Assignment

Q.1. Write short note on: operational vs Decision support system

Operational and Decision support system (DSS) play distinct roles in the realm of data warehousing and mining.
Operational systems focus on day-to-day transactions and routine processes within an organization. They are designed for efficient data processing, ensuring the smooth running of business activities. In contrast, Decision support systems are geared towards aiding decision making by providing analytical tools and insights derived from historical and current data.

Q.2. Write definition and features of data warehouse

Data warehousing:
Data warehousing is a process of collecting, storing and managing data from various sources to support business intelligence and decision making. It involves the

consolidation of data from different departments and systems into a centralized repository for analysis and reporting.

Key features:

1. Centralized Repository: Data warehouses serve as a central location for storing integrated and historical data from diverse sources.

2. Subject-oriented: Organized around specific subjects or themes, allowing users to focus on relevant data for analysis.

3. Time-variant: Captures and maintains historical data, enabling analysis of trends and changes over time.

4. Non-volatile: Once data is stored in the warehouse, it is not modified or deleted. This ensures consistency and traceability for historical analysis.

5. Optimized for query and reporting: Structured to facilitate complex queries and reporting, supporting business intelligence activities.

6. Data Transformation: Involves the

process of transforming raw data into a consistent and usable format for analysis, often through Extract, Transform, Load (ETL) processes.

7. Scalability: Designed to handle large volumes of data and accommodate the evolving needs of an organization.

8. Decision support system (DSS): Aids in decision-making by providing a comprehensive view of the data, promoting informed and strategic choices.

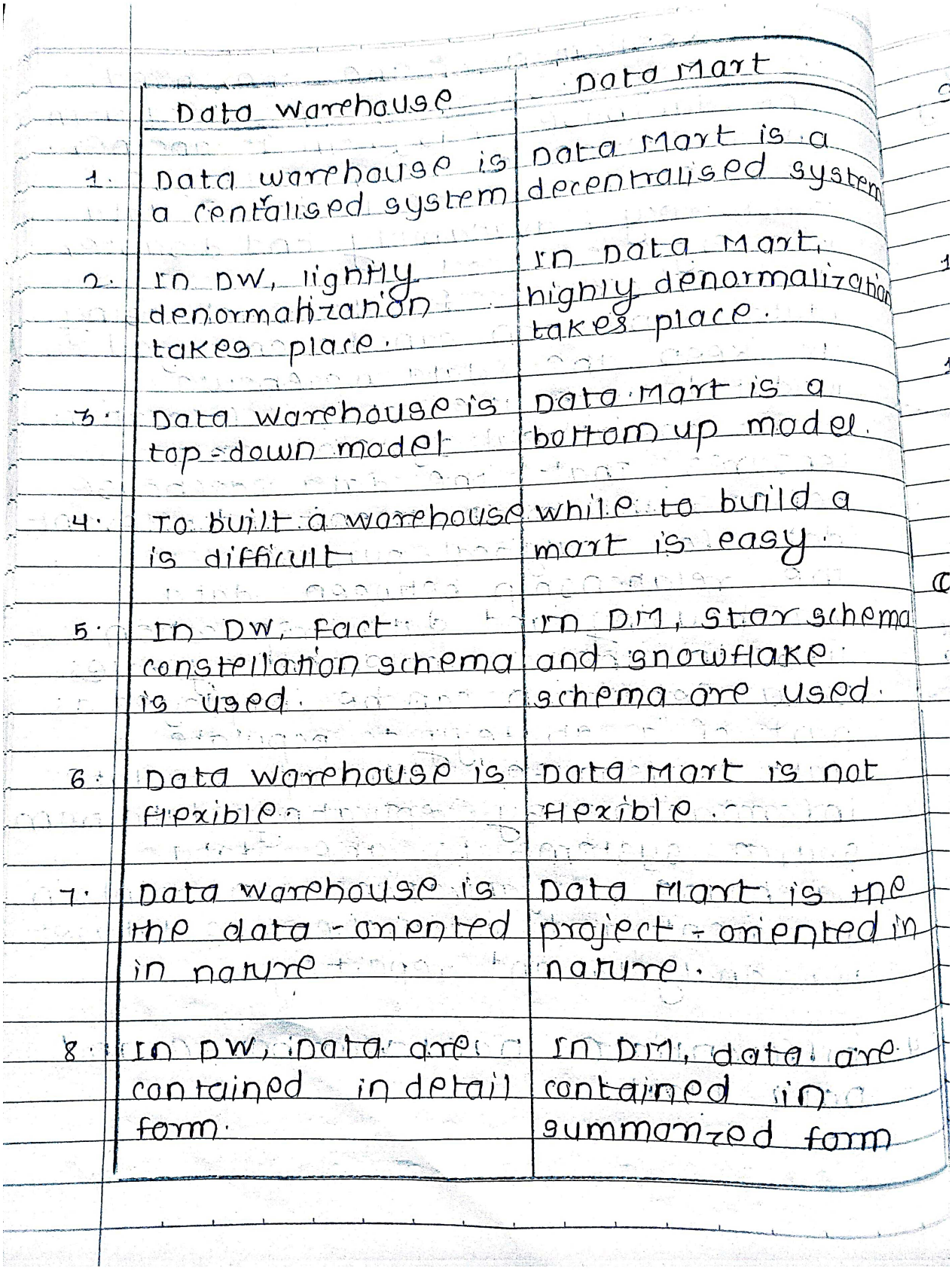
Q.3. What is relation between Data warehousing and Data replication.

