PART A

(PART A: TO BE REFFERED BY STUDENTS)

Experiment No. 01

Aim: Case study on building Data Warehouse/Data Mart.

a. Write detailed problem statement and design dimensional modelling
b. Creation of star and snowflake schema.
c. ETL operations

Prerequisite:

Database Management Concepts

Outcome:

After successful completion of this experiment students will be able to

1. Build a Data warehouse.

Theory:

A data warehouse is a centralized repository designed to store, manage and analyze large volumes of data from various sources.

Star Schema: The star schema is a simple and widely used data warehouse schema design. It consists of one or more fact tables referencing any number of dimension tables.

Fact table typically contains numerical measures or facts that business users want to analyze (e.g., sales amount, quantity sold). It also contains foreign key references to the dimension tables.)

Dimension Tables describe the dimensions (or perspectives) of the fact (eg. Time, location, product, etc.) Each dimension table has a primary key that is referred by the fact table.

In a star schema:

- The fact table sits at the center, surrounded by dimension tables radiating out like a star.
- It's denormalized, meaning each dimension is represented by a single
- Queries are generally simpler and can perform well for typical analytical queries.

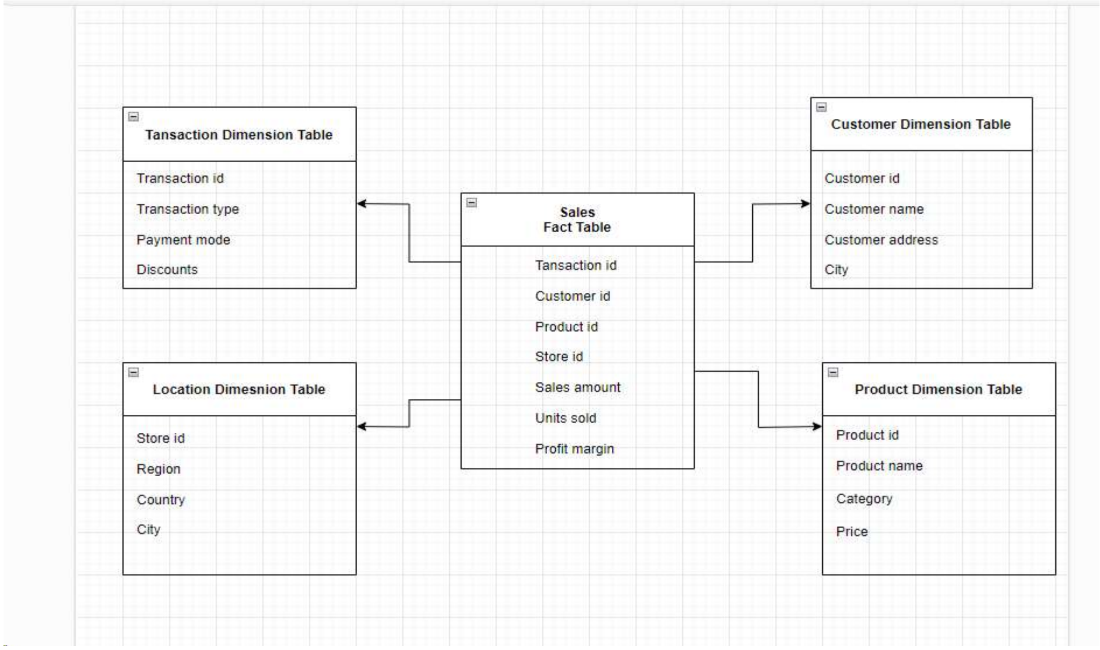


Fig: Star Schema - Sales

Snowflake Schema: It is a more complex variation of the star schema where dimensions are normalized into multiple related tables, which provide more flexibility in terms of data storage and data integrity.

Fact table is similar to that of the star schema. It contains measure and foreign keys to dimension tables.

Dimension Tables are normalized into multiple related tables. For instance, a dimension like "Product" might be split into sub-dimensions such as "Product Category" and "Product Subcategory", each represented by its own table.

