

Assignment - I

1. characteristics of olap system and explain types of olap operators in detail.

* online Analytics Processing Server (OLAP) is based on the multidimensional data model.

* It allows managers, and analytics to get an insight of the information, through fast, consistent, and interactive access of information.

Types of OLAP Servers:-

1. Relational OLAP

2. Multidimensional OLAP

3. Hybrid OLAP (HOLAP)

4. specialized SQL servers.

OLAP operations:-

→ Roll-up

→ Drill-down

→ slice and dice

→ pivot

Characteristics of OLAP system:-

The FASMI Test:-

It can represent the characteristics of OLAP application in a specific method, without dictating how it should be performed.

FASMI stands for Fast, Analysis, shared, Multidimensional, Information.

Fast - It defines that the system is targeted to produce most responses to users within about five seconds, with the understandable analysis taking no more than one second and very few taking more than 20 seconds.

Independent research in the Netherlands has shown that end-users consider that a process has declined if results are not received within 30 seconds, and they are suitable to hit 'ALT+CTRL+Delete' unless the system needs them that the report will take longer.

Analysis:- It defines that the system can manage with any business logic and statistical analysis that is appropriate for the application and the user. Although some preprogramming is required, it does not think it acceptable if all application definitions have to be completed using professional 4GL.

It is necessary to enable the user to represent new ad hoc calculations as a part of analysis and to report on the data in any desired method, without having a program, so it can exclude programs that do not enable the user.

Shared- It defines that the system implements all the security requirements for confidentially and, multiple write access is required, concurrent update areas at a suitable level. It is not all applications required users to write data back, but for the increasing numbers that does, the system must be able to handle several updates in an appropriate, secure manner. This is a major field of weakness in some OLAP products, which tend to consider that all OLAP applications will be read-only, with simple security controls.

Multidimensional- The system should support a multidimensional conceptual view of the data, including complete support for hierarchies and multiple hierarchies. It is not setting up a specific minimum number of dimensions that should be managed as it is too software dependent and most product seem to have enough for their target industry.

Information- Information is all of the data and derived data required, whether it is and however much is relevant for the software. We are measuring the capacity of several products in terms of how much input data can manage, not how many GB they take to save it.

Operations of OLAP:-

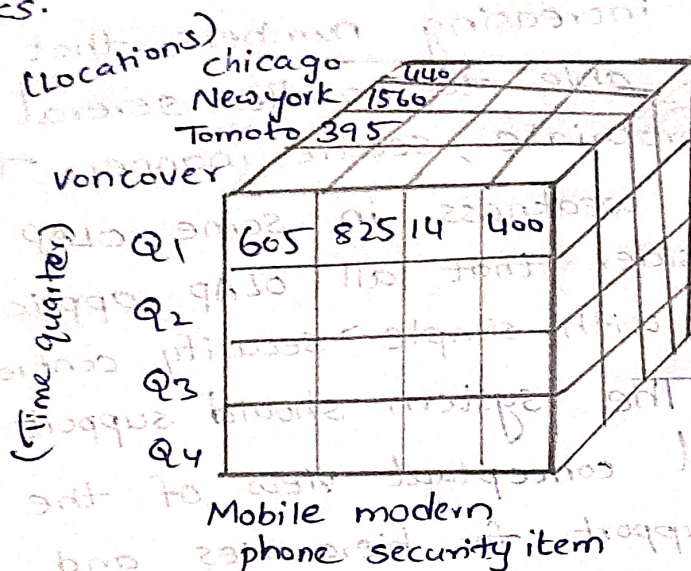
Roll-up:-

Roll up performs aggregation on a data cube in any one of the following ways:

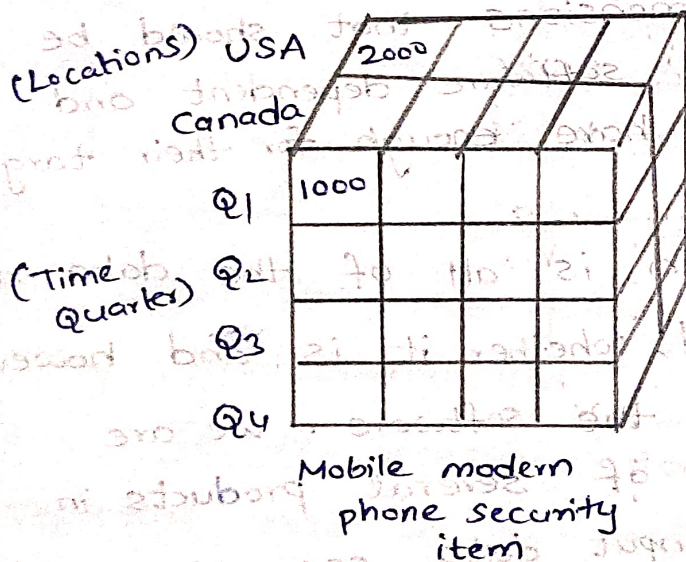
* By climbing up a concept hierarchy for a dimension.

* By dimension reduction.

The following diagram illustrates how roll-up works.



After Roll-up:-



* Roll-up is performed by climbing up a concept hierarchy for the dimension location.

* Initially the concept hierarchy was "street < city < province < country".

* On rolling up, the data is aggregated by

ascending the location hierarchy from the level of country.

- * The data is grouped into cities rather than countries.

