DWM CAT1 QB SOLUTIONS

a)Differentiate between:
i) Datamart and metadata ii) OLTP and OLAP

Data Warehouse	Data Mart
A Data Warehouse is a vast repository of information collected from various organizations or departments within a corporation.	A data mart is an only subtype of a Data Warehouses. It is architecture to meet the requirement of a specific user group.
It may hold multiple subject areas.	It holds only one subject area. For example, Finance or Sales.
It holds very detailed information.	It may hold more summarized data.
Works to integrate all data sources	It concentrates on integrating data from a given subject area or set of source systems.
In data warehousing, Fact constellation is used.	In Data Mart, Star Schema and Snowflake Schema are used.
It is a Centralized System.	It is a Decentralized System.
Data Warehousing is the data-oriented.	Data Marts is a project-oriented.

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Sr. No.	Category	OLAP (Online analytical processing)	OLTP (Online transaction processing)
1.	**Definition**	It is well-known as an online database query management system.	It is well-known as an online database modifying system.
2.	**Data source**	Consists of historical data from various Databases.	Consists of only of operational current data.
3.	**Method used**	It makes use of a data warehouse.	It makes use of a standard database management system (DBMS).
4.	**Application**	It is subject-oriented. Used for Data Mining, Analytics, Decisions making, etc.	It is application-oriented. Used for business tasks.
5.	**Normalized**	In an OLAP database, tables are not normalized.	In an OLTP database, tables are normalized (3NF).
6.	**Usage of data**	The data is used in planning, problem-solving, and decision-making.	The data is used to perform day-to-day fundamental operations.
7.	**Task**	It provides a multi-dimensional view of different business tasks.	It reveals a snapshot of present business tasks.
8.	**Purpose**	It serves the purpose to extract information for analysis and decision-making.	It serves the purpose to Insert, Update, and Delete information from the database.
9.	**Volume of data**	A large amount of data is stored typically in TB, PB	The size of the data is relatively small as the historical data is archived. For ex MB, GB
10.	**Queries**	Relatively slow as the amount of data involved is large. Queries may take hours.	Very Fast as the queries operate on 5% of the data.
11.	**Update**	The OLAP database is not often updated. As a result, data integrity is unaffected.	The data integrity constraint must be maintained in an OLTP database.
12.	**Backup and Recovery**	It only need backup from time to time as compared to OLTP.	Backup and recovery process is maintained rigorously
13.	**Processing time**	The processing of complex queries can take a lengthy time.	It is comparatively fast in processing because of simple and straightforward queries.
14.	**Types of users**	This data is generally managed by CEO, MD, GM.	This data is managed by clerks, managers.
15.	**Operations**	Only read and rarely write operation.	Both read and write operations.
16.	**Updates**	With lengthy, scheduled batch operations, data is refreshed on a regular basis.	The user initiates data updates, which are brief and quick.
17.	**Nature of audience**	Process that is focused on the customer.	Process that is focused on the market.
18.	**Database Design**	Design with a focus on the subject.	Design that is focused on the application.
19.	**Productivity**	Improves the efficiency of business analysts.	Enhances the user's productivity

b)What is OLAP? Define following with example :
i) ROLAP ii) MOLAP iii) HOLAP

Online Analytical Processing Server (OLAP) is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information. OLAP works with large amounts of data stored in a data warehouse.

OLAP cubes have two main purposes. The first is to provide business users with a data model more intuitive to them than a tabular model. This model is called a Dimensional Model.

The second purpose is to enable fast query response that is usually difficult to achieve using tabular models.

Relational OLAP

ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP includes the following –

- Implementation of aggregation navigation logic.
- Optimization for each DBMS back end.
- Additional tools and services.

Multidimensional OLAP

MOLAP uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the data set is sparse. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allows to store the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.

a)Explain in detail the components of Data warehouse system.

Components of Data Warehouse Architecture and their tasks :

1. Operational Source –

- An operational Source is a data source consists of Operational Data and External Data.
- Data can come from Relational DBMS like Informix, Oracle.

2. Load Manager –

- The Load Manager performs all operations associated with the extraction of loading data in the data warehouse.
- These tasks include the simple transformation of data to prepare data for entry into the warehouse.

3. Warehouse Manage –

- The warehouse manager is responsible for the warehouse management process.
- The operations performed by the warehouse manager are the analysis, aggregation, backup and collection of data, de-normalization of the data.

4. Query Manager –

- Query Manager performs all the tasks associated with the management of user queries.
- The complexity of the query manager is determined by the end-user access operations tool and the features provided by the database.

5. Detailed Data –

- It is used to store all the detailed data in the database schema.
- Detailed data is loaded into the data warehouse to complement the data collected.

6. Summarized Data –

- Summarized Data is a part of the data warehouse that stores predefined aggregations
- These aggregations are generated by the warehouse manager.

7. Archive and Backup Data –

- The Detailed and Summarized Data are stored for the purpose of archiving and backup.
- The data is relocated to storage archives such as magnetic tapes or optical disks.

8. Metadata –

- Metadata is basically data stored above data.
- It is used for extraction and loading process, warehouse, management process, and query management process.
- Metadata is data about the data which is needed by the users. It is used not only to instruct operators and users of the data warehouse about its status and the data held inside the data warehouse but also as a means of integration of incoming information and a tool to upgrade and refine the basic data warehouse model.

9. End User Access Tools –

- End-User Access Tools consist of Analysis, Reporting, and mining.
- By using end-user access tools users can link with the warehouse.

ETL

As mentioned above, ETL stands for Extract, Transform, Load. When DBAs want to move data from a data source into their data warehouse, this is the process they use. In

short, ETL converts data into a usable format so that once it’s in the data warehouse, it can be analyzed/queried/etc. For the purposes of this article, I won’t go into too much detail of how the entire ETL process works, but there are many different resources where you can learn about ETL.

b)Differentiate between Data Warehouse versus Operational DBMS

Operational Database	Data Warehouse
Operational systems are designed to support high-volume transaction processing.	Data warehousing systems are typically designed to support high-volume analytical processing (i.e., OLAP).
Operational systems are usually concerned with current data.	Data warehousing systems are usually concerned with historical data.
Data within operational systems are mainly updated regularly according to need.	Non-volatile, new data may be added regularly. Once Added rarely changed.
It is designed for real-time business dealing and processes.	It is designed for analysis of business measures by subject area, categories, and attributes.
It is optimized for a simple set of transactions, generally adding or retrieving a single row at a time per table.	It is optimized for extent loads and high, complex, unpredictable queries that access many rows per table.
It is optimized for validation of incoming information during transactions, uses validation data tables.	Loaded with consistent, valid information, requires no real-time validation.
It supports thousands of concurrent clients.	It supports a few concurrent clients relative to OLTP.
Operational systems are widely process-oriented.	Data warehousing systems are widely subject-oriented
Operational systems are usually optimized to perform fast inserts and updates of associatively small volumes of data.	Data warehousing systems are usually optimized to perform fast retrievals of relatively high volumes of data.
Data In	Data Out
Less Number of data accessed.	Large Number of data accessed.
Relational databases are created for on-line transactional Processing (OLTP)	Data Warehouse designed for on-line Analytical Processing (OLAP)

a)What is data model? Explain multidimensional data model in detail.