In a snowflake schema:

- Dimension tables are normalized, meaning they are structured into a hierarchy of related tables.
- It can save storage space by reducing redundancy in dimension tables.
- Query performance can sometimes be affected due to the increased number of joins needed to retrieve data.

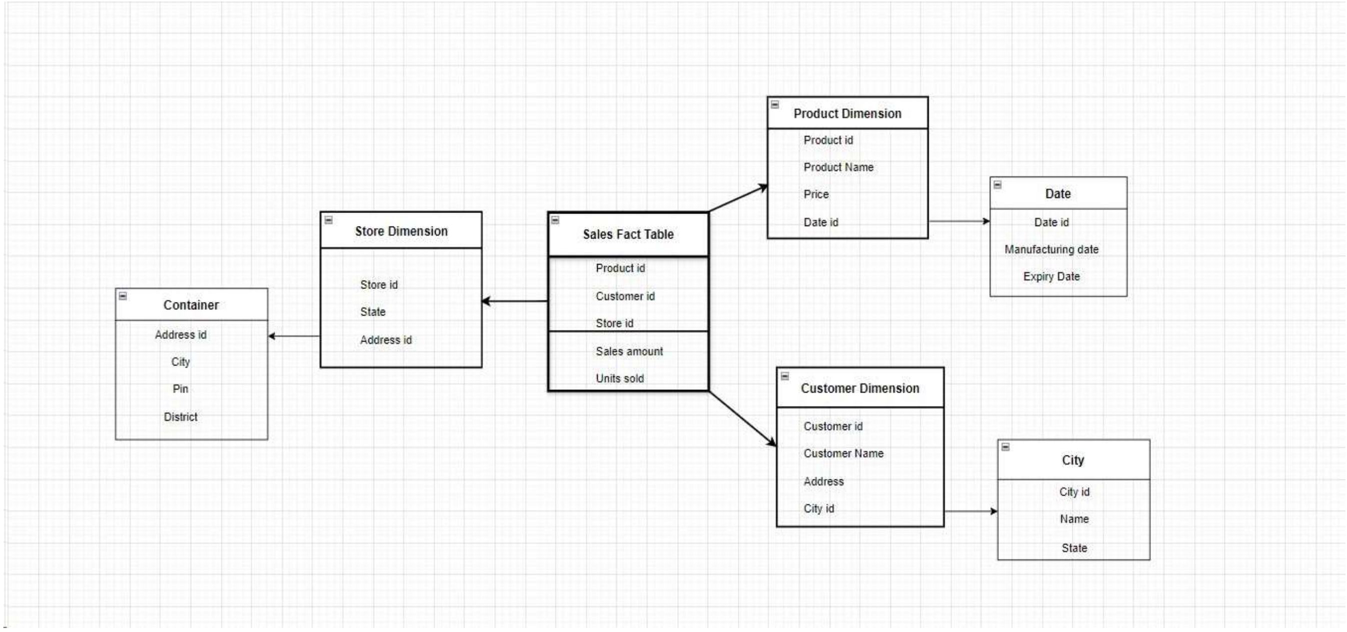


Fig: Snowflake Schema - Sales

ETL Process in Data Warehouse: 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:

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.