- * when roll-up is performed, one or more dimensions from the data cube are removed.

Drill-down:-

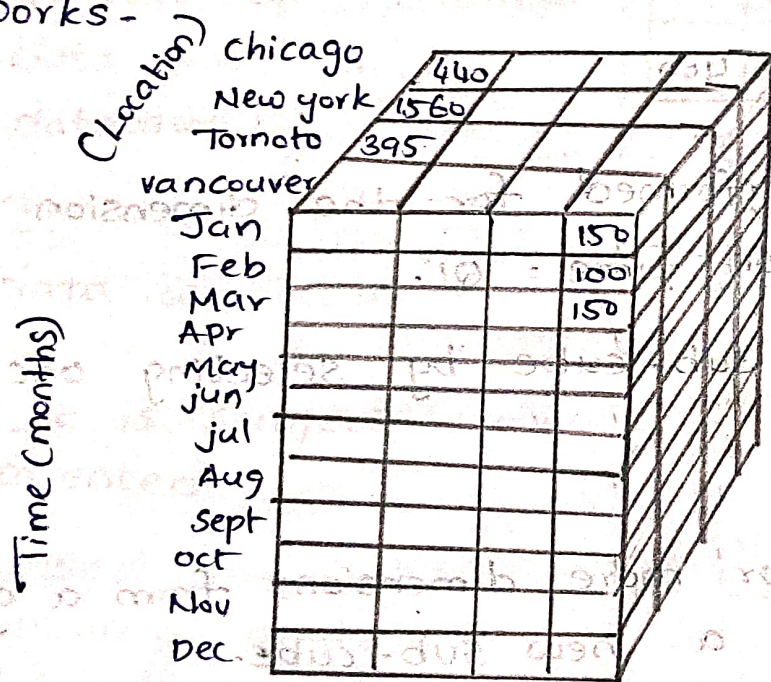
Drill-down is the reverse operation of roll-up.

It is performed by either of the following ways-

- * By stepping down a concept hierarchy for a dimension.

- * By introducing a new dimension.

The following diagram illustrates how drill-down works-



- * Drill-down is performed by stepping down a concept hierarchy for the dimension time.

- * Initially the concept hierarchy was "day < month < quarter & It year."

- * On drilling down, the time dimension is descended from the level of quarter to level of month.

ascending the location hierarchy from the level of country.

- * The data is grouped into cities rather than countries.
- * When roll-up is performed, one or more dimensions from the data cube are removed.

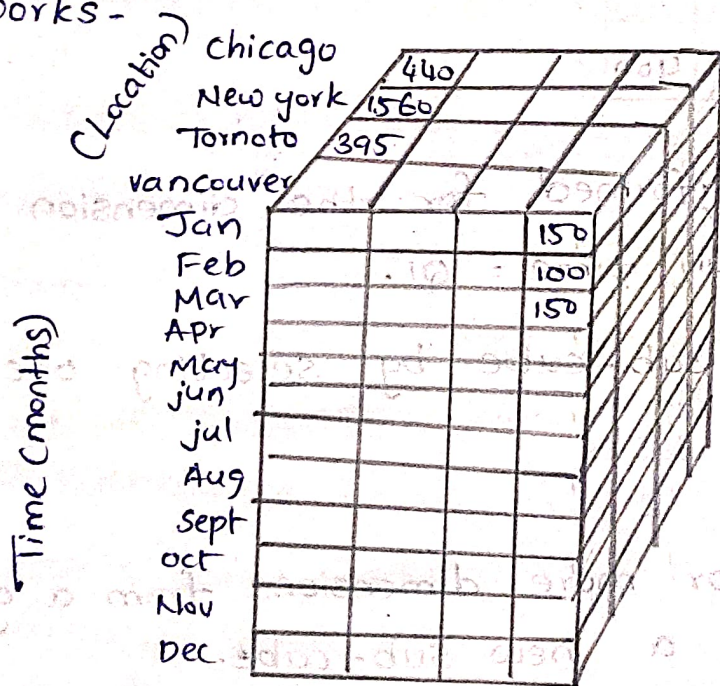
Drill-down:-

Drill-down is the reverse operation of roll-up.

It is performed by either of the following ways-

- * By stepping down a concept hierarchy for a dimension.
- * By introducing a new dimension.

The following diagram illustrates how drill-down works-



- * Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- * Initially the concept hierarchy was "day < month < quarter & It year."
- * On drilling down, the time dimension is descended from the level of quarter to level of month.

* when drill-down is performed, one or more dimensions from the data cube are added.

* It navigates the data from less detailed data to highly detailed data.

Slice:-

The slice operation selects one particular dimension from a given cube and provides a new sub-cube.

Consider the following diagram that shows how slice works:-

Chicago				
New York				
Toronto				
Vancouver	605	825	14	400

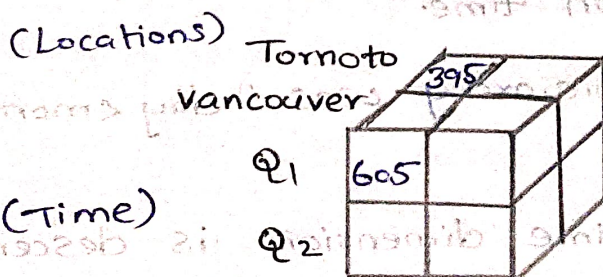
* Here slice is performed for the dimension "time" using the criterion time = "Q1".

