a) **Define Data warehouse how it is different from a database.**

The term Data Warehouse was defined by Bill Inmon in 1990, in the following way:

"A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process".

He defined the terms in the sentence as follows:

- Subject Oriented¬ - Data that gives information about a particular subject instead of about a company's ongoing operations
- Integrated -Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.
- Time-variant - All data in the data warehouse is identified with a particular time period.
- Non-volatile -Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business

b) **Define the term Data cleaning with example.**

Data cleaning is also known as scrubbing. The data deaning process detects and removes the errors and inconsistencies and improves the quality of the data. Data quality problems arise due to misspellings data entry. missing values or any other invalid data.

For example, if you conduct a survey and ask people for their phone numbers, people may enter their numbers in different formats.

c)List different Data cube computation methods.

Data Cube Computation Methods

- Multi-Way Array Aggregation.
- BUC.
- Star-Cubing.
- High-Dimensional OLAP.

D)Define the term Data mining.

Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions.

E)State Application of cluster analysis.

Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.

F) List Application of OLAP

- Accounting, forecasting, budgeting, cost, and profitability analysis and consolidation.
- Human resources, skill consolidation, labor scheduling, and optimization.
- Distribution, scheduling, and optimizat
- based analysis. ion.Marketing, churn, and market-

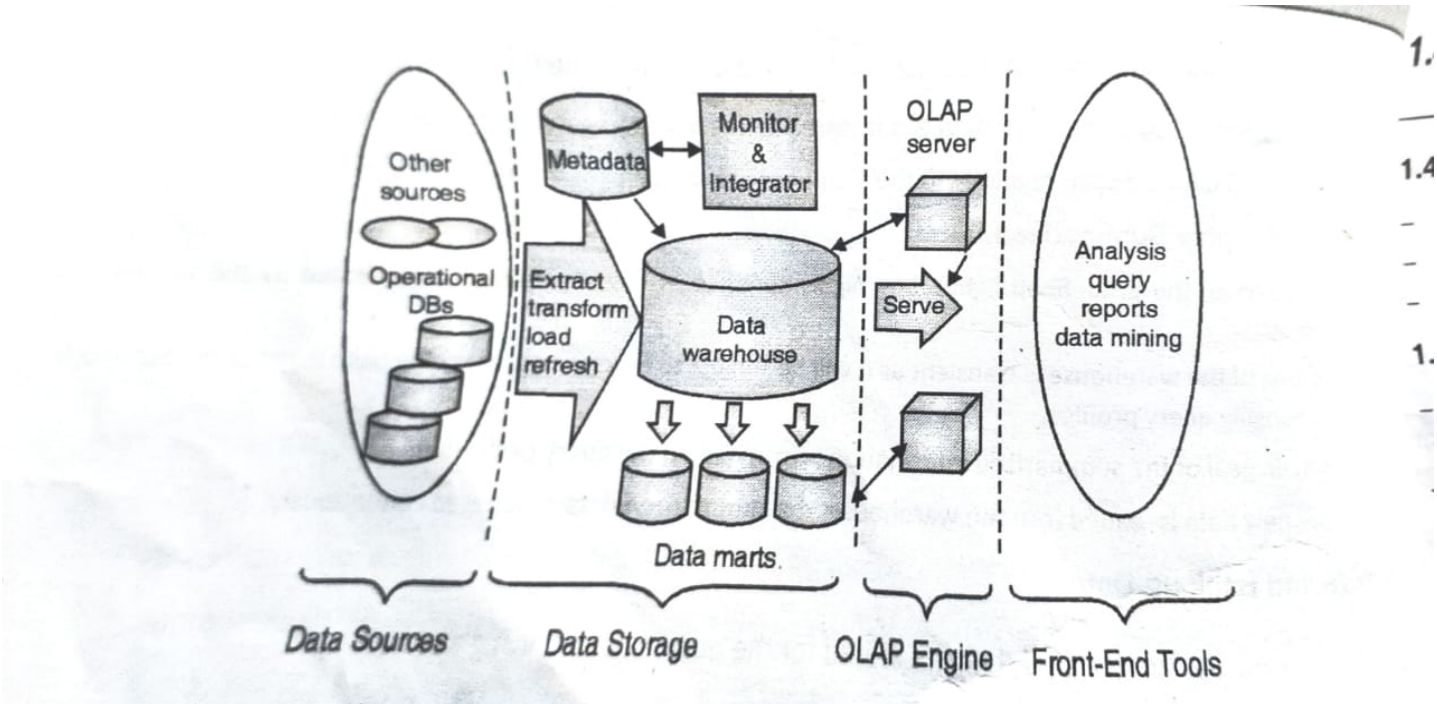
G) Define OLAP Data cube.