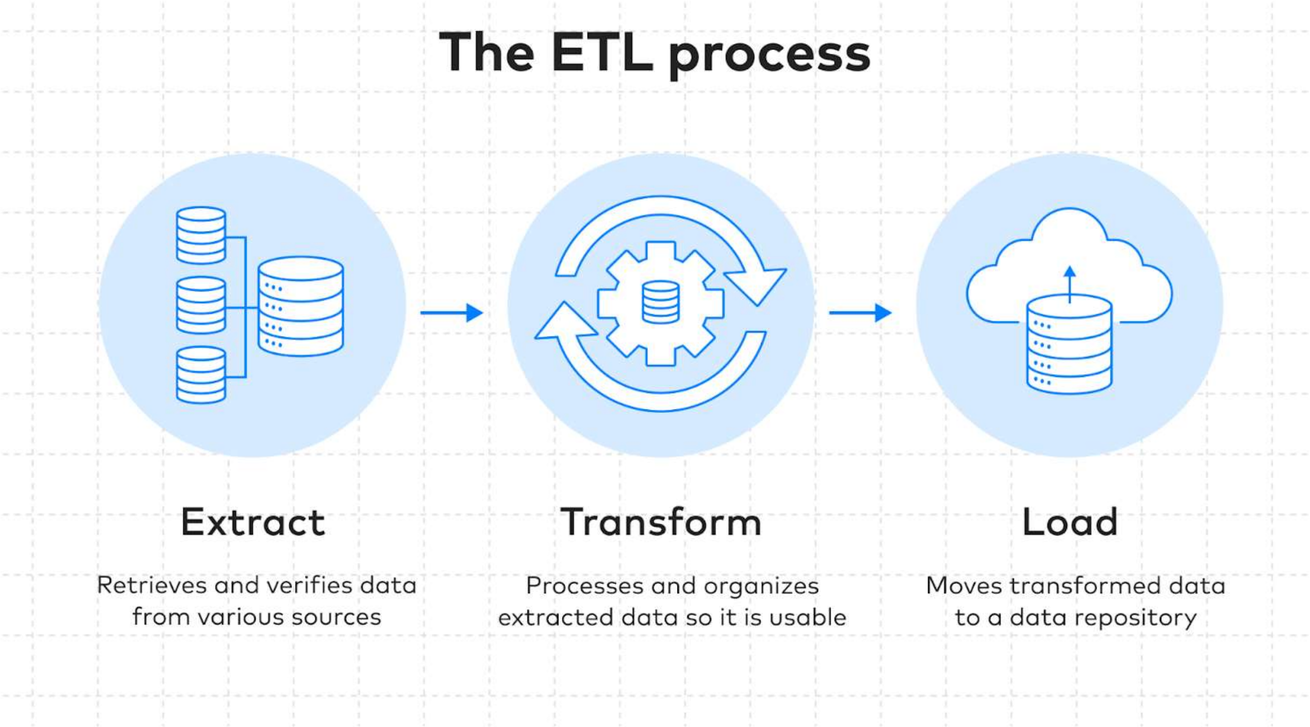

The ETL process
Extract
Retrieves and verifies data from various sources
Transform
Processes and organizes extracted data so it is usable
Load
Moves transformed data to a data repository

Fig: ETL Process

Advantages of ETL process in data warehousing:

1. Improved data quality: ETL process ensures that the data in the data warehouse is accurate, complete, and up-to-date.
2. Better data integration: ETL process helps to integrate data from multiple sources and systems, making it more accessible and usable.
3. Increased data security: ETL process can help to improve data security by controlling access to the data warehouse and ensuring that only authorized users can access the data.
4. Improved scalability: ETL process can help to improve scalability by providing a way to manage and analyze large amounts of data.
5. Increased automation: ETL tools and technologies can automate and simplify the ETL process, reducing the time and effort required to load and update data in the warehouse.

Disadvantages of ETL process in data warehousing:

1. High cost: ETL process can be expensive to implement and maintain, especially for organizations with limited resources.
2. Complexity: ETL process can be complex and difficult to implement, especially for organizations that lack the necessary expertise or resources.
3. Limited flexibility: ETL process can be limited in terms of flexibility, as it may not be able to handle unstructured data or real-time data streams.
4. Limited scalability: ETL process can be limited in terms of scalability, as it may not be able to handle very large amounts of data.
5. Data privacy concerns: ETL process can raise concerns about data privacy.

