

Q.1	a) Explain feature of Datawarehouse (5 Marks)
	<p>A data warehouse is a centralized repository for storing and managing large amounts of data from various sources for analysis and reporting. It is optimized for fast querying and analysis, enabling organizations to make informed decisions by providing a single source of truth for data. Data warehousing typically involves transforming and integrating data from multiple sources into a unified, organized, and consistent format.</p> <div data-bbox="501 479 1327 860" data-label="Diagram"> <pre> graph LR A[Characteristic Of Data Warehouse] --> B[Subject Oriented] A --> C[Integrated] A --> D[Time-Variant] A --> E[Non-Volatile] </pre> </div> <ol style="list-style-type: none"> <p>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.</p> <p>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.</p> <p>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.</p> <p>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.</p> <p>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.</p> <p>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</p>

	<p>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. In this, data is read-only and refreshed at particular intervals. This is beneficial in analysing historical data and in comprehension the functionality. It does not need transaction process, recapture and concurrency control mechanism. Functionalities such as delete, update, and insert that are done in an operational application are lost in data warehouse environment. Two types of data operations done in the data warehouse are:</p> <ul style="list-style-type: none"> • Data Loading • Data Access <ol style="list-style-type: none"> 1. Subject Oriented: Focuses on a specific area or subject such as sales, customers, or inventory. 2. Integrated: Integrates data from multiple sources into a single, consistent format. 3. Read-Optimized: Designed for fast querying and analysis, with indexing and aggregations to support reporting. 4. Summary Data: Data is summarized and aggregated for faster querying and analysis. 5. Historical Data: Stores large amounts of historical data, making it possible to analyze trends and patterns over time. 6. Schema-on-Write: Data is transformed and structured according to a predefined schema before it is loaded into the data warehouse. 7. Query-Driven: Supports ad-hoc querying and reporting by business users, without the need for technical support.
b)	<p>What is data processing? Explain the different methods for the data integration phase. (5 Marks)</p> <p>Data in its raw form is not useful to any organization. Data processing is the method of collecting raw data and translating it into usable information. It is usually performed in a step-by-step process by a team of data scientists and data engineers in an organization. The raw data is collected, filtered, sorted, processed, analyzed, stored, and then presented in a readable format.</p> <p>Data processing is essential for organizations to create better business strategies and increase their competitive edge. By converting the data into readable formats like graphs, charts, and documents, employees throughout the organization can understand and use the data.</p> <div data-bbox="715 1464 1104 1872" data-label="Diagram"> <pre> graph TD Collection --> Preparation Preparation --> Input Input --> Processing Processing --> Output Output --> Storage Storage --> Collection </pre> </div> <p>Generally, there are six main steps in the data processing cycle:</p> <p>Step 1: Collection</p>
Ans:	

	<p>The collection of raw data is the first step of the data processing cycle. The type of raw data collected has a huge impact on the output produced. Hence, raw data should be gathered from defined and accurate sources so that the subsequent findings are valid and usable. Raw data can include monetary figures, website cookies, profit/loss statements of a company, user behavior, etc.</p> <p>Step 2: Preparation</p> <p>Data preparation or data cleaning is the process of sorting and filtering the raw data to remove unnecessary and inaccurate data. Raw data is checked for errors, duplication, miscalculations or missing data, and transformed into a suitable form for further analysis and processing. This is done to ensure that only the highest quality data is fed into the processing unit.</p> <p>The purpose of this step to remove bad data (redundant, incomplete, or incorrect data) so as to begin assembling high-quality information so that it can be used in the best possible way for business intelligence.</p> <p>Step 3: Input</p> <p>In this step, the raw data is converted into machine readable form and fed into the processing unit. This can be in the form of data entry through a keyboard, scanner or any other input source.</p> <p>Step 4: Data Processing</p> <p>In this step, the raw data is subjected to various data processing methods using machine learning and artificial intelligence algorithms to generate a desirable output. This step may vary slightly from process to process depending on the source of data being processed (data lakes, online databases, connected devices, etc.) and the intended use of the output.</p> <p>Step 5: Output</p> <p>The data is finally transmitted and displayed to the user in a readable form like graphs, tables, vector files, audio, video, documents, etc. This output can be stored and further processed in the next data processing cycle.</p> <p>Step 6: Storage</p> <p>The last step of the data processing cycle is storage, where data and metadata are stored for further use. This allows for quick access and retrieval of information whenever needed, and also allows it to be used as input in the next data processing cycle directly.</p> <p>Now that we have learned what is data processing and its cycle, now we can look at the types.</p>
c)	What is hierarchical clustering? Explain divisive clustering . (5 Marks)
Ans:	<p>Hierarchical clustering is a connectivity-based clustering model that groups the data points together that are close to each other based on the measure of similarity or distance. The assumption is that data points that are close to each other are more similar or related than data points that are farther apart.</p> <p>A dendrogram, a tree-like figure produced by hierarchical clustering, depicts the hierarchical relationships between groups. Individual data points are located at the bottom of the dendrogram, while the largest clusters, which include all the data points, are located at the</p>

top. In order to generate different numbers of clusters, the dendrogram can be sliced at various heights.

The dendrogram is created by iteratively merging or splitting clusters based on a measure of similarity or distance between data points. Clusters are divided or merged repeatedly until all data points are contained within a single cluster, or until the predetermined number of clusters is attained.

We can look at the dendrogram and measure the height at which the branches of the dendrogram form distinct clusters to calculate the ideal number of clusters. The dendrogram can be sliced at this h Hierarchical Divisive clustering

It is also known as a top-down approach. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been split into singleton clusters.

Algorithm :

given a dataset (d1, d2, d3,dN) of size N

at the top we have all data in one cluster

the cluster is split using a flat clustering method eg. K-Means etc

repeat

choose the best cluster among all the clusters to split

split that cluster by the flat clustering algorithm

until each data is in its own singleton cluster

Hierarchical Divisive clustering - Geeksforgeeks

Hierarchical Divisive clustering

Computing Distance Matrix

While merging two clusters we check the distance between two every pair of clusters and merge the pair with the least distance/most similarity. But the question is how is that distance determined. There are different ways of defining Inter Cluster distance/similarity. Some of them are:

Min Distance: Find the minimum distance between any two points of the cluster.

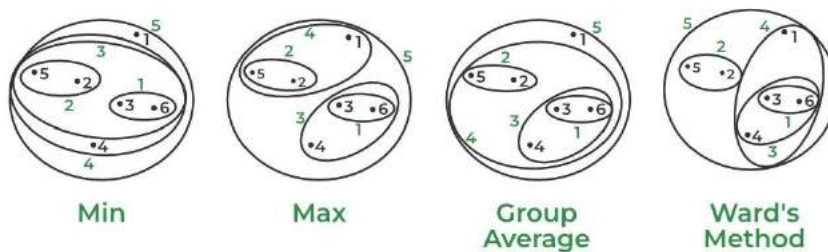
Max Distance: Find the maximum distance between any two points of the cluster.

Group Average: Find the average distance between every two points of the clusters.

Ward's Method: The similarity of two clusters is based on the increase in squared error when two clusters are merged.

For example, if we group a given data using different methods, we may get different results:

Distance Matrix Comparision in Hierarchical Clustering



d) Define metadata and explain the types of metadata. (5 Marks)

Metadata contains information about a data asset, such as properties, origin, history, location, creation, ownership, and versions. It offers this additional information about a data asset to inform users of the asset's meaning, and can be an important element for maintaining compliance with regulatory requirements.

For example, a digital image's metadata can include information related to the image's resolution, size, color depth, and time of creation. This information can be used for [data classification](#), labeling, organization, sorting, tracking, searching, and analysis.

