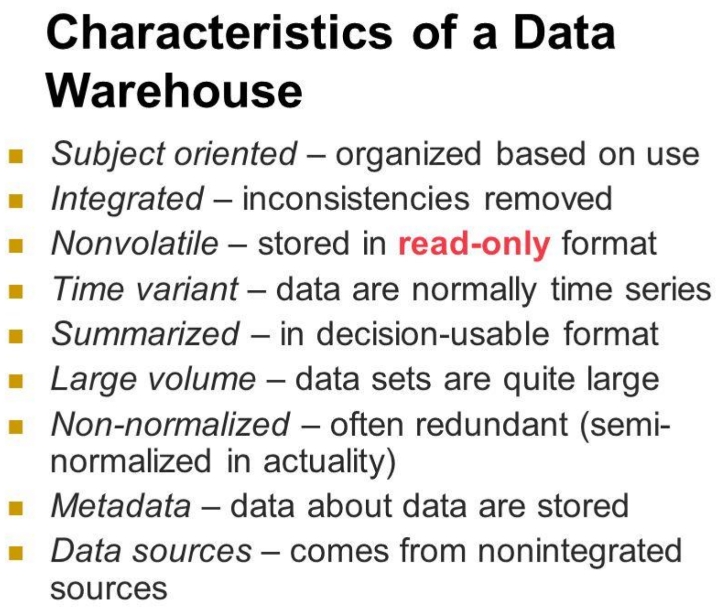
Data warehousing

Data warehousing is a method of organizing and compiling data into one database, whereas data mining deals with fetching important data from databases. Data mining attempts to depict meaningful patterns through a dependency on the data that is compiled in the data warehouse.

- A Data Warehouse (DW) is a relational database that is designed for query and analysis rather than transaction processing. It includes historical data derived from transaction data from single and multiple sources.
- A Data Warehouse provides integrated, enterprise-wide, historical data and focuses on providing support for decision-makers for data modeling and analysis.
- A Data Warehouse is a group of data specific to the entire organization, not only to a particular group of users.
- It is not used for daily operations and transaction processing but used for making decisions.

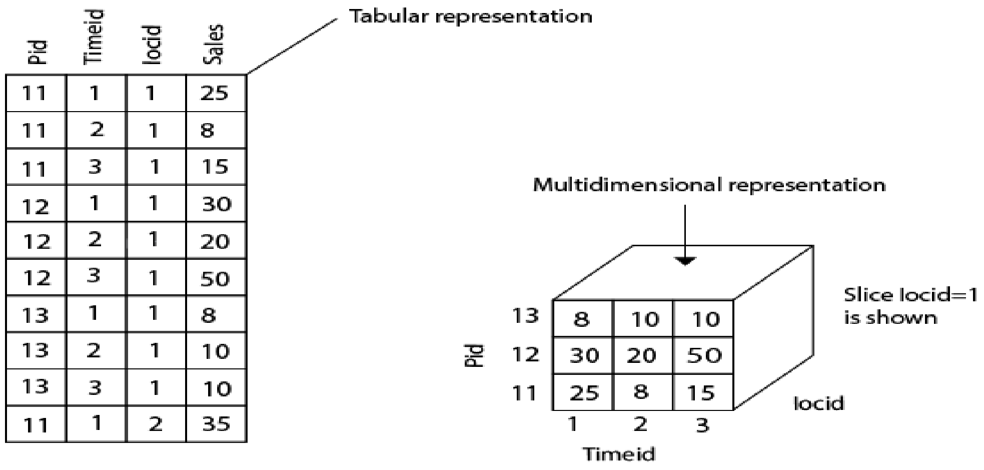


Characteristics of a Data Warehouse

- *Subject oriented* – organized based on use
- *Integrated* – inconsistencies removed
- *Nonvolatile* – stored in **read-only** format
- *Time variant* – data are normally time series
- *Summarized* – in decision-usable format
- *Large volume* – data sets are quite large
- *Non-normalized* – often redundant (semi-normalized in actuality)
- *Metadata* – data about data are stored
- *Data sources* – comes from nonintegrated sources

Multi-Dimensional Data Model?

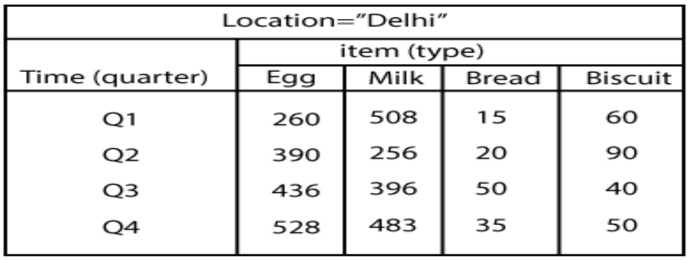
A multidimensional model views data in the form of a data-cube. A data cube enables data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.The dimensions are the perspectives or entities concerning which an organization keeps records. For example, a shop may create a sales data warehouse to keep records of the store's sales for the dimension time, item, and location. These dimensions allow the save 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



Pid	Timeid	locid	Sales
11	1	1	25
11	2	1	8
11	3	1	15
12	1	1	30
12	2	1	20
12	3	1	50
13	1	1	8
13	2	1	10
13	3	1	10
11	1	2	35

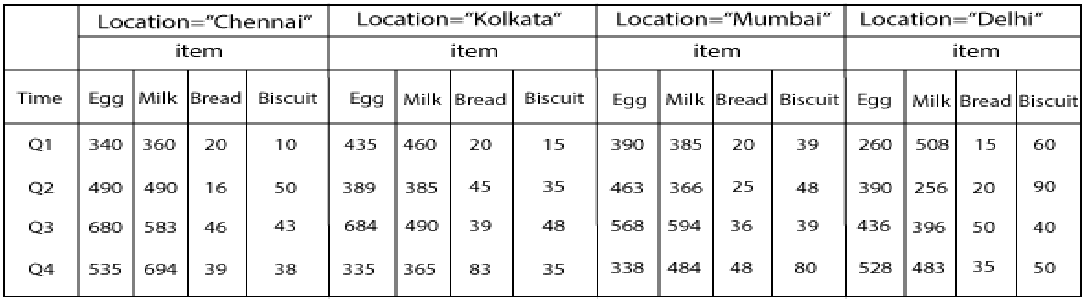
Pid \ Timeid	1	2	3
13	8	10	10
12	30	20	50
11	25	8	15

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).



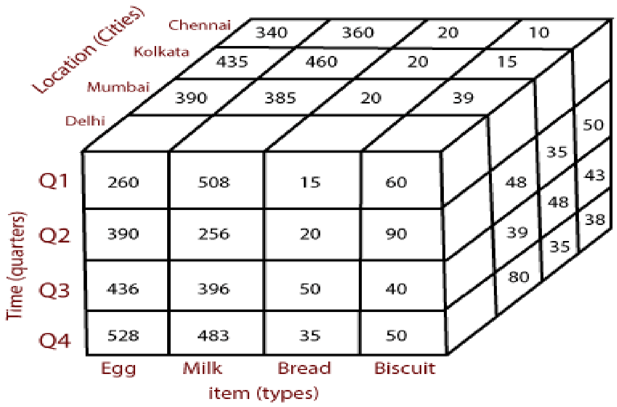
Location="Delhi"				
	item (type)			
Time (quarter)	Egg	Milk	Bread	Biscuit
Q1	260	508	15	60
Q2	390	256	20	90
Q3	436	396	50	40
Q4	528	483	35	50

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="Chennai"				Location="Kolkata"				Location="Mumbai"				Location="Delhi"			
	item				item				item				item			
Time	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit
Q1	340	360	20	10	435	460	20	15	390	385	20	39	260	508	15	60
Q2	490	490	16	50	389	385	45	35	463	366	25	48	390	256	20	90
Q3	680	583	46	43	684	490	39	48	568	594	36	39	436	396	50	40
Q4	535	694	39	38	335	365	83	35	338	484	48	80	528	483	35	50

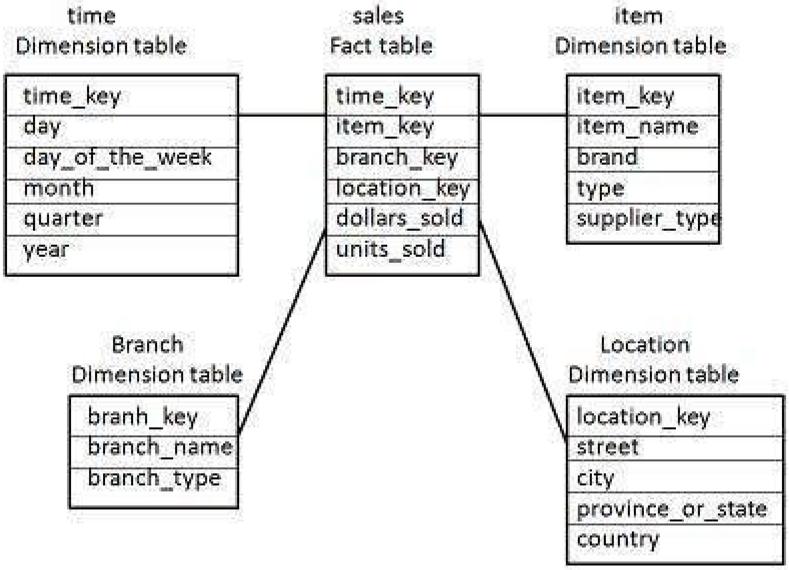
Conceptually, it may also be represented by the same data in the form of a 3D data cube, as shown in fig:



Schema is a logical description of the entire database. It includes the name and description of records of all record types including all associated data-items and aggregates. Much like a database, a data warehouse also requires to maintain a schema. A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema. In this chapter, we will discuss the schemas used in a data warehouse.

Star Schema

- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.

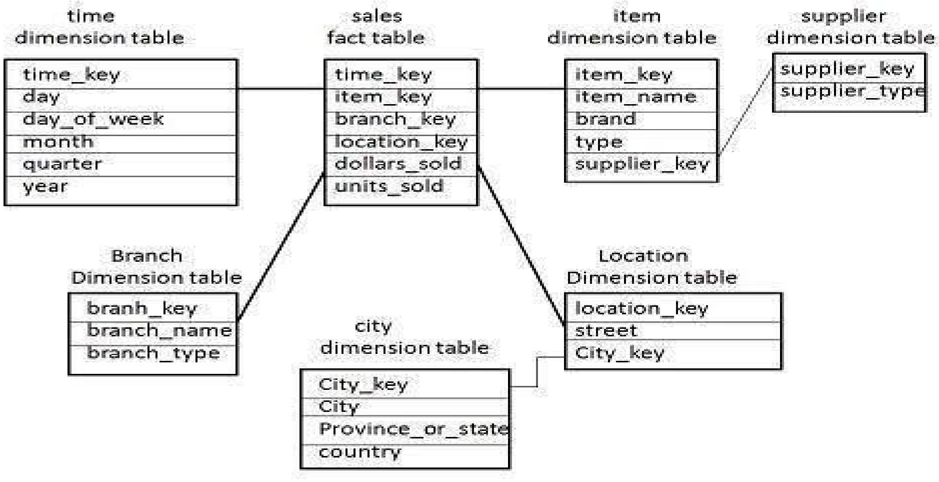


- There is a fact table at the center. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.