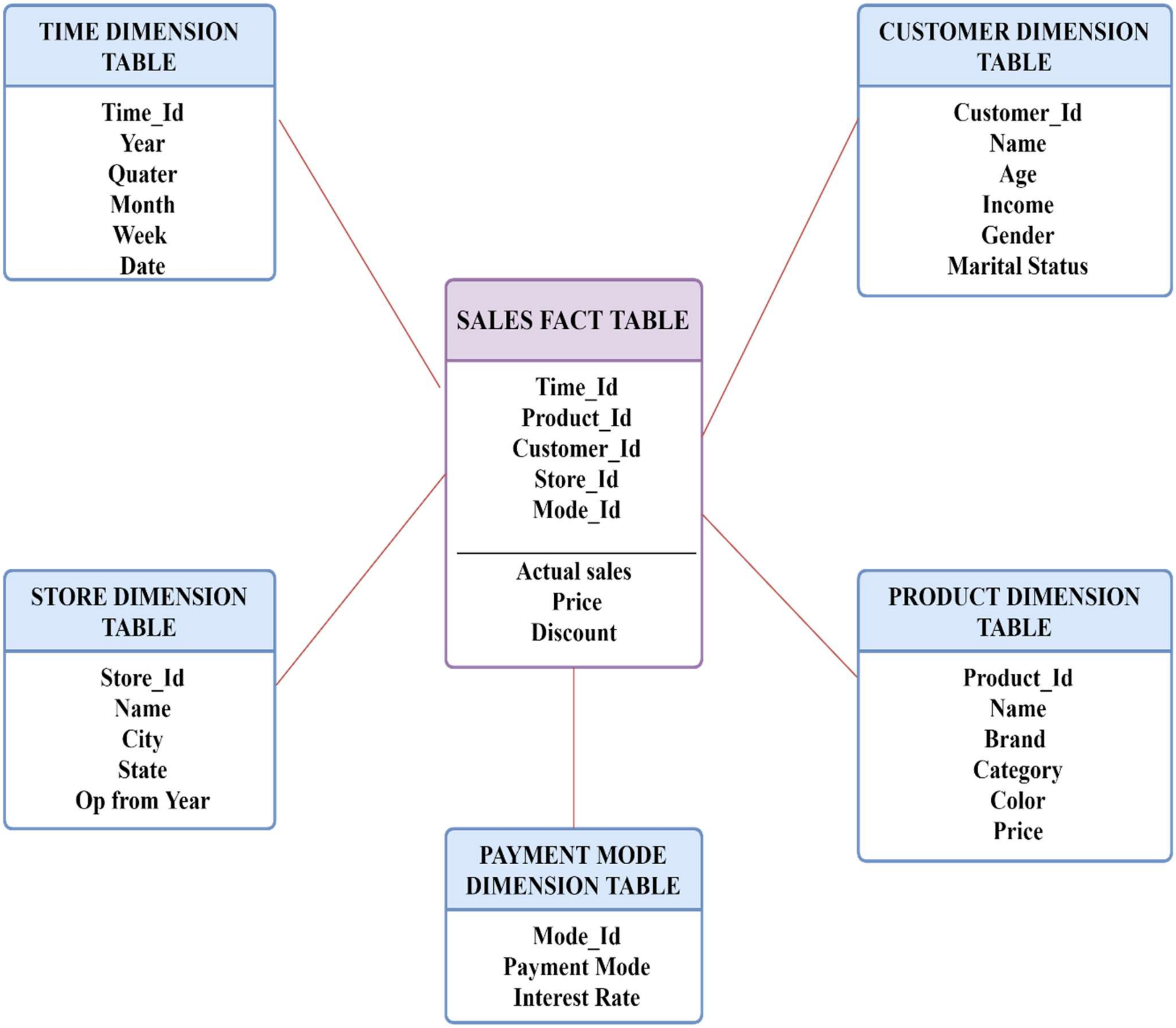
PART B

(PART B: TO BE COMPLETED BY STUDENTS)

