DWDM

1812401160 01
Vishwas Acharya

# Assignment - I

Page No. :
Date : / /

1) What is Data Warehouse? Explain the features of Data Warehouse.

Ans · Collections of databases that work together are called data warehouse.

· This makes it possible to integrate data from multiple databases & it is used to help individuals and organizations make better decisions.

· A database consists of one or more files that need to be stored on a computers.

· In large organizations, databases are typically not stored on the individual computers of employees but in a central system (server).

· As the number and complexity of database grows, we start referring to them together as a data warehouse.

· The ultimate goal of a database is not just to store data, but to help businesses make decisions based on that data.

· According to William H. Inmon, a leading architect in the construction of data warehouse systems, "A data warehouse is

a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process.

- Features of Data Warehouse

- Subject - Oriented
- Integrated
- Time - variant
- Nonvolatile.

- Subject - oriented

↳ A data warehouse is organized around major subjects, such as customer, supplier, product, and sales.

↳ Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data-warehouse focuses on the modeling and analysis of data for decision makers.

↳ Data warehouses typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process

- Integrated:

↳ A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records.

↳ Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.

- Time-Variant:

↳ Data are stored to provide information from a historical perspective (e.g., the past 5-10 years)

↳ Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

- Nonvolatile:

↳ A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment.

Page No. :
Date : / /

2. Explain the types of Data Warehouse.

Ans. The three main types of data warehouse are:

- Enterprise Data Warehouse:

↳ It is a centralized warehouse, which provides decision support service across the enterprise.

↳ It offers a unified approach to organizing and representing data.

↳ It also provides the ability to classify data according to the subject and give access according to those divisons.

- Operational Data Store:

↳ It is also called ODS, is data store required when neither data warehouse nor OLTP systems support organizations reporting needs.

↳ It is widely preferred for routine activities like storing records.

↳ In ODS, Data warehouse is refreshed in real time.

Page No. :
Date : / /

- **Data Mart:**

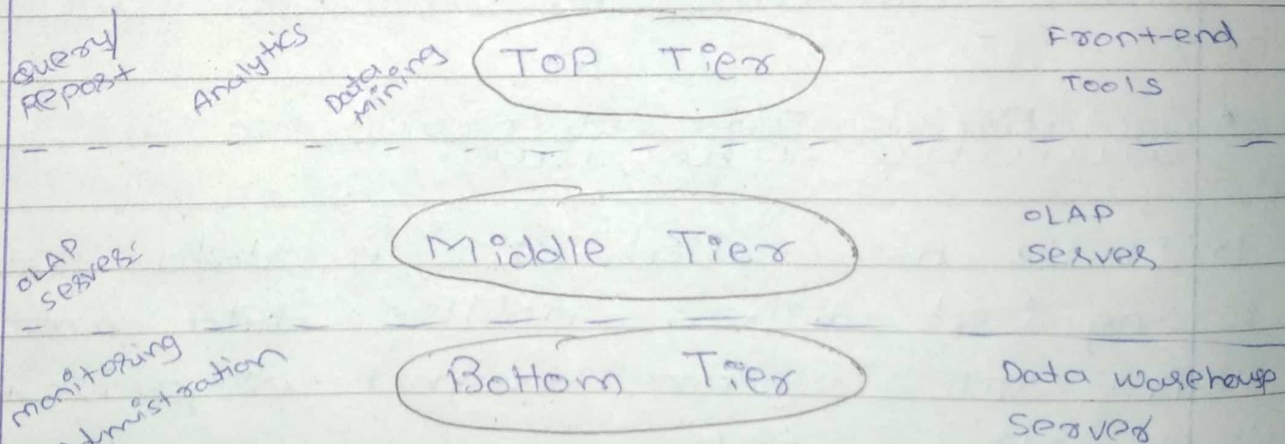
↳ A Data Mart is a subset of the data warehouse.

↳ It specially designed for specific segments like sales, finance, sales, or finance.

↳ In an independent data mart, data can collect directly from sources

3) Explain the Data Warehouse Architecture.

Ans · This architecture is divided into three types.




- **Bottom Tier**

↳ This tier is a warehouse database server that is almost always a relational database system.

↳ Back-end tools and utilities are used to

Page No. :
Date : / /

feed data into the bottom tier from operational databases or other external sources.