Data warehousing and data replication are related in the context of managing and ensuring the availability of data across an organization.
Data warehousing involves the collection, storage, and management of data from various sources of to support business intelligence and decision making processes. It centralize data into a single repository, making it easier for users to analyze and extract insights.

Page No.
Date

Data Replication, on the other hand, is the process of copying data from one database or system to another in real-time or near-real-time. This is often done to ensure data consistency, availability and disaster recovery.
In the context of data warehousing, data replication can be employed to keep the data warehouse updated with the latest information from operational systems. This ensures that the data warehouse reflects the most recent and relevant data for analytical purposes.
The relationship between data warehousing and data replication lies in their complementary roles. Data Replication can be employed as part of a strategy to populate a data warehouse with up-to-date information. By replicating data from source systems to the data warehouse, organizations can maintain a current and comprehensive dataset for analysis and reporting.

Q.4. Differentiate Data Warehouse and Data Mart.



	Data warehouse	Data Mart
1.	Data warehouse is a centalised system	Data Mart is a decentralised system
2.	In DW, lightly denormalization takes place.	In Data Mart, highly denormalization takes place.
3.	Data warehouse is top-down model	Data Mart is a bottom up model.
4.	To built a warehouse is difficult	while to build a mart is easy.
5.	In DW, Fact constellation schema is used.	In DM, star schema and snowflake schema are used.
6.	Data warehouse is flexible.	Data Mart is not flexible.
7.	Data warehouse is the data-oriented in nature	Data Mart is the project-oriented in nature.
8.	In DW, data are contained in detail form.	In DM, data are contained in summarized form

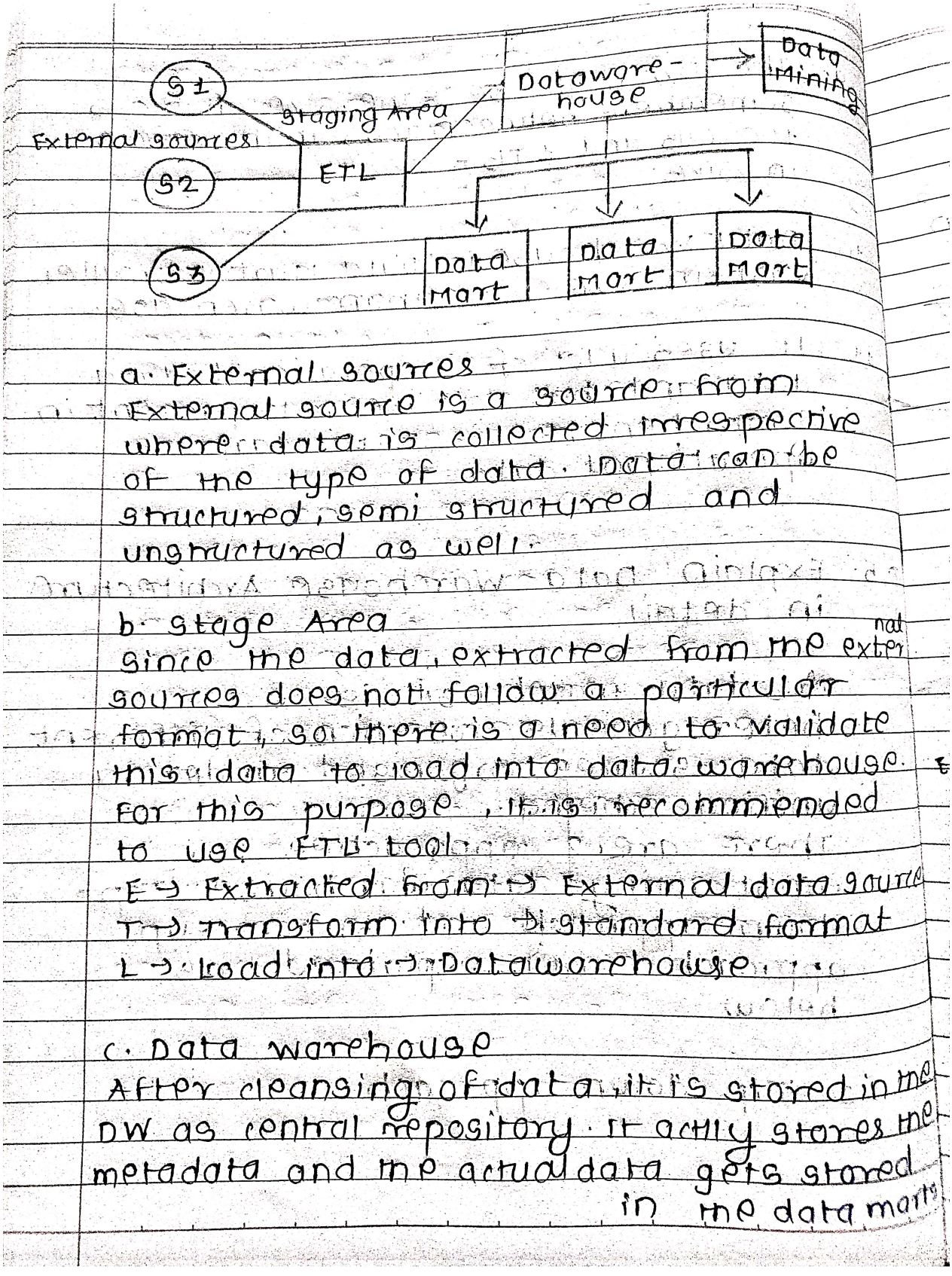
9.	The DW might be somewhere between 100 GB and 1 TB+ in size.	The size of DM is less than 100 GB.
10.	Data warehouse is vast in size	Data mart is smaller than warehouse.
11.	It uses a lot of data and has comprehensive operational data	operational data are not present in data mart