Note − Each dimension has only one dimension table and each table holds a set of attributes. For example, the location dimension table contains the attribute set {location_key, street, city, province_or_state,country}. This constraint may cause data redundancy. For example, "Vancouver" and "Victoria" both the cities are in the Canadian province of British Columbia. The entries for such cities may cause data redundancy along the attributes province_or_state and country.

Snowflake Schema

- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, nanamely item and supplier table.

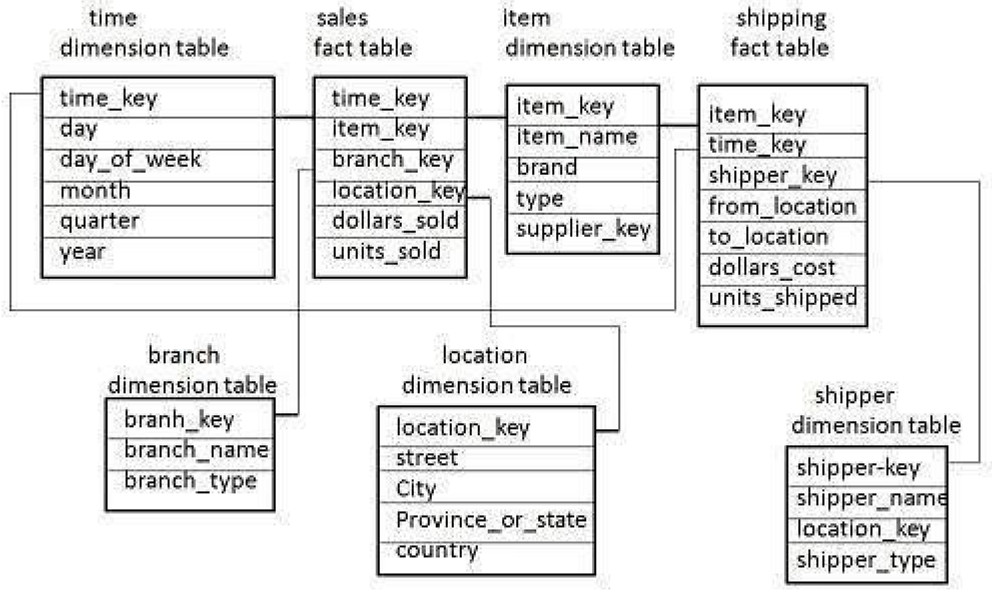


- Now the item dimension table contains the attributes item_key, item_name, type, brand, and supplier-key.
- The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.

Note − Due to normalization in the Snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.

Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping.

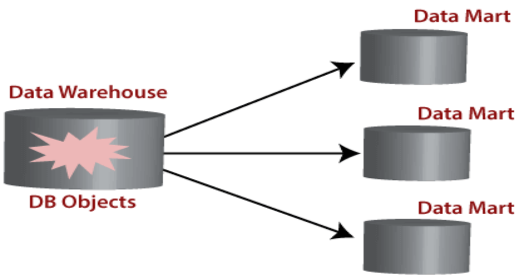


- The sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
- The shipping fact table also contains two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

What is Data Mart?

A **Data Mart** is a subset of a directorial information store, generally oriented to a specific purpose or primary data subject which may be distributed to provide business needs. Data Marts are analytical record stores designed to focus on particular business functions for a specific community within an organization. Data marts are derived from subsets of data in a data warehouse, though in the bottom-up data warehouse design methodology, the data warehouse is created from the union of organizational data marts.

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.



Reasons for creating a data mart

- Creates collective data by a group of users
- Easy access to frequently needed data
- Ease of creation

Types of Data Mart:
There are three types of data marts:
1. Dependent Data Mart – Dependent Data Mart is created by extracting the data from central repository, Datawarehouse. First data warehouse is created by extracting data (through ETL tool) from external sources and then data mart is created from data warehouse. Dependent data mart is created in top-down approach of datawarehouse architecture. This model of data mart is used by big organizations.
2. Independent Data Mart – Independent Data Mart is created directly from external sources instead of data warehouse. First data mart is created by extracting data from external sources and then datawarehouse is created from the data present in data mart. Independent data mart is designed in bottom-up approach of datawarehouse architecture. This model of data mart is used by small organizations and is cost effective comparatively.
3. Hybrid Data Mart – This type of Data Mart is created by extracting data from operational source or from data warehouse. 1Path reflects accessing data directly from external sources and 2Path reflects dependent data model of data mart.
Advantages of Data Mart:

Implementation of data mart needs less time as compared to implementation of datawarehouse as data mart is designed for a particular department of an organization.

Organizations are provided with choices to choose model of data mart depending upon cost and their business.

Data can be easily accessed from data mart.

It contains frequently accessed queries, so enable to analyse business trend.

Disadvantages of Data Mart:

Since it stores the data related only to specific function, so does not store huge volume of data related to each and every department of an organization like datawarehouse.

Creating too many data marts becomes cumbersome sometimes.

The steps for implementing data mart:

Designing

Designing is the first step in implementing data mart. Since, data mart stores data related to a particular topic, so this step includes identification of a subject or a topic related to which data mart will store data. Also it includes the sources to gather the information related to the subject and then designing logical and physical structures of data mart.

Building

Building is the second phase in implementing data mart. It includes building physical and logical structure of data mart which is designed in the first step. Physical Structure means constructing database so that data can be easily accessed from it and logical structure means outer schema.

Populating

Populating phase includes putting data into the data mart. Before putting data into data warehouse, there is a need to extract the data from the sources, to clean it and convert it into the correct format and then put the corrected data into the data mart. These steps are needed to perform so that data stored in the data mart is appropriate.

Accessing

Now the data mart is ready with its data. This is the time to access data from it by making requests related to query occurred. We can access data from data mart either through command line or GUI platform. Making query through GUI based platform is user friendly and used by many organisations comparatively.

Metadata

o Metadata are data about data

o When used in a data warehouse, metadata are the data that define warehouse objects

o Metadata are created for the data names and definitions of the given warehouse

o Additional metadata are created and captured for timestamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes

o For example, metadata are used as a directory to help the decision support system analyst locate the contents of the data warehouse, and as a guide to the data mapping when data are transformed from the operational environment to the data warehouse environment

o Metadata also serve as a guide to the algorithms used for summarization between the of current detailed data and the lightly summarized data, and between the lightly summarized data and the highly summarized data

Metadata should be stored and managed persistently (i.e., on disk)

<u>Types of Data Mining Models</u> –

Predictive Models

Descriptive Models

A predictive model constitutes prediction concern values of data using known results found from various data. Predictive modelling may be made based on the use of variant historical data. Predictive model data mining tasks comprise regression, time series analysis, classification, prediction.

The Predictive Model is known as Statistical Regression. It is a monitoring learning technique that Incorporates an explication of the dependency of few attribute values upon the values of other attributes In a similar item and the growth of a model that can predict these attribute values for recent cases.

Classification –

It is the act of assigning objects to one of several predefined categories. Or we can define classification as a learning function of a target function that sets each attribute to a predefined class label.

Regression –

It is used for appropriate data. It is a technique that verifies data values for a function. There are two types of regression –

1. Linear Regression is associated with the search for the optimal line to fit the two attributes so that one attribute can be applied to predict the other.

2. Multi-Linear Regression involves two or more than two attributes and data are fit to multidimensional space.

Time Series Analysis –

It is a set of data based on time. Time series analysis serves as an independent variable to estimate the dependent variable in time.

Prediction –

It predicts some missing or unknown values.

Description Model :

A descriptive model distinguishes relationships or patterns in data. Unlike Predictive Model, a descriptive model serves as a way to explore the properties of data being examined, not to predict new properties, clustering, summarization, associating rules, and sequence discovery are descriptive model data mining tasks.

Descriptive analytics Concentrate on the summarization and conversion of the data into significant information for monitoring and reporting.

Clustering –
It is the technique of converting a group of abstract objects into classes of identical objects.
Summarization –
It holds a set of data in a more in-depth, easy-to-understand form.
Associative Rules –
They find an exciting consistency or causal relationship between a large set of data objects.
Sequence –
It is the discovery of interesting patterns in the data is in relation to some objective or subjective measurement of how interesting it is.

Data warehouse maintenance systems must provide means to keep track of schema modifications as well as of instance modifications. On the schema level one needs operations for the Insertion, Deletion and Change of dimensions and categories. Category changes are for instance adding or deleting user defined attributes.

Nature Of Data

a data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization. An operational database undergoes frequent changes on a daily basis on account of the transactions that take place.

Security Of Data

At the warehouse stage, more groups than just the centralized data team will commonly have access. You must use data governance to safeguard certain pieces of sensitive information from being accessed by the wrong people in your organization. Many security regulations mandating data access rules have been passed, such as GDPR, and many companies have industry standard compliance rules that they adhere to as well, like SOC and HIPAA.

OnlineAnalyticalProcessing(OLAP)

o Data warehouses provide online analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data generalization and data mining

o Data warehouse systems serve users or knowledge workers in the role of data analysis and decision making

o Such systems can organize and present data in various formats in order to accommodate the diverse needs of different users are known as online analytical processing (OLAP) systems

Online Transaction Processing (OLTP)

Online operational database systems is to perform online transaction and a query processing are called online transaction processing (OLTP) systems

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

· The major distinguishing features of OLTP and OLAP are summarized as follows

Users and system orientation

An OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals

§ An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts

Data contents

An OLTP system manages current data that, typically, are too detailed to be easily used for decision making

An OLAP system manages large amounts of historic data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity

Database design

An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design

An OLAP system typically adopts either a star or a snowflake model and a subject-oriented database design View

An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historic data or data in different organizations

An OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization

OLAP systems also deal with information that originates from different organizations, integrating information from many data stores

Because of their huge volume, OLAP data are stored on multiple storage media

Access patterns

The access patterns of an OLTP system consist mainly of short, atomic transactions o

Such a system requires concurrency control and recovery mechanisms

However, accesses to OLAP systems are mostly read-only operations (because most data warehouses store historic rather than up-to-date information), although many could be complex queries

·Other features that distinguish between OLTP and OLAP systems include database size, frequency of operations, and performance metrics

Data modeling

Data modeling is the process of creating a simplified diagram of a software system and the data elements it contains, using text and symbols to represent the data and how it flows. Data models provide a blueprint for designing a new database or reengineering a legacy application.