* It will form a sub-cube by selecting one or more dimensions.

Dice:-

Dice selects two or more dimensions from a given cube and provides a new sub-cube.

Consider the following diagram that shows the dice operation:-



The dice operation on the cube based on the following selection criteria involves three dimensions-

1) (location = "Toronto" or "Vancouver") (time = "Q1" or "Q2")

2) Explain differences between OLAP and OLTP databases?

Difference between OLAP and OLTP:-

OLAP	OLTP
1) It is well-known as an online database query management system.	1) It is well-known as an online database modifying system.
2) Consists of historical data from various databases.	2) Consist of only operational current data.
3) It makes use of a data warehouse.	3) It makes use of a standard dbms.
4) It is subject oriented.	4) It is application oriented.
5) Used for data mining, analytics, decision making etc..	5) Used for business tasks.
6) In an OLAP database, tables are not normalized.	6) In an OLTP database, tables are normalized (3NF).
7) The data is used in planning, problem solving and decision making.	7) The data is used for performing day-to-day operations.

8) It provides multi-dimensional view of data.

9) It serves the purpose to extract information for analysis and decision-making.

10) A large amount of data is stored typically in TB, PB.

11) Relatively slow as the amount of data involved is large. Queries may take hours.

12) The OLAP database is not often updated. As a result, data integrity is unaffected.

13) It only needs backup from time to time as compared to OLTP.

14) The processing of complex queries can take a lengthy time.

8) It reveals a snapshot of present business tasks.

9) It serves the purpose to Insert, Update and delete information from the database.

10) The size of data is relatively small as the historical data is archived in MB and GB.

11) Very fast as the queries operate on 5% of the data.

12) The data integrity constraint must be maintained in an OLTP databases.

13) The backup and recovery process is maintained rigorously.

14) It is comparatively fast in processing because of simple and straight forward queries.

15) This data is generally managed by CEO, MD and GM	15) This data is managed by clerks, forex and managers.
16) Only read and rarely write operations.	16) Both read and write operations.
17) With lengthy, scheduled batch operations, data is refreshed on a regular basis.	17) The user initiates data updates, which are brief and quick.
18) The process is focused on the customer.	18) The process is focused on the market.
19) Design with a focus on subject.	19) Design that is focused on application.
20) Improves the efficiency of business analysts.	20) Enhances user's productivity.

3. Explain detail about data warehouse architecture of parallel processing and parallel dbms vendors.

A data warehouse is a system that combine data from multiple sources, organizes it under a single architecture, and helps organizes make better decisions. It simplifies data handling, storage and reporting, making analysis more efficient. Data warehouse architecture uses a structured framework to manage and store data

efficiently.

There are two common approaches to constructing a data warehouse:

1) Top-down approach

2) Bottom-up approach

Components of data warehouse architecture:-

A data warehouse architecture consists of several key components that work together to store, manage and analyze data.

External source: where data originates. Includes:

1) Structured (databases, spreadsheets)

2) Semi-structured (XML, JSON)

3) Unstructured (emails, images)

Staging Area: A temporary space where raw data is cleaned and validated before moving to the warehouse. ETL tools manage this process:

1) Extract (E): pulls raw data from sources.

2) Transform (T): standardizes and formats the data.

3) Load (L): Moves the data into the data warehouse.

Data warehouse:

A central storage for organized, cleansed data, including both raw data and metadata. Supports analysis, reporting and decision-making.

Data marts:

Smaller, focused sections of the data warehouse for specific teams (e.g., sales, marketing), enabling quick access to relevant data.

Data mining:

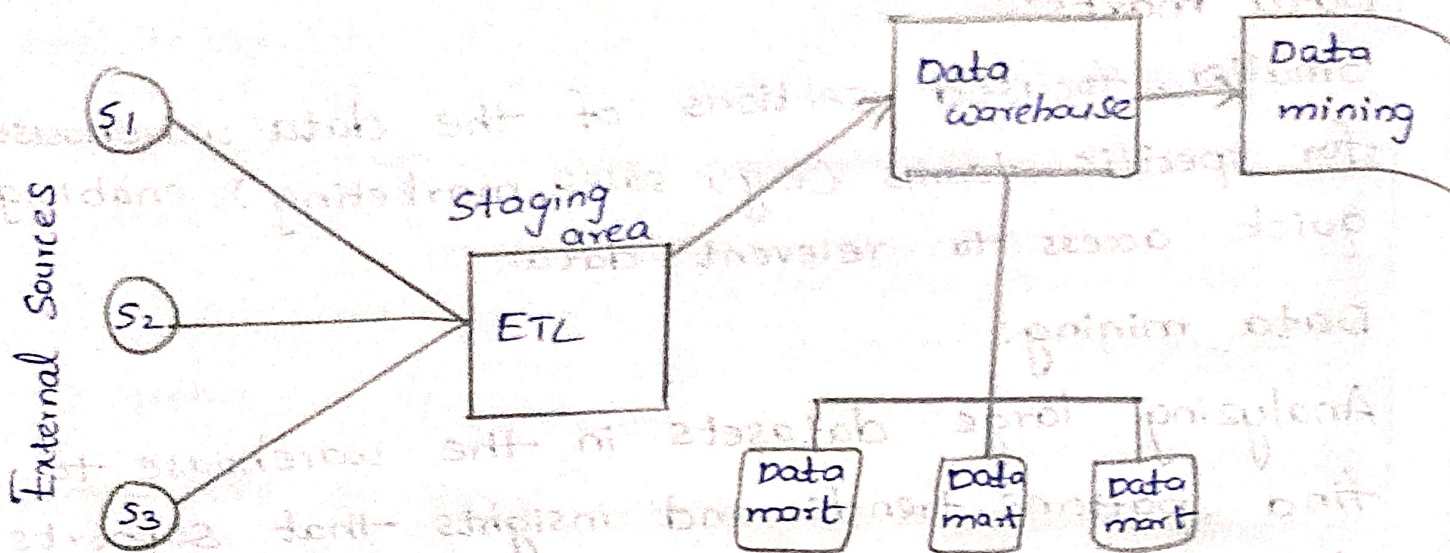
Analyzing large datasets in the warehouse to find patterns, trends and insights that supports decisions and improve operations.