An OLAP cube is a multi-dimensional array of data.[1] Online analytical processing (OLAP)[2] is a computer-based technique of analyzing data to look for insights. The term cube here refers to a multi-dimensional dataset, which is also sometimes called a hypercube if the

number of dimensions is greater than 3.

Q.2) Attempt any THREE of the following.

a) Explain three tier architecture of data warehousing.



1. Bottom Tier (Data Sources and Data Storage)

It is a warehouse database server, that is generally a RDBMS. Using Application Program interfaces (called as gateways), data is extracted from operational and external

Gateways like, ODBC(Open Database connection), OLE-DB (Open linking and embedding for database), 108C (Java Database Connection) is supported by underlying DBMS .

2. Middle Tier (OLAP Engine)

OLAP Engine is either implemented using ROLAP (Relational online Analytical Processing) or MOLAP(Multidimensional OLAP).

3.Top Tier (Front End Tools)

This tier is a client which contains query and reporting tools, Analysis tools, and /or data mining tools.

From the Architecture Point of view there are three data warehouse Models

1. Enterprise Warehouse

The information of the entire organization is collected related to various subjects in enterprise warehouse.

2. Data Mart

A subset of Warehouse that is useful to a specific group of users.

It can be categorized as Independent vs. dependent data

3. Virtual warehouse

A set of views over operational databases.

Only some of the possible sum Pmary views may be materialized.

B)list basic operation of OLAP Describe any one.

1) Consolidation or Roll Up

2)Drill-Down

3)Slicing and Dicing

4)Dice

5)Pivot/Rotate

6)Other OLAP operation

1) Consolidation or Roll Up:

-Multi-dimensional databases generally have hierarchies with respect to dimensions.

-Consolidation is rolling up or adding data relationship with respect to one or more dimensions. For example,adding up all product sales to get total City data.

-The roll up operation shown aggregates the data by city to the country by location hierarchy.

B)Define the term 1)OLAP 2)ROLAP 3)MOLAP 4)HOLAP

MOLAP - This is the more traditional way of OLAP analysis, In MOLAP, data is stored in a multidimensional cube. The storage is not in the relational database, but in proprietary formats.

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

HOLAP technologies attempt to combine the advantages of MOLAP and ROLAP. For summary-type information,

HOLAP leverages cube technology for faster performance.

When detail information is needed, HOLAP can "drill through" from the cube into the underlying relational data.

OLAP - OLAP (for online analytical processing) is software for performing multidimensional analysis at high speeds on large volumes of data from a data warehouse, data mart, or some other unified, centralized data store.

C) Describe any four Challenges of Data mining.

1. Security and Social Challenges

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

2. Noisy and Incomplete Data

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

errors in the instruments that measure the data.

3. Mining dependent on Level of Abstraction

Data Mining measure should be community-oriented in light of the fact that it permits clients to focus on example optimizing, presenting, and pattern finding for data mining dependent on brought results back.

4. Integration of Background Knowledge

Previous information might be utilized to communicate examples to express discovered patterns and to direct the exploration processes.

5. 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.

Q.3) Attempt any THREE of the following.

a) Compare OLAP and OLTP Systems.

OLTP	OLAP
It is an online transactional system and manages database modification.	It is an online data retrieving and data analysis system.
Insert, Update, Delete information from the database	Extract data for analyzing that helps in decision making.
OLTP and its transactions are the original source of data.	Different OLTPs database becomes the source of data for OLAP.
OLTP has short transactions	OLAP has long transactions

The processing time of a transaction is comparatively less in OLTP	The processing time of a transaction is comparatively more in OLAP.
Tables in OLTP database are normalized (3NF).	Tables in OLAP database are not normalized.
OLTP database must maintain data integrity constraint	OLAP database does not get frequently modified. Hence, data integrity is not affected

b) Explain Data Cleaning Process.