Q5. Explain Data Warehouse Architecture in detail.

A data warehouse is a heterogeneous collection of different data sources organised under a unified schema.

There are 2 approaches for construcing data warehouse: Top-down approach and Bottom-up approach are explained as below

1. Top-down approach



a. External sources
External source is a source from where data is collected irrespective of the type of data. Data can be structured, semi structured and unstructured as well.

b. Stage Area
Since the data, extracted from the external sources does not follow a particular format, so there is a need to validate this data to load into data warehouse. For this purpose, it is recommended to use ETL tool.
E → Extracted from → External data source
T → Transform into → standard format
L → Load into → Datawarehouse

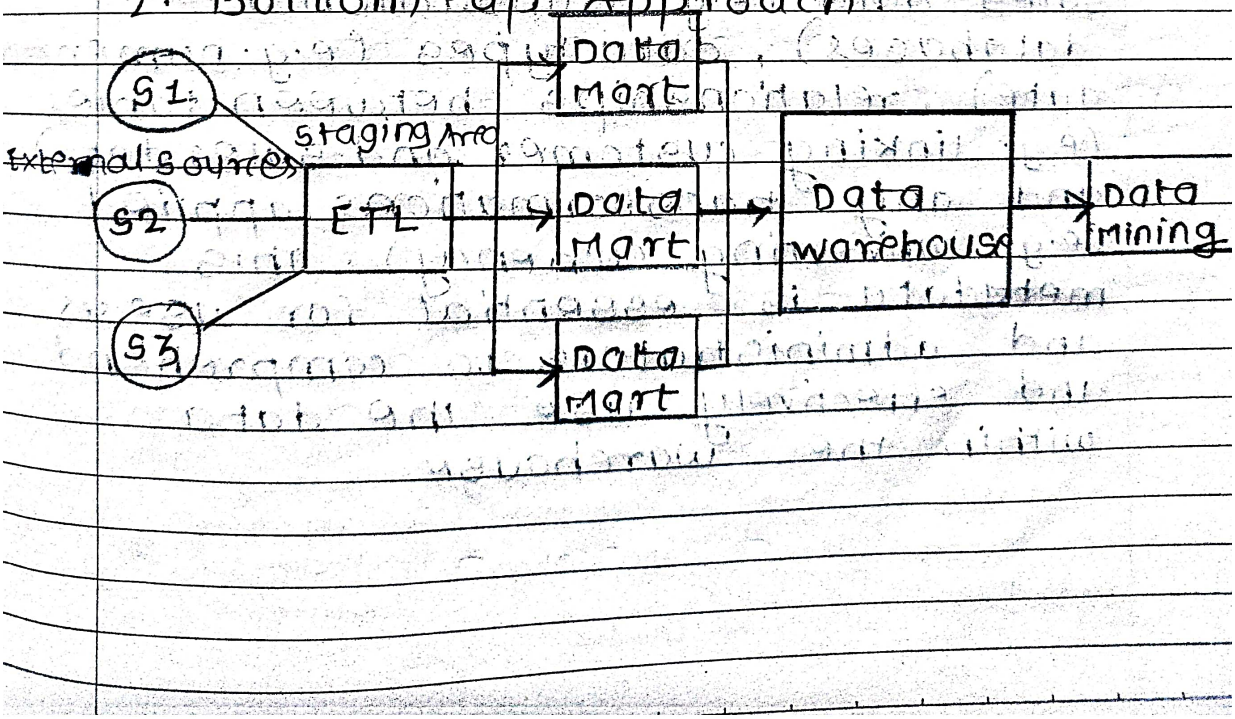
c. Data warehouse
After cleansing of data, it is stored in the DW as central repository. It actually stores the metadata and the actual data gets stored in the data marts

Page No.
Date

d. Data marts
Data mart is also a part of storage component. It stores the information of a particular function of an organization which is handled by single authority. There can be as many number of data marts depending upon the functions. DM contains subset of the data stored in DW.

e. Data mining
The practice of analysing the big data present in DW is DM. It is used to find the hidden patterns that are present in database or DW with help of algorithm of data mining.

