining is higher rivalry in the business market, numerous associations looking for esteem added answers for contend with different associations. With expanding finish in business and adjusting client points of view, associations are making colossal speculations to discover an answer that is fit for breaking down client and contender information to further develop intensity.

The essential wellspring of information is internet business sites, web-based media stages, distributed articles, study, and some more. The bigger piece of the produced

information is unstructured, which makes it trying and costly for the associations to examine with the assistance of individuals. This test coordinates with the remarkable development in information age has prompted the development of logical instruments. It isn't simply ready to deal with enormous volumes of text information yet in addition helps in dynamic purposes. Text mining programming enables a client to draw valuable data from a tremendous arrangement of information accessible sources.

## 2.6 Text Mining Process:

The text mining process incorporates the following steps to extract the data from the document.



**Figure 1 Process of Text Mining**

### Text transformation

A Text transformation is a strategy that is utilized to control the capitalization of the content

Here the two significant method of archive portrayal is given.

1. Bag of words

## 2.Vector Space

### **Text Pre-processing**

Pre-processing is a huge task and a critical step in Text Mining, Natural Language Processing (NLP), and information retrieval (IR). In the field of text mining, information pre-handling is utilized for extracting useful data and information from unstructured data. IR involves picking which records in an assortment ought to be recovered to satisfy the client's need.

### **Feature selection:**

Feature determination is a huge piece of data mining. It can be characterized as the process toward decreasing the contribution of handling or tracking down the fundamental data sources. This is additionally called as “variable selection”.

### **Data Mining:**

Presently, in this progression, the text mining strategy converges with the traditional interaction. Exemplary Data Mining systems are utilized in the primary information base.

### **Evaluate:**

Afterward, it evaluates the results. Once the result is evaluated, the result abandon.

### **Applications:**

Applications explained below:

### **Risk Management:**

Risk Management is a deliberate and sensible technique of examining, distinguishing, treating, and observing the dangers implied in any activity or process in associations. Inadequate danger examination is normally a main source of frustration. It is especially obvious in the monetary associations where appropriation of Risk Management Software dependent on text mining innovation can adequately

improve the capacity to decrease risks. It empowers the organization of millions of sources and petabytes of text documents and enabling to associate the information. It assists with getting to the suitable information at the ideal opportunity.

### **Customer Care Service:**

Text mining strategies, especially NLP, are discovering expanding importance in the field of customer care. Organizations are spending in text investigation programming to work on their general insight by getting to the text-based information from various sources like client input, overviews, client calls, and so on. The essential target of text analysis is to decrease the reaction time of the organizations and help to address the complaints of the customer quickly and productively

### **Business Intelligence:**

Organizations and business firms have begun to utilize text mining techniques as a significant part of their business intelligence. Other than giving huge experiences into customer behaviour and patterns, text mining procedures likewise support organizations to analyze the characteristics and weakness of their rival's along these lines, giving them an upper hand on the lookout.

### **Social Media Analysis:**

Web-based media investigation assists with following the online data, and there are various text mining tools designed particularly for execution of social media sites. These tools help to screen and decipher the content created by means of the web from the news, messages, online journals, and so on. Text mining devices can correctly examine the all-out no of posts, followers, and absolute no of preferences of your image on an online media stage that empowers you to comprehend the reaction of the people who are interacting with your image and content.

## **2.7 Text Mining Approaches in Data Mining:**

These are the following text mining approaches that are used in data mining.

### **2.7.1 Keyword-based Association Analysis:**

It gathers sets of keywords or terms that regularly happen together and thereafter find the affiliation relationship among them. To begin with, its pre-processes the text data by parsing, stemming, removing stop words, and so forth. When it pre-prepared the information, then, at that point it actuates association mining algorithms. Here, human exertion isn't needed, so the quantity of undesirable outcomes and the execution time is decreased.

### **2.7.2 Document Classification Analysis:**

#### **Automatic document classification:**

This investigation is utilized for the classification of the colossal number of online content archives like site pages, messages, and so forth. Text document varies with the grouping of social information as archive data sets are not coordinated by quality qualities sets.

### **2.8 Numericizing text:**

#### **2.8.1 Stemming algorithms**

A critical pre-processing step prior to requesting of information reports begins with the stemming of words. The expressions "stemming" can be characterized as a decrease of words to their roots. For instance, unique syntactic types of words and requested are something very similar. The main role of stemming is to guarantee a comparable word by text mining program.

#### **2.8.2 Support for different languages:**

There are some exceptionally language- dependent operations, for example, stemming, equivalents, the letters that are permitted in words. Accordingly, support for different dialects is significant.

#### **2.8.3 Exclude certain character:**

Excluding numbers, explicit characters, or series of characters, or words that are more limited or more than a particular number of letters should be possible before the requesting of the input records.

#### **2.8.4 Include lists, exclude lists (stop-words):**

A specific of words to be recorded can be described, and it is helpful when we need to look for a particular word. It likewise orders the information reports dependent on the frequencies with which those words happen. Moreover, "stop words," which means terms that are to be dismissed from the requesting can be portrayed. Regularly, a default list of English stop words fuses "the," "a," "since," and so on. These words are utilized in the individual language regularly yet convey next to no information in the record.