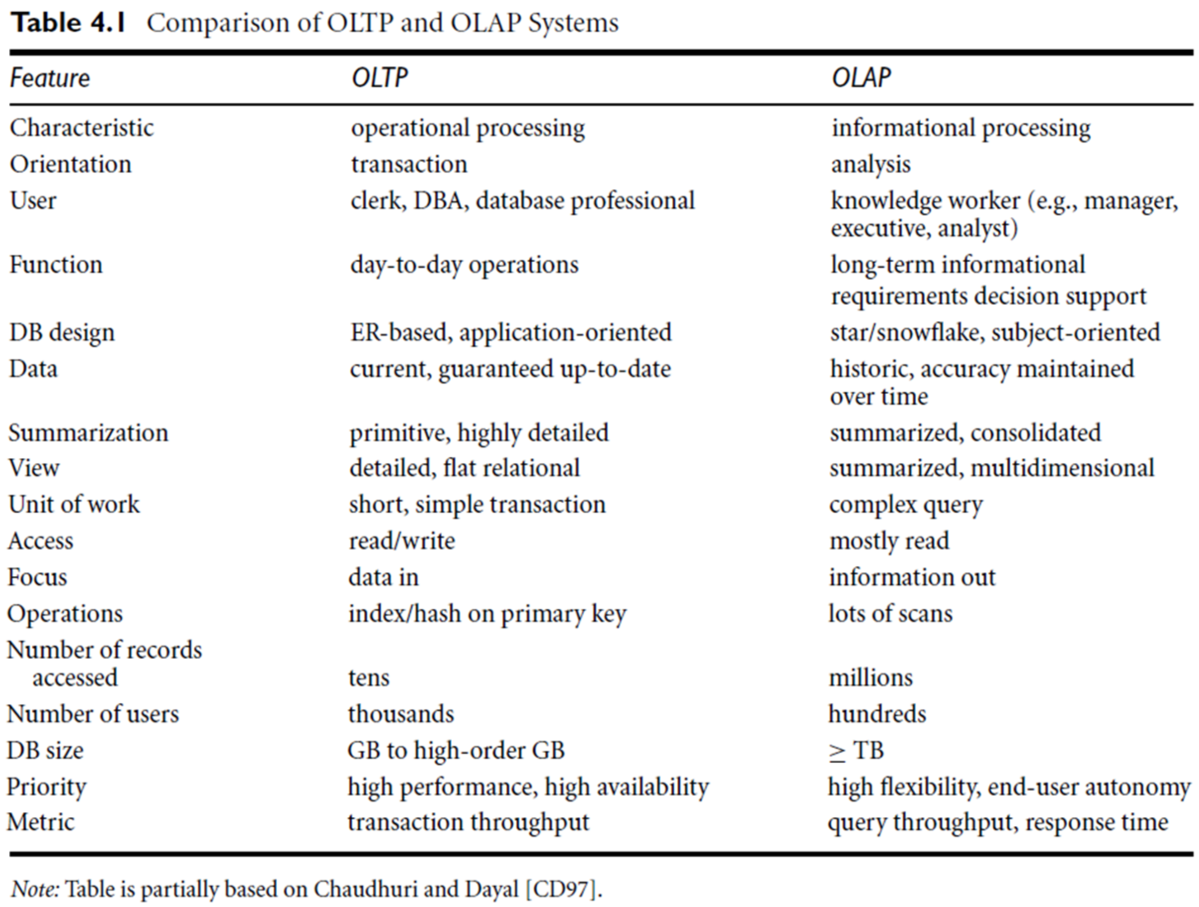


Table 4.1 Comparison of OLTP and OLAP Systems

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

Note: Table is partially based on Chaudhuri and Dayal [CD97].

· **DataWarehousing:AMultitiered Architecture**

A three-tier architecture

Bottom tier: Data warehouse server

§ The bottom tier is a warehouse database server that is almost always a relational database system

§ Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (e.g., customer profile information provided by external consultants)

§ These tools and utilities perform data extraction, cleaning, and transformation (e.g., to merge similar data from different sources into a unified format), as well as load and refresh functions to update the data warehouse

§ The data are extracted using application program interfaces known as gateways

§ A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server

· Data Warehousing: A Multitiered Architecture

o Bottom tier: Data warehouse server

Examples of gateways include ODBC (Open Database Connection) and OLEDB (Object Linking and Embedding

§ Database) by Microsoft and JDBC (Java Database Connection)

§ This tier also contains a metadata repository, which stores information about the data warehouse and its contents

· The middle tier: OLAP server

§ Implemented using either
§ A relational OLAP(ROLAP) model (i.e., an extended relational DBMS that maps operations on multidimensional data to standard relational operations)
§ A multidimensional OLAP (MOLAP) model (i.e., a special-purpose server that directly implements multidimensional data and operations)
· Top tier: Front-end tools
· Contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on)
· Data Warehouse Models
o From the architecture point of view, there are three data warehouse models
§ The enterprise warehouse
§ The data mart
§ The virtual warehouse

Enterprise warehouse
§ Collects all of the information about subjects spanning the entire organization
§ Provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope
§ Typically contains detailed data as well as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond
§ It may be implemented on traditional mainframes, computer super servers, or parallel architecture platforms.
§ Requires extensive business modeling and may take years to design and build
Data Warehouse Models
Data mart
§ Contains a subset of corporate-wide data that is of value to a specific group of users
§ The scope is confined to specific selected subjects
§ For example, a marketing data mart may confine its subjects to customer, item, and sales.
§ The data contained in data marts tend to be summarized
§ Usually implemented on low-cost departmental servers that are Unix/Linux or Windows based
§ The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years
§ Depending on the source of data, data marts can be categorized as independent or dependent
§ Independent data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area
§ Dependent data marts are sourced directly from enterprise data warehouses

Data Warehouse Models
o Virtual warehouse

§ A set of views over operational databases
§ For efficient query processing, only some of the possible summary views may be materialized
§ Easy to build but requires excess capacity on operational database servers
·The top-down approach to data warehouse development
§ Serves as a systematic solution and minimizes integration problems
§ Expensive, takes a long time to develop, and lacks flexibility due to the difficulty in achieving consistency and consensus for a common data model for the entire organization
·The bottom-up approach to data warehouse development
§ The bottom up approach to the design, development, and deployment of independent data marts provides flexibility, low cost, and rapid return of investment
§ Lead to problems when integrating various disparate data marts into a consistent enterprise data warehouse
· **Development of Data Warehousing**

A recommended method for the development of data warehouse systems is to implement the warehouse in an incremental and evolutionary manner, as shown in Figure 4.2.

First, a high-level corporate data model is defined within a reasonably short period (such as one or two months) that provides a corporate-wide, consistent
integrated view of data among different subjects and potential usages

This high-level model, although it will need to be refined in the further development of enterprise data warehouses and departmental data marts, will greatly reduce future integration problems
o Second, independent data marts can be implemented in parallel with the enterprise warehouse based on the same corporate data model set noted before

Development of Data Warehousing

o Third, distributed data marts can be constructed to integrate different data marts via hub servers
o Finally, a multitier data warehouse is constructed where the enterprise warehouse is the sole custodian of all warehouse data, which is then distributed to the various dependent data marts

Extraction, Transformation, and Loading

o Data warehouse systems use back-end tools and utilities to populate and refresh their data (Figure 4.1)
o These tools and utilities include the following functions:
§ Data extraction
§ which typically gathers data from multiple, heterogeneous, and external sources
· Data cleaning
§ which detects errors in the data and rectifies them when possible
· Data transformation
§ which converts data from legacy or host format to warehouse format
· Load

§ which sorts, summarizes, consolidates, computes views, checks integrity, and, builds indices and partitions

Refresh

§ which propagates the updates from the data sources to the warehouse

Metadata Repository

o Metadata are data about data

o When used in a data warehouse, metadata are the data that define warehouse objects

o Figure 4.1 showed a metadata repository within the bottom tier of the data warehousing architecture

o Metadata are created for the data names and definitions of the given warehouse Additional metadata are created and captured for timestamping any extracted data, the source of theextracted data, and missing fields that have been added by data cleaning or integration processes

o A metadata repository should contain the following:

§ A description of the data warehouse structure, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents

· **Metadata Repository**

o A metadata repository should contain the following:

§ Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails)

§ The algorithms used for summarization, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports

§ Mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control)

·

DATA MINING

Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called ***Knowledge Discovery in Database (KDD)***. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.

Types of Data Mining

Relational Database:

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization.

Data warehouses:

A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision- making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.

Data Repositories:

The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.

Object-Relational Database:

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc.

Transactional Database:

A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

Advantages of Data Mining

- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making process of an organization.
- It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Disadvantages of Data Mining

- There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.
- Many data mining analytics software is difficult to operate and needs advance training to work on.
- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

Application of Data Mining

Data Mining in Healthcare:
Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will

enhance health care services and reduce costs. Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Data mining also enables healthcare insurers to recognize fraud and abuse.

Data Mining in Market Basket Analysis:

Market basket analysis is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behavior of a buyer. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly. Using a different analytical comparison of results between various stores, between customers in different demographic groups can be done.

Data mining in Education:

Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments. EDM objectives are recognized as affirming student's future learning behavior, studying the impact of educational support, and promoting learning science. An organization can use data mining to make precise decisions and also to predict the results of the student. With the results, the institution can concentrate on what to teach and how to teach.

Data Mining in Manufacturing Engineering:

Knowledge is the best asset possessed by a manufacturing company. Data mining tools can be beneficial to find patterns in a complex manufacturing process. Data mining can be used in system-level designing to obtain the relationships between product architecture, product portfolio, and data needs of the customers. It can also be used to forecast the product development period, cost, and expectations among the other tasks.

Data Mining in CRM (Customer Relationship Management):

Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies. To get a decent relationship with the customer, a business organization needs to collect data and analyze the data. With data mining technologies, the collected data can be used for analytics.

Data Mining in Fraud detection:

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. Data mining provides meaningful patterns and turning data into information. An ideal fraud detection system should protect the data of all the users. Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent. A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

Data Mining Techniques

Data mining includes the utilization of refined data analysis tools to find previously unknown, valid patterns and relationships in huge data sets. These tools can incorporate statistical models, machine learning techniques, and mathematical algorithms, such as neural networks or decision trees. Thus, data mining incorporates analysis and prediction.

In recent data mining projects, various major data mining techniques have been developed and used, including association, classification, clustering, prediction, sequential patterns, and regression.

1. Classification:

This technique is used to obtain important and relevant information about data and metadata. This data mining technique helps to classify data in different classes.

1. Classification of Data mining frameworks as per the type of data sources mined:
   This classification is as per the type of data handled. For example, multimedia, spatial data, text data, time-series data, World Wide Web, and so on..
2. Classification of data mining frameworks as per the database involved:
   This classification based on the data model involved. For example. Object-oriented database, transactional database, relational database, and so on..
3. Classification of data mining frameworks as per the kind of knowledge discovered:
   This classification depends on the types of knowledge discovered or data mining functionalities. For example, discrimination, classification, clustering, characterization, etc. some frameworks tend to be extensive frameworks offering a few data mining functionalities together..

2. Clustering:

Clustering is a division of information into groups of connected objects. Describing the data by a few clusters mainly loses certain confine details, but accomplishes improvement. It models data by its clusters. Data modeling puts clustering from a historical point of view rooted in statistics, mathematics, and numerical analysis. From a machine learning point of view, clusters relate to hidden patterns, the search for clusters is unsupervised learning, and the subsequent framework represents a data concept. From a practical point of view, clustering plays an extraordinary job in data mining applications. For example, scientific data exploration, text mining, information retrieval, spatial database applications, CRM, Web analysis, computational biology, medical diagnostics, and much more.