2. Bottom-up Approach:



Q.6. Write short note on Metadata in data warehouse. Illustrate with suitable example.

Metadata in data warehousing refers to data that provides information about other data. It plays a crucial role in managing & understanding the contents of a data warehouse. Metadata includes details such as data source, data types, relationships, and transformations applied

For example, imagine a data warehouse that stores sales information. The metadata would include details like the sources of sales data (e.g. transactional databases), data types (e.g. numeric, date), relationships between tables (e.g. linking customer and sales data) and any transformations applied (e.g. converting currency). This metadata is essential for users and administrators to comprehend and effectively use the data within the warehouse

Q.7. Discuss ETL Process with labelled diagram

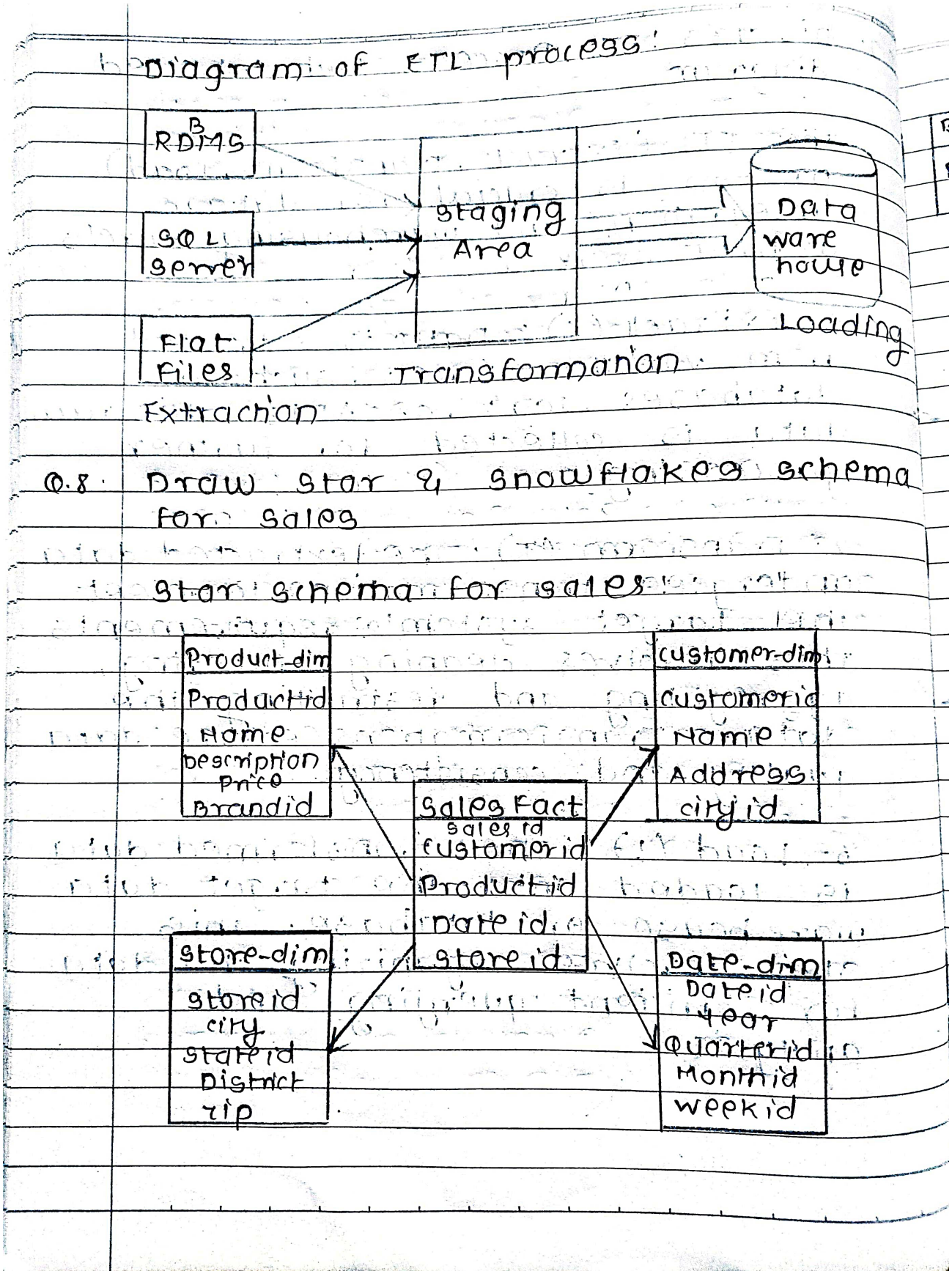
The ETL (Extract, Transform, Load) process is crucial for data integration and warehousing. Here's a brief overview:

1. Extract (E): Data is extracted from various sources, such as databases, logs, or APIs. This raw data is collected for further processing.