Data cleaning is a process by which inaccurate, poorly formatted, or otherwise messy data is organized and corrected. For example, if you conduct a survey and ask people for their phone numbers, people may enter their numbers in different formats. Before it can be used, those phone numbers need to be standardized so that they're all formatted the same . Data can be messy like this for lots of reasons. Addresses can be formatted inconsistently; records can get duplicated and need to be identified and reconciled; some records may use different terms, like

“Closed won” and “Closed Won” to represent what should be the same values; null values need to be handled correctly; and so on. An example: How data can be cleaned

Data can be cleaned in a number of ways. Sometimes, it’s done manually in SQL queries, in Python scripts, or in Excel. Sometimes, people use tools like Trifacta that are designed to programmatically clean data. And sometimes, it’s incorporated into ETL processes that clean data as they extract and load it into a warehouse.

Opinion: Data cleaning, data prep, and data modeling are all slightly different

Data cleaning often gets conflated with two other related terms: data prep, and data modeling. We think of these words as meaning three different, albeit overlapping, things.

C) explain Market basket analysis.

Market basket analysis is a modelling technique which is also called as affinity analysis, it helps identifying which items are likely to be purchased together.

The market-basket problem assumes we have some large number of items, e.g., "bread", "milk.", etc. Customers buy the subset of items as per their need and marketer gets the information that which things customers have taken

together. So the marketers use this information to put the items on different position.

For Example :If someone buys a packet of milk also tends to buy a bread at the same time. Milk=>Bread

Market basket analysis algorithms are straightforward; difficulties arise mainly in dealing with large amounts of transactional data, where after applying algorithm it may give rise to large number of rules which may be trivial in nature.

Market basket analysis is used in deciding the location of items inside a store, for e.g. if a customer buys a packet of bread he is more likely to buy a packet of butter too, keeping the bread and butter next to each other in a store would result in customers getting tempted to buy one item with the other.

Applications of Market Basket Analysis

- Credit card transactions done by a customer may be analysed.
- Phone calling patterns may be analysed.
- Fraudulent Medical insurance claims can be identified.

D) Explain Bitmap index in OLAP.

- It allows quick searching in data cubes.
- The bitmap index is an alternative representation of the record ID (RID) list.
- Each attribute is represented by distinct bit value.
- If attribute's domain consists of n values, then n bits are needed for each entry in the bitmap index.
- If the attribute value is present in the row then it is represented by 1 in the corresponding row of the bitmap index and all other bits for that row are set to 0.
- ADVANTAGES:
- Bitmap indexing is advantageous compared to hash and tree indices.
- useful for low-cardinality domains because comparison, join, and aggregation operations are then reduced to bit arithmetic, which substantially reduces the processing time.