3. Regression:

Regression analysis is the data mining process is used to identify and analyze the relationship between variables because of the presence of the other factor. It is used to define the probability of the specific variable. Regression, primarily a form of planning and modeling.

4. Association Rules:

This data mining technique helps to discover a link between two or more items. It finds a hidden pattern in the data set.

Association rules are if-then statements that support to show the probability of interactions between data items within large data sets in different types of databases. Association rule mining has several applications and is commonly used to help sales correlations in data or medical data sets.

5. Outer detection:

This type of data mining technique relates to the observation of data items in the data set, which do not match an expected pattern or expected behavior. This technique may be used in various domains like intrusion, detection, fraud detection, etc. It is also known as Outlier Analysis or Outilier mining.

KDD VS DATA MINING

KDD is the overall process of extracting knowledge from data, while Data Mining is a step inside the KDD process, which deals with identifying patterns in data.And Data Mining is only the application of a specific algorithm based on the overall goal of the

KDD process.KDD is an iterative process where evaluation measures can be enhanced, mining can be refined, and new data can be integrated and transformed to get different and more appropriate results.

DBMS VS DATA MINING

DBMS is a full-fledged system for housing and managing a set of digital databases. However Data Mining is a technique or a concept in computer science, which deals with extracting useful and previously unknown information from raw data. Most of the times, these raw data are stored in very large databases

Issue and Challanges in DATA mining

1. Security and Social Challenges

Dynamic techniques are done through data assortment sharing, so it requires impressive security. Private information about people and touchy information is gathered for the client's profiles, client standard of conduct understanding—illicit admittance to information and the secret idea of information turning into a significant issue.

2. Noisy and Incomplete Data

Data Mining is the way toward obtaining information from huge volumes of data. This present reality information is noisy, incomplete, and heterogeneous. Data in huge amounts regularly will be unreliable or inaccurate. These issues could be because of human mistakes blunders or errors in the instruments that measure the data.

3. Distributed Data

True data is normally put away on various stages in distributed processing conditions. It very well may be on the internet, individual systems, or even on the databases. It is essentially hard to carry all the data to a unified data archive principally because of technical and organizational reasons.

4. Complex Data

True data is truly heterogeneous, and it very well may be media data, including natural language text, time series, spatial data, temporal data, complex data, audio or video, images, etc. It is truly hard to deal with these various types of data and concentrate on the necessary information. More often than not, new apparatuses and systems would need to be created to separate important information.

5. Performance

The presentation of the data mining framework basically relies upon the productivity of techniques and algorithms utilized. On the off chance that the techniques and algorithms planned are not sufficient; at that point, it will influence the presentation of the data mining measure unfavorably.

6. Scalability and Efficiency of the Algorithms

The Data Mining algorithm should be scalable and efficient to extricate information from tremendous measures of data in the data set.

7. Improvement of Mining Algorithms

Factors, for example, the difficulty of data mining approaches, the enormous size of the database, and the entire data flow inspire the distribution and creation of parallel data mining algorithms.

8. Incorporation of Background Knowledge

In the event that background knowledge can be consolidated, more accurate and reliable data mining arrangements can be found. Predictive tasks can make more accurate predictions, while descriptive tasks can come up with more useful findings. Be that as it may, gathering and including foundation knowledge is an unpredictable cycle.

9. Data Visualization

Data visualization is a vital cycle in data mining since it is the foremost interaction that shows the output in a respectable way to the client. The information extricated ought to pass on the specific significance of what it really plans to pass on. However, ordinarily, it is truly hard to address the information in a precise and straightforward manner to the end-user. The output information and input data being very effective, successful, and complex data perception methods should be applied to make it fruitful.

Banking Industry

In the banking industry, concentration is given to risk management and policy reversal as well analyzing consumer data, market trends, government regulations and reports, and more importantly financial decision making.Most banks also use warehouses to manage the resources available on deck in an effective manner. Certain banking sectors utilize them for market research, performance analysis of each product, interchange and exchange rates, and to develop marketing programs.

Finance Industry

Similar to the applications seen in banking, mainly revolve around evaluation and trends of customer expenses which aids in maximizing the profits earned by their clients.

Consumer Goods Industry

They are used for prediction of consumer trends, inventory management, market and advertising research. In-depth analysis of sales and production is also carried out. Apart from these, information is exchanged business partners and clientele.

Healthcare

One of the most important sector which utilizes data warehouses is the Healthcare sector. All of their financial, clinical, and employee records are fed to warehouses as it helps them to strategize and predict outcomes, track and analyze their service feedback, generate patient reports, share data with tie-in insurance companies, medical aid services, etc.

Hospitality Industry

A major proportion of this industry is dominated by hotel and restaurant services, car rental services, and holiday home services. They utilize warehouse services to design and evaluate their advertising and promotion campaigns where they target customers based on their feedback and travel patterns.

Insurance

As the saying goes in the insurance services sector, “Insurance can never be bought, it can be only be sold”, the warehouses are primarily used to analyze data patterns and customer trends, apart from maintaining records of already existing participants. The design of tailor-made customer offers and promotions is also possible through warehouses.

Manufacturing and Distribution Industry

This industry is one of the most important sources of income for any state. A manufacturing organization has to take several make-or-buy decisions which can influence the future of the sector, which is why they utilize high-end OLAP tools as a part of data warehouses to predict market changes, analyze current business trends, detect warning conditions, view marketing developments, and ultimately take better decisions.

ASSOCIATION RULE

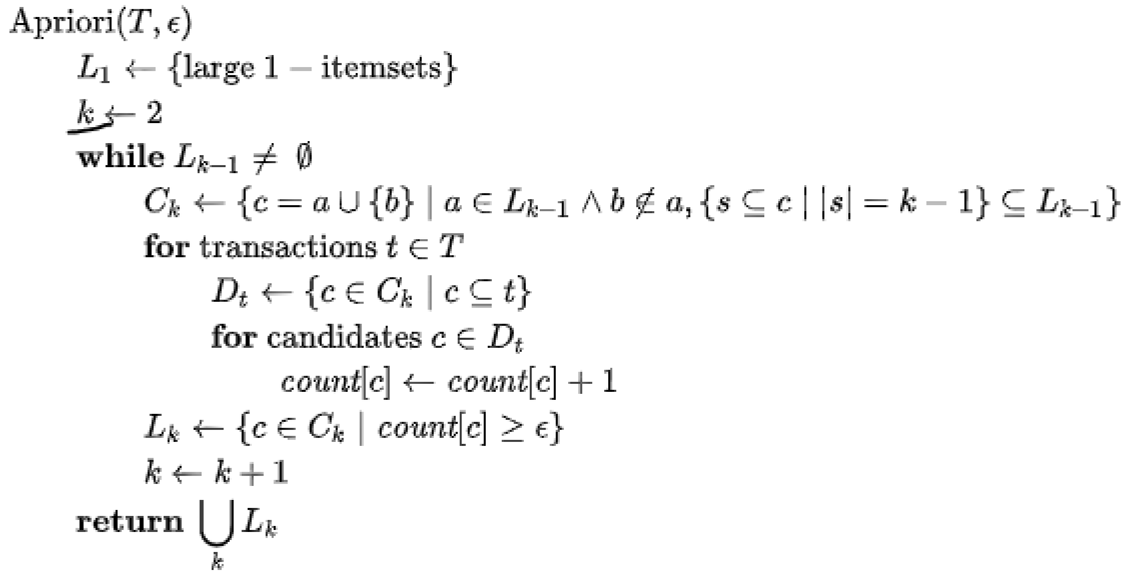
Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is Market Based Analysis.Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.

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.

APRIORI ALGORITHM

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.

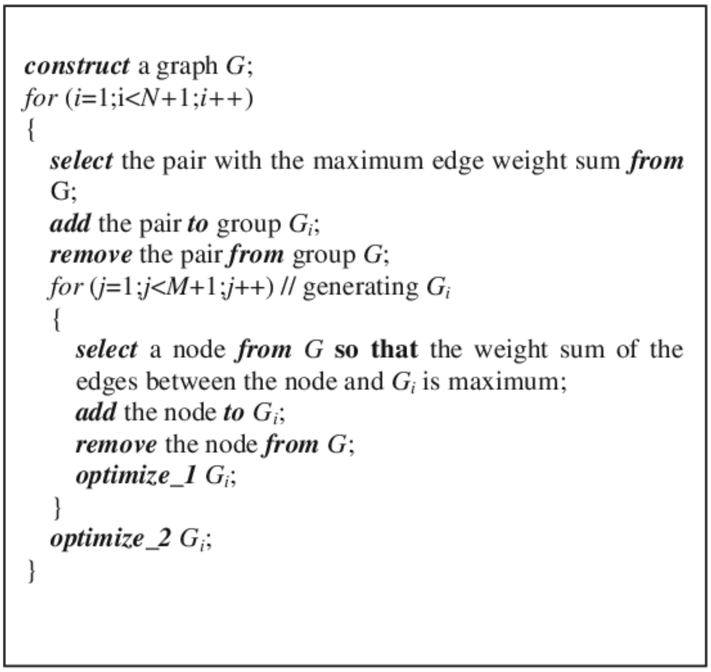
To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called Apriori property which helps by reducing the search space.



```
$\text{Apriori}(T, \epsilon)$
    $L_1 \leftarrow \{\text{large } 1 - \text{itemsets}\}$
    $k \leftarrow 2$
    while $L_{k-1} \neq \emptyset$
        $C_k \leftarrow \{c = a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a, \{s \subseteq c \mid |s| = k-1\} \subseteq L_{k-1}\}$
        for transactions $t \in T$
            $D_t \leftarrow \{c \in C_k \mid c \subseteq t\}$
            for candidates $c \in D_t$
                $count[c] \leftarrow count[c] + 1$
        $L_k \leftarrow \{c \in C_k \mid count[c] \geq \epsilon\}$
        $k \leftarrow k + 1$
    return $\bigcup_k L_k$
```

Partition Algorithm

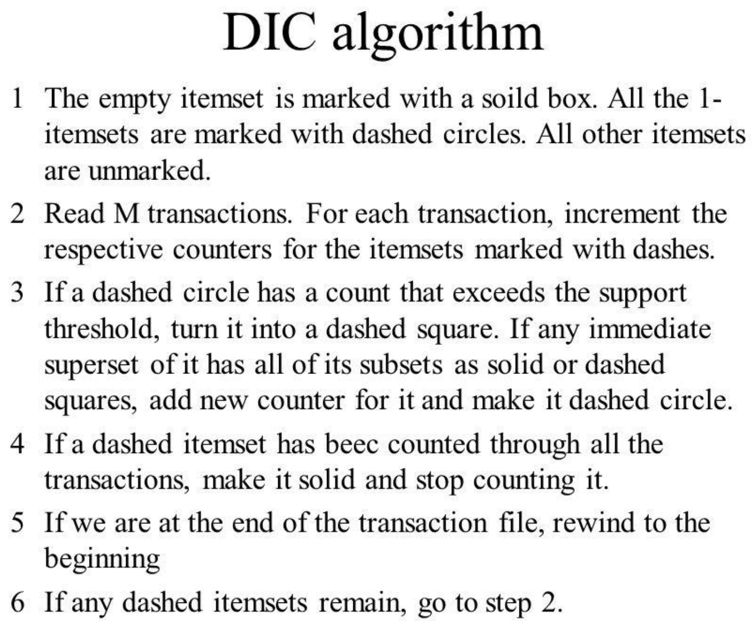
The logic is simple, we start from the leftmost element and keep track of index of smaller (or equal to) elements as i. While traversing, if we find a smaller element, we swap current element with arr[i]. Otherwise we ignore current element.



```
construct a graph $G$;
for ($i$=1;i<$N$+1;$i$++)
{
   select the pair with the maximum edge weight sum from
   G;
   add the pair to group $G_i$;
   remove the pair from group $G$;
   for ($j$=1;$j$<$M$+1;$j$++) // generating $G_i$
   {
      select a node from $G$ so that the weight sum of the
      edges between the node and $G_i$ is maximum;
      add the node to $G_i$;
      remove the node from $G$;
      optimize_1 $G_i$;
   }
   optimize_2 $G_i$;
}
```

DYANAMIC ITEMSET

Itemsets are marked in four different ways as they are counted: Solid box: confirmed frequent itemset - an itemset we have finished counting and exceeds the support threshold minsupp. Solid circle: confirmed infrequent itemset - we have finished counting and it is below minsupp.