### **2.9 What is Natural Language Processing (NLP)?**

Natural Language Understanding aides' machines "read" text (or another input such as speech) by mimicking the human ability to comprehend a characteristic language like English, Spanish or Chinese. Natural Language Processing incorporates both Natural Language Understanding and Natural Language Generation, which simulates the human capacity to make natural language text for example to sum up data or partake in a dialogue.

As an innovation, natural language handling has grown up in the course of recent years, with items, for example, Siri, Alexa and Google's voice search utilizing NLP to comprehend and react to user requests. Refined content mining applications have additionally been created in fields as different as clinical examination, risk management, customer care, insurance, and logical promoting.

Today's natural language processing systems can examine limitless measures of text-based information without weariness and in a steady, fair way. They can comprehend ideas inside complex settings and interpret ambiguities of language to remove key realities and connections or give outlines. Given the immense amount of unstructured information that is created each day, from electronic health records (EHRs) to social media posts, this type of computerization has gotten basic to investigating text-based information effectively.

### **2.9.1 Machine Learning and Natural Language Processing**

AI is a computerized reasoning (AI) innovation which furnishes frameworks with the capacity to consequently gain as a matter of fact without the requirement for explicit programming and can assist with tackling complex issues with exactness that can equal or even here and there outperform humans. However, AI requires well-curated contribution to prepare from, and this is ordinarily not accessible from sources, for example, electronic health records (EHRs) or logical writing where the greater part of the information is unstructured content.

When applied to EHRs, clinical preliminary records or full content writing, normal language preparing can remove the spotless, organized information expected to drive the high-level prescient models utilized in AI, subsequently lessening the requirement for costly, manual explanation of preparing information.

### **2.10 Big Data and the Limitations of Keyword Search**

While customary web crawlers like Google presently offer refinements like equivalents, auto-consummation and semantic inquiry (history and setting), by far most of query items just highlight the area of reports, leaving searchers with the issue of going through hours physically extricating the vital information by perusing singular records.

The restrictions of customary pursuit are compounded by the development in large information over the previous decade, which has helped increment the quantity of results returned for a solitary question by a web index like Google from many thousands to many millions.

### **2.11 Ontologies, Vocabularies and Custom Dictionaries**

Ontologies, vocabularies, and custom word references are incredible assets to help with search, information extraction and information incorporation. They are a critical part of numerous content mining devices, and give arrangements of key ideas, with names and equivalents regularly orchestrated in an order.

Web search tools, text examination devices and normal language preparing arrangements become considerably more impressive when conveyed with area explicit ontologies. Ontologies empower the genuine importance of the content to be seen, in any event, when it is communicated in an unexpected way (for example Tylenol versus Acetaminophen). NLP methods broaden the force of ontologies, for instance by permitting coordinating of terms with various spellings (Estrogen or Estrogen), and by considering setting ("SCT" can allude to the quality, "Secretin", or to "Step Climbing Test").

The detail of a metaphysics incorporates a jargon of terms and formal imperatives on its utilization. Venture prepared normal language handling requires scope of vocabularies, ontologies and related procedures to recognize ideas in their right setting:

- Thesauri, vocabularies, scientific categorizations and ontologies for ideas with known terms.
- Pattern-based methodologies for classes like estimations, transformations and compound names that can incorporate novel (concealed) terms.
- Domain-explicit, rule-based idea distinguishing proof, explanation, and change.
- Integration of client vocabularies to empower bespoke comment.
- Advanced search to empower the distinguishing proof of information ranges for dates, mathematical qualities, region, fixation, rate, span, length and weight.

## **2.12 Enterprise-Level Natural Language Processing**

The utilization of cutting-edge investigation addresses a genuine chance inside the drug and medical care ventures, where the test lies in choosing the proper

arrangement, and afterward carrying out it proficiently across the enterprise. Effective natural language handling requires various highlights that ought to be joined into any undertaking level NLP arrangement, and a portion of these are depicted below.

## **2.13 Analytical Tools**

There is immense assortment in archive piece and text-based setting, including sources, arrangement, language, and syntax. Handling this assortment requires a scope of procedures:

- Transformation of interior and outside record designs (for example HTML, Word, PowerPoint, Excel, PDF text, PDF picture) into a normalized accessible organization.
- The capacity to recognize, tag and search in explicit record segments (regions), for instance: centering an inquiry to eliminate commotion from a paper's reference area.

- Linguistic handling to distinguish the significant units inside text like sentences, thing, and action word assemblies with the connections between them.
- Semantic apparatuses that distinguish ideas inside the content like medications and sicknesses and standardize to ideas from standard ontologies. Notwithstanding center life science and medical care ontologies like MedDRA and MeSH, the capacity to add their own word references is a prerequisite for some associations.
- Pattern acknowledgment to find and recognize classifications of data, not effectively characterized with a word reference approach. These incorporate dates, mathematical data, biomedical terms (for example focus, volume, dose, energy) and quality/protein transformations.
- The capacity to handle inserted tables inside the content, regardless of whether designed utilizing HTML or XML, or as free content.

## **2.14 Scalability**

Text-mining difficulties differ in size, from infrequent admittance to a couple of archives to combined ventures over numerous storehouses and a great many records. An advanced regular language preparing arrangement should hence:

- Provide the capacity to run refined questions more than a huge number of records, every one of which might be a great many pages long.
- Handle vocabularies and ontologies containing a huge number of terms.
- Run on equal designs, regardless of whether standard multi-center, group or cloud.
- Provide a connector to run regular language handling in help situated conditions like ETL (Extract, Transform, Load), semantic enhancement and sign location, for instance: clinical danger observing in medical care.