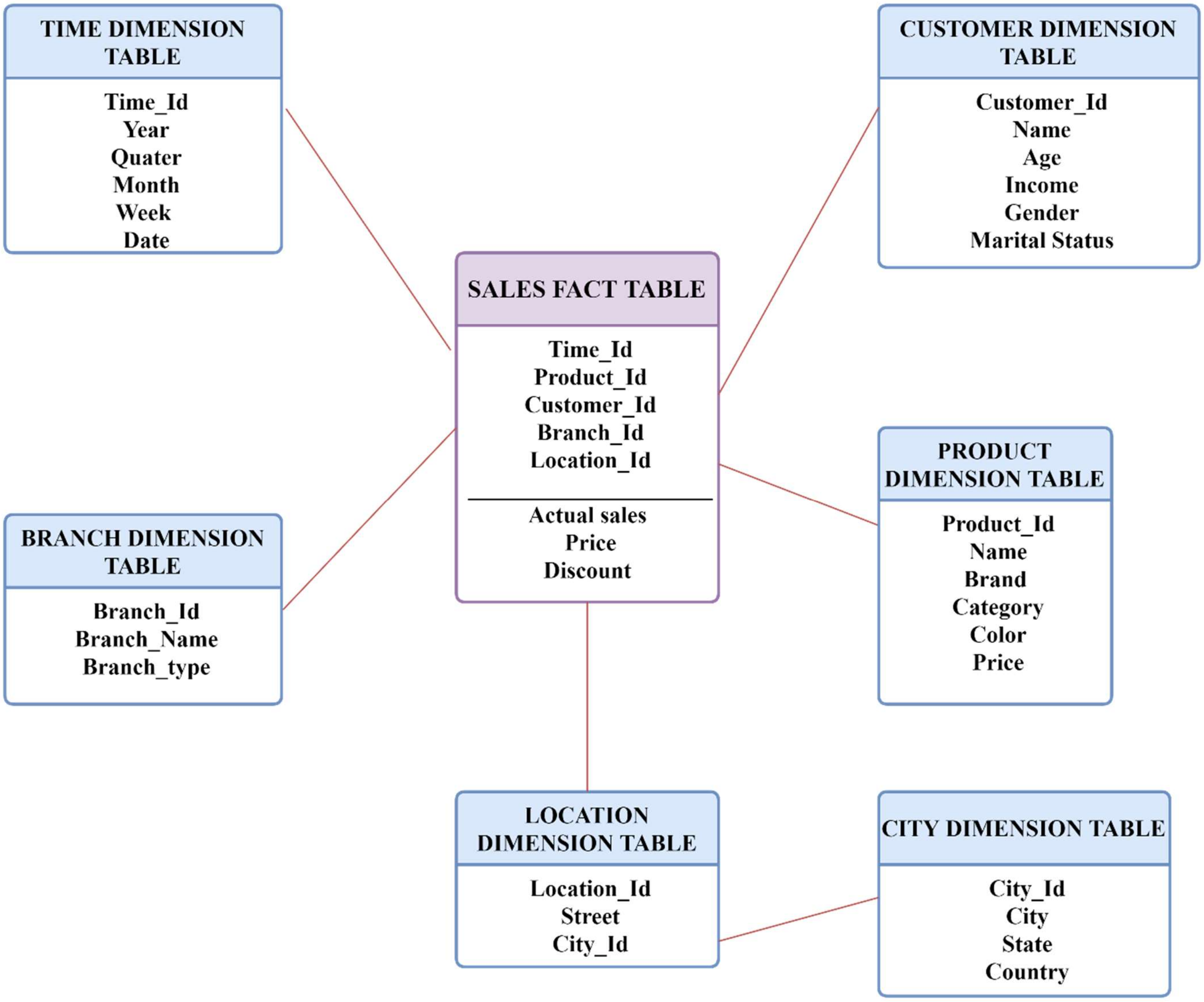
Roll. No: A07	Name: Niharika Seth
Class: AI & DS	Batch: A1
Date of Experiment: 19/07/2024	Date of Submission: 6/08/2024
Grade:	

B.1 Input and Output:

- Star Schema for Sales



- Snowflake Schema for Sales



B.2 Observations and learning:

Star schemas are simpler and generally preferred for querying and reporting due to their denormalized structure. Snowflake schemas are more normalized and reduce redundancy but can make querying more complex.

ETL operations can be complex and require careful design to ensure data quality and performance. The transformation stage is crucial for ensuring that the data conforms to the desired format and business rules.