Data models are visual representations of an enterprise's data elements and the connections between them. By helping to define and structure data in the context of relevant business processes, models support the development of effective information systems.

Data models describe how a database's logical structure is represented. Data models specify how data is linked to one another, as well as how it is handled and stored within the system.

* Multidimensional Data Models :—

- A multidemensional data model ; datawarehouse and OLAP tools are based on multidimensional data models. The multidimension data model is a method which is used for ordering data in the database.
- Multidimensional DM views data in the form of data cube.
- Data cube allows the data to be modelled and views in multiple dimensions. It is defined by dimensions and facts (branches). The dimension are perpectives or the entities with respect to which the organization wants to keep the data or records.

Example :

Take the example of data of the factory which sells the products per quater in location banglore The data is represent in the table given below.

	Location : Bangalore			
Time (Quarter)	Type of items			
	Jam	Bread	Sugar	Salt
Q1	350	389	35	50
Q2	260	528	50	90
Q3	483	256	20	60
Q4	436	396	15	40

In the above given presentation, the factory sales for the bangalore are, for the time dimension which is organised into quarter and the dimension of items, which is sorted a/c to the kind of items which is sold. The facts here are represented in Rupees (in thousands).

Now if we desire to view the data of the sales in a 3 dimensional table, then it is represented in the diagram given below.

Here the data of the sales is represent as a two dimensional table. Let's consider the data a/c to item, time and location

	location = "Kolkata"			location = "Delhi"			location = "mumbai"		
	item			item			item		
Time	Jam	Bread	Sugar	Jam	Bread	Sugar	Jam	Bread	Sugar
Q1	340	604	38	335	365	35	336	484	80
Q2	680	583	10	684	490	48	595	594	34
Q3	535	490	50	389	385	15	366	385	20

Fig: 3D data representation as 2D

The data can represented in the form 3D-conceptually, which is shown in the image below:

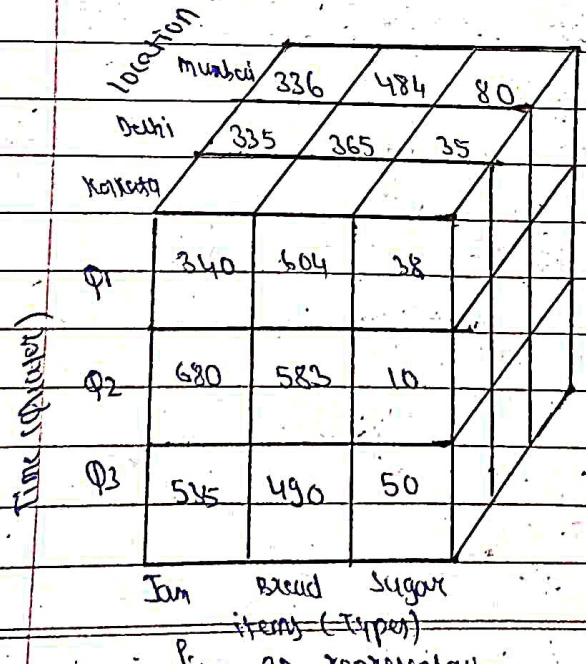


fig: 3D representation

Advantages of Multi Dimensional Data Model

The following are the advantages of a multi-dimensional data model :

- A multi-dimensional data model is easy to handle.
- It is easy to maintain.
- Its performance is better than that of normal databases (e.g. relational databases).
- The representation of data is better than traditional databases. That is because the multi-dimensional databases are multi-viewed and carry different types of factors.
- It is workable on complex systems and applications, contrary to the simple one-dimensional database systems.
- The compatibility in this type of database is an upliftment for projects having lower bandwidth for maintenance staff.

Disadvantages of Multi Dimensional Data Model

The following are the disadvantages of a Multi Dimensional Data Model :

- The multi-dimensional Data Model is slightly complicated in nature and it requires professionals to recognize and examine the data in the database.
- During the work of a Multi-Dimensional Data Model, when the system caches, there is a great effect on the working of the system.
- It is complicated in nature due to which the databases are generally dynamic in design.
- The path to achieving the end product is complicated most of the time.
- As the Multi Dimensional Data Model has complicated systems, databases have a large number of databases due to which the system is very insecure when there is a security break.

b)What are the characteristics of data warehouse?

Subject-oriented – A data warehouse is always a subject oriented as it delivers information about a theme instead of organization's current operations. It can be achieved on specific theme. That means the data warehousing process is proposed to handle with a specific theme which is more defined. These themes can be sales, distributions, marketing etc.
A data warehouse never put emphasis only current operations. Instead, it focuses on demonstrating and analysis of data to make various decision. It also delivers an easy and precise demonstration around particular theme by eliminating data which is not required to make the decisions.

Integrated – It is somewhere same as subject orientation which is made in a reliable format. Integration means founding a shared entity to scale the all similar data from the different databases. The data also required to be resided into various data warehouse in shared and generally granted manner.
A data warehouse is built by integrating data from various sources of data such that a mainframe and a relational database. In addition, it must have reliable naming conventions, format and codes. Integration of data warehouse benefits in effective analysis of data. Reliability in naming conventions, column scaling, encoding structure etc. should be confirmed. Integration of data warehouse handles various subject related warehouse.

Time-Variant – In this data is maintained via different intervals of time such as weekly, monthly, or annually etc. It founds various time limit which are structured between the large datasets and are held in online transaction process (OLTP). The time limits for data warehouse is wide-ranged than that of operational systems. The data resided in data warehouse is predictable with a specific interval of time and delivers information from the historical perspective. It comprises elements of time explicitly or implicitly. Another feature of time-variance is that once data is stored in the data warehouse then it cannot be modified, alter, or updated. Data is stored with a time dimension, allowing for analysis of data over time.