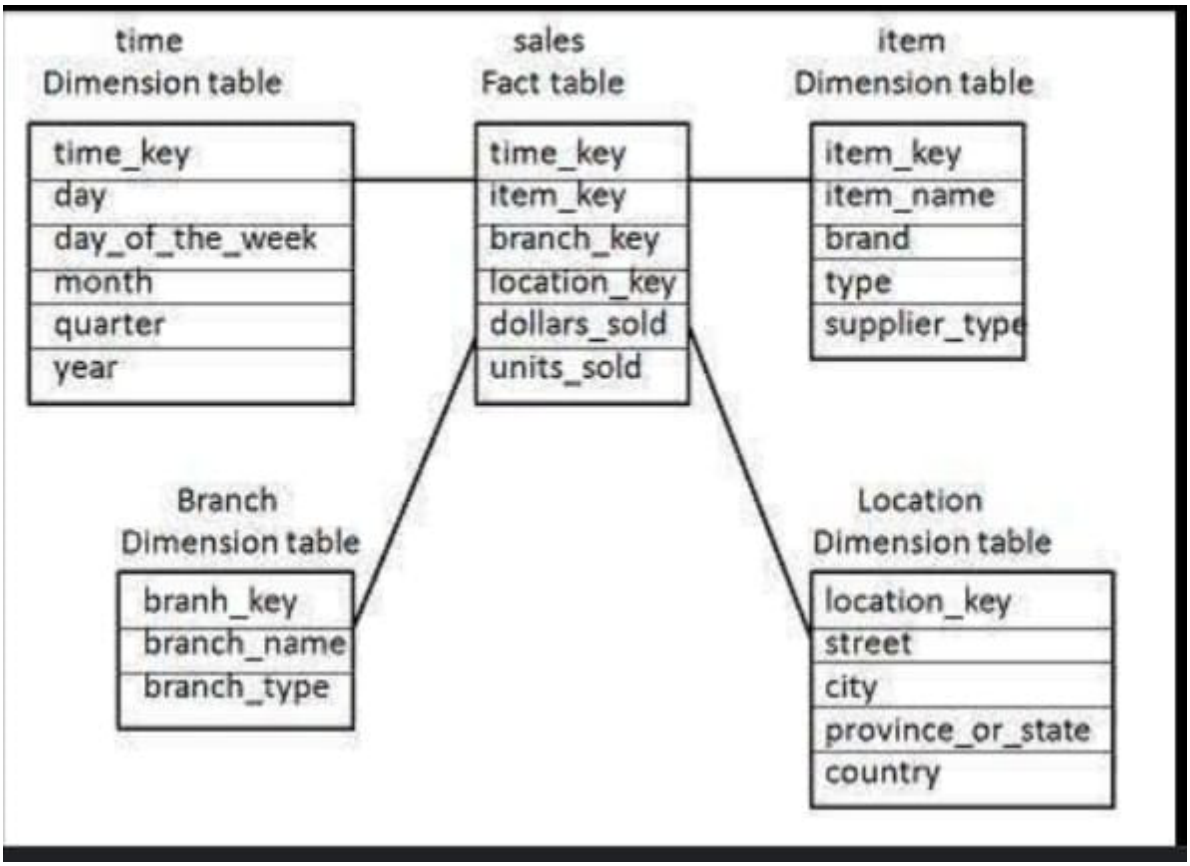
Q.4) Attempt any THREE of the following.

a) Differentiate between operational database system and data warehouse.

Operational System	Data Warehouse (DSS)
Application oriented	Subject oriented
Used to run business	Used to analyze business
Detailed data	Summarized and refined
Current up to date	Snapshot data
Isolated data	Integrated data
Repetitive access	Ad-hoc access
Clerical user	Knowledge user (manager)
Performance sensitive	Performance relaxed
Few records accessed at a time (tens)	Large volumes accessed at a time (millions)
Read/update access	Mostly read (batch update)
No data redundancy	Redundancy present
Database size 100 MB-100 GB	Database size 100GB - few terabytes

c) Draw star schema of a data warehouse for sales

considering Fact table Sales and dimensional tables as Time, Item, Branch and Location.



b) Describe the need of data preprocessing.

1.Real world data are generally

Incomplete: The data is said to be incomplete when certain attributes or attributes values are missing or onlyaggregate data is available.

Noisy: When the data contains errors or some outliers it is considered to be noisy data. Inconsistent: When the data contains differences in codes or names it is inconsistent data.

2. Major Tasks in data pre-processing

Data cleaning: This process consists of filling of missing values, smoothening noisy data, identifying and removingany outliers present and resolving inconsistencies.

Data Integration: This refers to integrating data from multiple sources likedatabases, data cubes, or files. Data transformation: Normalization and aggregation.

Data reduction: In data reduction the amount of data is reduced but same analytical results are produced.

Data discretization : Part of data reduction, replacing numerical attributes with nominal

C) Describe features of OLAP. Need for Online Analytical Processing

- OLAP or the On Line Analytical supports the multidimensional view of data.
- OLAP provides fast, steady, and proficient access to the various views of information.
- The complex queries can be processed
- It's easy to analyze information by processing complex queries on multidimensional views of data
- Data warehouse is generally used to analyse the information where huge amount of historical data is stored.
- Information in data warehouse is related to more than one dimension like sales, market trends, buying patterns,supplier, etc.

D) Describe Extraction, Transformation and Loading in datawarehousing.

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

- 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.
- 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.

❖ 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.