Numerous issues happen during the content mining interaction and impact the productivity and adequacy of dynamic. Intricacies can emerge at the middle of the road phase of text mining. In pre-handling stage different guidelines and guidelines are characterized to normalize the content that make text mining measure proficient. Prior to applying design investigation on the record there is a need to change over unstructured information into moderate structure however at this stage mining measure has its own confusions.

Another significant issue is a multilingual book refinement reliance that make issues. Just couple of apparatuses are accessible that help different dialects. Different calculations and methods are utilized freely to help multilingual content. Since various significant records endure outside the content mining measure in light of the fact that different devices don't uphold them. These issues make a ton of issues in information disclosure and dynamic cycle.

The utilization of equivalents, polysems and antonyms in the archives make issues (recondition) for the content mining devices that take both in a similar setting. It is hard to classify the archives when assortment of reports is huge and produced from different fields having a similar area. Shortened forms gives changed significance in various circumstance is likewise a major issue. Differing ideas of granularity change the setting of text as indicated by the condition and space information. There is need to depict rules as indicated by the field that will be utilized as a norm nearby and can be inserted in text mining instruments as a module.

It involves loads of exertion and time to create and send modules in all fields independently. Words having same spelling yet give different significance, for instance, fly and fly. Text mining apparatuses considered both as comparative while

one is action word and other is thing. Syntactic standards as indicated by the nature and setting is as yet an open issue in the field of text mining.

The issue of TR can be officially defined as to distinguish a subset of important reports to a question from an assortment of records. There are two systems to execute this objective: (1) direct choice; and (2) backhanded determination through positioning.

By and large, positioning is liked and more basic since pertinence involves degree and regardless of whether we can choose the right archives, it's as yet attractive to rank them. Therefore, most existing examination in data recovery has accepted that the objective is to foster a decent positioning capacity. We will cover various approaches to rank reports later. These are additionally called recovery models.

All recovery frameworks have some normal segments. One of them is the tokenizer, which has to do with planning a book to a flood of tokens/terms. This has to do with the more broad issue of addressing text in the framework in some structure so we can coordinate with a question with a report. The overwhelming procedure for text portrayal is to address a book as a "sack of terms". Tokenization has to do with deciding the terms.

A term can be a solitary work, an expression, or n-grams of characters (i.e., a succession of n characters). One usually utilized method in preparing a language like English is stemming, which maps semantically related words, for example, "PCs", "PC", "figure", and "calculation" everything appears a similar root structure (e.g., "process"). This methodology regularly helps yet doesn't generally. It is as yet an open inquiry whether one ought to do stemming, and the appropriate response exceptionally relies upon explicit applications.

A fundamental part of any recovery model is the criticism instrument. That is, the point at which the client will pass judgment on archives and mark some as significant others as non-applicable, the framework ought to have the option to gain from such guides to further develop search precision. This is called significance input. Client contemplations have shown; notwithstanding, a client is regularly reluctant to make such decisions, raising worries about the commonsense worth of pertinence input. Pseudo criticism (likewise called daze/programmed input) essentially expects some highest level reports to be applicable, hence doesn't need a client to mark records.

Pseudo criticism has likewise been demonstrated to be successful by and large, however it might hurt execution for certain questions. Naturally, pseudo criticism approach depends on term co-events in the highest-level records to dig for related terms to the question terms. These new terms can be utilized to extend an inquiry and increment review. Pseudo input may likewise further develop exactness through enhancing the first inquiry terms with new related terms and allotting more precise loads to question terms.

## **2.17 List of References**

1. Text Mining: Applications and Theory by Michael W. Berry, Jacob Kogan
2. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data by  
Book by James Sanger and Ronen Feldman
3. An Introduction to Text Mining: Research Design, Data Collection, and Analysis  
Book by Gabe Ignatow and Rada Mihalcea

## **2.18 Bibliography**

1. Zanini, Nadir & Dhawan, Vikas. (2015). Text Mining: An introduction to theory and some applications. *Research Matters*. 38-44.
2. Anton Kanev, Stuart Cunningham, Terekhov Valery, "Application of formal grammar in text mining and construction of an ontology", *Internet Technologies and Applications (ITA)* 2017, pp. 53-57, 2017.

3. Wang Huiqin, Lin Weiguo, "Analysis of the Art of War of Sun Tzu by Text Mining Technology", *Computer and Information Science (ICIS) 2018 IEEE/ACIS 17th International Conference on*, pp. 626-628, 2018.
4. Avik Sarkar, Md. Sharif Hossen, "Automatic Bangla Text Summarization Using Term Frequency and Semantic Similarity Approach", *Computer and Information Technology (ICCIT) 2018 21st International Conference of*, pp. 1-6, 2018.
5. Anoud Shaikh, Naeem Ahmed Mahoto, Mukhtiar Ali Unar, "Bringing Shape to Textual Data – A Feasible Demonstration", *Mehran University Research Journal of Engineering and Technology*, vol. 38, pp. 901, 2019.
6. Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem, Khaled Shaalan, *Intelligent Natural Language Processing: Trends and Applications*, vol. 740, pp. 373, 2018.
7. Text Mining: Applications and Theory by Michael W. Berry, Jacob Kogan
8. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data by  
Book by James Sanger and Ronen Feldman
9. An Introduction to Text Mining: Research Design, Data Collection, and Analysis  
Book by Gabe Ignatow and Rada Mihalcea
10. Text Mining in Practice with R Book by Ted Kwartler

11. Natural Language Processing and Text Mining Book by Steve R. Poteet

## **2.19 Chapter End Exercises**

1. Explain Text Mining.
2. What are the various text mining Techniques?
3. Briefly explain issues in text mining field.
4. Short notes on: Ontologies, Vocabularies and Custom Dictionaries
5. What is Natural Language Processing?
6. Explain Text mining Process.