Types of Metadata

Structural Metadata

Structural metadata contains useful information that helps in the establishment of object relationships. The characteristics of structural metadata include:

1. Enables people to comprehend and successfully employ the data resource.
2. Describes the hierarchical structures that exist between various data resources.

Ans:

3. Improves the display and navigation of obtained information using page-turning software. It depends on how customer images are delivered and stored.

Descriptive Metadata

Descriptive metadata is important for identifying and distinguishing a data resource. It includes details regarding the data's context and content. Descriptive metadata is structured and frequently follows one or more established standard schemes. At the system level, descriptive metadata enables users to search for and obtain information.

Preservation Metadata

Metadata can assist the process of sustaining a digital object or file. This data may be needed to [access a file](#). Metadata preserves a digital file or object from beginning to end. One of the common patterns is Preservation Metadata Implementation Strategies. It emphasizes preservation and maintenance fundamentals.

Provenance Metadata

Provenance metadata gives useful information about a data resource's origins. The characteristics of provenance metadata include:

1. The data, such as data ownership, transformations, consumption, and archival, facilitates monitoring a resource's lifecycle.
2. Provenance metadata is established when a new version of a set of data is created. It describes the relationship among various versions of data objects.
3. Users can query the relationship among versions and provide fine or coarse-grained provenance information on collected data.

Definitional Metadata

Definitional metadata is data that establishes a shared lexicon for comprehending the significance of the data. Semantic and schematic are metadata types. Textual vocabularies can describe organized and unstructured data semantically. Schemas display database data in a structured format.

Administrative Metadata

Administrative metadata describes a file's constraints. Administrators can restrict file access using this data. Administrative metadata provides complete details regarding data. Users can manage a variety [of data files](#).

Administrative metadata is analogous to a basic version of data. Even if a data set is incredibly complex, its metadata will be significantly more detailed. Therefore, administrative metadata is concerned with control, specifically managing and simplifying these complex elements.

Q.2	
a)	Explain association rule mining and multilevel association rules giving example of multidimensional association rules. (10 Marks)
Ans:	<p>MULTILEVEL ASSOCIATION RULES:</p> <ul style="list-style-type: none"> • Association rules generated from mining data at multiple levels of abstraction are called multiple-level or multilevel association rules. • Multilevel association rules can be mined efficiently using concept hierarchies under a support-confidence framework. • Rules at high concept level may add to common sense while rules at low concept level may not be useful always. <ul style="list-style-type: none"> ◦ Using uniform minimum support for all levels: • When a uniform minimum support threshold is used, the search procedure is simplified. • The method is also simple, in that users are required to specify only one minimum support threshold. • The same minimum support threshold is used when mining at each level of abstraction. • For example, in Figure, a minimum support threshold of 5% is used throughout. • (e.g. for mining from “computer” down to “laptop computer”). • Both “computer” and “laptop computer” are found to be frequent, while “desktop computer” is not. • Using reduced minimum support at lower levels: <ul style="list-style-type: none"> ◦ Each level of abstraction has its own minimum support threshold. ◦ The deeper the level of abstraction, the smaller the corresponding threshold is. ◦ For example in Figure, the minimum support thresholds for levels 1 and 2 are 5% and 3%, respectively. ◦ In this way, “computer,” “laptop computer,” and “desktop computer” are all considered frequent. <p>Multilevel Association rule consists of alternate search strategies and Controlled level cross filtering:</p> <p>1. Alternate Search Strategies:</p> <ul style="list-style-type: none"> • Level by level independent: <ul style="list-style-type: none"> ◦ Full breadth search. ◦ No background knowledge in pruning. ◦ Leads to examine lot of infrequent items. • Level-cross filtering by single item: <ul style="list-style-type: none"> ◦ Examine nodes at level i only if node at level (i-1) is frequent. ◦ Misses frequent items at lower level abstractions (due to reduced support). • Level-cross filtering by k-item set: <ul style="list-style-type: none"> ◦ Examine k-itemsets at level i only if k-itemsets at level (i-1) is frequent. ◦ Misses frequent k-itemsets at lower level abstractions (due to reduced support). • Controlled Level-cross filtering by single item: <ul style="list-style-type: none"> ◦ A modified level-cross filtering by single item. ◦ Sets a level passage threshold for every level. • Allows the inspection of lower abstractions even if its ancestor fails to satisfy min_sup threshold.

MULTIDIMENSIONAL ASSOCIATION RULES:

1. In Multi dimensional association:

- Attributes can be categorical or quantitative.
- Quantitative attributes are numeric and incorporates hierarchy.
- Numeric attributes must be discretized.
- Multi dimensional association rule consists of more than one dimension:

Eg: buys(X,"IBM Laptop computer")buys(X,"HP Inkjet Printer")

2. Three approaches in mining multi dimensional association rules:

1. Using static discretization of quantitative attributes.

- Discretization is static and occurs prior to mining.
- Discretized attributes are treated as categorical.
- Use apriori algorithm to find all k-frequent predicate sets (this requires k or k+1 table scans).
- Every subset of frequent predicate set must be frequent.
- Eg: If in a data cube the 3D cuboid (age, income, buys) is frequent implies (age, income), (age, buys), (income, buys) are also frequent.
- Data cubes are well suited for mining since they make mining faster.
- The cells of an n-dimensional data cuboid correspond to the predicate cells.

2. Using dynamic discretization of quantitative attributes:

- Known as mining Quantitative Association Rules.
- Numeric attributes are dynamically discretized.
- Eg: age(X,"20..25") \wedge income(X,"30K..41K")buys (X,"Laptop Computer")

	Age=20	Age=21	Age=22	Age=23	Age=24	Age=25
Income,38 to 41						
Income,34 to 37						
Income,30 to 33						

GRID FOR TUPLES

3. Using distance based discretization with clustering.

This is dynamic discretization process that considers the distance between data points.

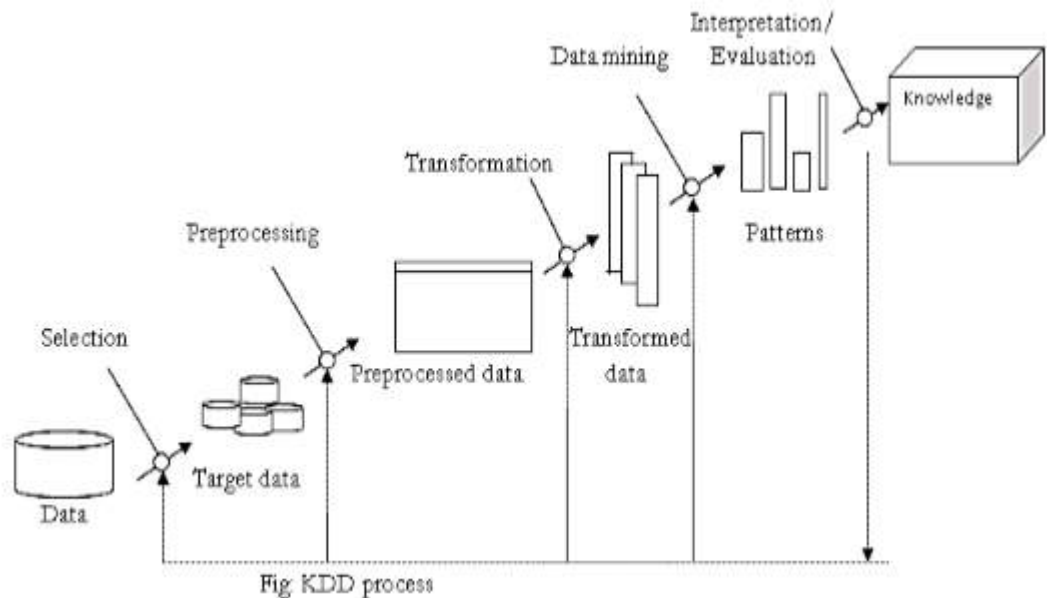
- It involves a two step mining process:
 - Perform clustering to find the interval of attributes involved.
 - Obtain association rules by searching for groups of clusters that occur together.
- The resultant rules may satisfy:
 - Clusters in the rule antecedent are strongly associated with clusters of rules in the consequent.
 - Clusters in the antecedent occur together.
 - Clusters in the consequent occur together.

b)

Give Data mining as a step in KDD. Give the architecture of typical data mining system. (10 Marks)

KDD Process (Knowledge Discovery in Database):

- The term KDD refers to the broad process of finding knowledge in data, and emphasizes the high level application of particular data mining methods.
- The goal of the KDD process is to extract knowledge from data in the context of large databases.