DIC algorithm

1 The empty itemset is marked with a soild box. All the 1-itemsets are marked with dashed circles. All other itemsets are unmarked.

2 Read M transactions. For each transaction, increment the respective counters for the itemsets marked with dashes.

3 If a dashed circle has a count that exceeds the support threshold, turn it into a dashed square. If any immediate superset of it has all of its subsets as solid or dashed squares, add new counter for it and make it dashed circle.

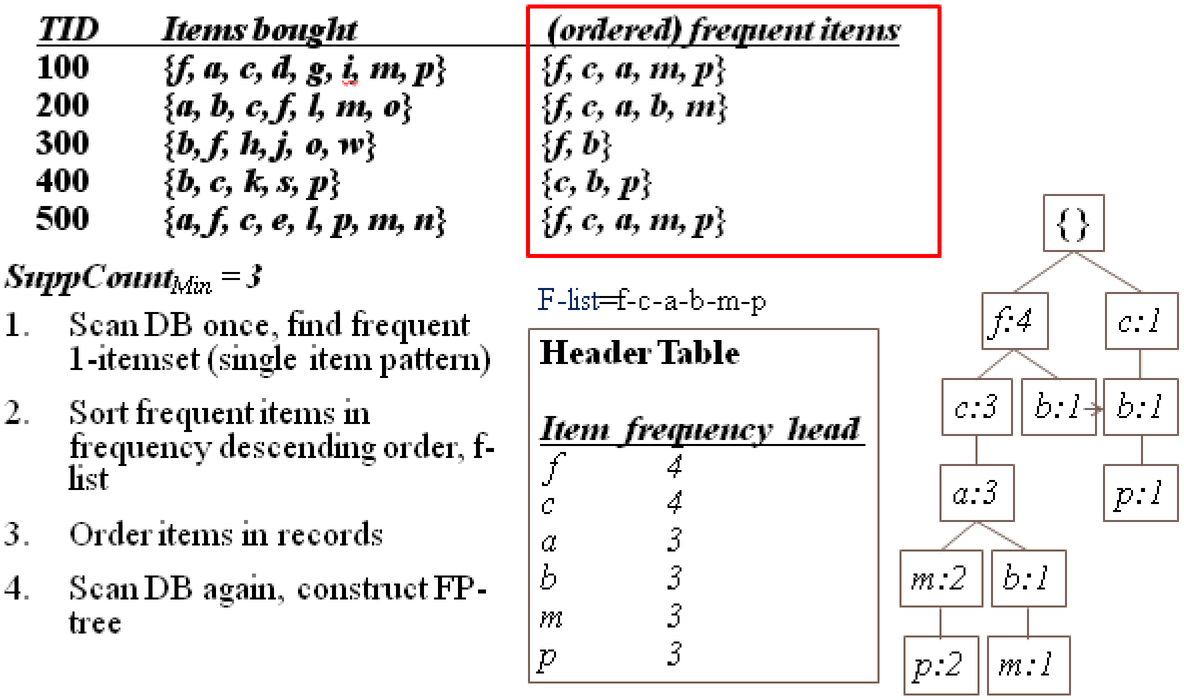
4 If a dashed itemset has beec counted through all the transactions, make it solid and stop counting it.

5 If we are at the end of the transaction file, rewind to the beginning

6 If any dashed itemsets remain, go to step 2.

FP GROWTH ALGO

FP-growth is an improved version of the Apriori Algorithm which is widely used for frequent pattern mining(AKA Association Rule Mining). It is used as an analytical process that finds frequent patterns or associations from data sets.



TID	Items bought	(ordered) frequent items
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

$SuppCount_{Min} = 3$

1. Scan DB once, find frequent 1-itemset (single item pattern)
2. Sort frequent items in frequency descending order, f-list
3. Order items in records
4. Scan DB again, construct FP-tree

F-list=f-c-a-b-m-p

Header Table

Item	frequency	head
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	

Generalized association rule
Generalized association rule extraction [3] is a widely used exploratory technique that allows discovering hidden correlations among data. By evaluating a taxonomy (is-a hierarchy) over data items, items can be aggregated according to different granularity levels. The aggregated concepts are called generalized items.
for example, the jacket, coat, mittens, and hat items. Outerwear might be their corresponding generalized item. Thus, generalized items and itemsets provide a high level view of the patterns hidden in the analyzed data. They have been profitably exploited in different application domains (e.g., market-basket analysis [3], [23], network traffic domain [4]) to provide a high level abstraction of the mined knowledge.

CLUSTURING TECHNIQUE

Clustering is an undirected technique used in data mining for identifying several hidden patterns in the data without coming up with any specific hypothesis. The reason behind using clustering is to identify similarities between certain objects and make a group of similar ones.

Points to Remember :

One group is treated as a cluster of data objects

- In the process of cluster analysis, the first step is to partition the set of data into groups with the help of data similarity, and then groups are assigned to their respective labels.
- The biggest advantage of clustering over-classification is it can adapt to the changes made and helps single out useful features that differentiate different groups.

Applications of cluster analysis :

- It is widely used in many applications such as image processing, data analysis, and pattern recognition.
- It helps marketers to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.
- It can be used in the field of biology, by deriving animal and plant taxonomies, identifying genes with the same capabilities.
- It also helps in information discovery by classifying documents on the web.

Requirement of clustering in data mining :

The following are some points why clustering is important in data mining.

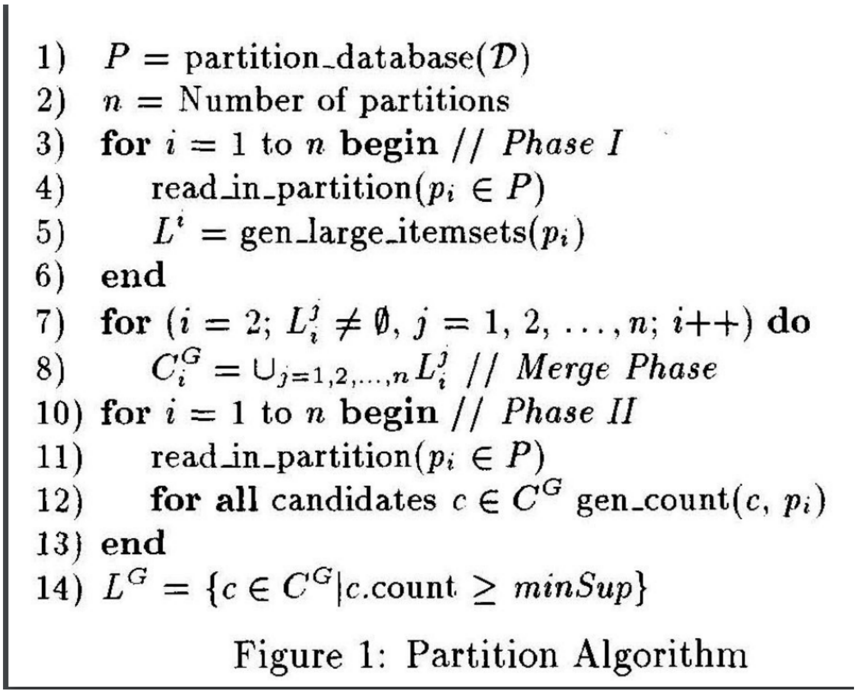
- **Scalability –**
  we require highly scalable clustering algorithms to work with large databases.
- **Ability to deal with different kinds of attributes –**
  Algorithms should be able to work with the type of data such as categorical, numerical, and binary data.
- **Discovery of clusters with attribute shape –**
  The algorithm should be able to detect clusters in arbitrary shape and it should not be bounded to distance measures.
- **Interpretability –**
  The results should be comprehensive, usable, and interpretable.

- **High dimensionality –**
  The algorithm should be able to handle high dimensional space instead of only handling low dimensional data.

Partitioning Method:

- This clustering method classifies the information into multiple groups based on the characteristics and similarity of the data. Its the data analysts to specify the number of clusters that has to be generated for the clustering methods.
- In the partitioning method when database(D) that contains multiple(N) objects then the partitioning method constructs user-specified(K) partitions of the data in which each partition represents a cluster and a particular region. There are many algorithms that come under partitioning method some of the popular ones are K-Mean, PAM(K-Mediods), CLARA algorithm (Clustering Large Applications) etc.

The Partition Algorithm executes in two phases: ◃ Phase I: the algorithm logically divides the database into a number of non-overlapping partitions. The partitions are considered one at a time and all large itemsets for that partition are generated. At the end of phase I, these large itemsets are merged to generate a set of all potentially large itemsets. ◃ Phase II: the actual supports for these itemsets are generated and the large itemsets are identified.