- 3.0 Objectives
- 3.1 Introduction
- 3.2 An Overview
  - 3.2.1 What is Information Retrieval?
  - 3.2.2 What is an IR Model?
  - 3.2.3 Components of Information Retrieval/ IR Model
- 3.3 Difference Between Information Retrieval and Data Retrieval
- 3.4 User Interaction with Information Retrieval System
- 3.5 Past, Present, and Future of Information Retrieval
  - 3.5.1 IR on the Web
  - 3.5.2 Why is IR difficult?
- 3.6 Functional Overview
  - 3.6.1 Item Normalization
  - 3.6.2 Selective Dissemination (Distribution, Spreading) of Information
  - 3.6.3 Document Database Search
  - 3.6.4 Multimedia Database Search
- 3.7 Application areas within IR
  - 3.7.1 Cross language retrieval
  - 3.7.2 Speech/broadcast retrieval
  - 3.7.3 Text Categorization
  - 3.7.4 Text Summarization
  - 3.7.5 Structured document element retrieval
- 3.8 Web Information Retrieval Models
  - 3.8.1 Vector Model
  - 3.8.2 Vector space model
  - 3.8.3 Probabilistic Model
- 3.9 Let us Sum Up
- 3.10 List of References

### 3.11 Bibliography

### 3.12 Chapter End Exercises

## Chapter 3 – Information Retrieval

### 3.0 Objectives

- Knowledge about Information Retrieval
- Various IR models
- Components of IR
- User Interaction
- Application Area

### 3.1 Introduction

An IR framework can address, store, sort out, and access data things. A bunch of catchphrases are needed to look. Catchphrases are the thing individuals are looking for in web crawlers. These keywords sum up the portrayal of the data.

### 3.2 An Overview

The framework looks more than billions of records put away on large number of PCs. A spam channel, manual or programmed implies are given by email program to grouping the sends so it very well may be put straightforwardly into specific envelopes. For instance, Information Retrieval can be the point at which a client enters an inquiry into the framework.

Not just custodians, proficient searchers, and so on draw in themselves in the action of data recovery however these days countless individuals take part in IR consistently when they use web indexes. Data Retrieval is accepted to be the prevailing type of Information access. The IR framework helps the clients in discovering the data they require yet it doesn't expressly return the responses to the inquiry. It tells in regard to the presence and area of reports that may comprise of the necessary data. Data recovery additionally stretches out help to clients in perusing or sifting record assortment or preparing a bunch of recovered reports.

### **3.2.1 What is Information Retrieval?**

IR characterized as a product program that arrangements with the association, stockpiling, recovery, and assessment of data from record storehouses, especially literary data. Data Retrieval is the action of getting material that can generally be recorded on an unstructured nature for example typically text which fulfills a data need from inside huge assortments which is put away on PCs.

### **3.2.2 What is an IR Model?**

An Information Retrieval (IR) model chooses and positions the archive that is needed by the client, or the client has requested as an inquiry. The reports and the inquiries are addressed likewise, so that record determination and positioning can be formalized by a coordinating with work that profits a recovery status esteem (RSV) for each archive in the assortment. A considerable lot of the Information Retrieval frameworks address report substance by a bunch of descriptors, called terms, having a place with a jargon V. An IR model decides the inquiry report coordinating with work as per four fundamental methodologies:

### **3.2.3 Components of Information Retrieval/ IR Model**

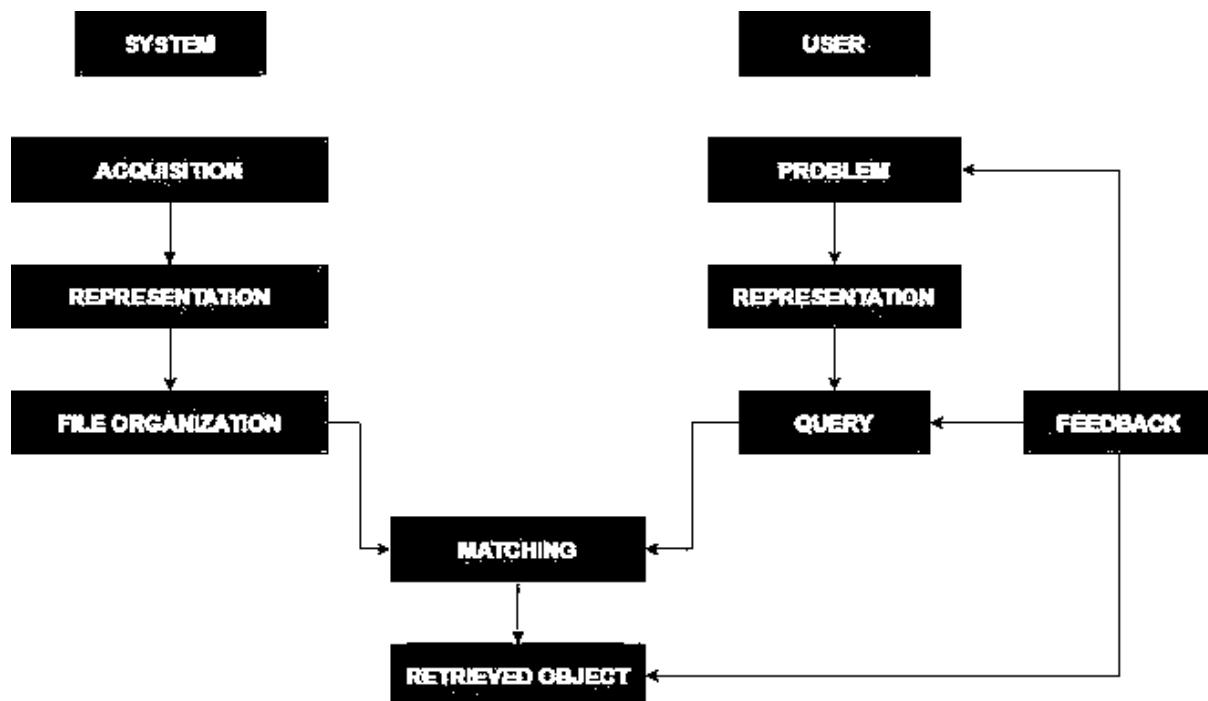


Figure 1: IR Model

### **Acquisition:**

In this progression, the choice of records and different items from different web assets that comprise of text-based archives happens. The necessary information is gathered by web crawlers and put away in the data set.

### **Representation:**

It comprises of ordering that contains free-text terms, controlled jargon, manual and programmed strategies also. model: Abstracting contains summing up and Bibliographic portrayal that contains creator, title, sources, information, and metadata.

### **File Organization:**

There are two sorts of document association strategies. for example Consecutive: It contains records by archive information. Altered: It contains term by term, rundown of records under each term. Mix of both.

### **Query:**

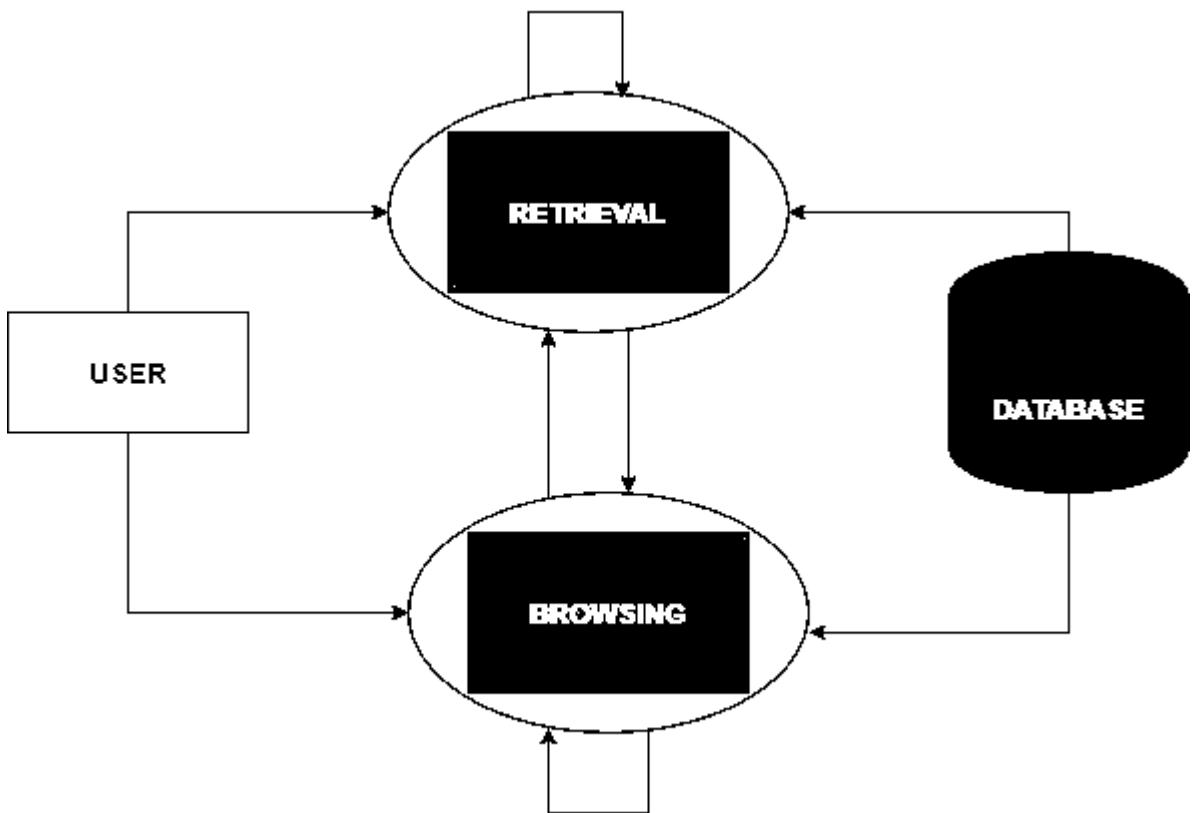
An IR cycle begins when a client enters an inquiry into the framework. Questions are formal explanations of data needs, for instance, search strings in web search tools. In data recovery, an inquiry doesn't remarkably recognize a solitary article in the assortment. All things being equal, a few articles might coordinate with the inquiry, maybe with various levels of pertinence.

