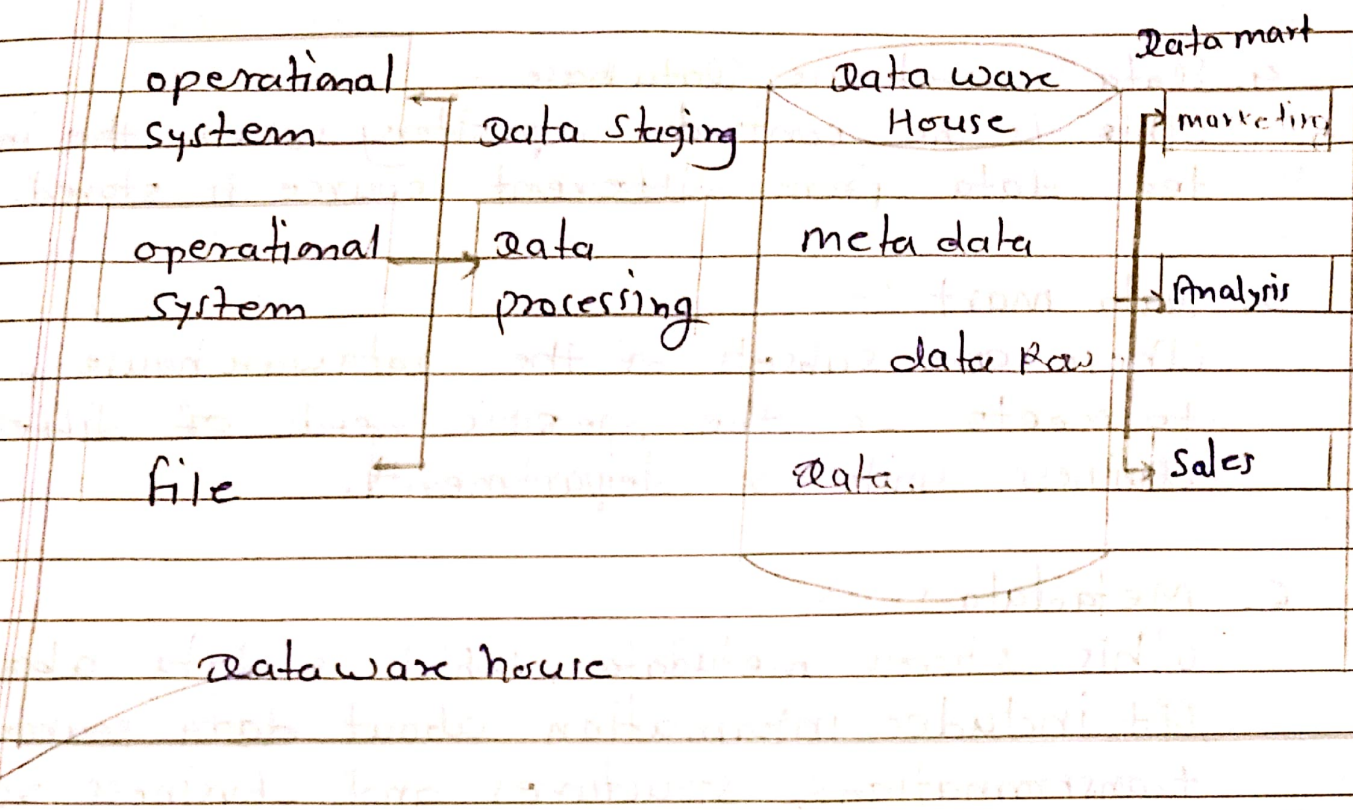
Assignment 2.

Q1 Explain the main component of a typical Data warehouse architecture and their roles?

→



- A typical data warehouse architecture consists of several key components, each playing a curcial rol in the data management and analytic process:

1. Source system:-
   These are the operational database and externa data source where row data originales there include CRM, ERP and Flat file.