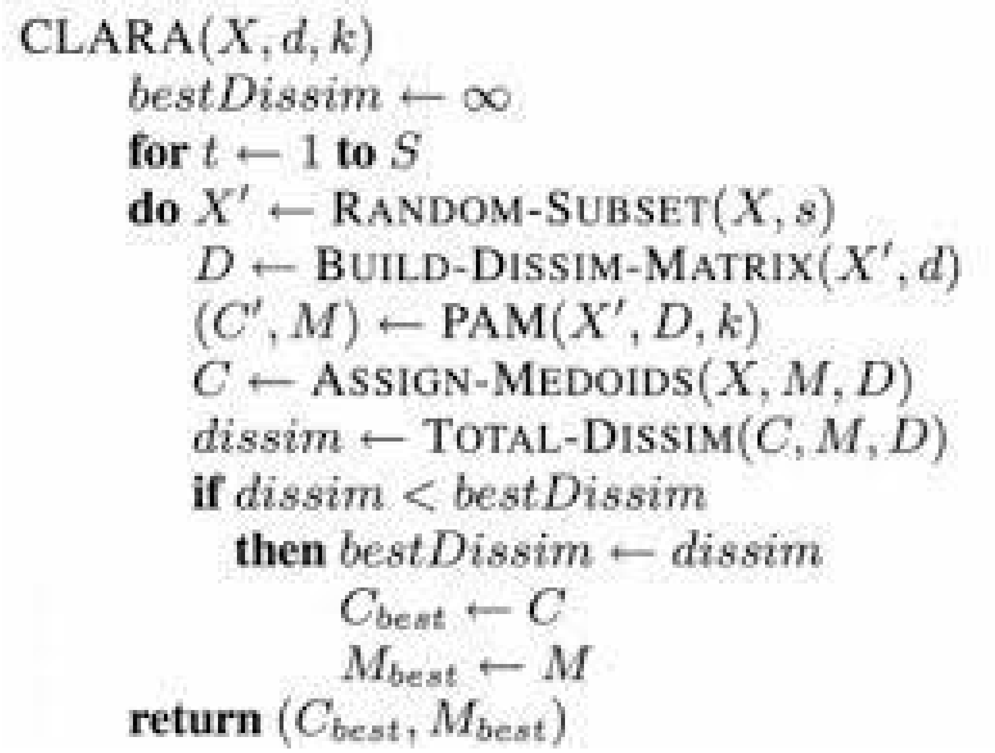


```
$P$ = partition_database($\mathcal{D}$)
$n$ = Number of partitions
for $i = 1$ to $n$ begin // Phase I
    read_in_partition($p_i \in P$)
    $L^i$ = gen_large_itemsets($p_i$)
end
for ($i = 2$; $L_i^j \neq \emptyset$, $j = 1, 2, \ldots, n$; $i$++) do
    $C_i^G = \cup_{j=1,2,\ldots,n} L_i^j$ // Merge Phase
for $i = 1$ to $n$ begin // Phase II
    read_in_partition($p_i \in P$)
    for all candidates $c \in C^G$ gen_count($c$, $p_i$)
end
$L^G = \{c \in C^G | c.\text{count} \geq minSup\}$
```

Figure 1: Partition Algorithm

CLARA

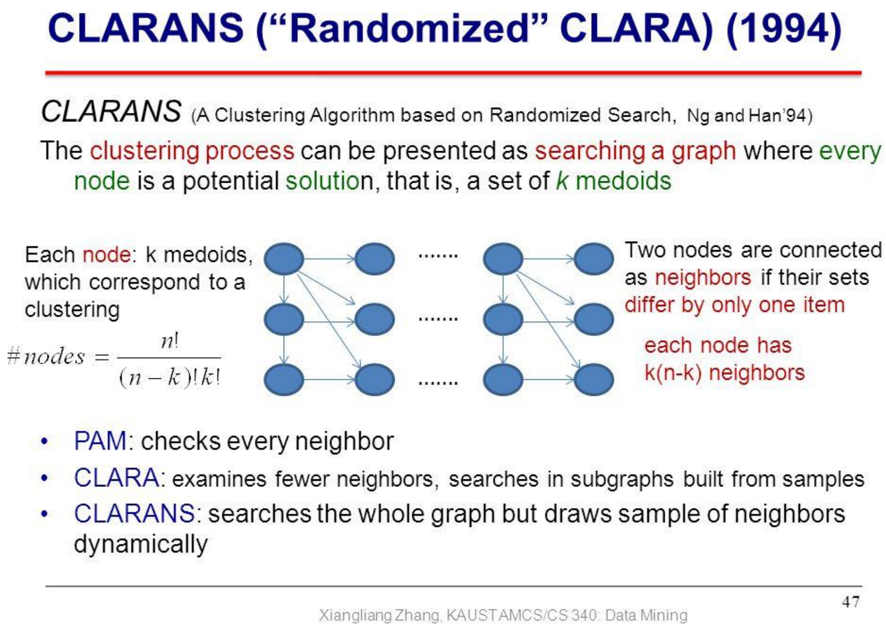
CLARA (Clustering LARge Applications) relies on the sampling approach to handle large data sets. Instead of finding medoids for the entire data set, CLARA draws a small sample from the data set and applies the PAM algorithm to generate an optimal set of medoids for the sample.



```
CLARA($X, d, k$)
    $bestDissim \leftarrow \infty$
    for $t \leftarrow 1$ to $S$
    do $X' \leftarrow$ RANDOM-SUBSET($X, s$)
        $D \leftarrow$ BUILD-DISSIM-MATRIX($X', d$)
        $(C', M) \leftarrow$ PAM($X', D, k$)
        $C \leftarrow$ ASSIGN-MEDOIDS($X, M, D$)
        $dissim \leftarrow$ TOTAL-DISSIM($C, M, D$)
        if $dissim < bestDissim$
          then $bestDissim \leftarrow dissim$
               $C_{best} \leftarrow C$
               $M_{best} \leftarrow M$
    return $(C_{best}, M_{best})$
```

CLARANS

CLARANS is a partitioning method of clustering particularly useful in spatial data mining. We mean recognizing patterns and relationships existing in spatial data (such as distance-related, direction-relation or topological data, e.g. data plotted on a road map) by spatial data mining



CLARANS ("Randomized" CLARA) (1994)

CLARANS (A Clustering Algorithm based on Randomized Search, Ng and Han'94)

The clustering process can be presented as searching a graph where every node is a potential solution, that is, a set of *k* medoids

Each node: k medoids, which correspond to a clustering

$$\#nodes = \frac{n!}{(n-k)!k!}$$

Two nodes are connected as neighbors if their sets differ by only one item

each node has k(n-k) neighbors

- PAM: checks every neighbor
- CLARA: examines fewer neighbors, searches in subgraphs built from samples
- CLARANS: searches the whole graph but draws sample of neighbors dynamically

Xiangliang Zhang, KAUST AMCS/CS 340: Data Mining 47

Hierarchical clustering

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.

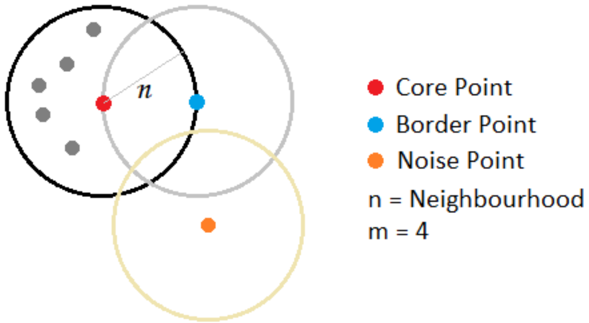
Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

The hierarchical clustering technique has two approaches:

1. Agglomerative: Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. Divisive: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

DBSCAN algorithm

DBSCAN stands for density-based spatial clustering of applications with noise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers). The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.



DBSCAN CLUSTERING

Abhijit Annaldas

BIRCH in Data Mining

BIRCH (balanced iterative reducing and clustering using hierarchies) is an unsupervised data mining algorithm that performs hierarchical clustering over large data sets. With modifications, it can also be used to accelerate k-means clustering and Gaussian mixture modeling with the expectation-maximization algorithm. An advantage of BIRCH is its ability to incrementally and dynamically cluster incoming, multi-dimensional metric data points to produce the best quality clustering for a given set of resources (memory and time constraints). In most cases, BIRCH only requires a single scan of the database.Basic clustering algorithms like K means and agglomerative clustering are the most commonly used clustering algorithms. But when performing clustering on very large datasets, BIRCH and DBSCAN are the advanced clustering algorithms useful for performing precise clustering on large datasets. Moreover, BIRCH is very useful because of its easy implementation. BIRCH is a clustering algorithm that clusters the dataset first in small summaries, then after small summaries get clustered. It does not directly cluster the dataset. That is why BIRCH is often used with other clustering algorithms; after making the summary, the summary can also be clustered by other clustering algorithms.

It is provided as an alternative to MinibatchKMeans. It converts data to a tree data structure with the centroids being read off the leaf. And these centroids can be the final cluster centroid or the input for other cluster algorithms like Agglomerative Clustering.

CURE(Clustering Using Representatives)

It is a hierarchical based clustering technique, that adopts a middle ground between the centroid based and the all-point extremes. Hierarchical clustering is a type of clustering, that starts with a single point cluster, and moves to merge with another cluster, until the desired number of clusters are formed.

It is used for identifying the spherical and non-spherical clusters.

It is useful for discovering groups and identifying interesting distributions in the underlying data.

Instead of using one point centroid, as in most of data mining algorithms, CURE uses a set of well-defined representative points, for efficiently handling the clusters and eliminating the outliers.

Categorical Clusturing

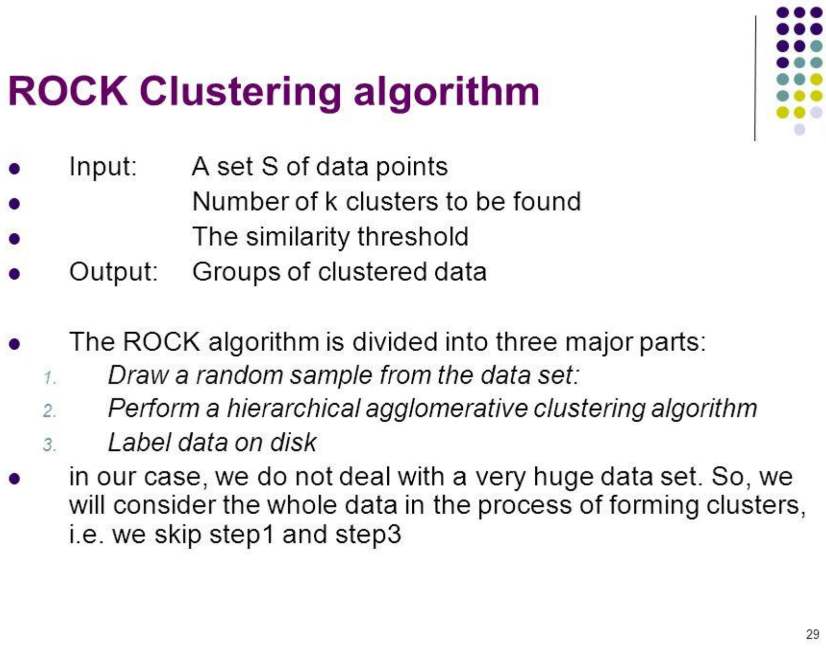
Data clustering is informally defined as the problem of partitioning a set of objects into groups, such that the objects in the same group are similar, while the objects in different groups are dissimilar. Categorical data clustering refers to the case where the data objects are defined over categorical attributes. A categorical attribute is an attribute whose domain is a set of discrete values that are not inherently comparable. That is, there is no single ordering or inherent distance function for the categorical values, and there is no mapping from categorical to numerical values that is semantically meaningful.

STIRR

Raghavan [12] proposed an algorithm called STIRR (Sieving Through Iterated Relational Reinforcement), for clustering of categorical data. It converts dataset into weighted graph and propagates these weights in iterative manner; this corresponds to a similarity measure based on co-occurrence of values in the dataset.

ROCK

The ROCK algorithm is divided into three general parts: Obtaining a random sample of data. Performing clustering on the data using the link agglomerative approach. A goodness measure is used to determine which pair of points is merged at each step.