- The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:
 1. Developing an understanding of:
 - The application domain
 - The relevant prior knowledge
 - The goals of end user
 2. Creating a target data set:
 - Selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed.
 3. Data cleaning and preprocessing:
 - Removal of noise or outliers.
 - Strategies for handling missing data fields.
 4. Data reduction and projection:
 - Finding useful features to represent the data depending on the goal of the task.
 5. Choosing the data mining task:
 - Deciding whether the goal of the KDD process is classification, regression, clustering, etc.
 6. Choosing the data mining algorithm:
 - Selecting methods to be used for searching the pattern in the data.

Ans:

- Deciding which models and parameters may be appropriate.
 - Matching a particular data mining method with the overall criteria of the KDD process.
7. Data mining:
 - Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.
 8. Interpreting mined patterns
 9. Consolidating discovered knowledge

Architecture of Typical Data mining system

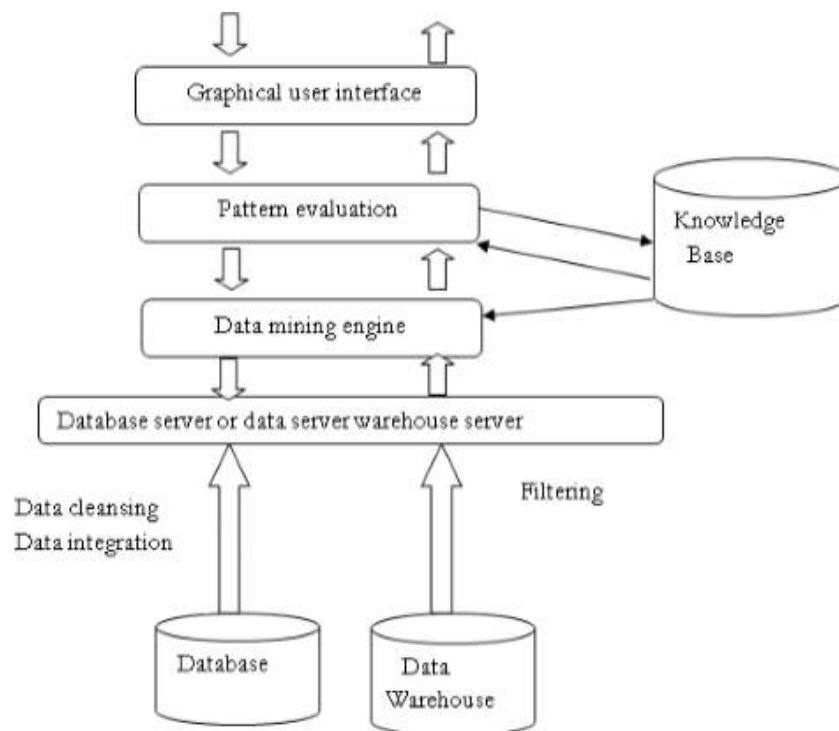
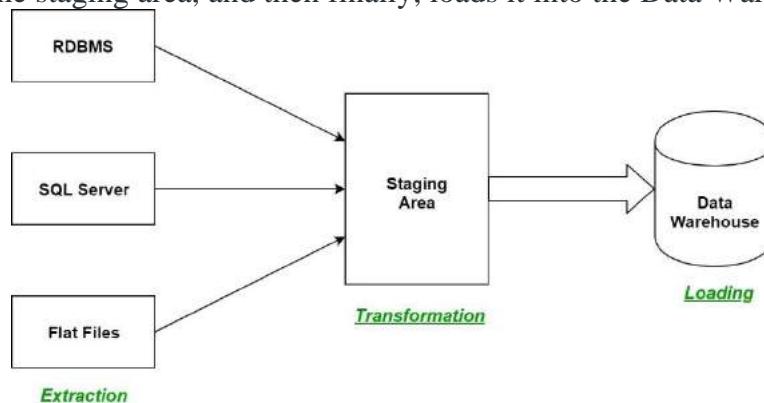


Fig. Architecture of Typical Data mining system

- Architecture of a typical data mining system may have the following major components as shown in fig:
 1. Database, data warehouse, or other information repository:
 - This is information repository.
 - Data cleaning and data integration techniques may be performed on the data.
 2. Databases or data warehouse server:
 - It fetches the data as per the users' requirement which one need for data mining task.
 3. Knowledge base:
 - This is used to guide the search, and gives the interesting and hidden patterns from data.
 4. Data mining engine:

	<ul style="list-style-type: none"> ○ It performs the data mining task such as characterization, association, classification, cluster analysis etc. <p>5. Pattern evaluation module:</p> <ul style="list-style-type: none"> ○ It is integrated with the mining module and it give the search of only the interesting patterns. <p>6. Graphical user interface:</p> <ul style="list-style-type: none"> ○ This module is used to communicate between user and the data mining system and allow users to browse databases or data warehouse schemas.
Q.3)	
A	Explain Extraction and Transformation in ETL process. (10 Marks)
	<ol style="list-style-type: none"> 1. ETL stands for Extract, Transform, Load and it is a process used in data warehousing to extract data from various sources, transform it into a format suitable for loading into a data warehouse, and then load it into the warehouse. The process of ETL can be broken down into the following three stages: 2. Extract: The first stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets, and flat files. This step involves reading data from the source systems and storing it in a staging area. 3. Transform: In this stage, the extracted data is transformed into a format that is suitable for loading into the data warehouse. This may involve cleaning and validating the data, converting data types, combining data from multiple sources, and creating new data fields. 4. Load: After the data is transformed, it is loaded into the data warehouse. This step involves creating the physical data structures and loading the data into the warehouse. 5. The ETL process is an iterative process that is repeated as new data is added to the warehouse. The process is important because it ensures that the data in the data warehouse is accurate, complete, and up-to-date. It also helps to ensure that the data is in the format required for data mining and reporting. <p>Additionally, there are many different ETL tools and technologies available, such as Informatica, Talend, DataStage, and others, that can automate and simplify the ETL process.</p> <p>ETL is a process in Data Warehousing and it stands for Extract, Transform and Load. It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area, and then finally, loads it into the Data Warehouse system.</p>  <pre> graph LR RDBMS[RDBMS] --> StagingArea[Staging Area] SQLServer[SQL Server] --> StagingArea FlatFiles[Flat Files] --> StagingArea StagingArea -- Transformation --> DataWarehouse[(Data Warehouse)] DataWarehouse -- Loading --> DataWarehouse </pre> <p>Let us understand each step of the ETL process in-depth:</p>

1. **Extraction:**

The first step of the ETL process is extraction. In this step, data from various source systems is extracted which can be in various formats like relational databases, No SQL, XML, and flat files into the staging area. It is important to extract the data from various source systems and store it into the staging area first and not directly into the data warehouse because the extracted data is in various formats and can be corrupted also. Hence loading it directly into the data warehouse may damage it and rollback will be much more difficult. Therefore, this is one of the most important steps of ETL process.