2. Transform (T): The extracted data undergoes transformations to meet the target system's requirements. This involves cleaning, filtering, aggregating and restructuring the data. Transformations ensure data quality and consistency.

3. Load (L): The transformed data is loaded into the target data ware house or database. This step involves organizing the data for efficient querying and analysis.

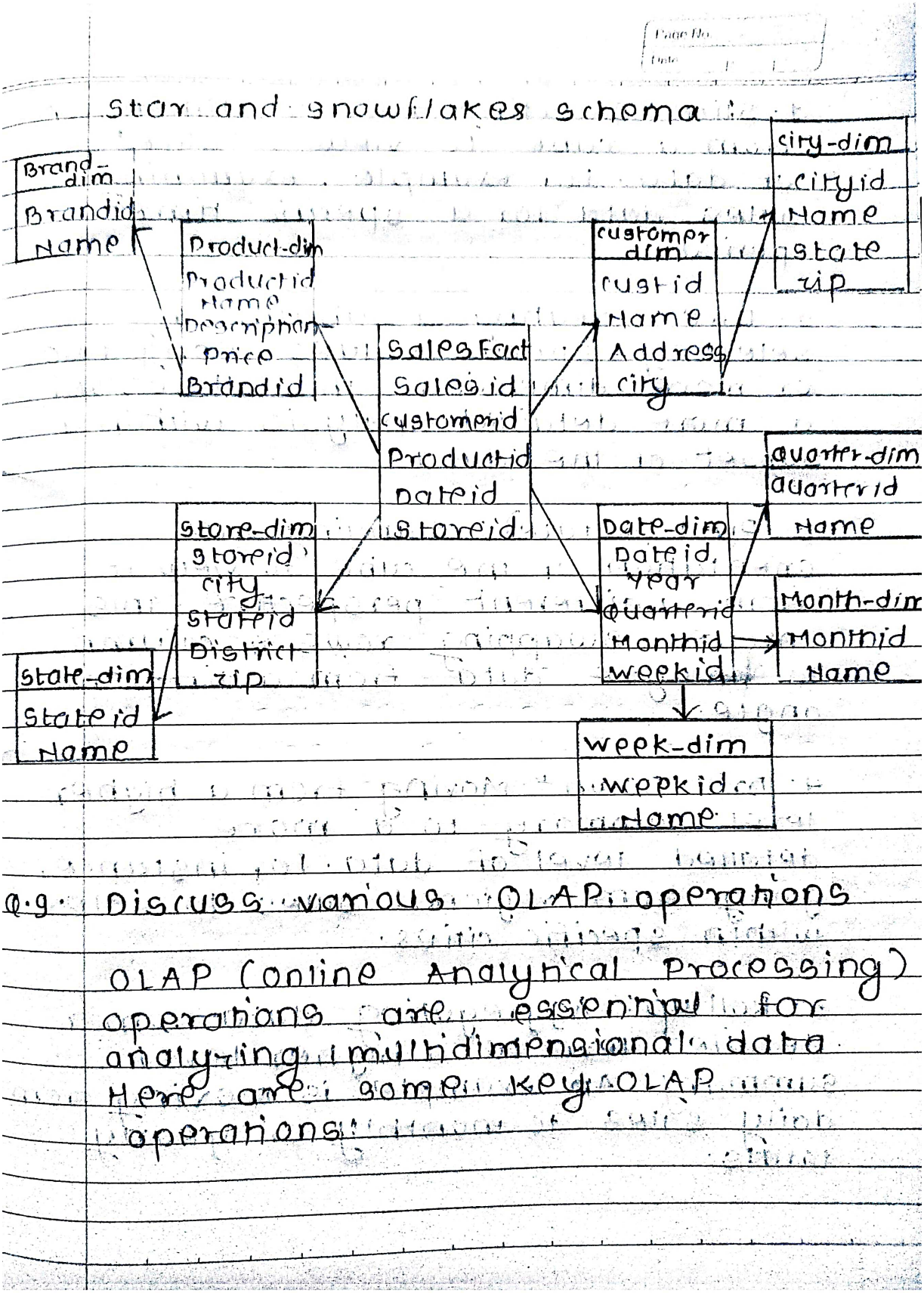
Diagram of ETL process:



Q.8 Draw star & snowflakes schema for sales

Star schema for sales:

Star and snowflakes schema



Q.g. Discuss various OLAP operations

OLAP (online Analytical Processing) operations are essential for analyzing multidimensional data. Here are some key OLAP operations:

1. Slice : Selecting a single dimension from a cube to view a 'slice' of data. For example, examining sales data for a specific time period.