Top-Down approach :-

The Top-down approach, introduced by Bill Inmon, is a method for designing data warehouses that starts by building a centralized, company wide data warehouse. This central repository acts as the single source of truth for managing and analyzing data across the organization. It ensures data consistency and provides a strong foundation for decision making.

Working of Top-down approach :-

- 1) Central data warehouse: The process begins with creating a comprehensive data warehouse where data from various sources is collected, integrated and stored. This involves ETL process to clean and transform the data.
- 2) Specialized Data Marts: Once the central warehouse is established, smaller, department-specific data marts are built. These data marts pull information from the main data warehouse, ensuring access across departments.



Advantages:-

- 1) Consistent view
- 2) High data consistency
- 3) Simplified maintenance
- 4) Scalable architecture
- 5) Stronger governance
- 6) Less data duplication
- 7) Better reporting
- 8) Improved Integration

Disadvantages:-

- 1) High cost & Time
- 2) Complex setup
- 3) Low flexibility
- 4) Limited user input
- 5) Data delays
- 6) Unclear data ownership

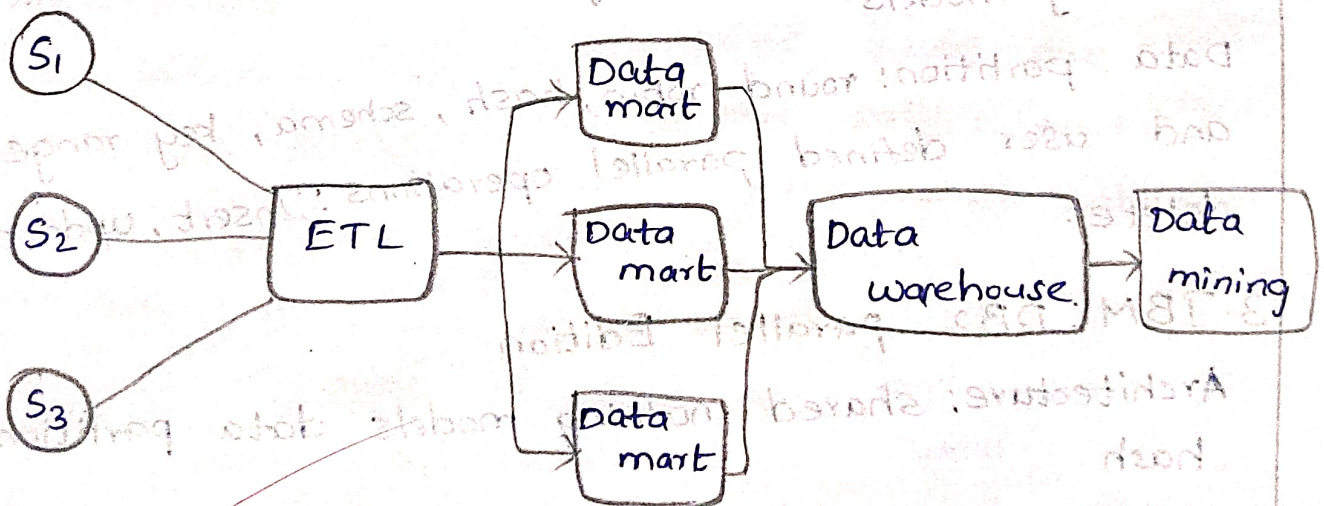
Bottom-up approach:-

The bottom-up approach, popularized by Ralph Kimball, takes more flexible and incremental path to designing data warehouses. Instead of starting with a

a central data warehouse, it designs by building small, department specific data marts that cater to the immediate needs of individual teams, such as sales or finance. These data marts are later integrated to form larger, unified data warehouse.

working of bottom-up approach:-

- 1) Department - specific data marts: The process starts with creating data marts for individual departments or specific business functions. These data marts are designed to meet immediate data analysis and reporting needs, allowing departments to gain quick insights.
- 2) Integration into a data warehouse: Over time, these data marts are connected and consolidated to create a unified data warehouse. The integration ensures consistency and provides a comprehensive view of organization's data.



Advantage:-

- 1) Faster reporting
- 2) step-by-step development
- 3) User-centric
- 4) highly flexible

5) quick results

Disadvantages:-

- 1) Inconsistent views
- 2) Data silos
- 3) Integration difficulty.
- 4) Redundant efforts
- 5) Harder to manage

parallel DBMS Vendors:-

1. Oracle: Parallel Query option (PQO) architecture
shared disk arch.

Data partition: key, range, hash, round robin

parallel operation: hash joins, scan and sort

2. Informix:- extended parallel server (xps)

Architecture: shared memory, shared disk and shared nothing models.