Non-Volatile – As the name defines the data resided in data warehouse is permanent. It also means that data is not erased or deleted when new data is inserted. It includes the mammoth quantity of data that is inserted into modification between the selected quantity on logical business. It evaluates the analysis within the technologies of warehouse. Data is not updated, once it is stored in the data warehouse, to maintain the historical data.

① Subject Oriented :→

A data warehouse is organized around major subject such as customers, suppliers, product and sales rather than concentrating on the day to day operations and transaction's processing of an organization.

② Integrated :→

A Data warehouse is usually constructed by integrating multiple heterogenous sources such as relational database (ex. RDBMS, file, excel & etc.) DB, flat file and online transactional record.

③ Time - Variant :→ ~~(data can not be modified)~~

Data are stored to provide information from a historical perspective (the past 5-10 years).

④ Non - Volatile :→ (do not modified)

A Data warehouse is always a physically separate store data, transform from the application data found in the operational environment. Due to this separation the data warehouse does not require transaction processing, recovery and concurrency control mechanism.

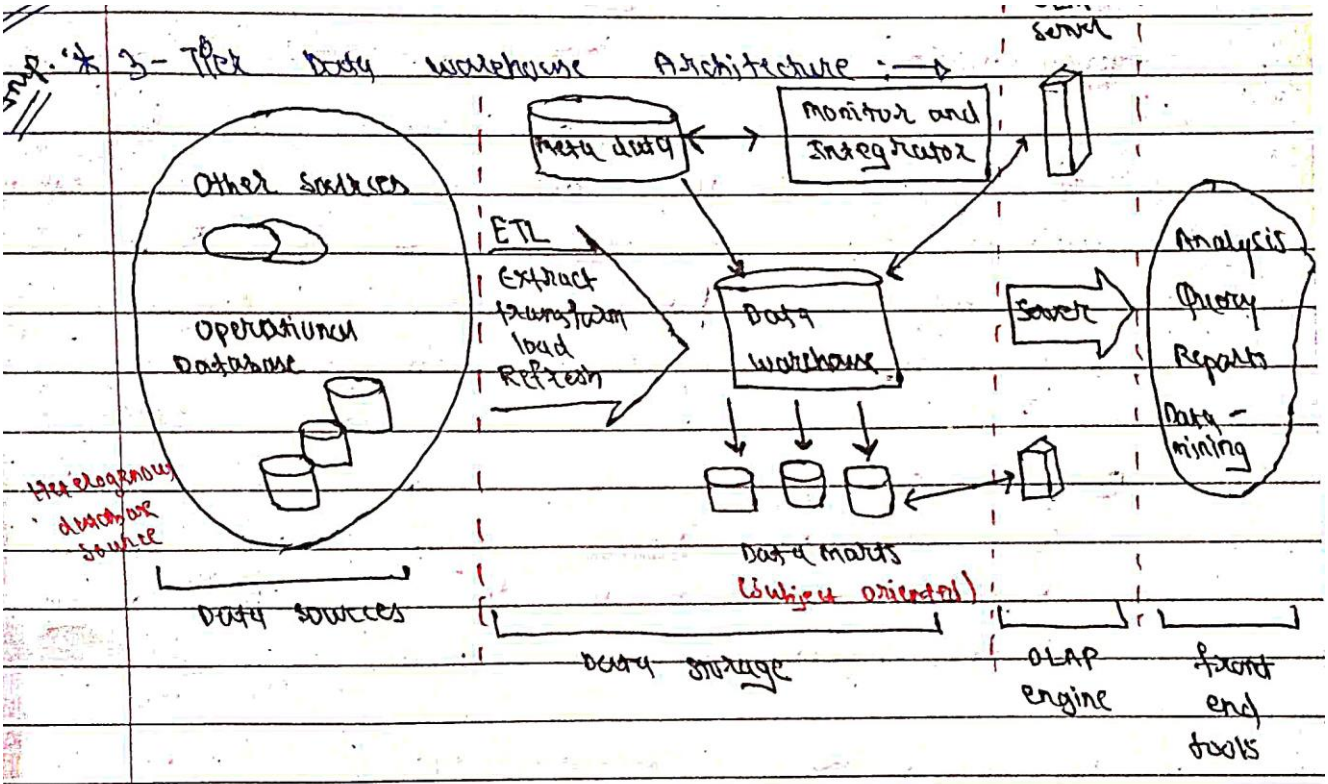
a) Define data warehouse. Draw the architecture of data warehouse and explain the three tiers in detail.

A data warehouse is a centralized storage system that allows for the storing, analyzing, and interpreting of data in order to facilitate better decision-making.

A data warehouse is a large collection of business data used to help an organization make decisions.

* Data warehouse and characteristics :→

- A data warehouse is a subject oriented, integrated, Non-volatile and time variant collection of data in support of manggements decision making process.



Data warehouse usually have three levels (tier) architecture that includes :

① Bottom tier that consist of data warehouse server which is almost an RDBMS. It may include several specialize data marts and meta-data repository. Data from operational DB's and external sources (such as user profile data provided by external consultants) are extracted using application program interfaces called a gateway.

A gateway is provided by the underlined DBMS and allows customer programs to generate SQL code to be executed at a server example of gateways contains ODBC (open database connection / connectivity) & JDBC (Java database connectivity.

② Middle Tier :

which consist of an OLAP server for fast query of the data warehouse. There can be different types of OLAP servers ie. ROLAP (relational OLAP server), MOLAP (multi dimensional OLAP server), HOLAP (Hybrid OLAP server).

③ A Top tier : that contains front-end tools for displaying result provided by OLAP, as well as additional tools for data mining of the OLAP generated data.

b) Why Do We Need Data Warehouses?

Data Warehouse is needed for the following reasons:

1. **Business User:** Business users require a data warehouse to view summarized data from the past. Since these people are non-technical, the data may be presented to them in an elementary form.
2. **Store historical data:** Data Warehouse is required to store the time variable data from the past. This input is made to be used for various purposes.
3. **Make strategic decisions:** Some strategies may be depending upon the data in the data warehouse. So, data warehouse contributes to making strategic decisions.
4. **For data consistency and quality:** Bringing the data from different sources at a commonplace, the user can effectively undertake to bring the uniformity and consistency in data.
5. **High response time:** Data warehouse has to be ready for somewhat unexpected loads and types of queries, which demands a significant degree of flexibility and quick response time.