↳ These tools and utilities perform data extraction, cleaning, and transformation, as well as load and refresh functions to update the data warehouse.

↳ The data are extracted using application program interfaces known as gateways.

↳ A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server.

↳ Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection)

↳ This tier also contains a metadata (data about data) repository, which stores information about the data warehouse and its contents.

- Middle Tier

↳ The middle tier is an OLAP (online

Page No. :
Date : / /

Analytical Processing Server) that is typically implemented using either

- A relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations

- A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

- Top Tier

↳ The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools

4) Give the difference between OLTP vs OLAP.

Ans · First of all,

OLAP = On-Line Analytical Processing

OLTP = On-Line Transaction Processing

↳ Now, the difference between OLTP and OLAP is given further.

| OLTP | OLAP |
|---|---|
| • Many Short Transactions (Queries + Updates) | • Long Transactions (complex Queries) |
| • Examples<br>↳ Update account balance<br>↳ Enroll in course<br>↳ Add book to shopping cart | • Examples<br>↳ Report total sales for each department in each month<br>↳ Identify top-selling books<br>↳ Count classes with fewer than 10 students |
| • Queries touch small amount of data (one record or few records) | • Queries touch large amout of data |
| • Updates are frequent | • Updates are infrequent |

-> Now, let's see the functionality difference.

| Functionality | OLTP | OLAP |
|---|---|---|
| characteristic | Operational Processing informational processing | Transactional Analysis |
| Orientation | Transaction | Analysis |
| Users | Clerk, DBA, database professional | Knowledge workers (eg., manager, executive, analyst) |
| Function | day-to-day operations | long-term information requirements, decision support |

| Functionality | OLTP | OLAP |
|---|---|---|
| DB design | ER based, application-oriented | Star/snowflake, subject-oriented |
| Data | Current; guaranteed up-to-date | Historical; accuracy maintained over time |
| Summarization | Primitive, highly detailed | Summarized, consolidated |
| View | Detailed, flat relational | Summarized, multidimensional |
| Unit of work | Short, simple transaction | Complex query |
| Access | Read/write | Mostly read |

5) Explain Data Warehouse Schema

Ans
- Data warehouse environment usually transforms the relational data model into some special architectures.

- There are many schema models designed for data warehousing but the most commonly used are:

↳ Star Schema
↳ Snowflake Schema
↳ Fact constellation (Group of star, Collection of fact tables) Schema.

- The determination of which schema model should be used for a data warehouse based upon the analysis of project requirements, accessible tools and project team preferences.

Page No. :
Date : / /

- **STAR Schema**

- The star schema architecture is the simplest data warehouse schema.

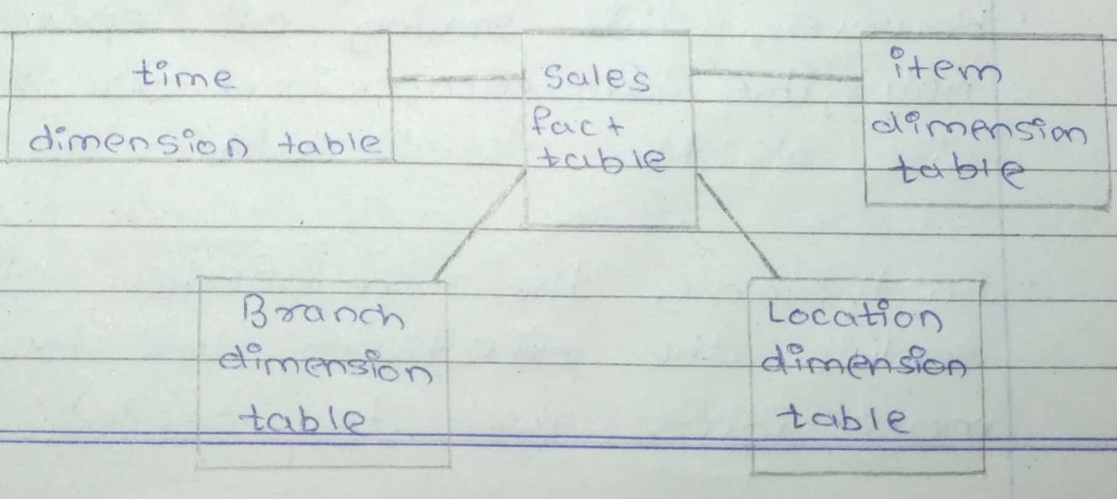
- It is called a star schema because the diagram resembles a star, with points radiating from a center.