Data partition: round robin, hash, schema, key range
and user defined parallel operations: Insert, update, delete.

3. IBM: DB2 parallel Edition

Architecture: shared nothing models data partition: hash

parallel operations: INSERT, update, delete, load recovery, index creation, backup, table recognition.

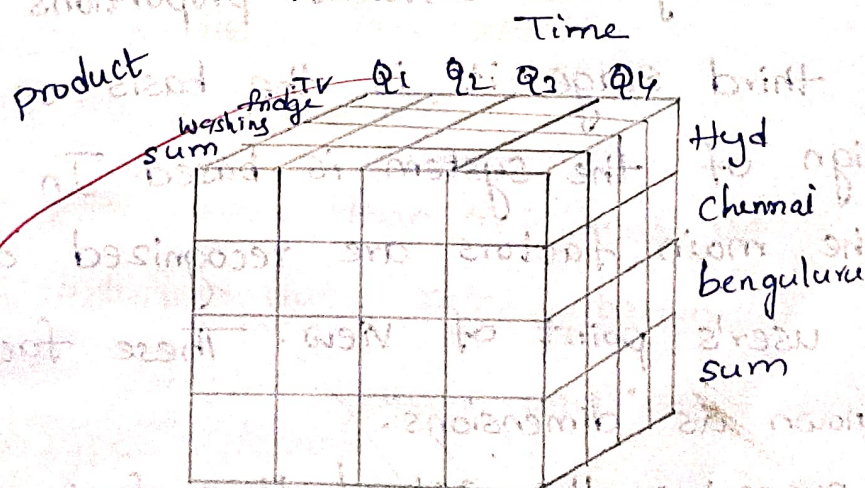
4. Explain multi-dimensional data model and various datawarehouse schema for decision support?

Multi Dimensional Data model:-

A multidimensional data model is defined as a model that allows data to be organized and viewed in multiple dimensions, such as product, time and location.

Features :-

- * It allows users to ask analytical questions associated with multiple dimensions, which help us know market or business trends.
- * OLAP (Online analytical processing) and data warehousing uses multi dimensional databases.
- * It represents data in the form of data cubes. Data cubes allows to model and view the data from many dimensionals and perspectives.
- * It is defined by dimensions and facts and is represented by a fact table. Facts are numerical measures and fact tables contain measures to the related dimensional tables or names of the facts.



working of multidimensional data model:-

The following stages should be followed by every project for building a multi dimensional data model:

Stage 1: Assembling data from the client

In first stage, a multidimensional data model collects correct data from the client. Mostly, software professionals provide simplicity to the client about the range of data with can be guided with the selected technology and collect the complete data in detail.

Stage 2: Grouping different segments of the system

In the second stage, the multi-dimensional data model recognizes and classifies all the data to the respective section they belong to and also builds it problem-free to apply step-by-step

Stage 3: Noticing the different proportions

In the third stage, it is the basis on which the design of the system is based. In this stage, the main factors are recognized according to the user's point of view. These factors are also known as 'dimensions'.

Stage 4: preparing the actual-time factors and their respective qualities

In the fourth stage, the factors which are recognized in the previous step are used further for identifying the related qualities. These qualities are also known as "attributes" in the database.

Stage 5: Finding the actuality of factors which are listed previously and their qualities

In the fifth stage, a multidimensional data model separates and differentiates the actuality from the factors which are collected by it. These actually play a significant role in the arrangement of multi-dimensional data model.

Stage 6: Building the schema to place the data, with respect to the information collected from the steps above

In this sixth stage, on the basis of the data which was collected previously, a schema is built.

Advantage:-

- 1) Easy to handle
- 2) Simple to maintain
- 3) Better performance than relational db.
- 4) More intuitive data representation.
- 5) Handless complex systems and applications well.

Disadvantage:-

- 1) Requires skilled professionals

In the fourth stage, the factors which are recognized in the previous step are used further for identifying the related qualities. These qualities are also known as "attributes" in the database.

stage 5: Finding the actuality of factors which are listed previously and their qualities

In the fifth stage, a multidimensional data model separates and differentiates the actuality from the factors which are collected by it. These actually play a significant role in the arrangement of multi-dimensional data model.

stage 6: Building the schema to place the data, with respect to the information collected from the steps above

In this sixth stage, on the basis of the data which was collected previously, a schema is built.

Advantage:-

- 1) Easy to handle
- 2) Simple to maintain
- 3) Better performance than relational db.
- 4) More intuitive data representation.
- 5) Handless complex systems and applications well.