6. Data Warehouse gives the ability to quickly run analysis on huge volumes of datasets.
7. If there is any change in the structure of the data available in the operational or transactional Databases. It will not break the business reports running on top of it because they are not directly connected to BI tools or Reporting tools.
8. Cloud Data Warehouse (such as Amazon Redshift and Google BigQuery) offer an added advantage that you need not invest in them upfront. Instead, you pay as you go as the size of your data increases. You can refer to this article on Amazon Redshift vs Google BigQuery for a comparison of the two.
9. When companies want to make the data available for all, they will understand the need for Data Warehouse. You can expose the data within the company for analysis. While you do so you can hide certain sensitive information (such as PII – Personally Identifiable Information about your customers, or Partners).
10. There is always the need for Data Warehouse as the complexity of queries increases and users need faster query processing. Because the transactional Databases are built to store a store in a normalized form whereas fast query processing can be achieved by denormalized data that is available in Data Warehouse

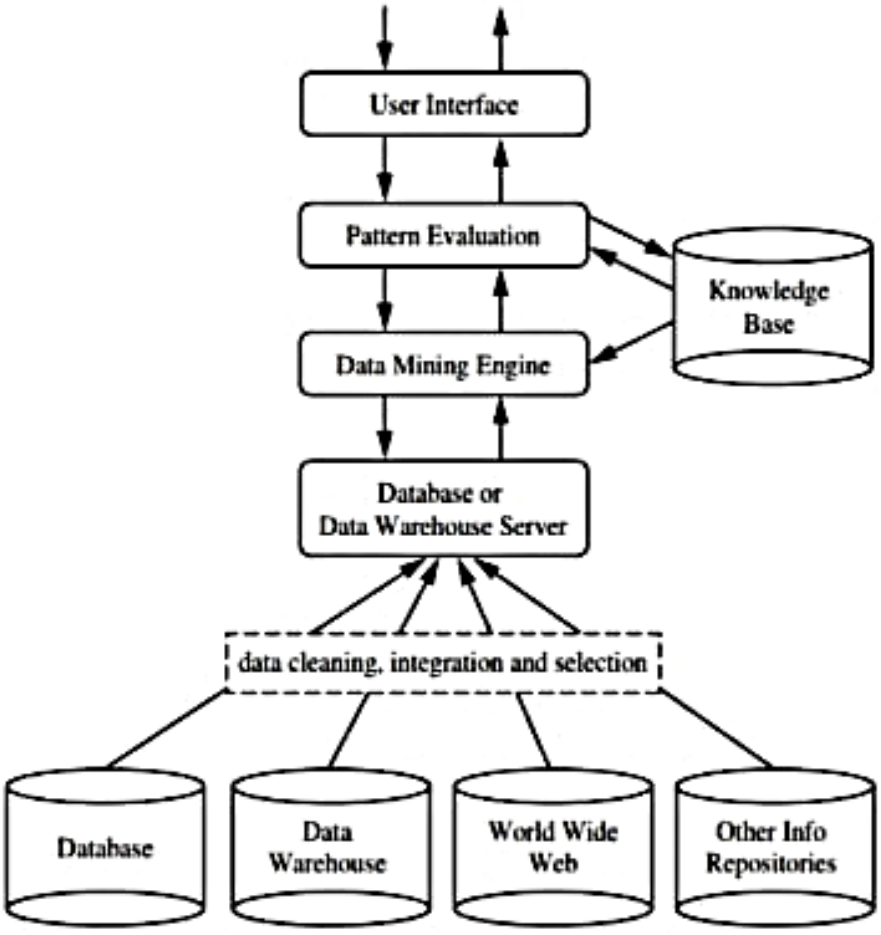
a) As a Bank manager how would you decide whether to give loan to an applicant or not by using DM strategies.

b) Discuss different OLAP tools.

a) What are major components of a typical data mining? Draw architecture of data mining system and explain it.

The major components of data mining are as follows –

- **Databases** – This is one or a set of databases, data warehouses, spreadsheets, and another type of data repository where data cleaning and integration techniques can be implemented.
- **Data warehouse server** – This component fetches the relevant records based on users request from a data warehouse.
- **Knowledge base** – It is a knowledge domain that is employed for discovering interesting patterns.
- **Data mining engine** – It uses a functional module that is used to perform tasks including classification, association, cluster analysis, etc.
- **Pattern evaluation module** – This component uses interestingness measures that communicate with data mining structure to target the search towards interesting patterns.
- **User interface** – This interface enables users to interact with the system by describing a data mining function or a query through the graphical user interface.



Knowledge Base: Knowledge Base is an important part of the data mining engine that is quite beneficial in guiding the search for the result patterns. Data mining engines may also sometimes get inputs from the knowledge base. This knowledge base may contain data from user experiences. The objective of the knowledge base is to make the result more accurate and reliable.

Pattern Evaluation Modules: They are responsible for finding interesting patterns in the data and sometimes they also interact with the database servers for producing the result of the user requests.

Data Mining Engine: It is one of the core components of the data mining architecture that performs all kinds of data mining techniques like association, classification, characterization, clustering, prediction, etc.

Data Sources: Database, World Wide Web(WWW), and data warehouse are parts of data sources. The data in these sources may be in the form of plain text, spreadsheets, or other forms of media like photos or videos. WWW is one of the biggest sources of data.

