



## **Data cube and Data warehouse Architecture**

### 1. What are the uses of multifeature cubes?

multifeature cubes , which compute complex queries involving multiple dependent aggregates at multiple granularity. These cubes are very useful in practice. Many complex data mining queries can be answered by multifeature cubes without any significant increase in computational cost, in comparison to cube computation for simple queries with standard data cubes.

### 2. Compare OLTP and OLAP Systems?

If an on-line operational database systems is used for efficient retrieval, efficient storage and management of large amounts of data, then the system is said to be on-line transaction processing. Data warehouse systems serves users (or) knowledge workers in the role of data analysis and decision-making. Such systems can organize and present data in various formats. These systems are known as on-line analytical processing systems.

### 3. What is data warehouse metadata?

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

### 4. Explain the differences between star and snowflake schema.

The dimension table of the snowflake schema model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space.

### 5. In the context of data warehousing what is data transformation?

In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following: Smoothing, Aggregation, Generalization, Normalization, Attribute construction.

## 6. Define Slice and Dice operation

The slice operation performs a selection on one dimension of the cube resulting in a sub cube. The dice operation defines a sub cube by performing a selection on two (or) more dimensions.

## 7. List the characteristics of a data warehouse.

There are four key characteristics which separate the data warehouse from other major operational systems: 1. Subject Orientation: Data organized by subject 2. Integration: Consistency of defining parameters 3. Non-volatility: Stable data storage medium 4. Time-variance: Timeliness of data and access terms.

## 8. What are the various sources for data warehouse?

Handling of relational and complex types of data: Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. Mining information from heterogeneous databases and global information systems:

Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases

## 9. What is bitmap indexing?

The bitmap indexing method is popular in OLAP products because it allows quick searching in data cubes. The bitmap index is an alternative representation of the record ID (RID) list.

## 10. What is data warehouse?

A data warehouse is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making . (or) A data warehouse is a subject-oriented, time-variant and nonvolatile collection of data in support of management's decision-making process.

## 11. Differentiate fact table and dimension table.

Fact table contains the name of facts (or) measures as well as keys to each of the related dimensional tables. A dimension table is used for describing the dimension. (e.g.) A dimension table for item may contain the attributes item\_name, brand and type.

## 12. Briefly discuss the schemas for multidimensional databases.

Stars schema: The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the

data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension. Snowflakes schema: The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake. Fact Constellations: 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

13. How is a data warehouse different from a database? How are they similar?

Data warehouse is a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision-making. A relational databases is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes(columns or fields) and usually stores a large set of tuples(records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values. Both are used to store and manipulate the data.

14. What is descriptive and predictive data mining?

Descriptive data mining, which describes data in a concise and summarative manner and presents interesting general properties of the data. predictive data mining, which analyzes data in order to construct one or a set of models and attempts to predict the behavior of new data sets. Predictive data mining, such as classification, regression analysis, and trend analysis.

15. List out the functions of OLAP servers in the data warehouse architecture.

The OLAP server performs multidimensional queries of data and stores the results in its multidimensional storage. It speeds the analysis of fact tables into cubes, stores the cubes until needed, and then quickly returns the data to clients.

16. Differentiate data mining and data warehousing.

data mining refers to extracting or “mining” knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named “knowledge mining from data,” A data warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount.

17. What do you understand about knowledge discovery?

people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as a process and an iterative sequence of the following steps: 1. Data cleaning (to remove noise and inconsistent data) 2. Data integration (where multiple data sources may be combined) 3. Data selection (where data relevant to the analysis task are retrieved from the database) 4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance) 5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns) 6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures) 7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)