2. **Transformation:**

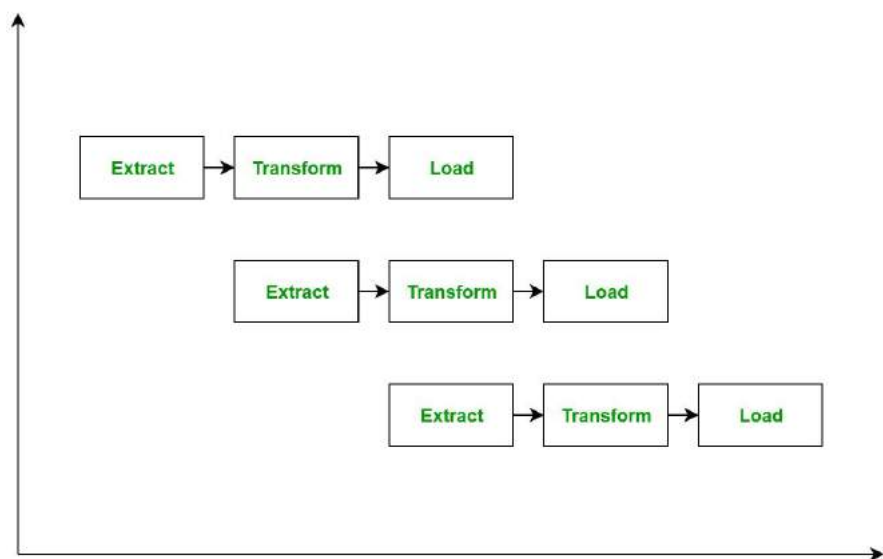
The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format. It may involve following processes/tasks:

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

3. **Loading:**

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

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



ETL Tools: Most commonly used ETL tools are **Hevo**, Sybase, Oracle Warehouse builder, CloverETL, and MarkLogic.

Data Warehouses: Most commonly used Data Warehouses are **Snowflake**, Redshift, BigQuery, and Firebolt.

b)	illustrate multidimensional association rules with suitable example.(10 Marks)
Ans:	
	<p>Multidimensional Association Rules :</p> <p>Multidimensional association rules involve analyzing the relationships between three or more variables. These rules are useful for discovering more complex relationships between items, such as the association between a customer's age, gender, and purchasing habits. Multidimensional association rules can be represented as "if A and B, then C" statements, where A and B are the antecedents and C is the consequent. For example, a multidimensional association rule could be "If a customer is female, over 30 years old, and has previously purchased skincare products, they are likely to also purchase anti-aging products."</p> <p>In Multi dimensional association rule Qualities can be absolute or quantitative.</p> <ul style="list-style-type: none"> • Quantitative characteristics are numeric and consolidates order. • Numeric traits should be discretized. • Multi dimensional affiliation rule comprises of more than one measurement. • Example –buys(X, "IBM Laptop computer")buys(X, "HP Inkjet Printer") <p>Approaches in mining multi dimensional affiliation rules :</p> <p>Three approaches in mining multi dimensional affiliation rules are as following.</p> <ol style="list-style-type: none"> 1. Using static discretization of quantitative qualities : <ul style="list-style-type: none"> • Discretization is static and happens preceding mining. • Discretized ascribes are treated as unmitigated. • Use apriori calculation to locate all k-regular predicate sets(this requires k or k+1 table outputs). Each subset of regular predicate set should be continuous. <p>Example –</p> <p>If in an information block the 3D cuboid (age, pay, purchases) is continuous suggests (age, pay), (age, purchases), (pay, purchases) are likewise regular.</p> <p>Note –</p> <p>Information blocks are appropriate for mining since they make mining quicker. The cells of an n-dimensional information cuboid relate to the predicate cells.</p> 2. Using powerful discretization of quantitative traits : <ul style="list-style-type: none"> • Known as mining Quantitative Association Rules. • Numeric properties are progressively discretized. <p>Example –:</p> <p>age(X, "20..25") \wedge income(X, "30K..41K")buys (X, "Laptop Computer")</p> 3. Grid FOR TUPLES : <p>Using distance based discretization with bunching –</p> <p>This id dynamic discretization measure that considers the distance between information focuses. It includes a two stage mining measure as following.</p> <ul style="list-style-type: none"> • Perform bunching to discover the time period included. • Get affiliation rules via looking for gatherings of groups that happen together. <p>The resultant guidelines may fulfill –</p> <ul style="list-style-type: none"> • Bunches in the standard precursor are unequivocally connected with groups of rules in the subsequent. • Bunches in the forerunner happen together. • Bunches in the ensuing happen together.

Q.4	
a)	Define classification, issues of classification and explain Naïve Bayesian classification with example (10 Marks)
	<p>It is a classification technique based on Bayes' Theorem with an independence assumption among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.</p> <p>The Naïve Bayes classifier is a popular supervised machine learning algorithm used for classification tasks such as text classification. It belongs to the family of generative learning algorithms, which means that it models the distribution of inputs for a given class or category. This approach is based on the assumption that the features of the input data are conditionally independent given the class, allowing the algorithm to make predictions quickly and accurately.</p> <p>In statistics, naive Bayes classifiers are considered as simple probabilistic classifiers that apply Bayes' theorem. This theorem is based on the probability of a hypothesis, given the data and some prior knowledge. The naive Bayes classifier assumes that all features in the input data are independent of each other, which is often not true in real-world scenarios. However, despite this simplifying assumption, the naive Bayes classifier is widely used because of its efficiency and good performance in many real-world applications.</p> <p>Moreover, it is worth noting that naive Bayes classifiers are among the simplest Bayesian network models, yet they can achieve high accuracy levels when coupled with kernel density estimation. This technique involves using a kernel function to estimate the probability density function of the input data, allowing the classifier to improve its performance in complex scenarios where the data distribution is not well-defined. As a result, the naive Bayes classifier is a powerful tool in machine learning, particularly in text classification, spam filtering, and sentiment analysis, among others.</p> <p>For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features,</p>

all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

An NB model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes' Theorem:

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

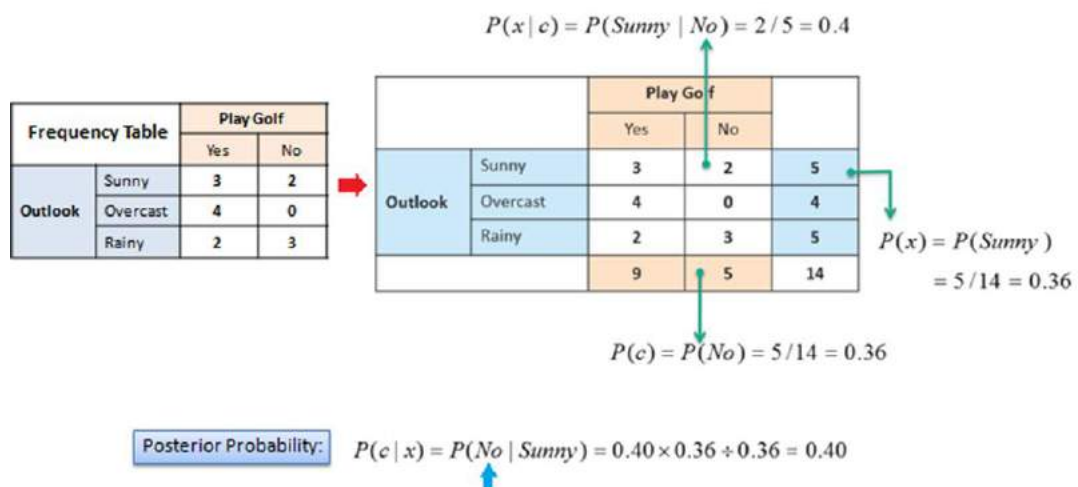
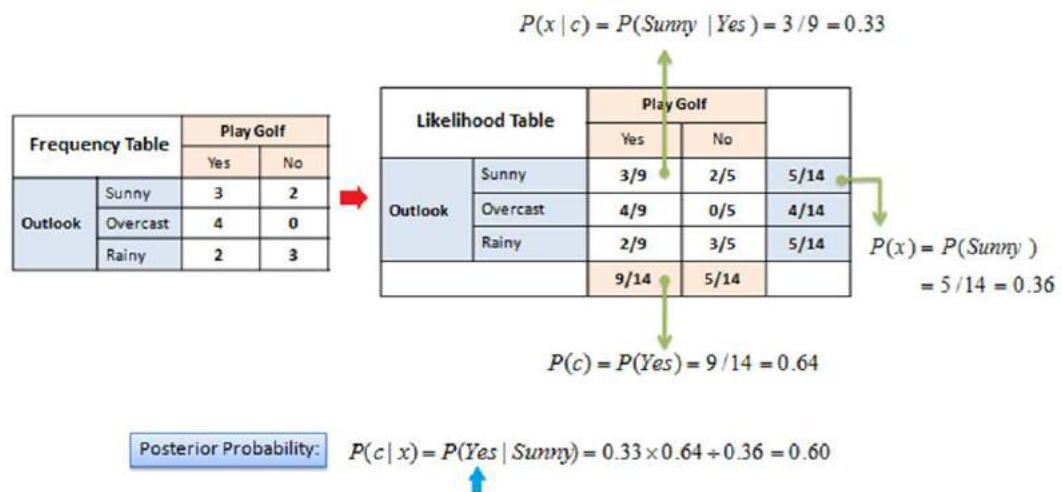
P(B) is Marginal Probability: Probability of Evidence.