2. Dice : Creating a subcube by selecting specific values along two or more dimensions. This allows for a more detailed analysis within a subset of the data.

3. Pivot (Rotate) : Changing the orientation of the cube to view it from a different perspective. This involves swapping rows & columns to analyze data from a different angle.

4. Drill Down : Moving from a higher level summary to a more detailed level of data. For instance, going from regional sales to sales within specific cities.

5. Roll up : Aggregating data from a detailed level to a higher level summary. An example is moving from daily sales to monthly or yearly totals.

Page No.
Date

1 Discuss different steps involve in data preprocessing

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transformation 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

Steps include in data preprocessing :-

- Data cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Data Compression

i) Data cleaning : This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, imputation, removal, and transformation

ii) Data Integration : This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling

data with different format, structure, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

iii) Data Transformation: This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, standardization is used to transform the data to have zero mean and unit variance. Discrentization is used to convert continuous data into discrete categories.

iv) Data Reduction: This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction.

v) Data Compression: This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purpose. It can be done using techniques such as wevlet compression, JPEG compression and gzip compression.

2 Define Data preprocessing

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.

3. Explain data mining Functionalities and technologies used

Data Mining Techniques : Data mining uses algorithms and various othe techniques to convert large Collection of data into useful output. the most popular types of data mining techniques include association rules, classification, clustering, decision trees, k-Nearest Neighbor, neural networks, and predictive

i) Association rules - It also reffered to as market basket analysis, search for relationship between variables. This relationship in itself creates additional value within the data set as it strives to link pieces of data

ii) Classification - uses predefined classes to assign to objects. These classes describe the characteristics of items or represent what the data points have in common with each other. This data mining technique allows the underlying data to be more neatly categorized and summarized across similar Features or product lines

iii) Clustering - It is similar to classification. However, clustering identifies similarities between objects, then groups those items based on what makes them different from other items. While classification may result in groups such as "shampoo," "Conditioner," "soap" and "toothpaste," clustering may identify groups such as "hair care" and "dental health".

iv) Decision trees - are used to classify or predict an outcome based on a set list of criteria or decisions. A decision tree is used to ask for the input of series of cascading question that sort the dataset based on the responses given. Sometimes depicted as a tree-like visual, a decision tree allows for specific direction and user input when drilling deeper into the data.

Page No.
Date

v) k-Nearest Neighbor (kNN) - It is an algorithm that classified data based for its proximity to other data. The basis for kNN is rooted in the assumption that data points that are close to each other are more similar to each other than other bits of data. This non-parametric, supervised techniques is used to predict the features of group based on individual data points.

vi) Neural networks - It process data through the use of nodes. These nodes are comprised of inputs, weights, and an output. Data is mapped through supervised learning, similar to how the human brain is interconnected. This model can be programmed to give theshold values to determine a model's accuracy.

vii) Predictive analysis - It strives to leverage historical information to built graphical or mathematical models to forecast future outcomes. overlapping with regression analysis, this techniques aims to support an unknown figure in the Future based on current data on hand.

4. Explain integration and data mining system with data warehouse.

Data integration in data mining refers to the process of combining data from multiple sources into a single, unified view. This can involve cleaning and transforming the data, as well as resolving any inconsistencies or conflicts that may exist between the different sources. The goal of data integration is to make the data more useful and meaningful for the purposes of analysis and decision making. techniques used in data integration include data warehousing, ETL (extract, transform, load) processes, and data federation.

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 data integration approaches are formally defined as triple <G, S, M> where,
G stands for the global schema,
S stands for heterogeneous source of schema,
M stands for mapping between the queries

Page No.
Date

of sources and global search
- Issues in Data integration:
i) Data quality
ii) Data semantics
iii) Data Heterogeneity
iv) Data privacy and security
v) Scalability
vi) Data Governance
vii) Performance
viii) Integration with existing System
ix) Complexity

5. Explain what kind of data can be mined

In data mining, we can mine different types of data to find useful patterns and insights. Here are some example of the kind of data we can mine

i) Numerical Data - This includes data like numbers, such as sales figures, temperatures or ages

ii) Categorical Data - These are labes or categories that can fall into, like colours, types of products, or job tittles

iii) Text data - This involves mining information from text, such as emails, social media posts, or product reviews

iv) Image data – Data mining can also be applied to images, Finding patterns in pictures, like identifying faces or objects.

v) Temporal Data – This type of data include information collected over time, like stock prices weather data, or website traffic

vi) Spatial data – Spatial data involves location on earth, like GPS, Coordinates. addresses or maps.