### **3.3 Difference Between Information Retrieval and Data Retrieval**

Information Retrieval	Data Retrieval
The software the program that	Data retrieval deals with obtaining data from

Information Retrieval	Data Retrieval
deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual information.	a database management system such as ODBMS. It is A process of identifying and retrieving the data from the database, based on the query provided by user or application.
Retrieves information about a subject.	Determines the keywords in the user query and retrieves the data.
Small errors are likely to go unnoticed.	A single error object means total failure.
Not always well structured and is semantically ambiguous.	Has a well-defined structure and semantics?
Does not provide a solution to the user of the database system.	Provides solutions to the user of the database system.
The results obtained are approximate matches.	The results obtained are exact matches.
Results are ordered by relevance.	Results are unordered by relevance.
It is a probabilistic model.	It is a deterministic model.

### 3.4 User Interaction with Information Retrieval System



**Figure 2 : User Interaction with Information Retrieval System**

#### The User Task:

The data initially should be converted into an inquiry by the client. In the data recovery framework, there is a bunch of words that pass on the semantics of the data that is required though, in an information recovery framework, a question articulation is utilized to pass on the limitations which are fulfilled by the items. Model: A client needs to look for something yet winds up looking with something else. This implies that the client is perusing and not looking. The above figure shows the collaboration of the client through various errands.

#### Logical View of the Documents:

Quite a while past, records were addressed through a bunch of file terms or catchphrases. These days, present day PCs address archives by a full arrangement of words which lessens the arrangement of agent watchwords. This should be possible by disposing of key words for example articles and connectives. These tasks are text activities. These content tasks diminish the intricacy of the report portrayal from full content to set off list terms.

### **3.5 Past, Present, and Future of Information Retrieval**

#### **1. Early Developments:**

**As there was an increment in the requirement for a great deal of data, it became important to construct information designs to get quicker access. The file is the information structure for quicker recovery of data. Over hundreds of years manual order of chains of importance was accomplished for records.**

#### **2. Data Retrieval in Libraries:**

**Libraries were quick to receive IR frameworks for data recovery. In original, it comprised, mechanization of past advances, and the pursuit depended on creator name and title. In the subsequent age, it included looking by subject heading, catchphrases, and so on in the third era, it comprised of graphical interfaces, electronic structures, hypertext highlights, and so on**

#### **3. The Web and Digital Libraries:**

**It is less expensive than different wellsprings of data, it gives more noteworthy admittance to networks because of computerized correspondence, and it gives free admittance to distribute on a bigger medium.**

##### **3.5.1 IR on the Web**

- No stable document collection (spider, crawler)
- Invalid document, duplication, etc.
- Huge number of documents (partial collection)
- Multimedia documents

- Great variation of document quality
- Multilingual problem

### **3.5.2 Why is IR difficult?**

- Vocabularies mismatching – Synonymy: e.g., car vs. automobile – Polysemy: table
- Queries are ambiguous, they are partial specification of user's need
- Content representation may be inadequate and incomplete
- The user is the ultimate judge, but we don't know how the judge judges...
  - The notion of relevance is imprecise, context- and user dependent

### **3.6 Functional Overview**

A Total Information Storage and Retrieval System is composed of four major functional processes:

- 1) Item Normalization
- 2) Selective Dissemination of Information (i.e., "Mail")
- 3) Archival Document Database Search, and an Index
- 4) Database Search along with the Automatic File Build process that supports Index Files.

#### **3.6.1 Item Normalization:**

The initial phase in any incorporated framework is to standardize the approaching things to a standard arrangement. Thing standardization gives consistent rebuilding of the thing. Extra activities during thing standardization are expected to make an accessible information structure: recognizable proof of preparing tokens (e.g., words), portrayal of the tokens, and stemming (e.g., eliminating word endings) of the tokens.

The handling tokens and their portrayal are utilized to characterize the accessible content from the complete got text. shows the standardization cycle. Normalizing the information takes the diverse outer organizations of information and plays out the interpretation to the arrangements worthy to the framework. A framework might have

a solitary configuration for all things or permit numerous organizations. One illustration of normalization could be interpretation of unknown dialects into Unicode. Each language has an alternate inner twofold encoding for the characters in the language. One standard encoding that covers English, French, Spanish, and so forth is ISO-Latin.

To help the clients in creating lists, particularly the expert indexers, the framework gives an interaction called Automatic File Build (AFB). Multi-media adds an additional measurement to the standardization interaction. As well as normalizing the printed input, the multi-media input likewise should be normalized. There are a ton of alternatives to the principles being applied to the standardization. In the event that the information is video the reasonable advanced norms will be either MPEG-2, MPEG-1, AVI or Real Media. MPEG (Motion Picture Expert Group) guidelines are the most all-inclusive principles for better video where Real Media is the most well-known norm for lower quality video being utilized on the Internet. Sound guidelines are normally WAV or Real Media (Real Audio). Pictures differ from JPEG to BMP.

