NAAN MUDHALVAN

BIG DATA ANALYTICS

ASSESSMENT – VII

Data Warehouseing

Homework

1. What is Data warehouse? List the types of data warehouse architecture.

A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. The data within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications.A data warehouse centralizes and consolidates large amounts of data from multiple sources. Its analytical capabilities allow organizations to derive valuable business insights from their data to improve decision-making. Over time, it builds a historical record that can be invaluable to data scientists and business analysts. Because of these capabilities, a data warehouse can be considered an organization's "single source of truth."

The types of Data Warehouse Architecture are:

1.1 Data Warehouse Architecture: Basic
1.2 Data Warehouse Architecture: With Staging Area
1.3 Data Warehouse Architecture: With Staging Area and Data Marts

2. What does OLAP stand for?

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.

Most business data have multiple dimensions—multiple categories into which the data are broken down for presentation, tracking, or analysis. For example, sales figures might have several dimensions related to location (region, country, state/province, store), time (year, month, week, day), product (clothing, men/women/children, brand, type), and more.

But in a data warehouse, data sets are stored in tables, each of which can organize data into just two of these dimensions at a time. OLAP extracts data from multiple relational data sets and reorganizes it into a multidimensional format that enables very fast processing and very insightful analysis.

3. What does OLAP stand for?

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.

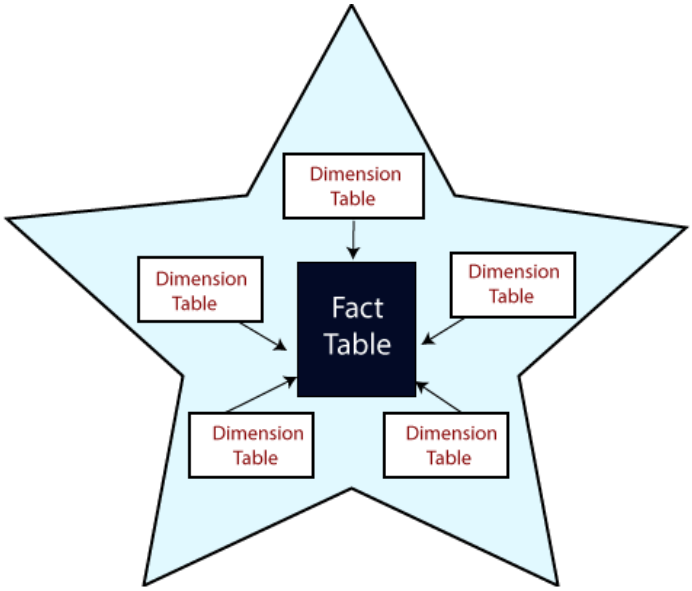
Most business data have multiple dimensions—multiple categories into which the data are broken down for presentation, tracking, or analysis. For example, sales figures might have several dimensions related to location (region, country, state/province, store), time (year, month, week, day), product (clothing, men/women/children, brand, type), and more.

But in a data warehouse, data sets are stored in tables, each of which can organize data into just two of these dimensions at a time. OLAP extracts data from multiple relational data sets and reorganizes it into a multidimensional format that enables very fast processing and very insightful analysis.

4. What is star schema?

A star schema is the elementary form of a dimensional model, in which data are organized into **facts** and **dimensions**. A fact is an event that is counted or measured, such as a sale or log in. A dimension includes reference data about the fact, such as date, item, or customer.

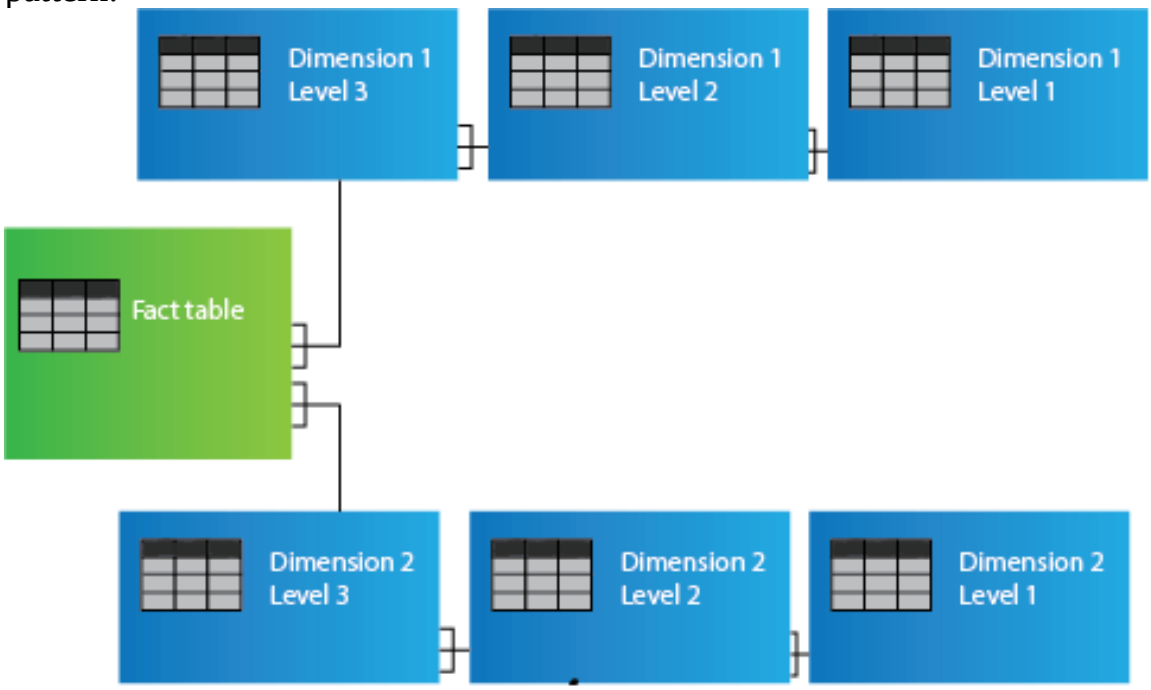
A star schema is a relational schema where a relational schema whose design represents a multidimensional data model. The star schema is the explicit data warehouse schema. It is known as **star schema** because the entity-relationship diagram of this schemas simulates a star, with points, diverge from a central table. The center of the schema consists of a large fact table, and the points of the star are the dimension tables.