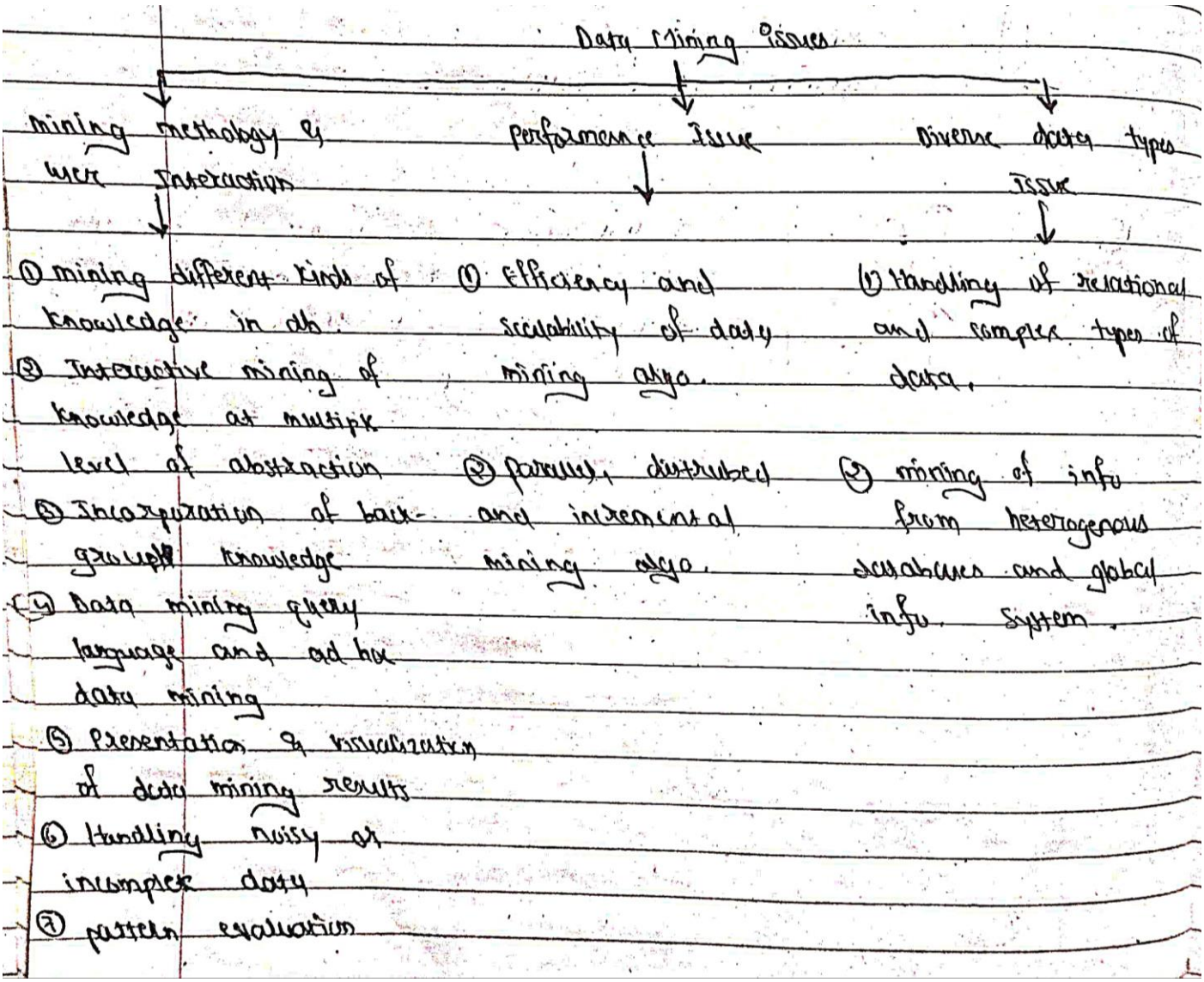
Graphic User Interface: Since the user cannot fully understand the complexity of the data mining process so graphical user interface helps the user to communicate effectively with the data mining system.

b) Explain data mining functionalities in detail.

There are various data mining functionalities which are as follows –

- **Data characterization** – It is a summarization of the general characteristics of an object class of data. The data corresponding to the user-specified class is generally collected by a database query. The output of data characterization can be presented in multiple forms.
- **Data discrimination** – It is a comparison of the general characteristics of target class data objects with the general characteristics of objects from one or a set of contrasting classes. The target and contrasting classes can be represented by the user, and the equivalent data objects fetched through database queries.
- **Association Analysis** – It analyses the set of items that generally occur together in a transactional dataset. There are two parameters that are used for determining the association rules –
    - It provides which identifies the common item set in the database.
    - Confidence is the conditional probability that an item occurs in a transaction when another item occurs.
- **Classification** – Classification is the procedure of discovering a model that represents and distinguishes data classes or concepts, for the objective of being able to use the model to predict the class of objects whose class label is anonymous. The derived model is established on the analysis of a set of training data (i.e., data objects whose class label is common).
- **Prediction** – It defines predict some unavailable data values or pending trends. An object can be anticipated based on the attribute values of the object and attribute values of the classes. It can be a prediction of missing numerical values or increase/decrease trends in time-related information.
- **Clustering** – It is similar to classification but the classes are not predefined. The classes are represented by data attributes. It is unsupervised learning. The objects are clustered or grouped, depends on the principle of maximizing the intraclass similarity and minimizing the intraclass similarity.
- **Outlier analysis** – Outliers are data elements that cannot be grouped in a given class or cluster. These are the data objects which have multiple behaviour from the general behaviour of other data objects. The analysis of this type of data can be essential to mine the knowledge.
- **Evolution analysis** – It defines the trends for objects whose behaviour changes over some time.

a)Explain various major issues and challenges in data mining in detail.



Data mining systems face a lot of **data mining challenges** and issues in today's world some of them are:

1. Mining methodology and user interaction issues
2. Performance issues
3. Issues relating to the diversity of database types

1. Mining methodology and user interaction issues:

i. Mining different kinds of knowledge in databases:

Different user - different knowledge - different way. That means different client want a different kind of information so it becomes difficult to cover vast range of data that can meet the client requirement.

ii. Interactive mining of knowledge at multiple levels of abstraction:

Interactive mining allows users to focus the search for patterns from different angles. The data mining process should be interactive because it is difficult to know what can be discovered within a database.

iii. Incorporation of background knowledge:
Background knowledge is used to guide discovery process and to express the discovered patterns.

iv. Query languages and ad hoc mining:
Relational query languages (such as SQL) allow users to pose ad-hoc queries for data retrieval. The language of data mining query language should be in perfectly matched with the query language of data warehouse.

v. Handling noisy or incomplete data:
In a large database, many of the attribute values will be incorrect. This may be due to human error or because of any instruments fail. Data cleaning methods and data analysis methods are used to handle noise data.

2. Performance issues
i. Efficiency and scalability of data mining algorithms:
To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.

ii. Parallel, distributed, and incremental mining algorithms:
The huge size of many databases, the wide distribution of data, and complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel.

3. Issues relating to the diversity of database types:
i. Handling of relational and complex types of data:
There are many kinds of data stored in databases and data warehouses. It is not possible for one system to mine all these kind of data. So different data mining system should be construed for different kinds data.

ii. Mining information from heterogeneous databases and global information systems:
Since data is fetched from different data sources on Local Area Network (LAN) and Wide Area Network (WAN).The discovery of knowledge from different sources of structured is a great challenge to data mining.