Disadvantage:-

- 1) Requires skilled professionals
- 2) Complex structure

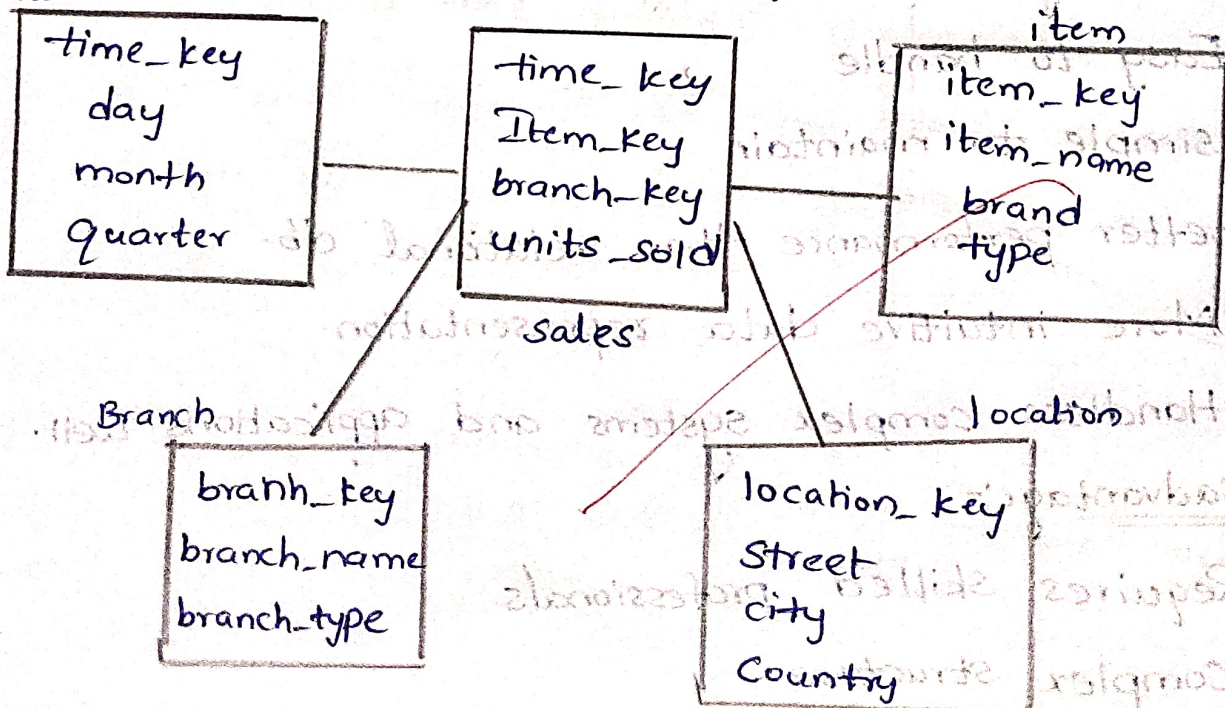
- 3) System performance drops if cache fails
- 4) Dynamic and harder to design
- 5) Longer path to final output

Data warehouse schema:-

Schema is a logical description of the entire database. It includes the name and description of records of all record types including all associative data items and aggregates. Much like a database, a data warehouse also require to maintain a schema. A database uses relational model, while a data warehouse uses star, snowflake, and fact constellation schema.

Star schema:-

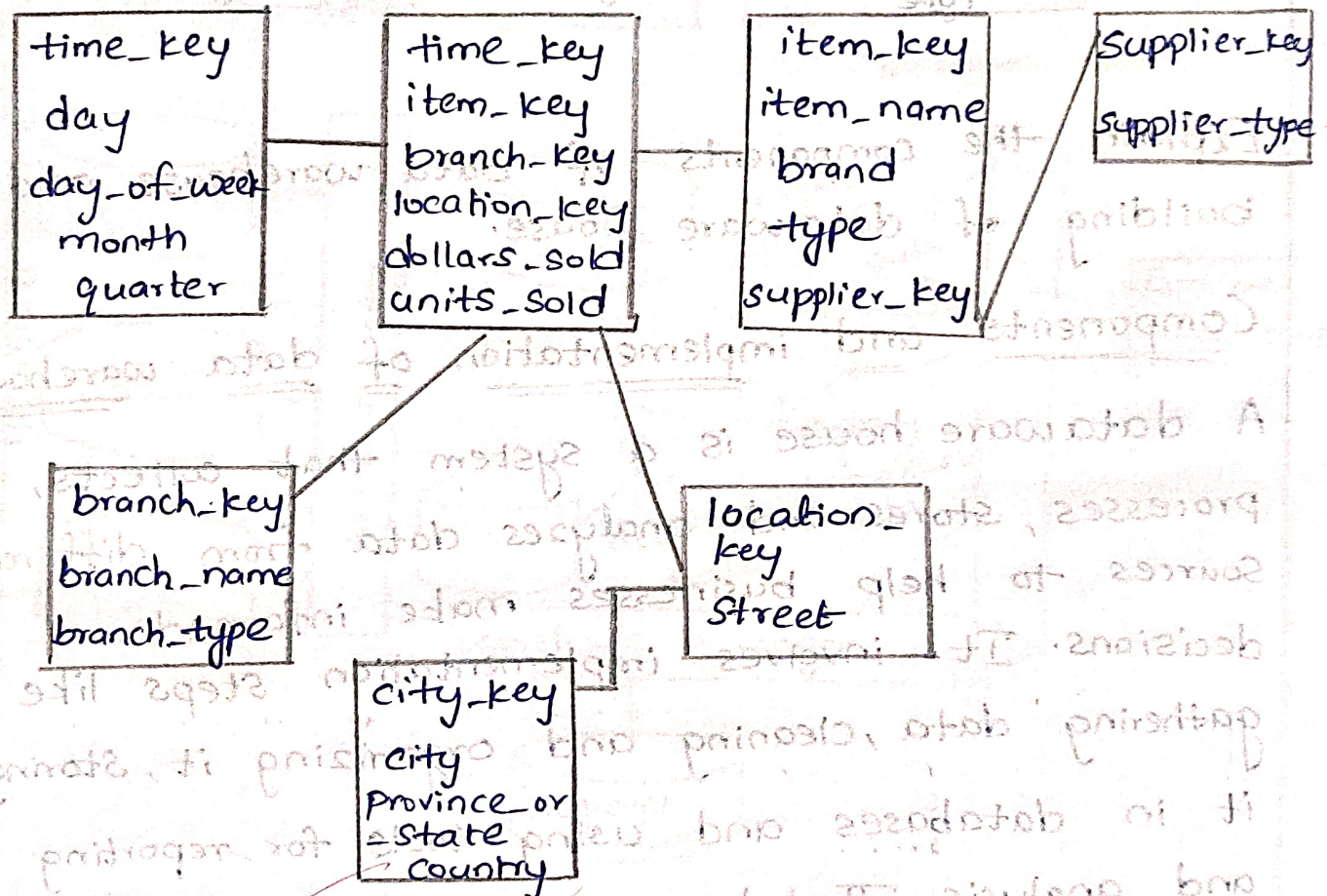
- * Each dimension in a star schema is represented with only one-dimension table.
- * This dimension table contains the set of attributes.
- * The following diagram shows the sales data of company with respect to the four dimensions, namely time, item, branch and location.