Dimension Table
Dimension Table
Dimension Table
Fact Table
Dimension Table
Dimension Table

Star Schema

5. What is snowflake ?

A snowflake schema is equivalent to the star schema. “A schema is known as a snowflake if one or more dimension tables do not connect directly to the fact table but must join through other dimension tables.”The snowflake schema is an expansion of the star schema where each point of the star explodes into more points. It is called snowflake schema because the diagram of snowflake schema resembles a snowflake. Snowflaking is a method of normalizing the dimension tables in a STAR schemas. When we normalize all the dimension tables entirely, the resultant structure resembles a snowflake with the fact table in the middle.Snowflaking is used to develop the performance of specific queries. The schema is diagramed with each fact surrounded by its associated dimensions, and those dimensions are related to other dimensions, branching out into a snowflake pattern.

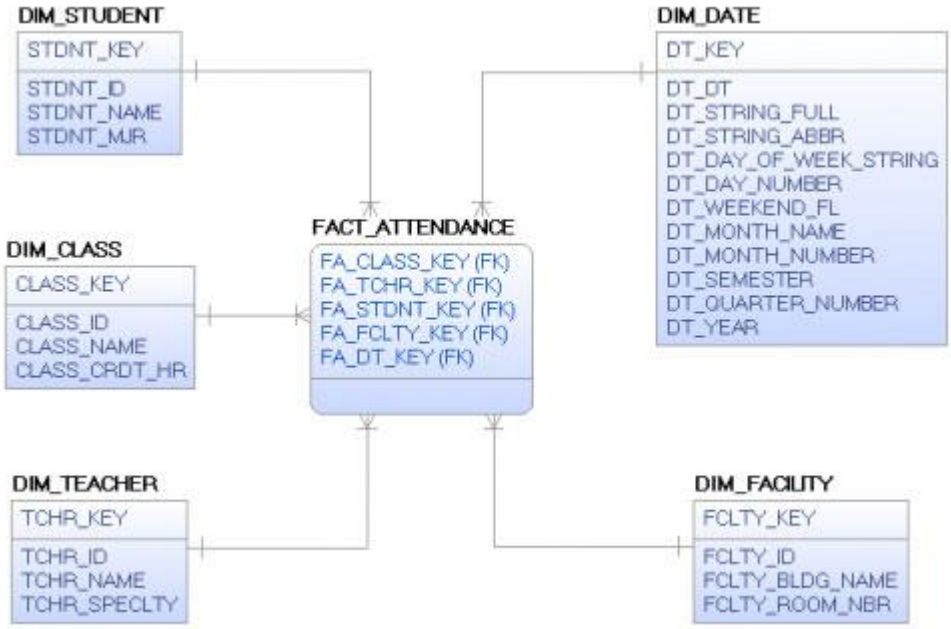


Snowflake Schema

6. Define fact-less fact .

Factless facts are those fact tables that have no measures associated with the transaction. Factless facts are a simple collection of dimensional keys which define the transactions or describing condition for the time period of the fact.

Eg: FACT_ATTENDANCE is an amalgamation of the DATE_KEY, the STUDENT_KEY, and the CLASS_KEY



7. What do you understand by dimensional modelling?

Data warehouse modeling is the process of designing the schemas of the detailed and summarized information of the data warehouse. The goal of data warehouse modeling is to develop a schema describing the reality, or at least a part of the fact, which the data warehouse is needed to support.
Data warehouse modeling is an essential stage of building a data warehouse for two main reasons. Firstly, through the schema, data warehouse clients can visualize the relationships among the warehouse data, to use them with greater ease. Secondly, a well-designed schema allows an effective data warehouse structure to emerge, to help decrease the cost of implementing the warehouse and improve the efficiency of using it.Data modeling in data warehouses is different from data modeling in operational database systems. The primary function of data warehouses is to support DSS processes. Thus, the objective of data warehouse modeling is to make the data warehouse efficiently support complex queries on long term information.

8. What is a data Mart?

A data mart is a simple form of data warehouse focused on a single subject or line of business. With a data mart, teams can access data and gain insights faster, because they don't have to spend time searching within a more complex data warehouse or manually aggregating data from different sources.