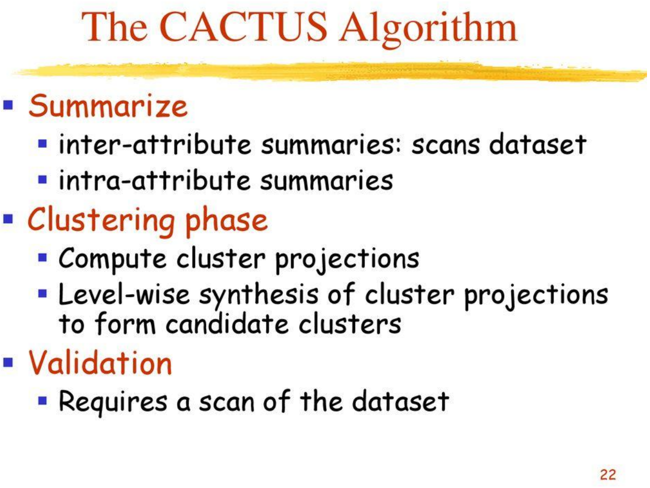
ROCK Clustering algorithm

- Input: A set S of data points
- Number of k clusters to be found
- The similarity threshold
- Output: Groups of clustered data

- The ROCK algorithm is divided into three major parts:
  1. *Draw a random sample from the data set:*
  2. *Perform a hierarchical agglomerative clustering algorithm*
  3. *Label data on disk*
- in our case, we do not deal with a very huge data set. So, we will consider the whole data in the process of forming clusters, i.e. we skip step1 and step3

29

CACTUS



The CACTUS Algorithm

- Summarize
  - inter-attribute summaries: scans dataset
  - intra-attribute summaries
- Clustering phase
  - Compute cluster projections
  - Level-wise synthesis of cluster projections to form candidate clusters
- Validation
  - Requires a scan of the dataset

22

Web Mining is the process of Data Mining techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.

Applications of Web Mining:

Web mining helps to improve the power of web search engine by classifying the web documents and identifying the web pages.

It is used for Web Searching e.g., Google, Yahoo etc and Vertical Searching e.g., FatLens, Become etc.

Web mining is used to predict user behavior.

Web mining is very useful of a particular Website and e-service e.g., landing page optimization.

Web Content Mining:
Web content mining is the application of extracting useful information from the content of the web documents. Web content consist of several types of data – text, image, audio, video etc. Content data is the group of facts that a web page is designed. It can provide effective and interesting patterns about user needs. Text documents are related to text mining, machine learning and natural language processing. This mining is also known as text mining. This type of mining performs scanning and mining of the text, images and groups of web pages according to the content of the input.

Web Structure Mining:
Web structure mining is the application of discovering structure information from the web. The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Structure mining basically shows the structured summary of a particular website. It identifies relationship between web pages linked by information or direct link connection. To determine the connection between two commercial websites, Web structure mining can be very useful.

Web Usage Mining:
Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets. And these patterns enable you to understand the user behaviors or something like that. In web usage mining, user access data on the web and collect data in form of logs. So, Web usage mining is also called log mining.

Comparison Between Data mining and Web mining:

Points	Data Mining	Web Mining
Definition	Data Mining is the process that attempts to discover pattern and hidden knowledge in large data sets in any system.	Web Mining is the process of data mining techniques to automatically discover and extract information from web documents.
Application	Data Mining is very useful for web page analysis.	Web Mining is very useful for a particular website and e-service.

Target Users	Data scientist and data engineers.	Data scientists along with data analysts.
Access	Data Mining is access data privately.	Web Mining is access data publicly.
Structure	In Data Mining get the information from explicit structure.	In Web Mining get the information from structured, unstructured and semi-structured web pages.
Problem Type	Clustering, classification, regression, prediction, optimization and control.	Web content mining, Web structure mining.
Tools	It includes tools like machine learning algorithms.	Special tools for web mining are Scrapy, PageRank and Apache logs.
Skills	It includes approaches for data cleansing, machine learning algorithms. Statistics and probability.	It includes application level knowledge, data engineering with mathematical modules like statistics and probability.

TEXT MINING

Text mining is a process of extracting useful information and nontrivial patterns from a large volume of text databases. There exist various strategies and devices to mine the text and find important data for the prediction and decision-making process. The selection of the right and accurate text mining procedure helps to enhance the speed and the time complexity also. This article briefly discusses and analyzes text mining and its applications in diverse fields.

The conventional process of text mining as follows:

Gathering unstructured information from various sources accessible in various document organizations, for example, plain text, web pages, PDF records, etc.

Pre-processing and data cleansing tasks are performed to distinguish and eliminate inconsistency from the data. The data cleansing process makes sure to capture the genuine text, and it is performed to eliminate stop words stemming (the process of identifying the root of a certain word and indexing the data.

Processing and controlling tasks are applied to review and further clean the data set.

Pattern analysis is implemented in Management Information System.

Information processed in the above steps is utilized to extract important and applicable data for a powerful and convenient decision-making process and trend analysis.

1. Spatial Data Mining :
Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. In spatial data mining analyst use geographical or spatial information to produce business intelligence or other results. Challenges involved in spatial data mining include identifying patterns or finding objects that are relevant to research project.
2. Temporal Data Mining :
Temporal data refers to the extraction of implicit, non-trivial and potentially useful abstract information from large collection of temporal data. It is concerned with the analysis of temporal data and for finding temporal patterns and regularities in sets of temporal data tasks of temporal data mining are –

Data Characterization and Comparison
Cluster Analysis
Classification
Association rules
Prediction and Trend Analysis
Pattern Analysis

Temporal data mining has led to a new way of interacting with a temporal database and specifying queries at a much more abstract level than say, temporal structured query language permits. It also facilities data exploration for problems that are due to multiple and multi-dimensionality.
The basic goal of temporal classification is to predict temporally related fields in a temporal database based on other fields. The problem, in general, is cast as deciding the general value of the temporal variable being predicted given the different fields, the training data in which the target variable is given for each observation, and a set of assumptions representing one's prior knowledge of the problem. Temporal classification techniques are associated with the complex problem of density estimation.

Difference between Spatial and Temporal Data Mining :

SNO.	Spatial data mining	Temporal data mining
1.	It requires space.	It requires time.

2.	Spatial mining is the extraction of knowledge/spatial relationship and interesting measures that are not explicitly stored in spatial database.	Temporal mining is the extraction of knowledge about occurrence of an event whether they follow Cyclic , Random ,Seasonal variations etc.
3.	It deals with spatial (location , Geo-referenced) data.	It deals with implicit or explicit Temporal content , from large quantities of data.
4.	Spatial databases reverses spatial objects derived by spatial data. types and spatial association among such objects.	Temporal data mining comprises the subject as well as its utilization in modification of fields.
5.	It includes finding characteristic rules, discriminant rules, association rules and evaluation rules etc.	It aims at mining new and unknown knowledge, which takes into account the temporal aspects of data.
6.	It is the method of identifying unusual and unexplored data but useful models from spatial databases.	It deals with useful knowledge from temporal data.
7.	Examples – Determining hotspots , Unusual locations.	Examples – An association rule which looks like – "Any Person who buys a car also buys steering lock". By temporal aspect this rule would be – " Any person who buys a car also buys a steering lock after that ".

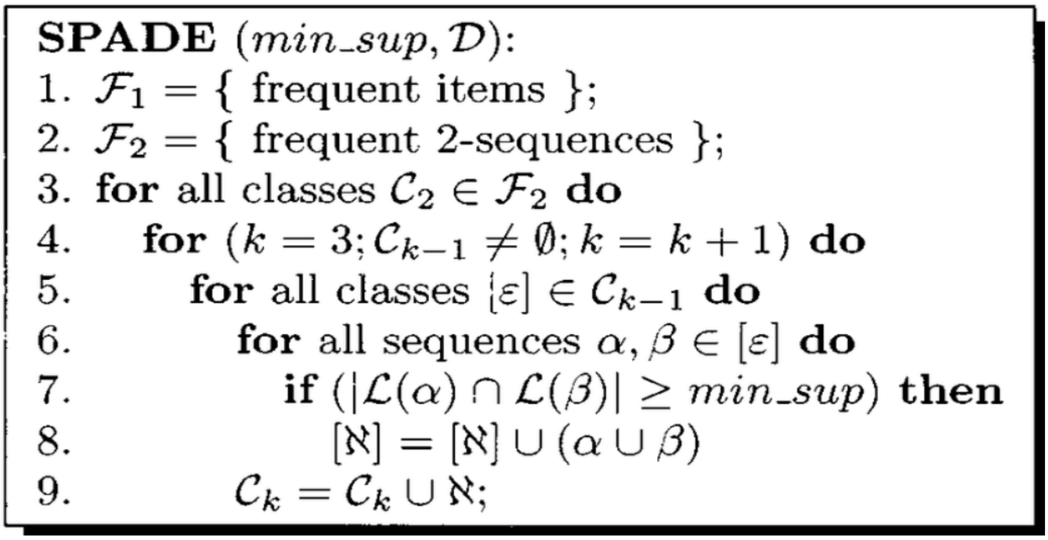
Generalized Sequential Pattern(GSP)

GSP is a very important algorithm in data mining. It is used in sequence mining from large databases. Almost all sequence mining algorithms are basically based on a prior algorithm. GSP uses a level-wise paradigm for finding all the sequence patterns in the data. It starts with finding the frequent items of size one then passes that as input to the next iteration of the GSP algorithm. The database is passed multiple times to this algorithm. In each iteration, GSP removes all the non-frequent itemsets. This is done based on a threshold frequency which is called support. Only those itemsets are kept whose frequency is greater than the support count. After the first pass, GSP finds all the frequent sequences of length-1 which are called

1-sequences. This makes the input to the next pass, it is the candidate for 2-sequences. At the end of this pass, GSP generates all frequent 2-sequences, which makes the input for candidate 3-sequences. The algorithm is recursively called until no more frequent itemsets are found.