- The center of the star consists of fact table and the points of the star are the dimension tables.

- Usually the fact tables in a star schema are in third normal form (3NF) whereas dimensional tables are de-normalized.

- Despite the fact that the star schema is the simplest architecture, it is most commonly used nowadays and is recommended by oracle.

◉ Figure →



- Given diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.

- There is a fact table at the centes. It contains the keys to each of four dimensions.

- The fact table also contains the attributes, namely dollars sold and units sold.

- Snowflake Schema

- The snowflake schema architecture is a more complex variation of the star schema used in a data warehous, becaus the tables which describe the dimensions are normalized.

- This table is easy to maintain and saves storage space.

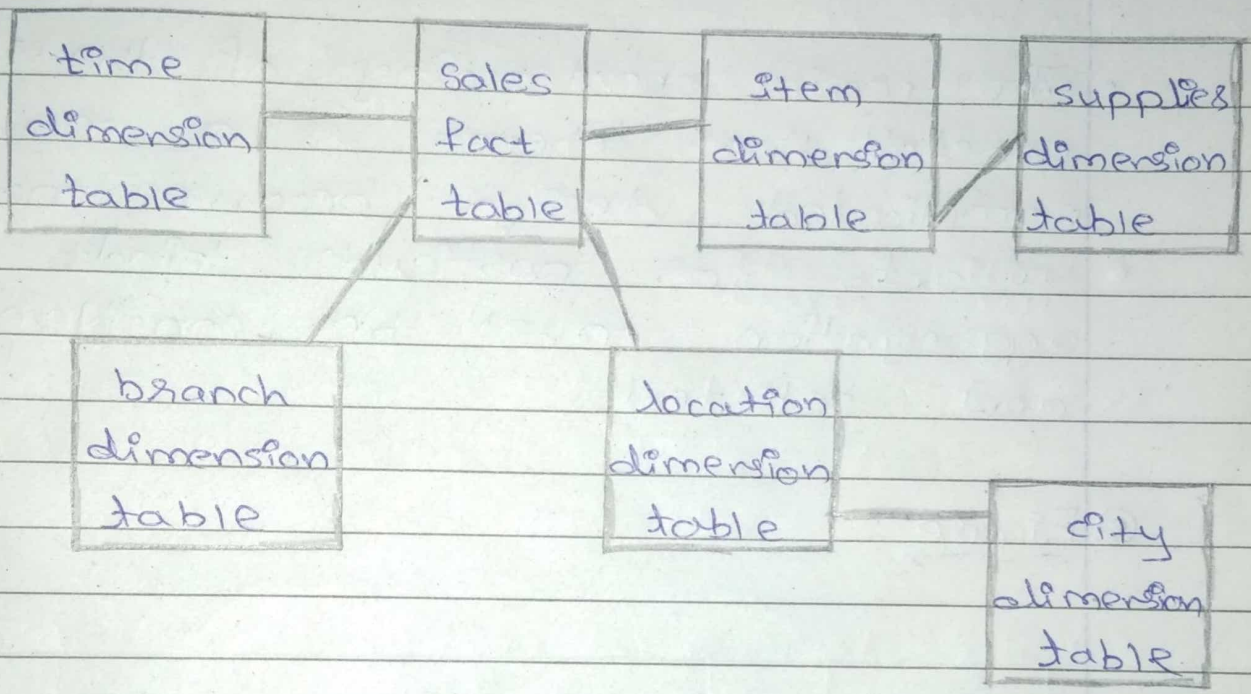
- However, this saving of space is negligible in comparison to the typical size of the fact table.

- Furthermore, the snowflake structure can reduce the effectivess of browsing, since more joins will be needed to execute

a query.

- Hence, although the snowflake schema reduces redundancy, it is not as popular as the star schema in data warehouse design.

◉ Figure →




- Fact Constellation Schema

- Sophisticated applications may require multiple fact tables to share dimension tables.