Example 1:

We use the same simple Weather dataset here.

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

The posterior probability can be calculated by first, constructing a frequency table for each attribute against the target. Then, transforming the frequency tables to likelihood tables and finally use the Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction



The likelihood tables for all four predictors.

Frequency Table				Likelihood Table			
		Play Golf				Play Golf	
		Yes	No			Yes	No
Outlook	Sunny	3	2	Outlook	Sunny	3/9	2/5
	Overcast	4	0		Overcast	4/9	0/5
	Rainy	2	3		Rainy	2/9	3/5
		Play Golf				Play Golf	
		Yes	No			Yes	No
Humidity	High	3	4	Humidity	High	3/9	4/5
	Normal	6	1		Normal	6/9	1/5
		Play Golf				Play Golf	
		Yes	No			Yes	No
Temp.	Hot	2	2	Temp.	Hot	2/9	2/5
	Mild	4	2		Mild	4/9	2/5
	Cool	3	1		Cool	3/9	1/5
		Play Golf				Play Golf	
		Yes	No			Yes	No
Windy	False	6	2	Windy	False	6/9	2/5
	True	3	3		True	3/9	3/5

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(\text{Yes} | X) = P(\text{Rainy} | \text{Yes}) \times P(\text{Cool} | \text{Yes}) \times P(\text{High} | \text{Yes}) \times P(\text{True} | \text{Yes}) \times P(\text{Yes})$$

$$P(\text{Yes} | X) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529 \rightarrow 0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

$$P(\text{No} | X) = P(\text{Rainy} | \text{No}) \times P(\text{Cool} | \text{No}) \times P(\text{High} | \text{No}) \times P(\text{True} | \text{No}) \times P(\text{No})$$

$$P(\text{No} | X) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.02057 \rightarrow 0.8 = \frac{0.02057}{0.02057 + 0.00529}$$

- b) Find the mean, median, mode, midrange, variance of data
13,15,16,16,19,20,20,21,22,25,26,26,26,30,33,36,40,45,46,52,52,70 (10 Marks)

a) Mean:-

$$\text{Mean} = \frac{\text{Sum of the number}}{\text{Total no. of observations}}$$

$$= \frac{13+15+16+16+19+20+20+21+22+25+26+26+26+30+33+36+40+45+46+52+52+70}{22}$$

$$= 30.40$$

$$\text{Mean} = 30.40$$

b) Median:

arrange data in ascending order

$$\text{Median} = \frac{26+26}{2} = 26$$

c) Mode:

The only number which appears multiple times

16, 20, 26, 52

d) midrange:-

subtract the smallest number from the largest number

$$= 70 - 13$$

$$= 57$$

$$\therefore \text{Range} = 57$$

e) Variance

Data score	Data - mean	Difference from mean
13	13 - 30.40	- 17.4
15	15 - 30.40	- 15.4
16	16 - 30.40	
16	16 - 30.40	
19	19	
20	20	
20	20	
21	21	
22	22	
25	25	
26	26	

Next we square each of these differences and then sum them

Difference	Difference squared
- 17.4	302.76
- 15.4	237.16
.	
.	
.	

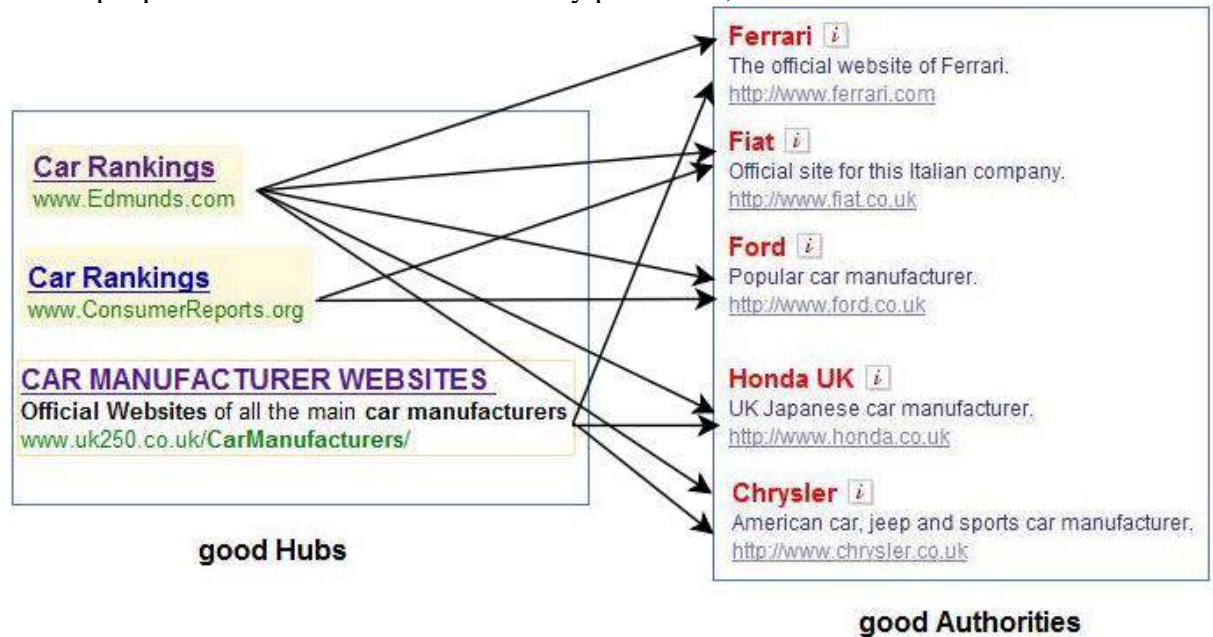
The sum of the squares is
next we find the "mean" of this sum

$$\text{Variance} = \frac{\quad}{22} = 210.696$$

Ans:

Q.5)	Explain HITS algorithm and illustrate its working. (10 Marks)
	<p>In the same time that PageRank was being developed, Jon Kleinberg a professor in the Department of Computer Science at Cornell came up with his own solution to the Web Search problem. He developed an algorithm that made use of the link structure of the web in order to discover and rank pages relevant for a particular topic. HITS (<i>hyperlink-induced topic search</i>) is now part of the Ask search engine (www.Ask.com).</p> <p>One of the interesting points that he brought up was that the human perspective on how a search process should go is more complex than just compare a list of query words against a list of documents and return the matches. Suppose we want to buy a car and type in a general query phrase like "the best automobile makers in the last 4 years", perhaps with the intention to get back a list of top car brands and their official web sites. When you ask this question to your friends, you expect them to be able to understand that automobile means car, vehicle, and that automobile is a general concept that includes vans, trucks, and other type of cars. When you ask this question to a computer that is running a text based ranking algorithm, things might be very different. That computer will count all occurrences of the given words in a given set of documents, but will not do intelligent rephrasing for you. The list of top pages we get back, while algorithmically correct, might be very different than what expected. One problem is that most official web sites are not enough self descriptive. They might not advertise themselves the way general public perceives them. Top companies like Hunday, Toyota, might not even use the terms "automobile makers" on their web sites. They might use the term "car manufacturer" instead, or just describe their products and their business.</p> <p>What is to be done in this case? It would be of course great if computers could have a dictionary or ontology, such that for any query, they could figure out sinonimes, equivalent meanings of phrases. This might improve the quality of search, nevertheless, in the end, we would still have a text based ranking system for the web pages. We would still be left with the initial problem of sorting the huge number of pages that are relevant to the different meanings of the query phrase. We can easily convince ourselves that this is the case. Just remember one of our first examples, about a page that repeats the phrase "automobile makers = cars manufacturers = vehicle designers" a billion times. This web page would be the first one displayed by the query engine. Nevertheless, this page contains practically no usable information.</p> <p>The conclusion is that even if trying to find pages that contain the query words should be the starting point, a different ranking system is needed in order to find those pages that are authoritative for a given query. Page <i>i</i> is called an authority for the query "automobile makers" if it contains valuable information on the subject. Official web sites of car manufacturers, such as www.bmw.com, HyundaiUSA.com, www.mercedes-benz.com would be authorities for this search. Commercial web sites selling cars might be authorities on the subject as well. These are the ones truly relevant to the given query. These are the ones that the user expects back from the query engine. However, there is a second category of pages relevant to the process of finding the authoritative pages, called hubs. Their role is to advertise the authoritative pages. They contain useful links towards the authoritative pages. In other words, hubs point the search engine in the "right direction". In real life, when you buy a car, you are more inclined to purchase it from a certain dealer that your friend recommends. Following the analogy, the authority in this case would be the car dealer, and the hub would be your friend. You trust your friend, therefore you trust what your friend recommends. In the world wide web,</p>

hubs for our query about automobiles might be pages that contain rankings of the cars, blogs where people discuss about the cars that they purchased, and so on.



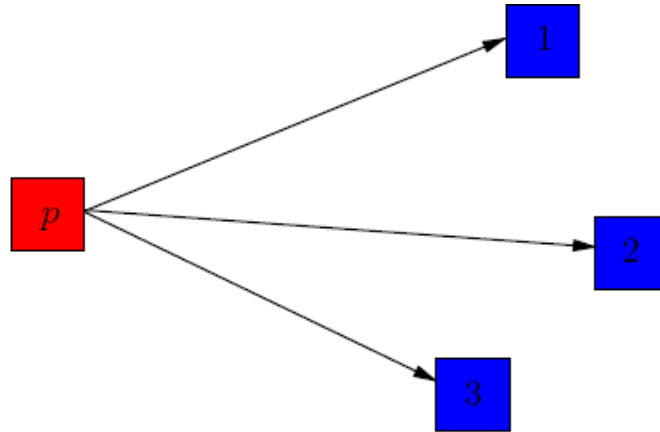
Query: Top automobile makers

Jon Kleinberg's algorithm called **HITS** identifies good authorities and hubs for a topic by assigning two numbers to a page: an authority and a hub weight. These weights are defined recursively. A higher authority weight occurs if the page is pointed to by pages with high hub weights. A higher hub weight occurs if the page points to many pages with high authority weights.

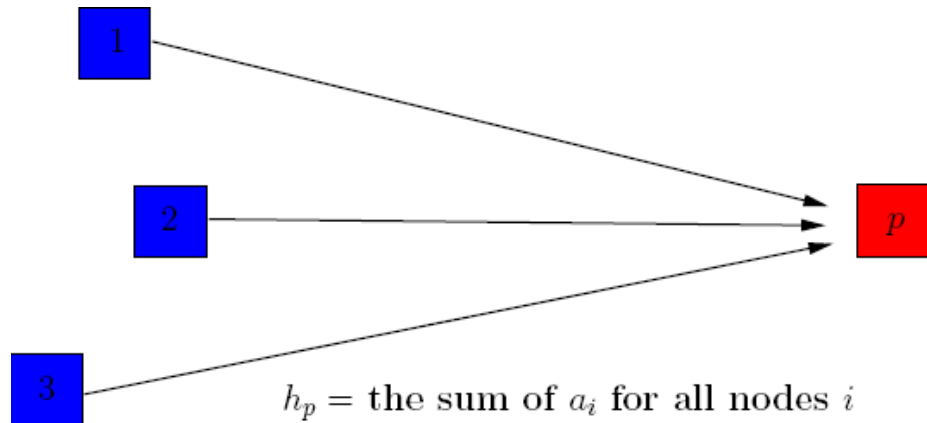
In order to get a set rich in both hubs and authorities for a query Q , we first collect the top 200 documents that contain the highest number of occurrences of the search phrase Q . These, as pointed out before may not be of tremendous practical relevance, but one has to start somewhere. Kleinberg points out that the pages from this set called root (R_Q) are essentially very heterogeneous and in general contain only a few (if any) links to each other. So the web subgraph determined by these nodes is almost totally disconnected; in particular, we can not enforce Page Rank techniques on R_Q .

Authorities for the query Q are not extremely likely to be in the root set R_Q . However, they are likely to be pointed out by at least one page in R_Q . So it makes sense to extend the subgraph R_Q by including all edges coming from or pointing to nodes from R_Q . We denote by S_Q the resulting subgraph and call it the *seed* of our search. Notice that S_Q we have constructed is a reasonably small graph (it is certainly much smaller than the 30 billion nodes web graph!). It is also likely to contain a lot of authoritative sources for Q . The question that remains is how to recognize and rate them? Heuristically, authorities on the same topic should have a lot of common pages from S_Q pointing to them. Using our previous terminology, there should be a great overlap in the set of hubs that point to them.

From here on, we translate everything into mathematical language. We associate to each page i two numbers: an authority weight a_i , and a hub weight h_i . We consider pages with a higher a_i number as being better authorities, and pages with a higher h_i number as being better hubs. Given the weights $\{a_i\}$ and $\{h_i\}$ of all the nodes in S_Q , we dynamically update the weights as follows:



$a_p =$ the sum of h_i for all nodes i pointing to p



$h_p =$ the sum of a_i for all nodes i pointed to by p

A good hub increases the authority weight of the pages it points. A good authority increases the hub weight of the pages that point to it. The idea is then to apply the two operations above alternatively until equilibrium values for the hub and authority weights are reached.