Data mining techniques can be used on various types of data to uncover hidden patterns, relationships or trends that can be help business make better decision or understand their data better.

19/3/2024

Assignment - 02.

Q1. Explain market basket analysis with example & application

→ A data mining technique that is used to uncover purchase patterns in any retail setting is known as Market Basket Analysis. In simple terms Basically, Market Basket analysis in data mining to analyze the combination of products which been bought together.

This is a technique that gives careful study of purchase 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 ~~complain~~ companies, and data mining techniques helps to achieve this analysis task.

Example :-

1). 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.

2). Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining

3). Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior

Applications ~~for~~ of Market Basket Analysis

1) Retail :- Market Basket research is frequently used in retail sector to examine consumer buying patterns and inform decisions about product placements, pricing pricing tactics.

2) Ecommerce :- Market Basket analysis can help online merchants better understand the customer buying habits and make data-driven decisions about product recommendations and targetted advertising campaigns.

3). Finance :- It can be used to evaluate investor behavior and forecast the types of investment items that investors will likely buy in the future.

Page No.
Date

Types of Market Basket Analysis -

1) Descriptive Market Basket analysis :-
This sort of analysis looks for patterns & connections in data that exist between the components of a market basket. This type of study is mostly used to understand consumer behavior, including what products are purchased in combination and what the most typical item combinations
2) Predictive Market Basket Analysis.
3) Differential Market Basket Analysis.

Q. Discuss association and corelation rule with example.

→ Association Rule -

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

TID	Items.
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs.
3	Milk, Diaper, Beer, Coke.
4	Bread, Milk, Diaper, Beer.
5	Bread, Milk, Diaper, Coke.

1) Support Count (σ) - Frequency of occurrence of a itemset.
2) Frequent Interest - An item whose support is greater or equal.
3) Association Rule - An implication expression of form $X \rightarrow Y$ where X & Y are any 2 itemsets.

Page No.
Date

Q3. Define Frequent item set.

→ Frequent item sets, also known as association rules are a fundamental concepts in association rule mining, which is a technique used ~~by~~ in data mining to discover relationships between items in a dataset. The goal of association rule mining is to identify relationships between items in dataset that occur frequently together.

A frequent item set is a set of items that occur together frequently in a dataset. The frequency of an item set is measured by support count, which is the number of transactions or records in dataset that can contain the item set.

Q4. Write Application of market basket analysis.

→ 1) Retail - Market Basket & research is frequently used in retail sector to examine ~~customer~~ consumer buying patterns and inform decisions about product placement, inventory management, & pricing tactics.

2) E-commerce - Market basket analysis can help online merchants better understand the customer buying habits and make data-driven decisions about product recommendations and targeted advertising campaigns.

3) Finance - Market basket analysis can be used to evaluate investors behavior and forecast the types of investments items that investors will likely buy in future.

4) Telecommunications - To evaluate consumer behavior and make data-driven decisions, about which goods & services to provide, telecommunications business might employ market basket analysis.

5) Manufacturing - To evaluate consumer behavior & make data-driven decisions about which products to produce & which materials to employ in product^n process

Q9. Write short note on multilevel & multidimensional association rule.

→ Multilevel Association Rule-

Association rules created from mining information at different degrees of reflection are called various level or staggered association rules. Multilevel association rules can be mined effectively utilizing idea progressions under a help certainty system.

Approaches to multilevel association rule mining:

1). Uniform Support
2). Reduced support
3). Group-based support.

Multidimensional Association Rule-

Multi dimensional association rule Qualities can be absolute or quantitative. Quantitative characteristics are numeric and consolidated order. Numeric traits should be discretized. Multidimensional affiliation rule comprises of more than one measurement.

Approaches to multidimensional association rule:

1). Using static discretization of quantitative qualities.
2). Using powerful discretization of quantitative traits.
3). Using distance based discretization with bunching.

Page No.
Date

Q8. Explain apriori algorithm with example.

→ Apriori algorithm refers to algorithm which is used to calculate association rules between objects. It means how two or more objects are related to one another. In other words, we can say that apriori algorithm is an association rule leaning that analyzes people who bought product A also bought product B.

The primary objective of apriori algorithm is to create association rule between different objects.

Example.

TID	Rice	Pulse	Oil	Milk	Apple
T_1	1	1	1	0	0
T_2	0	1	1	1	0
T_3	0	0	0	1	1
T_4	1	1	0	1	0
T_5	1	1	1	0	1
T_6	1	1	1	1	1

Step 1.

Product	Frequency.
Rice (R)	4
Pulse (P)	5
Oil (O)	4
Milk (M)	4

min support = 3.

Step 2.

Item Set	Frequency.
Rice Pulse	4
Rice Oil	3
Rice Milk	2
Pulse Oil	4
Pulse Milk	3

=>

Item Set	Frequency.
Rice Pulse	4
Rice Oil	3
Pulse Oil	4
Pulse Milk	3

Step 3

Rice Pulse Oil	4
Rice Pulse Milk	2
Pulse Oil Milk	3

Item Set	Frequency
Rice Pulse Oil	4
Pulse Oil Milk	3

⟹ Most frequent Item set is Rice Pulse Oil.