The following cycle is to parse the thing into coherent sub-divisions that have importance to the client. This cycle, called "Drafting," is apparent to the client and used to build the accuracy of a hunt and streamline the showcase. A run of the mill thing is partitioned into zones, which might cover and can be various levelled, like Title, Author, Abstract, Main Text, Conclusion, and References. The drafting data is passed to the handling token recognizable proof activity to store the data, permitting searches to be confined to a particular zone. For instance, on the off chance that the client is keen on articles talking about "Einstein" the inquiry ought to exclude the Bibliography, which could incorporate references to articles composed by "Einstein." Systems decide words by separating input images into 3 classes:

- 1) Valid word symbols
- 2) Inter-word symbols
- 3) Special processing symbols.

A word is characterized as a bordering set of word images limited by between word images. In numerous frameworks between word images are non-accessible and ought to be painstakingly chosen. Instances of word images are alphabetic

characters and numbers. Instances of conceivable between word images are spaces, periods and semicolons. The specific meaning of a between word image is subject to the parts of the language area of the things to be prepared by the framework. For instance, a punctuation might be of little significance if by some stroke of good luck utilized for the possessive case in English yet may be basic to address unfamiliar names in the information base. Then, a Stop List/Algorithm is applied to the rundown of potential handling tokens.

The target of the Stop work is to save framework assets by disposing of from the arrangement of accessible preparing tokens those that have little worth to the framework. Given the critical expansion in accessible modest memory, stockpiling and handling power, the need to apply the Stop capacity to preparing tokens is diminishing. Instances of Stop calculations are: Stop all numbers more prominent than "999999" (this was chosen to permit dates to be accessible) Stop any handling token that has numbers and characters intermixed.

### **3.6.2 Selective Dissemination (Distribution, Spreading) of Information**

The Selective Dissemination of Information (Mail) Process gives the ability to powerfully analyse recently got things in the data framework against standing proclamations of interest of clients and convey the thing to those clients whose assertion of interest coordinates with the substance of the thing. The Mail interaction is made out of the hunt cycle, client explanations of interest (Profiles) and client mail records. As everything is gotten, it is prepared against each client's profile. A profile contains an ordinarily wide hunt proclamation alongside a rundown of client mail records that will get the report if the pursuit explanation in the profile is fulfilled. Particular Dissemination of Information has not yet been applied to sight and sound sources.

### **3.6.3 Document Database Search**

The Document Database Search Process gives the capacity to an inquiry to look against all things got by the framework. The Document Database Search measure is made out of the inquiry cycle, client entered questions (normally specially appointed inquiries) and the record data set which contains all things that have been gotten, handled and put away by the framework. Regularly things in the Document Database don't change (i.e., are not altered) once got.

File Database Search When a note set in stone to be of interest, a client might need to save it for future reference. This is in actuality documenting it. In a data framework this is refined by means of MRCET-IT Page 11 the record interaction. In this interaction the client can legitimately store a thing in a record alongside extra file terms and distinct content the client needs to connect with the thing. The Index Database Search Process gives the ability to make records and search them.

There are 2 classes of index files:

- 1) Public Index files
- 2) Private Index files

Each client can have at least one Private Index documents prompting an extremely enormous number of records. Every Private Index record references just a little subset of the complete number of things in the Document Database. Public Index records are kept up with by proficient library administrations staff and regularly file each thing in the Document Database. There are few Public Index records. These records approach records (i.e., arrangements of clients and their advantages) that permit anybody to look or recover information. Private Index documents commonly have extremely restricted admittance records

### **3.6.4 Multimedia Database Search**

According to a framework point of view, the multi-media information isn't sensibly its own information structure, however an expansion to the current designs in the Information Retrieval System. Relationship to Database Management Systems From a common-sense stance, the joining of DBMS's and Information Retrieval Systems is vital. Business data set organizations have effectively coordinated the two kinds of frameworks. One of the initial business data sets to coordinate the two frameworks into a solitary view is the INQUIRE DBMS. This has been accessible for more than fifteen years. A more current model is the ORACLE DBMS that presently offers an imbedded ability called CONNECTIS, which is an educational recovery framework that utilizes a complete thesaurus which gives the premise to produce "topics" for a specific thing. The INFORMIX DBMS can connection to RetrievalWare to give reconciliation of organized information and data alongside capacities related with Information Retrieval Systems.

## **3.7 Application areas within IR**

- Cross language retrieval
- Speech/broadcast retrieval
- Text categorization
- Text summarization
- Structured document element retrieval (XML)

### **3.7.1 Cross language retrieval**

Cross-lingual Information Retrieval is the assignment of recovering important data when the report assortment is written in an alternate language from the client question.

### **3.7.2 Speech/broadcast retrieval**

Speech search is worried about the recovery of spoken substance from assortments of discourse or sight and sound information. The key difficulties raised by discourse search are ordering through a suitable interaction of discourse acknowledgment and productively getting to explicit substance components inside spoken information.

### **3.7.3 Text Categorization**

**Text categorization** (a.k.a. **text classification**) is the errand of allocating predefined classifications to free-message archives. It can give reasonable perspectives on archive assortments and has significant applications in reality. For instance, reports are commonly coordinated by subject classifications (points) or geological codes; scholastic papers are regularly grouped by specialized spaces and sub-areas; patient reports in medical services associations are frequently listed from various perspectives, utilizing scientific categorizations of sickness classes, sorts of surgeries, protection repayment codes, etc.

### **3.7.4 Text summarization**

Text summarization is the issue of making a short, exact, and familiar rundown of a more extended book document. Automatic text outline strategies are enormously expected to address the always developing measure of text information accessible online to both better assist with finding important data and to devour applicable data quicker.