Major Challenges In Data Mining

Transforming data into organized information is not an easy process. There are many challenges in data mining.

Below are some of these Challenges listed and briefly explained:

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.

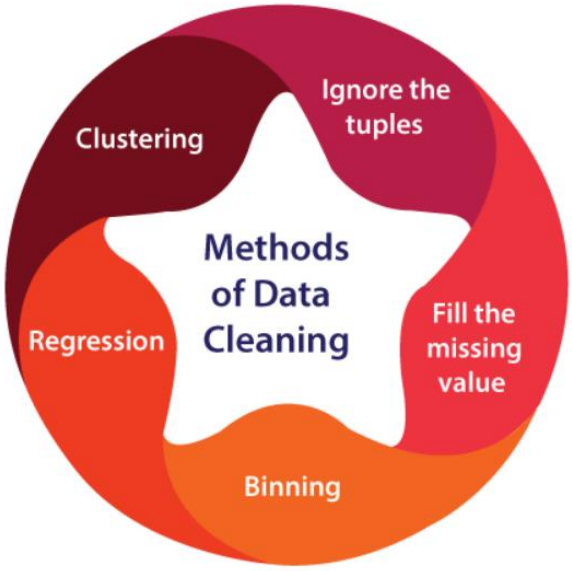
8. Incorporation of Background Knowledge

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.

b) Discuss: i) Data cleaning ii) Data integration

Data cleaning is a crucial process in Data Mining. Data cleaning is fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

Generally, data cleaning reduces errors and improves data quality. Data mining is a key technique for data cleaning. Data cleaning is not simply about erasing information to make space for new data, but rather finding a way to maximize a data set's accuracy without necessarily deleting information.



Steps of Data Cleaning

1. **Remove duplicate or irrelevant observations**
2. **Fix structural errors**
3. **Filter unwanted outliers**
4. **Handle missing data**
5. **Validate and QA**

DATA INTEGRATION:

Data Integration is a data preprocessing technique that combines data from multiple heterogeneous data sources into a coherent data store and provides a unified view of the data. These sources may include multiple data cubes, databases, or flat files.

The goal of data integration is to make it easier to access and analyze data that is spread across multiple systems or platforms, in order to gain a more complete and accurate understanding of the data.

Data integration can be challenging due to the variety of data formats, structures, and semantics used by different data sources. Different data sources may use different data types, naming conventions, and schemas, making it difficult to combine the data into a single view. Data integration typically involves a combination of manual and automated processes, including data profiling, data mapping, data transformation, and data reconciliation.

There are mainly 2 major approaches for data integration – one is the “tight coupling approach” and another is the “loose coupling approach”.

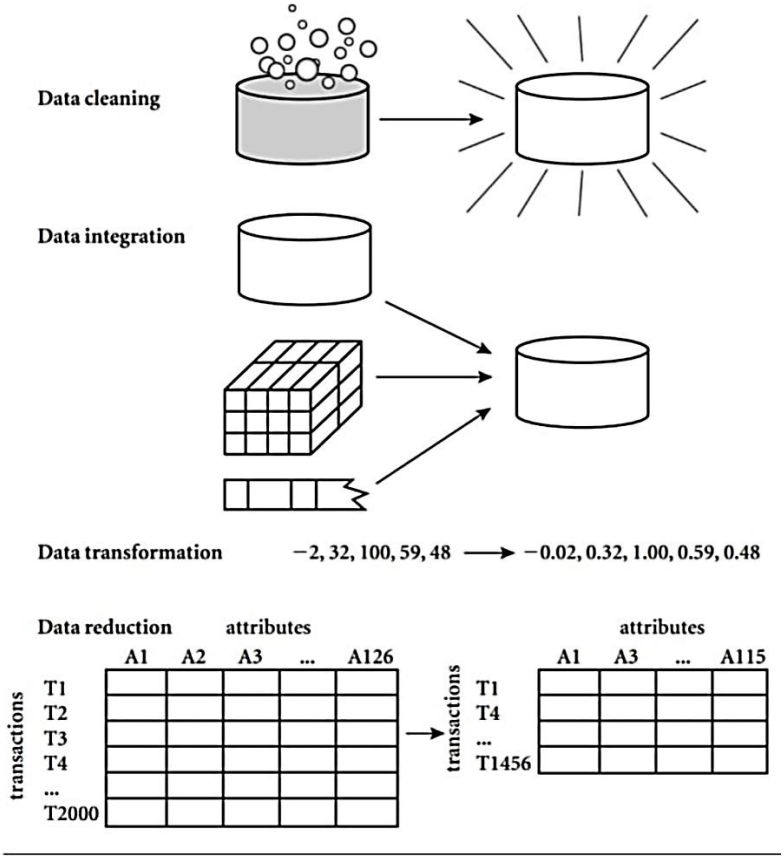
Tight Coupling:
This approach involves creating a centralized repository or data warehouse to store the integrated data. This approach is also known as data warehousing, and it enables data consistency and integrity, but it can be inflexible and difficult to change or update.

Loose Coupling:
This approach involves integrating data at the lowest level, such as at the level of individual data elements or records. This approach is also known as data federation, and it enables data flexibility and easy updates, but it can be difficult to maintain consistency and integrity across multiple data sources.

a)Explain Data preprocessing in detail.

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.Some common steps in data preprocessing include:

- Data cleaning: this step involves identifying and removing missing, inconsistent, or irrelevant data. This can include removing duplicate records, filling in missing values, and handling outliers.
- Data integration: this step involves combining data from multiple sources, such as databases, spreadsheets, and text files. The goal of integration is to create a single, consistent view of the data.
- Data transformation: this step involves converting the data into a format that is more suitable for the data mining task. This can include normalizing numerical data, creating dummy variables, and encoding categorical data.
- Data reduction: this step is used to select a subset of the data that is relevant to the data mining task. This can include feature selection (selecting a subset of the variables) or feature extraction (extracting new variables from the data).
- Data discretization: this step is used to convert continuous numerical data into categorical data, which can be used for decision tree and other categorical data mining techniques.