Q 7. Write apriori algorithm with min support 2.

TID	Item Sets
T_1	A, B
T_2	B, D
T_3	B, C
T_4	A, B, D
T_5	A, C
T_6	B C
T_7	A, C
T_8	A, B, C, E
T_9	A, B, C.

⟹ C_1

Itemset	Support
A	6
B	7
C	6
D	2
E	1

L_1

Itemset	Support
A	6
B	7
C	6
D	2

Page No.
Date

~~Example~~ -

Correlation Analysis in a data mining is a statistical technique for determining the strength of a link between two variables. It is used to detect patterns and trends in data and to forecast future occurrences.

Types of Correlation -

1) Positive Correlation - Positive correlation indicates that two variables have a direct relationship. As one variable increases, the other variable also increases.
2) Negative Correlation - Negative correlation indicates that two variables have an inverse relationship. As one variable increases, the other variable decreases.
3) Zero correlation - Zero correlation indicates that there is no relationship between two variables. The change in one variable do not affect the other variable.

Applications of Correlation Analysis

1). E-commerce and Finance - Help in analyzing the economic trends by understanding relations between suply & demand
2). Business Analytics - Helps in making better decisions for company and provides valuable insights
3) Weather forecast - Analyzing the correlation between different variables so as to predict weather.

Assignment :04

Q1. What is clustering?

→ The task of grouping data points based on their similarity with each other is called Clustering. This method is defined under the branch of Unsupervised learning, which aims at gaining insights from ~~labelled~~ unlabelled data points that is, unlike supervised learning we don't have a target variable.

Clustering aims at forming groups of homogeneous data points from a heterogeous dataset. It evaluates similarity based on metric like Euclidean distance, Cosine similarity, Manhattan distance, etc. and then group points with highest score together

Q2. What is dendrogram?

→ A dendrogram is a tree or branch diagram that visually shows relationship between similar objects. Each of branches of the tree represents a category of class, while the entire tree diagram shows the hierarchy relationship between all classess or branches.

Q3. List major clustering method.

→ i). Centroid-based Clustering.
ii). Density-based Clustering.
iii). Distribution-based Clustering.
iv). Hierarchical Clustering.

Q4) Differentiate between agglomerative and divisive method.

→

Agglomerative	Divisive
i) Divide the data points into different clusters & then aggregate them as distance decreases.	i). Combine all the data points as a single cluster and divide them as distance increases
ii) It is bottom-up approach.	ii). It is top-down approach.
iii). It starts clustering by treating individual data points as a single cluster, then it is merged continuously based on similarity	iii) It starts by considering all data points into a big single cluster and later on splitting them into smaller heterogeneous clusters.
iv). It is good at identifying small clusters.	iv). It is good at identifying large cluster.

Q5 Explain hierarchy clustering method.

-> A Hierarchy clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data point as a seperate cluster Then, it repeatedly executes subsequent steps:

1) Identify 2 clusters which can be closet together
2) Merge the 2 maximum comparable clusters. We need to continue these steps until all clusters are merged togethes

In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called Dendrogram, graphically represents this hierarchy and is an inverted tree that describes order in which factors are merged or clusters are broken up.

Hierarchical clustering is a method of cluster analysis in data mining that creates hierarchical representation of clusters in a dataset. The

Advantages -

i) Ability to handle non-convex clusters and clusters of different sizes and densities.
ii) Ability to handle missing data and noisy data

Disadvantages :-

i) Need for ~~creatt~~ criterion to stop clustering process & determine final number of clusters.
ii) Computation cost and memory requirements of method can be high especially for large datasets

Q6. Explain partioning clustering method

→ Partioning clustering method classifies classifies the information into multiple groups based on characteristics and similarity of the data Its the data analysts to specify number of clusters that has to be generated for clustering methods.

In partioning method when database (D) that contains multiple (N) objects that then partitioning method constructs user specified (K) partitions of data in which each partition represents a cluster and a particular region.

There are many algorithms that come under partioning method some of popular ones are k-Mean. PAM (k-Medoids), CLARA algorithm (Clustering large Applications) etc.

Q7. Explain grid based clustering with example.

→ Grid based clustering methods use a multi-resolu grid data structure. It quantizes object areas into a finite number of cells that form a grid structure on which all of operations for clustering are implemented. The benefit of the method is its quick processing time, which is generally independent of numbers of data objects still dependent on only multiple cells in each dimension in quantized space.

An instance of grid-based approach involves STING which explores statistical data stored in grid cells, WaveCluster which clusters objects using a wavelet transform approach and CLIQUE which defines a grid and density-based approach for clustering in high-dimensional data space.

For Example:-