Snow flake schema:-

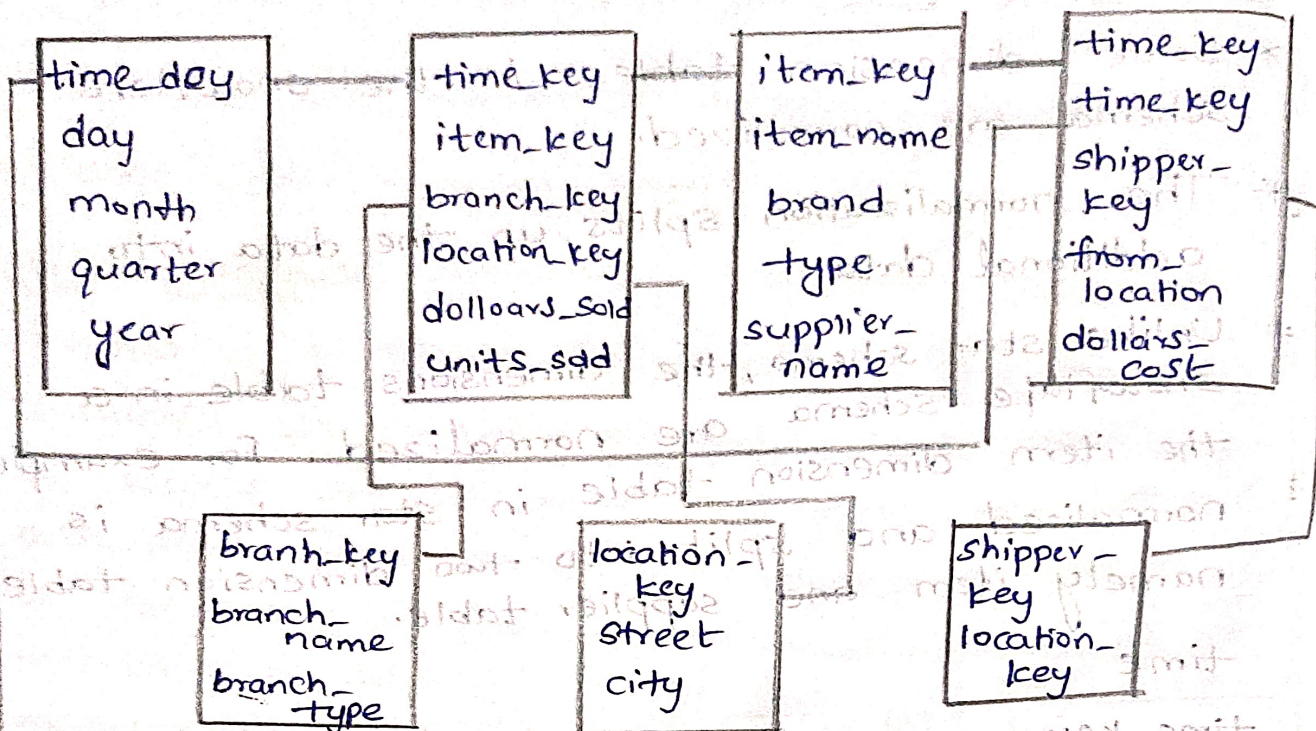
- * Some dimension tables in the snowflake schema are normalized.
- * The normalization splits up the data into additional data.
- * Unlike star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension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