Let A be the adjacency matrix of the graph S_Q and denote the authority weight vector by v and the hub weight vector by u , where

$$v = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \quad \text{and} \quad u = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix}$$

Let us notice that the two update operations described in the pictures translate

$$\begin{cases} v = A^t \cdot u \\ u = A \cdot v \end{cases}$$

to:

If we consider that the initial weights of the nodes

$$u_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{and} \quad v_0 = A^t \cdot \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

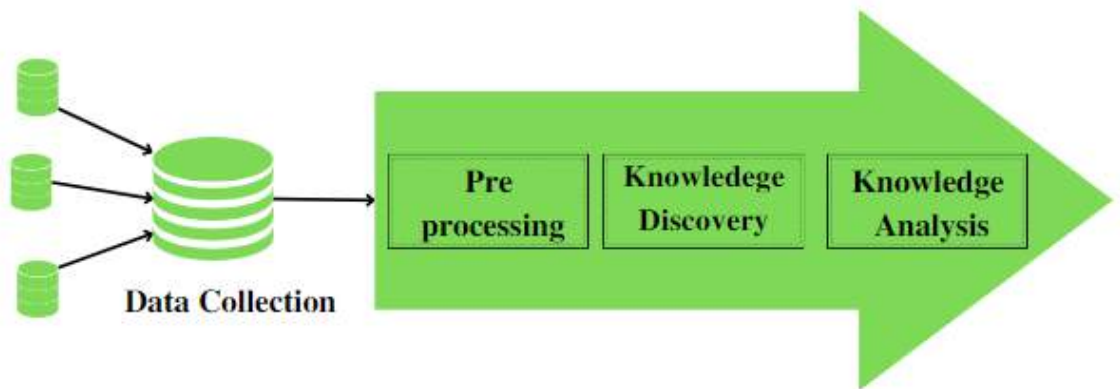
are

then, after k steps we get the system:

$$\begin{cases} v_k = (A^t \cdot A) \cdot v_{k-1} \\ u_k = (A \cdot A^t) \cdot u_{k-1} \end{cases}$$

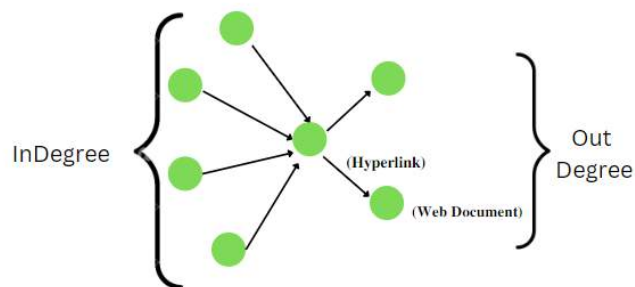
b) Explain Web structure mining in detail. (10 Marks)

Web Structure Mining is one of the three different types of techniques in Web Mining. In this article, we will purely discuss about the Web Structure Mining. Web Structure Mining is the technique of discovering structure information from the web. It uses graph theory to analyze the nodes and connections in the structure of a website.



Depending upon the type of Web Structural data, Web Structure Mining can be categorised into two types:

1.Extracting patterns from the hyperlink in the Web: The Web works through a system of hyperlinks using the [hyper text transfer protocol \(http\)](http). Hyperlink is a structural component that connects the web page according to different location. Any page can create a hyperlink of any other page and that page can also be linked to some other page. the intertwined or self-referral nature of web lends itself to some unique network analytical algorithms. The structure of Web pages could also be analyzed to examine the pattern of hyperlinks among pages.



2. Mining the document structure. It is the analysis of tree like structure of web page to describe HTML or XML usage or the tags usage . There are different terms associated with Web Structure Mining :

- **Web Graph:** Web Graph is the directed graph representing Web.
- **Node:** Node represents the web page in the graph.
- **Edge(s):** Edge represents the hyperlinks of the web page in the graph (Web graph)
- **In degree(s):** It is the number of hyperlinks pointing to a particular node in the graph.
- **Degree(s):** Degree is the number of links generated from a particular node. These are also called the Out Degrees.

All these terminologies will be more clear by looking at the following diagram of Web Graph:

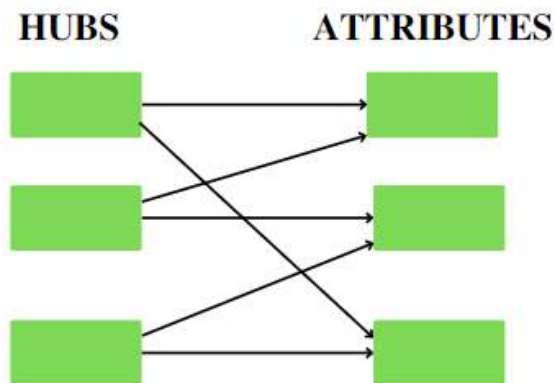
Example of Web Structure Mining:

One of the techniques is the [Page rank Algorithm](#) that the **Google** uses to rank its web pages. The rank of a page is dependent on the number of pages and the quality of links pointing to the target node.

So, we can say that the Web Structure Mining is the type of Mining that can be performed either at the **document level** (intra-page) or at the **hyperlink level** (inter-page). The research done at the hyperlink level is called as Hyperlink Analysis. the Hyperlink Structure can be used to retrieve useful information on the Web.

Web structure Mining basically has two main approaches or there are two basic strategic models for successful websites:

- Page rank : refer [Page Rank](#)
- Hubs and Authorities



Hubs And Attributes

- **Hubs:** These are pages with large number of interesting links. They serve as a hub or a gathering point, where people visit to access a variety of information. More focused sites can aspire to become a hub for the new emerging areas. The pages on website themselves could be analyzed for quality of content that attracts most users.
- **Authorities:** People usually gravitate towards pages that provide the most complete and authentic information on a particular subject. This could be factual information, news, advice, etc. these websites would have the most number of inbound links from other websites.

Applications of Web Structure Mining:

- Information retrieval in social networks.
- To find out the relevance of each web page.
- Measuring the completeness of Websites.
- Used in Search engines to find the relevant information.

Q.6)

a)

What is clustering ? Explain K-means clustering algorithm. Suppose the data is {2,4,10,12,3,20,11,25} Consider k=2, cluster the given data using above algorithm. (10 Marks)

Suppose the data is
 $\{2, 4, 10, 12, 13, 20, 11, 25\}$

Iteration 1

M_1, M_2 are the two randomly selected centroids/ means where

$$M_1 = 4, M_2 = 11$$

and the initial clusters are

$$C_1 = \{4\}, C_2 = \{11\}$$

Calculate the Euclidean distance as

$$D = [x, a] = \sqrt{(x-a)^2}$$

D_1 is the distance from M_1

D_2 is the distance from M_2

Datapoint	D_1	D_2	Cluster
2	2	9	C_1
4	0	7	C_1
10	6	1	C_2
12	8	1	C_2
3	1	8	C_1
20	16	9	C_2
11	07	0	C_2
25	21	14	C_2

As we can see in the above table

2 datapoints are added to cluster C_1 & other datapoints added to cluster C_2

Therefore

$$C_1 = \{2, 4, 3\}$$

$$C_2 = \{10, 12, 20, 11, 25\}$$

Iteration 2

Calculate new mean of datapoints in C_1 and

Ans:

Therefore

$$M_1 = (2+3+4)/3 = 3$$

$$M_2 = (10+12+20+30+25)/5 = 17.6$$

Calculating distance and updating clusters based on table below

Datapoint	D1	D2	Cluster
2	1	16	C1
4	1	14	C1
3	0	15	C1
10	7	8	C1
12	9	6	C2
20	17	2	C2
11	08	7	C2
25	22	7	C2

New cluster

$$C_1 = \{2, 3, 4, 10\}$$

$$C_2 = \{12, 20, 11, 25\}$$

Iteration 3

Calculate new mean of datapoints in C1 and C2

Iteration 3

Calculate new mean of datapoints in C1 & C2

Therefore

$$M_1 = (2+3+4+10)/4 = 4.75$$

$$M_2 = (12+20+11+25)/4 = 17$$

Calculating distance and updating clusters based on table below

START WRITING HERE
(Begin answer for each question on a new page)

datapoints	D1	D2	cluster
2	2.75	15	C1
4	0.75	13	C1
3	1.75	14	C1
10	5.25	7	C1
12	7.25	5	C2
20	15.25	3	C2
11	6.25	6	C2
25	20.25	8	C2

As we can see that the data points in the cluster C1 & C2 in iteration 2 & 3 are same.