Both schema design and ETL processes should be optimized for performance to handle large datasets efficiently.

B.3 Conclusion:

In conclusion, the experiment successfully demonstrated the process of building a Data warehouse/Data mart, including problem identification, dimensional modeling, schema design and ETL operations. The implementation of both star and snowflake schemas provided flexibility in meeting different analytical needs while the ETL processes ensured data quality and efficient data loading.

B.4 Question of Curiosity:

Q.1) What is the difference between a star schema and a snowflake schema?

Ans:

Sr. No.	Star Schema	Snowflake Schema
1.	In star schema, the fact tables and the dimension tables are contained.	In snowflake schema, the fact tables, dimension tables as well as sub dimension tables are contained.
2.	Star schema is a top-down model.	Snowflake Schema is a bottom-up model.
3.	Star schema uses more space.	Snowflake Schema uses less space.
4.	It takes less time for the execution of queries.	It takes more time than star schema for the execution of queries.
5.	In star schema, Normalization is not used.	In snowflake schema, both normalization and denormalization are used.
6.	The query complexity of star schema is low.	The query complexity of snowflake schema is higher than star schema.
7.	It has high data redundancy.	It has low data redundancy.
8.	It’s understanding is very simple.	It’s understanding is difficult.
9.	It has less number of foreign keys.	It has more number of foreign keys.
10.	It’s design is very simple.	It’s design is complex.

Q.2) How does effective ETL contribute to the success of a data warehouse?

Ans: Effective ETL contributes to the success of a data warehouse in several ways which are as follows:

1. Data Accuracy: Ensures that the data loaded into the data warehouse is accurate, consistent, and free from errors, leading to reliable analysis and reporting.
2. Data Integration: Integrates data from multiple source systems into a unified format, allowing for comprehensive analysis across different data sources.
3. Timeliness: Transforms and loads data in a timely manner, providing up-to-date information for decision-making and reporting.
4. Data Quality: Applies cleaning and transformation processes to improve data quality, including handling missing values, correcting inconsistencies, and standardizing data formats.
5. Scalability: Supports the ability to handle increasing volumes of data efficiently, ensuring that the data warehouse can grow with the organization’s needs.

6. Performance Optimization: Enhances performance by optimizing data extraction, transformation, and loading processes, resulting in faster query and reporting times.
7. Consistency: Maintains consistency in data definitions and formats across the warehouse, making it easier for users to interpret and analyze data.
8. Error Handling: Implements error handling and recovery procedures to address issues during ETL processes, reducing the risk of data discrepancies and system failures.
9. Business Rules Application: Enforces business rules and transformations to ensure that the data in the warehouse aligns with organizational requirements and standards.
10. Improved Decision-Making: Provides a reliable and accurate data foundation that supports informed and strategic decision-making across the organization.

Q.3) What is dimensional modeling, and why is it used in data warehousing? What are the key components of a dimensional model?

Ans: Dimensional modeling is a design methodology used in data warehousing that focuses on structuring data in a way that optimizes query performance and ease of use for end users. It involves organizing data into fact tables and dimension tables.

- Fact tables contain quantitative data, such as sales figures or transaction counts, and are typically the central point of analysis.
- Dimension tables provide descriptive context to these facts, such as product names, customer details, or time periods.
- This design approach is used because it simplifies data retrieval, making it intuitive for users to query and analyze data.
- By structuring data into a clear, denormalized format, dimensional modeling facilitates fast querying and efficient reporting, which is essential for decision-making and business intelligence applications.

The key components of dimensional model are as follows:

1. Fact Tables: Central tables in the schema that contain quantitative data for analysis, such as sales revenue or transaction counts. They typically include measures and foreign keys to dimension tables.
2. Dimension Tables: Tables that contain descriptive attributes related to the fact data, such as customer details, product descriptions, or time periods. They provide context to the measures in the fact tables.
3. Measures: Numeric data that can be aggregated, such as sales amount, quantity sold, or profit.
4. Attributes: Descriptive fields in dimension tables that provide additional detail about the dimensions, such as product name, customer address, or date.

****