2. ETL (Extract Transform Load) Tools:-
   These tools extract data from source systems, transform it into a suitabal format and load it into the data warehouse.
   ETL Process handles data clearing integration, etc

3 Data Staging Area :-
This is a temporary storage area where data is held before it is transformed and loaded into the data warehouse.

4. Data warehouse Database :-
This is the central repository where the integrated data from different source is stored.

5. Data mart :-
These are subsets of the datawarehouse, tailored to meets or the specific needs of different business units or departments.

6. Metadata :-
This stores metadata, which is data about data. It includes information about data source, transformations, structures and business rules.

7. OLAP (Online Analytical Process) Tools :-
These tools enables to perform multiple-dimensional analysis of the data stored in the data warehouse.

Q2. What is difference between the data staging area and the data presentation area in a data warehouse architecture? How do they contribute to the overall functionality of the system.

Data staging area	Data presentation area
• The data staging area is a temporary storage area where row data from various sources is collected.	• The data presentation area is where processed data is stored and made available for querying.
• It serves as a workspace for ETL (Extract Transform Load) processes.	• It is designed to provide efficient data access data.
• Row data is extracted from different source system, which can include flate files, RDBs or snowfalk schema.	• stores integrated and cleaned data in structured format, often in a star database,
• Data Transformed into a suitable format which can include aggregation normalization.	• Organize data is a way that is understables and usable by business uses often through dashboard reported and visualizatio

• Contribution :-

* Data Staging area :-
Ensures data quality and consistency before it is loaded into the warehouse.

• Provides a controlled environment for complex data processing tasks without impacting operational system.

* Data Presentation Area :- facilitates easy and efficient access to data for decision-making process.

- Enhances the performance of queries by provided optimized structures and indexing.

Q3 Describe the concept of data normalization and denormalization in the context of data warehouse architecture. What are the benefits and challenges of each approach?

→ * Data Normalization :-

- Normalization involves organizing data into tables and columns to reduces redundancy and improve data integrity. It follows a set of rules called normal forms.
- It reduces data redundancy improved data integrity, easier data maintence, and minitized update anomalies
- Can lead to complex queries, may impact query performance due to the need for multiple table joins.

* Data Denormalization :-

- Denormalization involves combining tables to reduce the number of joins required for queries. It increases redundancy to optimize need performance.
- Improved query performance, simple queries, faster data retrieval for read-heavy operations.
- Increased data redundancy, potential for data anomalias, more complex data updates and maintance.

Q4. In a Data warehouse architecture, what is the purpose of an OLAP? How does it differ from the OLTP (online Transactional Processing) layer, and why it is important for analytical tasks?

→ OLAP (Online Analytical Processing) layer :-

* Purpose :-

The OLAP layer is designed to support complex analytical queries and multi-dimensional analysis. It allows users to analyze data across multiple dimensional.

* Importance :-

Essential for decission-making and strategic planning, enables users to gains insights from large volumes of data, supports advanced analytical functional such as forecasting and trend analysis.

* Difference from OLTP :-

- Optimized for transactions processing, focuses on data integrity and consistency, handles a large numbers of short online transactions, supports day-to-day operations.
- Optimized for data analysis, focuses on query performance and data aggregation, handles complex queries and large volumes of data supports business intelligence and decision making.

Q.5. Discuss the various data distribution strategies used in Data warehouse architecture. Compare and contrasts hash-based distribution and range-based distribution, highlighting their respective advantage and potential use case.

→

Hash-based Distribution :-

Concept - Data is distributed across nodes based on a hash function applied to a distribution key. Each value of the key is assigned to a specific node.

Advantages :-

- Even distribution of data
- Prevents data skew.
- Efficient for equality-based queries.

* Range-based Distribution :-

Concepts - Data is distributed across nodes based on predefined ranges of the distribution key. Each range is assigned to a specific node.

Advantages :-

- Efficient for range-based queries
- Allow for localized queries processing
- Can optimize data storage for specific access patterns.