Q.5) Attempt any TWO of the following. (12 Marks)

a)Explain multidimensional Data model? How it is used in data warehouse.

A multidimensional model views data in the form of a data -cube. A data cube enables data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts. The dimensions are the perspectives or entities concerning which an organization keeps records.

- Use

-The multi-Dimensional Data Model is a method which is used for ordering data in the database along with good arrangement and assembling of the contents in the database.

-The Multi Dimensional Data Model allows customers to interrogate analytical questions associated with market or business trends, unlike relational databases which allow

customers to access data in the form of queries. They allow users to rapidly receive answers to the requests which they made by creating and examining the data comparatively fast.

b) Explain top down and bottom up design approach of data warehouse.

- Top-Down Design Model:

In the top-down model, an overview of the system is formulated without going into detail for any part of it. Each part of it then refined into more details, defining it in yet more details until the entire specification is detailed enough to validate the model.

Advantages:

-Breaking problems into parts help us to identify what needs to be done.

-At each step of refinement, new parts will become less complex and therefore easier to solve.Parts of the solution may turn out to be reusable.

-Breaking problems into parts allows more than one person to solve the problem.

- Bottom-Up Design Model:

In this design, individual parts of the system are specified in detail. The parts are linked to form larger components, which are in turn linked until a complete system is formed. Object-oriented language such as C++ or java uses a bottom-up approach where each object is identified first.

Advantage:

-Make decisions about reusable low-level utilities then decide how there will be put together to create high-level construct.

-The contrast between Top-down design and bottom-up design.

c) List clustering Methods explain any two.

* Basic Clustering Method

-A good clustering method will produce high quality clusters with:

° High intra-class

similarity

°Low inter-class similarity

- Major clustering methods can be classified into the following categories:

1.Partitioning methods: In Partitioning based approach,

various partitions are created and then they are evaluated based on certain criteria. 2. Hierarchical methods: The set of data objects are decomposed hierarchically using certain cr

Method	General characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape. – Distance-based. – May use mean or medoid (etc.) to represent cluster center. – Effective for small to medium sized data sets.
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels). – Cannot correct erroneous merges or splits. – May incorporate other techniques like micro-clustering or consider object "linkages".

Q.6) Attempt any TWO of the following. (12 Marks)

a) Explain Data preprocessing technique in data mining.

b) Explain Apriori algorithms for frequent itemset using candidate generation.

*Frequent Itemsets

-An itemset X is frequent if X's support is no less than a minimum support threshold.

-A frequent itemset is a set of items that appears at least in a pre-specified number of transactions. Frequentitemsets are typically used to generate association rules.

-Consider a data set S, frequent itemset in S are those

items that appear in at least a fraction s of the basket, wheres is a chosen constant with a value of 0.01 or 1%.

-To find frequent itemsets one can use the monotonicity principle or a-priori trick which is given as,If a set of items say S is frequent then all its subsets are also frequent.

-The procedure to find frequent itemsets:

-A level wise search may be conducted to find the frequent -1 items(set of size 1), then proceed to find frequent2 items and so on.

-Next search for all maximal frequent itemsets.

c) Explain steps involved in KDD process with diagram.

1.Developing an understanding of

-The application domain

-The relevant prior knowledge

-The goals of the 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 pre-processing

-Noise or outliers are removed.

-Essential information is collected for modelling or accounting for noise.

-Missing data fields are handled by using appropriate strategies.

-Time sequence information and changes are maintained.

4.Data reduction and projection

-Based on the goal of the task, useful features are found to represent the data.

-The number of variables may be effectively reduced using methods like dimensionality reduction or transformation. Invariant representations for the data may also be found out.

5.Choosing the data mining task

-Selecting the appropriate Data mining tasks like classification, clustering, regression based on the goal of the KDD process.

6. Choosing the data mining algorithm(s)

-Pattern search is done using the appropriate Data Mining method(s).

-A decision is taken on which models and parameters may be appropriate.

-Considering the overall criteria of the KDD process a match for the particular data mining method is done.

7.Data mining

-Using a representational form or other representations like classification, rules or trees, regression clustering for searching patterns of interest.

8 Interpreting mined patterns

9. Consolidating discovered knowledge