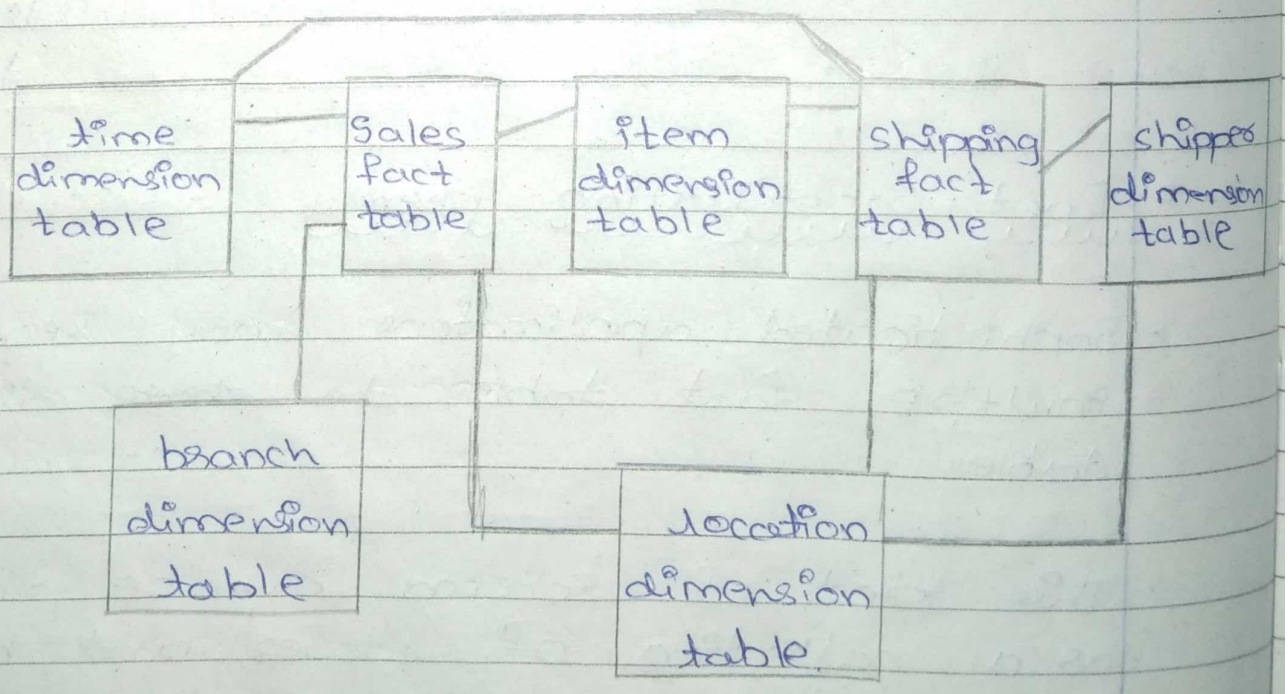
- This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.

- A fact constellation schema allows dimension tables to be shared between fact tables.

- For example, the dimensions tables for time, item, and location are shared between both the sales and shipping fact tables.

- The main shortcoming of the fact constellation schema is a more complicated design because many variants for particular kinds of aggregation must be considered and selected.

◉ Figure →



Page No. :
Date : / /

6) Explain the OLAP Operations in detail.

Ans All the OLAP operations are given below:

- Roll up
- Drill Down
- Slice
- Dice
- Pivot (Rotate)

↳ Now, let's discuss it in detail

- Roll up

↳ It is also called as drill-up or aggregation operation.

↳ It perform aggregation (formation) on a data cube by following ways:

- By climbing up a concept hierarchy for a dimension
- By dimension reduction

↳ 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 city to 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.

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

↳ Initially the concept hierarchy was "day < month < quarter < year."

↳ On drilling down, the time dimension is descended from the level of

quarter to the 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.

↳ Here slice is performed for the dimension "time" using the criterion time = "Q1", time = "Q2", time = "Q3" etc.

↳ It will form a new 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.

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

- (location = "Toronto" or "Vancouver")
- (time = "Q1" or "Q2")
- (item = "Mobile" or "Modem")

- **Pivot**

↳ It is a technique of changing one dimension operation to another.

↳ The pivot operation is also known as rotation

↳ It rotates the data axes in presentation of data.