$C1 = \{2, 3, 4, 10\}$
 $C2 = \{12, 11, 20, 25\}$

It means that none of the data points has moved to other cluster.

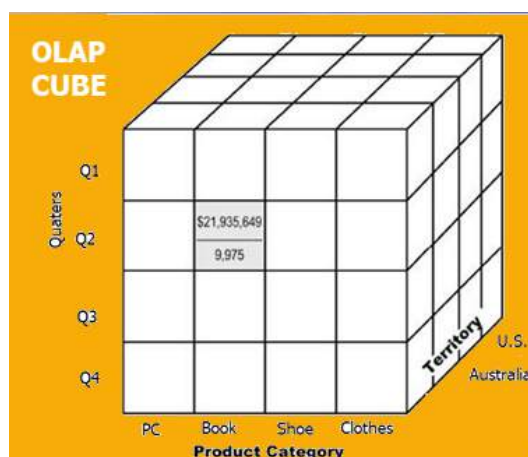
b) Illustrate with various operations and examples of OLAP cube. (10 Marks)

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

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

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

OLAP cube



OLAP Cube

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

Ans:

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

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

How does it work?

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

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

Basic analytical operations of OLAP

Four types of analytical OLAP operations are:

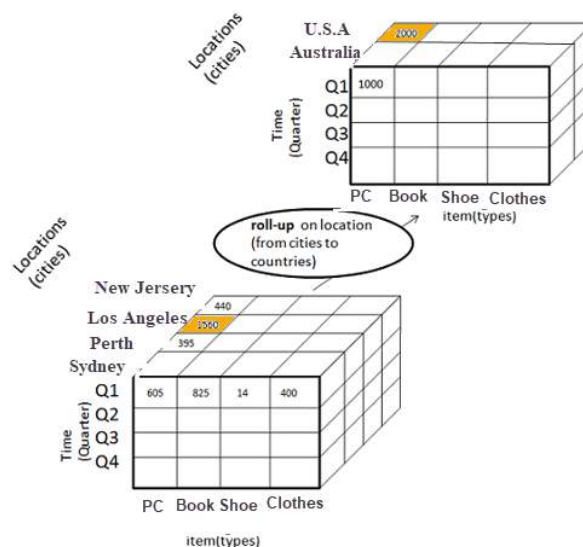
1. Roll-up
2. Drill-down
3. Slice and dice
4. Pivot (rotate)

1) Roll-up:

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

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

Consider the following diagram



Roll-up operation in OLAP

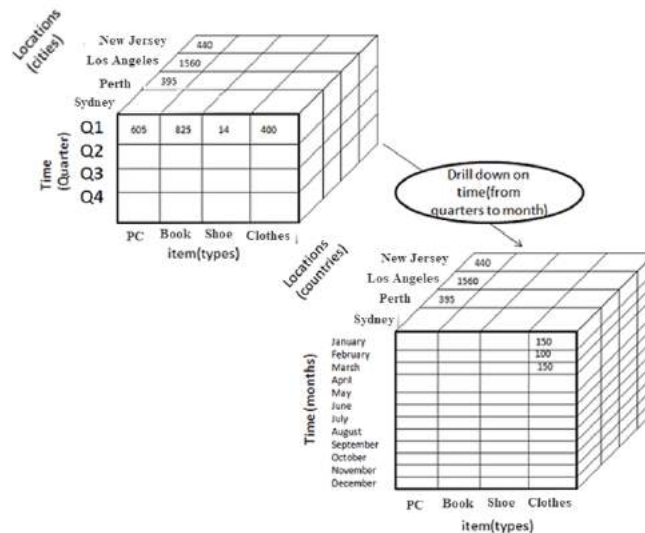
- In this example, cities New jersey and Lost Angles and rolled up into country USA
- The sales figure of New Jersey and Los Angeles are 440 and 1560 respectively. They become 2000 after roll-up

- In this aggregation process, data location hierarchy moves up from city to the country.
- In the roll-up process at least one or more dimensions need to be removed. In this example, Cities dimension is removed.

2) Drill-down

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

- Moving down the concept hierarchy
- Increasing a dimension



Drill-down operation in OLAP

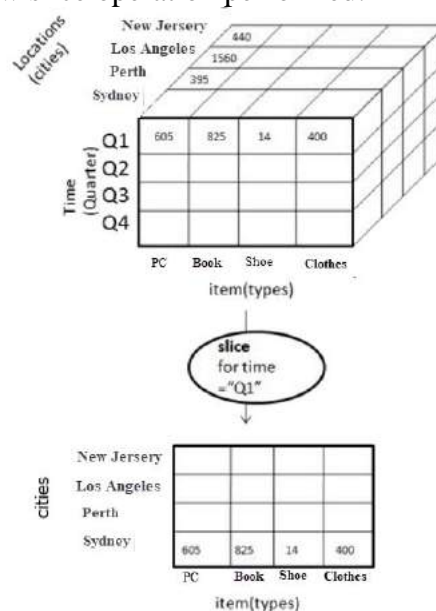
Consider the diagram above

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

3) Slice:

Here, one dimension is selected, and a new sub-cube is created.

Following diagram explain how slice operation performed:

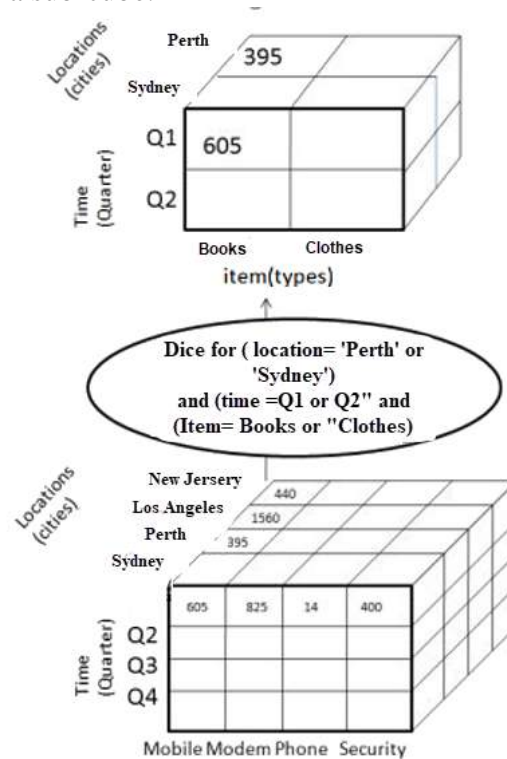


Slice operation in OLAP

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

Dice:

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

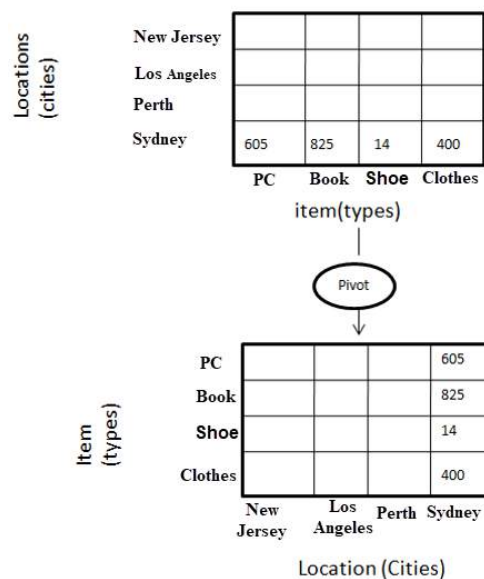


Dice operation in OLAP

4) Pivot

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

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



Pivot operation in OLAP