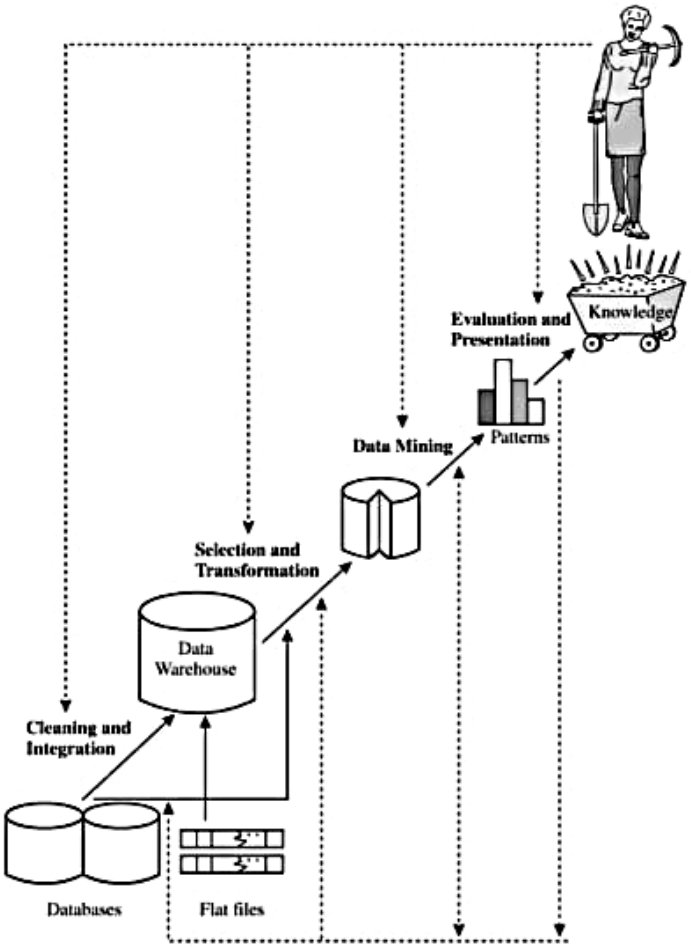
b) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

is depicted in Figure 1.4 and consists of an iterative sequence of the following steps:

1. **Data cleaning** (to remove noise and inconsistent data)
2. **Data integration** (where multiple data sources may be combined)[1]
3. **Data selection** (where data relevant to the analysis task are retrieved from the database)
4. **Data transformation** (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)[2]
5. **Data mining** (an essential process where intelligent methods are applied in order to extract data patterns)
6. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on some **interestingness measures**; Section 1.5)
7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. Note that according to this view, data mining is only one step in the entire process, albeit an essential one because it uncovers hidden patterns for evaluation.

We agree that data mining is a step in the knowledge discovery process. However, in industry, in media, and in the database research milieu, the term data mining is becoming more popular than the longer term of knowledge discovery from data. Therefore, in this book, we choose to use the term data mining. We adopt a broad view of data mining functionality: data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories.



a)Write any two applications of data mining.

Market Basket Analysis: Market Basket Analysis is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies and data mining techniques helps to achieve this analysis task. Example:

- Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
- Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Education: For analyzing the education sector, data mining uses Educational Data Mining (EDM) method. This method generates patterns that can be used both by learners and educators. By using data mining EDM we can perform some educational task:

- Predicting students admission in higher education
- Predicting students profiling
- Predicting student performance
- Teachers teaching performance
- Curriculum development
- Predicting student placement opportunities

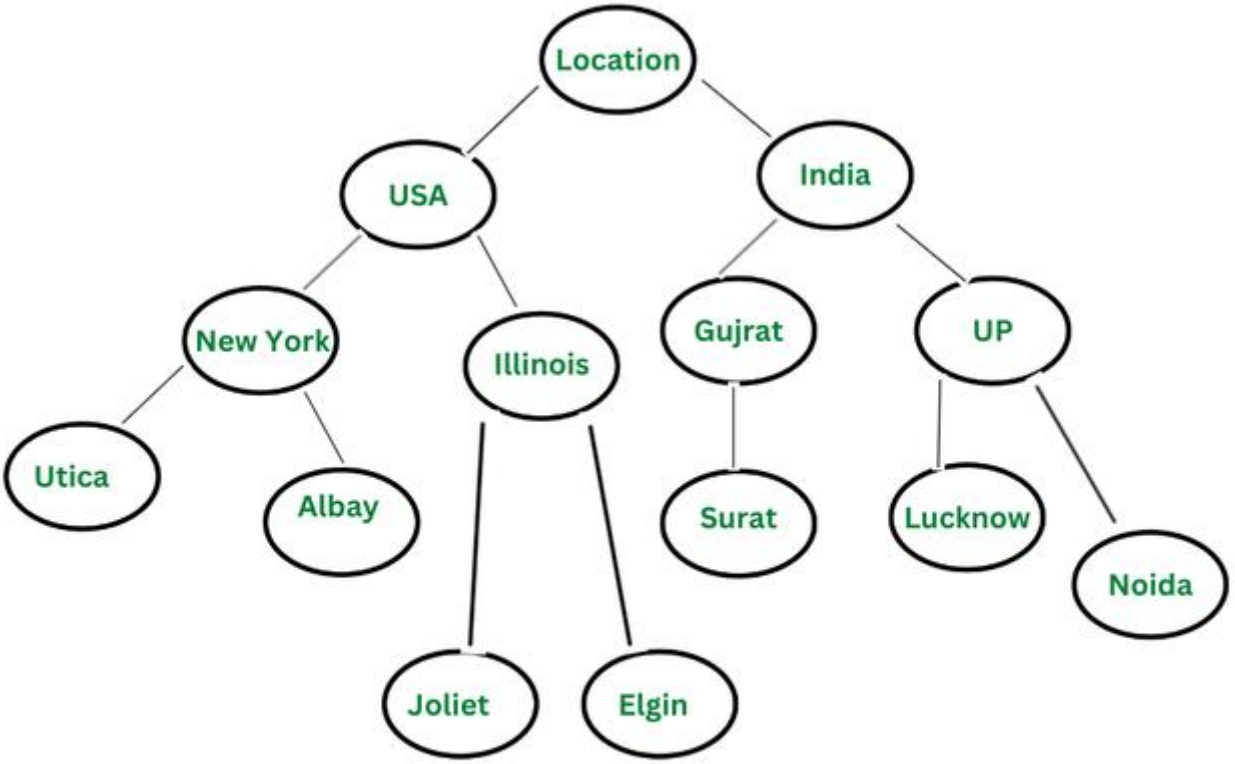
Financial/Banking Sector: A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product.

- Credit card fraud detection.
- Identify ‘Loyal’ customers.
- Extraction of information related to customers.
- Determine credit card spending by customer groups.