SPADE

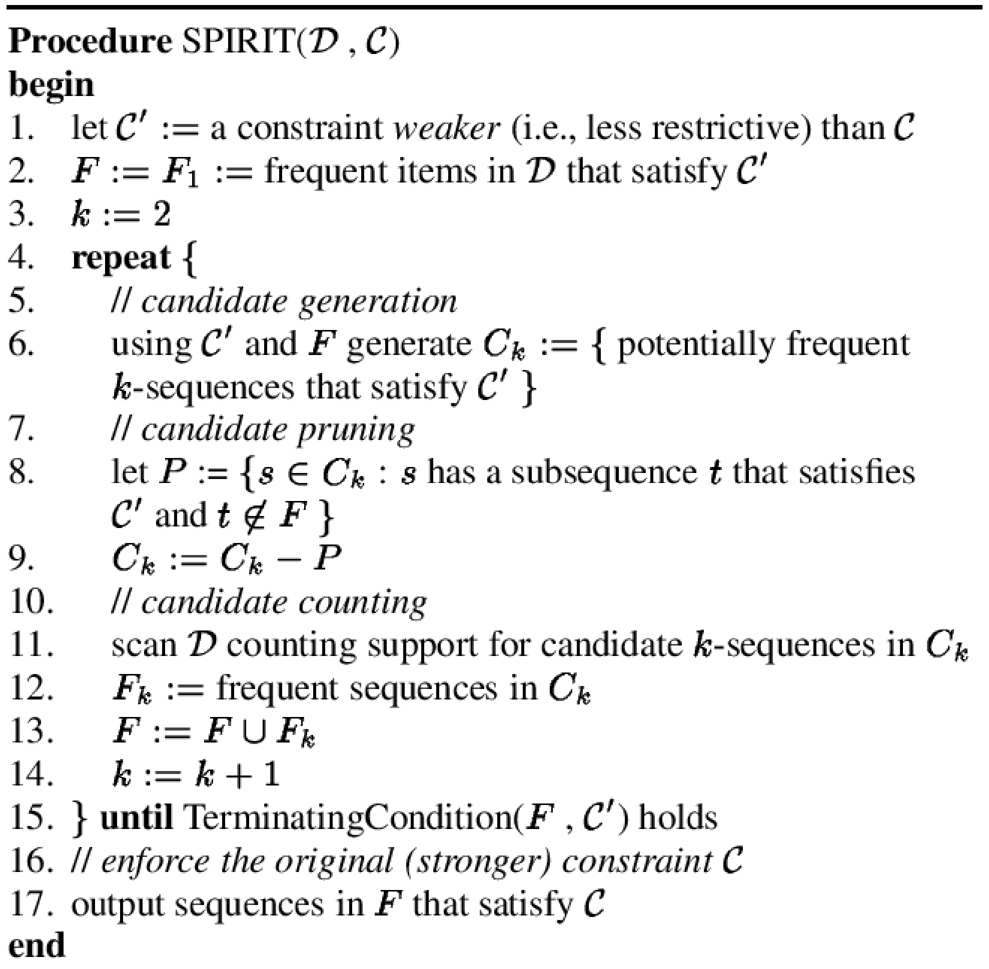
An algorithm to Frequent Sequence Mining is the SPADE (Sequential PAttern Discovery using Equivalence classes) algorithm. It uses a vertical id-list database format, where we associate to each sequence a list of objects in which it occurs.



```
SPADE ($min\_sup, \mathcal{D}$):
$\mathcal{F}_1$ = { frequent items };
$\mathcal{F}_2$ = { frequent 2-sequences };
for all classes $\mathcal{C}_2 \in \mathcal{F}_2$ do
    for ($k = 3; \mathcal{C}_{k-1} \neq \emptyset; k = k + 1$) do
        for all classes $[\varepsilon] \in \mathcal{C}_{k-1}$ do
            for all sequences $\alpha, \beta \in [\varepsilon]$ do
                if ($|\mathcal{L}(\alpha) \cap \mathcal{L}(\beta)| \geq min\_sup$) then
                    $[\aleph] = [\aleph] \cup (\alpha \cup \beta)$
            $\mathcal{C}_k = \mathcal{C}_k \cup \aleph$;
```

SPIRIT

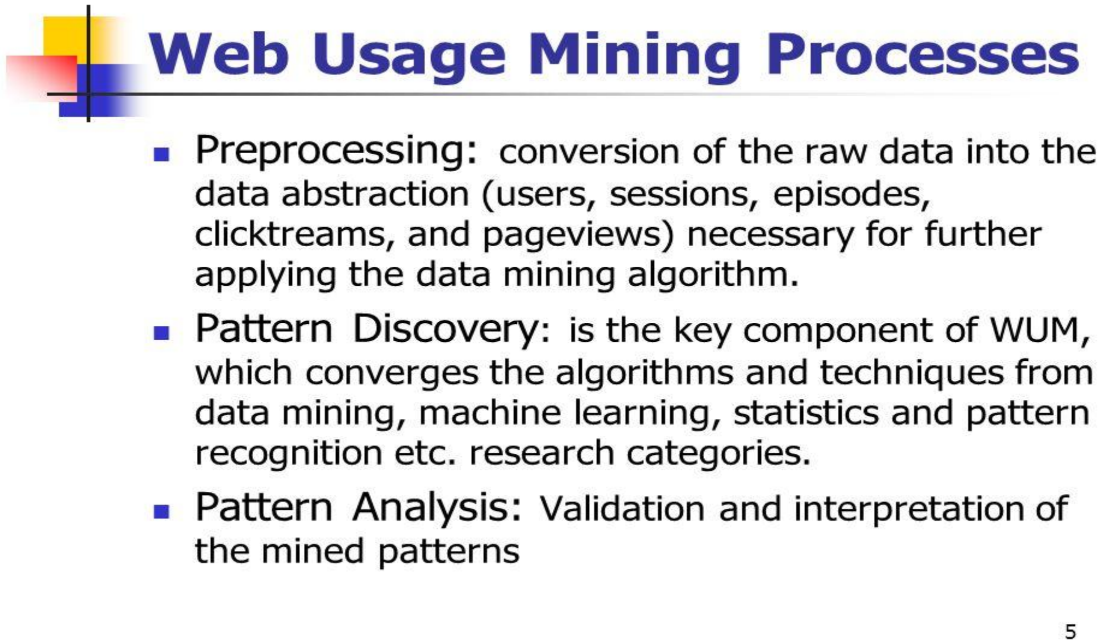
SPIRIT [Garofalakis 1999] is a family of apriori-based algorithms that uses a regular language to constrain the mining process. The core of the algorithm is similar to GSP, and its main difference resides on the candidate generation step, which creates candidates that potentially satisfy the constraint.



```
Procedure SPIRIT($\mathcal{D}$, $\mathcal{C}$)
begin
let $\mathcal{C}'$ := a constraint weaker (i.e., less restrictive) than $\mathcal{C}$
$F := F_1$ := frequent items in $\mathcal{D}$ that satisfy $\mathcal{C}'$
$k := 2$
repeat {
    // candidate generation
    using $\mathcal{C}'$ and $F$ generate $C_k := \{$ potentially frequent
    $k$-sequences that satisfy $\mathcal{C}'$ $\}$
    // candidate pruning
    let $P := \{s \in C_k : s$ has a subsequence $t$ that satisfies
    $\mathcal{C}'$ and $t \notin F \}$
    $C_k := C_k - P$
    // candidate counting
    scan $\mathcal{D}$ counting support for candidate $k$-sequences in $C_k$
    $F_k$ := frequent sequences in $C_k$
    $F := F \cup F_k$
    $k := k + 1$
} until TerminatingCondition($F$, $\mathcal{C}'$) holds
// enforce the original (stronger) constraint $\mathcal{C}$
output sequences in $F$ that satisfy $\mathcal{C}$
end
```

WUM

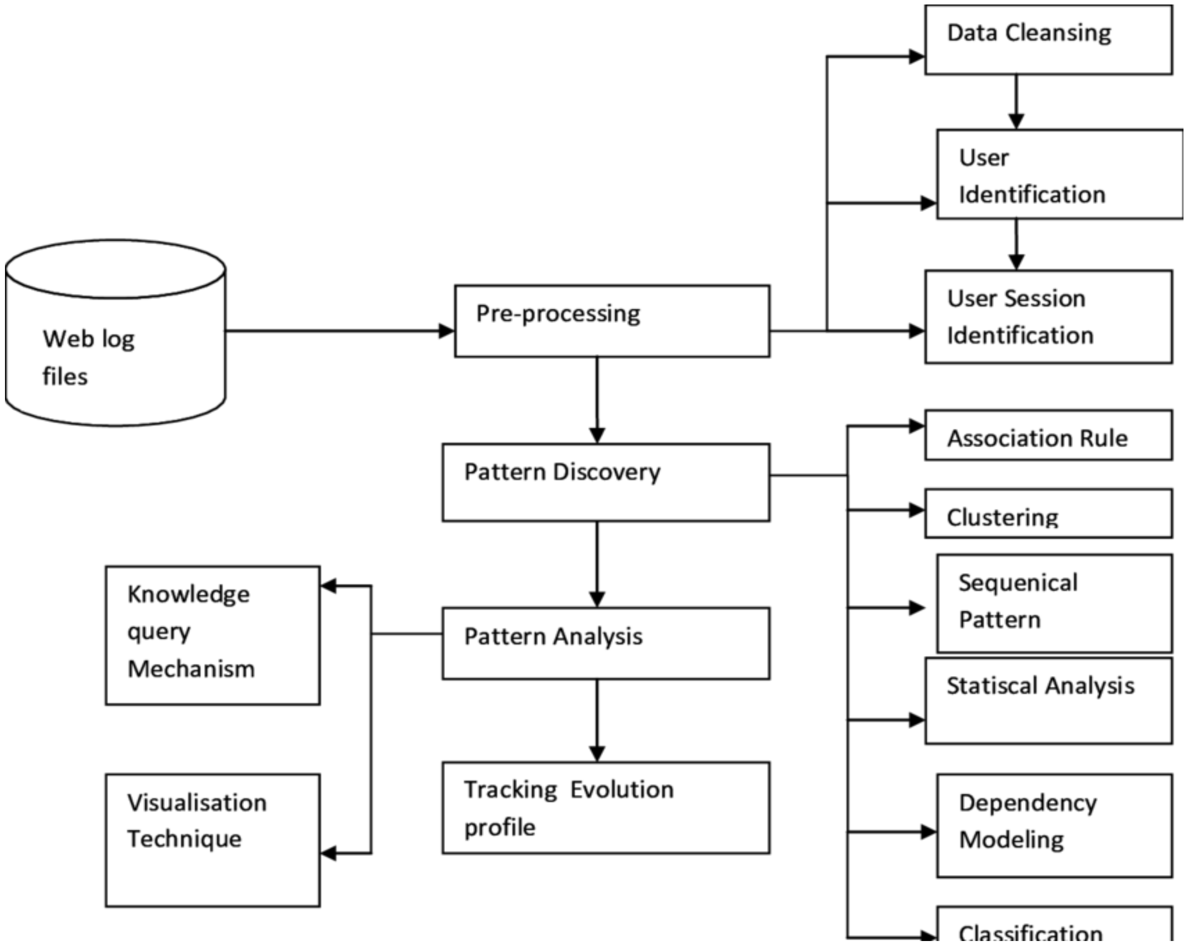
Data mining on web log files is called Web Usage Mining (WUM). User clustering based on access patterns is an important part of WUM. Most papers just consider web pages hits, but ignore the succession of pages during user clustering.



Web Usage Mining Processes

- Preprocessing: conversion of the raw data into the data abstraction (users, sessions, episodes, clicktreams, and pageviews) necessary for further applying the data mining algorithm.
- Pattern Discovery: is the key component of WUM, which converges the algorithms and techniques from data mining, machine learning, statistics and pattern recognition etc. research categories.
- Pattern Analysis: Validation and interpretation of the mined patterns

5


Data Cleansing
User Identification
Web log files
Pre-processing
User Session Identification
Association Rule
Pattern Discovery
Clustering
Sequenical Pattern
Knowledge query Mechanism
Pattern Analysis
Statiscal Analysis
Visualisation Technique
Tracking Evolution profile
Dependency Modeling
Classification