### **3.7.5 Structured document element retrieval**

Structured document retrieval is Organized record recovery is worried about the recovery of archive parts. The design of the record, regardless of whether unequivocally given by an increase language or inferred, is abused to decide the most pertinent archive sections to return as answers to a given question. The recognized most applicable record pieces would themselves be able to be utilized to decide the most pertinent reports to return as answers to the given question.

## **3.8 Web Information Retrieval Models**

### **3.8.1 Vector Model**

- Index terms are assigned positive and non-binary weights
  - The index terms in the query are also weighted

$$d_j = (w_1, j, w_2, j, \dots, w_t, j) = q(w_{1,q}, w_{2,q}, \dots, w_{t,q})$$

- Term weights are used to compute the degree of similarity between documents and the user query
- Then, retrieved documents are sorted in decreasing order

#### **Advantages**

- Its term-weighting scheme improves retrieval performance
- Its partial matching strategy allows retrieval of documents that approximate the query conditions
- Its cosine ranking formula sorts the documents according to their degree of similarity to the query

#### **Disadvantage**

- The assumption of mutual independence between index terms

### **3.8.2 Vector space model**

- Vector space = all the keywords encountered
- Document D =  $\langle a_1, a_2, a_3, \dots, a_n \rangle$   $a_i$  = weight of  $t_i$  in D

$$\text{Query } Q = \langle b_1, b_2, b_3, \dots, b_n \rangle \quad b_i = \text{weight of } t_i \text{ in } Q$$

$$\bullet R(D, Q) = \text{Sim}(D, Q)$$

### **3.8. 3 Probabilistic Model**

- Introduced by Roberston and Sparck Jones, 1976 – Binary independence retrieval (BIR) model
- Idea: Given a user query  $q$ , and the ideal answer set  $R$  of the relevant documents, the problem is to specify the properties for this set
  - Assumption (probabilistic principle): the probability of relevance depends on the query and document representations only; ideal answer set  $R$  should maximize the overall probability of relevance
  - The probabilistic model tries to estimate the probability that the user will find the document  $d_j$  relevant with ratio  $P(d_j \text{ relevant to } q)/P(d_j \text{ non relevant to } q)$

### **3.9 Let us Sum Up**

- Information retrieval (IR) is the science of searching for information in documents
- Three types of Information Retrieval (IR) models- Classical IR Model , Non-Classical IR Model, Alternative IR Model
- Boolean Model — This model required information to be translated into a Boolean expression and Boolean queries
- Vector Space Model — This model takes documents and queries denoted as vectors and retrieves documents depending on how similar they are
- Probability Distribution Model — In this model, the documents are considered as distributions of terms and queries are matched based on the similarity of these representations.
- Probabilistic Models — The probabilistic model is rather simple and takes the probability ranking to display results.

### **3.10 List of References**

- .1. Stefan Buettcher, Charles L. A. Clarke, Gordon V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines*, The MIT Press, 2010.
2. Ophir Frieder “*Information Retrieval: Algorithms and Heuristics: The Information Retrieval Series*”, 2nd Edition, Springer, 2004.
3. Manu Konchady, “*Building Search Applications: Lucene, Ling Pipe*”, and First Edition, Gate Mustru Publishing, 2008.

### **4. Introduction to information retrieval - Book by Christopher D. Manning, Hinrich Schütze, and Prabhakar Raghavan**

**5. Introduction to Modern Information Retrieval - Book by Gobinda G. Chowdhury**

**3.11 Bibliography**

1. Sanjib Kumar Sahu, D. P. Mahapatra, R. C. Balabantaray, "Analytical study on intelligent information retrieval system using semantic network", *Computing Communication and Automation (ICCCA) 2016 International Conference on*, pp. 704-710, 2016.
2. Federico Bergenti, Enrico Franchi, Agostino Poggi, *Collaboration and the Semantic Web*, pp. 83, 2012.

**3. 1. Introduction to information retrieval - Book by Christopher D. Manning, Hinrich Schütze, and Prabhakar Raghavan**

**4.. Information Retrieval: Implementing and Evaluating Search Engines - Book by Charles L. A. Clarke, Gordon Cormack, and Stefan Büttcher**

**5. Introduction to Modern Information Retrieval - Book by Gobinda G. Chowdhury**

6. Stefan Buettcher, Charles L. A. Clarke, Gordon V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines*, The MIT Press, 2010.
7. Ophir Frieder "Information Retrieval: Algorithms and Heuristics: The Information Retrieval Series ", 2nd Edition, Springer, 2004.
8. Manu Konchady, "Building Search Applications: Lucene, Ling Pipe", and First Edition, Gate Mustru Publishing, 2008.

### **3.12 Chapter End Exercises**

1. Define information retrieval
2. What are the applications of IR?
3. What are the components of IR?
4. How to AI applied in IR systems?
5. Give the functions of information retrieval system.
6. List the issues in information retrieval system.
7. Discuss the impact of IR on the web
8. Define indexing & document indexing.
9. What are the three classic models in information retrieval system?
10. Explain Boolean Model.
11. Explain Vector Model
12. Why the Classic IR might lead to poor retrieval?
13. Draw the flow diagram for relevance feedback query processing model
14. Give brief notes about user Relevance feedback method and how it is used in query expansion
15. What is the use of Link analysis?