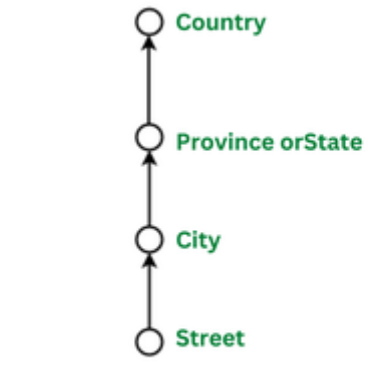
b)Explain the concept Hierarchies in detail using example.

In data mining, the concept of a concept hierarchy refers to the organization of data into a tree-like structure, where each level of the hierarchy represents a concept that is more general than the level below it. This hierarchical organization of data allows for more efficient and effective data analysis, as well as the ability to drill down to more specific levels of detail when needed. The concept of hierarchy is used to organize and classify data in a way that makes it more understandable and easier to analyze. The main idea behind the concept of hierarchy is that the same data can have different levels of granularity or levels of detail and that by organizing the data in a hierarchical fashion, it is easier to understand and perform analysis.

Example:



Concept Hierarchy for Dimension Location



Hierarchical Structure for Dimension Location

Explanation:

As shown in the above diagram, it consists of a concept hierarchy for the dimension location, where the user can easily retrieve the data. In order to evaluate it easily the data is represented in a tree-like structure. The top of the tree consists of the main dimension location and further splits into various sub-nodes. The root node is located, and it further splits into two nodes countries ie. USA and India. These countries are further then splitted into more sub-nodes, that represent the province states ie. New York, Illinois, Gujarat, UP. Thus the concept hierarchy as shown in the above example organizes the data into a tree-like structure and describes and represents in more general than the level below it.

The hierarchical structure represents the abstraction level of the dimension location, which consists of various footprints of the dimension such as street, city, province state, and country.

a) Enlist various types of data in cluster analysis.

Types Of Data In Cluster Analysis Are:

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

	1	0	*sum*
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

Let p=a+b+c+d

Simple matching coefficient (invariant, if the binary variable is symmetric):

$$d(i, j) = \frac{b + c}{a + b + c + d}$$

Jaccard coefficient (noninvariant if the binary variable is asymmetric):

$$d(i, j) = \frac{b + c}{a + b + c}$$

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.

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

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 xif by their rank,

$$r_{if} \in \{1, \ldots, M_f\}$$

By mapping the range of each variable onto [0, 1] by replacing the i-th object in the f-th variable by,

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

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-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.

b) Explain tree induction algorithm for building decision tree.

6.3.1 Decision Tree Induction

During the late 1970s and early 1980s, J. Ross Quinlan, a researcher in machine learning, developed a decision tree algorithm known as **ID3** (Iterative Dichotomiser). This work expanded on earlier work on *concept learning systems*, described by E. B. Hunt, J. Marin, and P. T. Stone. Quinlan later presented **C4.5** (a successor of ID3), which became a benchmark to which newer supervised learning algorithms are often compared. In 1984, a group of statisticians (L. Breiman, J. Friedman, R. Olshen, and C. Stone) published the book *Classification and Regression Trees* (**CART**), which described the generation of binary decision trees. ID3 and CART were invented independently of one another at around the same time, yet follow a similar approach for learning decision trees from training tuples. These two cornerstone algorithms spawned a flurry of work on decision tree induction.

ID3, C4.5, and CART adopt a greedy (i.e., nonbacktracking) approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. Most algorithms for decision tree induction also follow such a top-down approach, which

Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition D.

Input:

- Data partition, D, which is a set of training tuples and their associated class labels;
- *attribute_list*, the set of candidate attributes;
- *Attribute_selection_method*, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split point* or *splitting subset*.

Output: A decision tree.

Method:

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let Dj be the set of data tuples in D satisfying outcome j; // a partition
(12)     if Dj is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(Dj, attribute_list) to node N;
     endfor
(15) return N;
```

a)Discuss typical requirements of clustering in data mining

- **Scalability** – We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes** – Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape** – The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality** – The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space. A database or a data warehouse can include multiple dimensions or attributes. Some clustering algorithms are good at managing low-dimensional data, containing only two to three dimensions.
- **Ability to deal with noisy data** – Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** – The clustering results should be interpretable, comprehensible, and usable.

b) Explain 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 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.

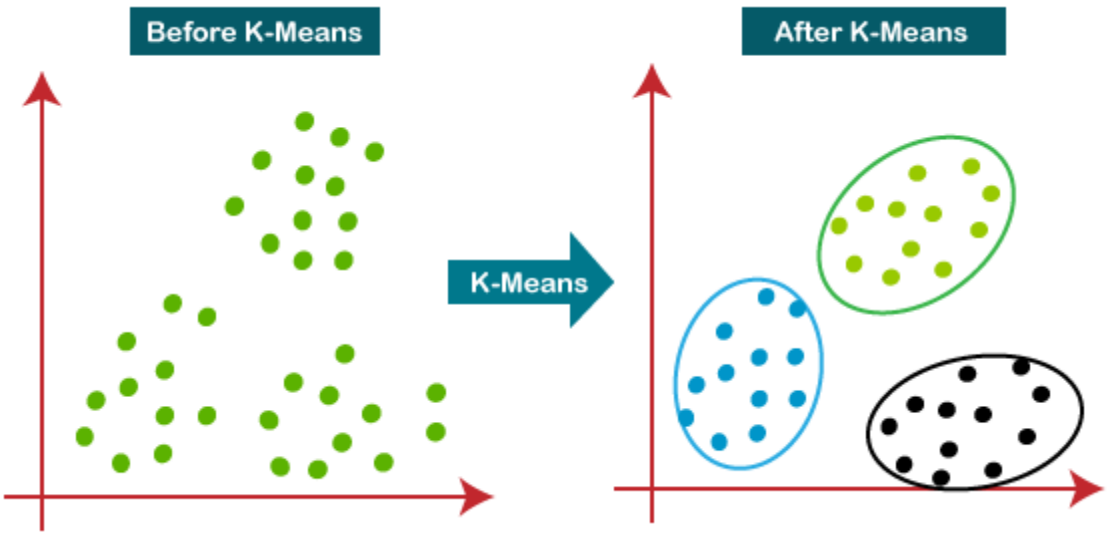
The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

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.

The below diagram explains the working of the K-means Clustering Algorithm:



How does the K-Means Algorithm Work?

The working of the K-Means algorithm is explained in the below steps:

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 means 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.