time



Fact constellation schema:-

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



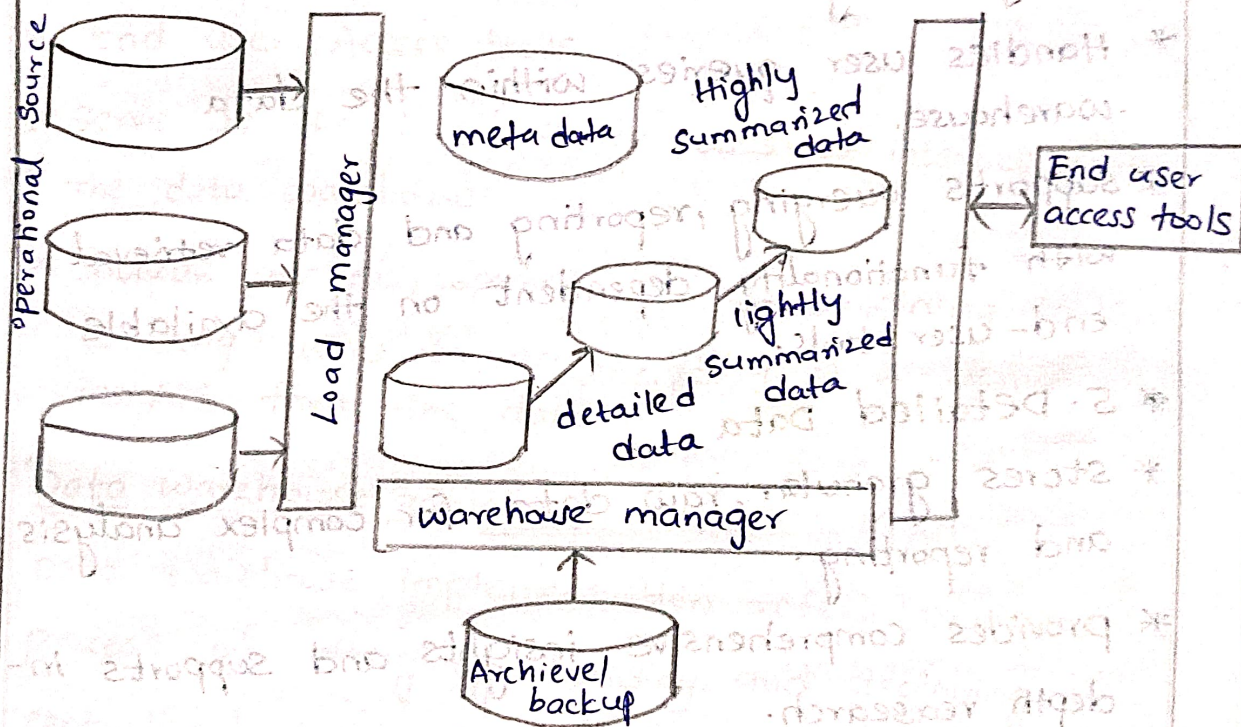
5. Explain the components of data warehouse and building of dataware house.

Components and implementation of data warehouse:

A dataware house is a system that collects, processes, stores and analyzes data from different sources to help businesses make informed decisions. It involves implementation steps like gathering data, cleaning and organizing it, storing it in databases and using tools for reporting and analysis. The key components include data sources, ETL for processing, storage for keeping structured data, metadata for data organisation and query tools for analysis.

components of data warehouse:-

A data warehouse is an integrated system that consolidates data from operational system and external sources, providing valuable insights for decision-making. It includes various components that work together to store, manage and access data efficiently. The data moves from the data source area through the staging area to the presentation server. The entire process is better known as ETL.



1. operational source systems

- * provide raw data from internal systems (e.g., relational databases like informix, oracle) and external sources

- * Serve as the primary input for the data warehouse.

2. Load Manager:-

- * Manages the ETL (Extract, transform, load) process for data extraction and transformation.
- * prepare data for entry into the warehouse, ensuring it meets the required format.

3. warehouse manager

- * Oversees data storage, aggregation and analysis within the data warehouse.
- * Handles task like data denormalization, backup, collection and optimization for better performance.

4. Query manager

- * Handles user queries within the data warehouse.
- * supports querying, reporting and data retrieval, with functionality dependent on the available end-user tools.

* 5. Detailed Data

- * stores granular, raw data for complex analysis and reporting.
- * provides comprehensive insights and supports in-depth research.

6. Summarized data

- * stores predefined aggregations of detailed data for faster queries and reports.
- * provides high-level insights to assist in decision-making.

7. Archive and Backup data

- * Ensures data integrity and disaster recovery through regular backups and archival storage.
- * Archives, older, less frequently accessed data for future use or compliance

8. Metadata

- * Metadata contains information about data structure, source and transformation processes.
- * Supports the ETL process, warehouse management and querying by providing essential context for data.

9. End user Access tools

- * Serve as the interface for users to interact with the data warehouse.
- * Include analysis, reporting and data mining tools, enabling users to access, query and derive insights from the data.

Data Warehouse implementation:-

Data warehouse implementation refers to the process of designing, building and deploying a centralized system that stores, manages and integrates data from various sources to support business analysis and decision-making. The purpose of implementing a data warehouse is to provide organizations with a unified repository of historical data, enabling efficient querying,

reporting and data analysis.

Steps in data warehouse implementation:-

1. Planning and requirement gathering:- Understand business needs, set objectives and determine hardware/software requirements.
2. Data modeling and design:- choose the appropriate schema to optimize data organization and query performance.
3. ETL process: Extract, clean, transform and load data, ensuring consistency and accuracy, with customized ETL tools.
4. Database design and architecture: Design the physical architecture, including storage, indexing and optimization for efficient performance.
5. Data warehouse development: Create tables, views and other objects, ensuring scalability for future data growth.
6. Testing and validation:- Verify data accuracy and performance, ensuring the system meets requirements and resolving any discrepancies.
7. Deployment and maintenance:- Deploy the system, address issues, apply updates and maintain continuous data integration.