**1) What is data modelling?**

Data modelling is the process of creating a model for the data to store in a database. It is a conceptual representation of data objects, the association between different data objects, and the rules.

**2) Explain various types of data models**

There are mainly three different types of data models:

Conceptual: Conceptual data model defines what should the system contain. This model is typically created by business stakeholders and data architects. The purpose is to organize, scope, and define business concepts and rules.

Logical: Defines how the system should be implemented regardless of the DBMS. This model is typically created by data architects and business analysts. The purpose is to develop a technical map of rules and data structures.

Physical: This data model describes how the system will be implemented using a specific DBMS system. This model is typically created by DBA and developers. The purpose is the actual implementation of the database.

**Explain the fact and fact table**

The fact represents quantitative data. For example, the net amount which is due. A fact table contains numerical data as well as foreign keys from dimensional tables.

**List out various design schema in data modelling**

There are two different types of data modelling schemes schemas: 1) Star Schema, and 2) Snowflake Schema.

**When should you consider denormalization?**

Denormalization is used when there is a lot of involvement of the table while retrieving data. It is used to construct a data warehouse.

**Explain dimension and attribute**

Dimensions represent qualitative data. For example, product, class, plan, etc. A dimension table has textual or descriptive attributes. For example, the product category and product name are two attributes of the product dimension table.

**What is the fact less fact?**

Fact less fact is a table having no fact measurement. It contains only the dimension keys.

**What is in-memory analytics?**

In-memory analytics is a process of caching the database in RAM.

**What is the difference between OLTP and OLAP?**

**Define data sparsity**

Data sparsity is a term used for how much data you have for entity/ dimension of the model.

**What is metadata?**

Metadata describes the data about data. It shows what type of data is actually stored in the database system.

**What is data mart?**

A data mart is a condensed version of a data warehouse and is designed for use by a specific department, unit, or set of users in an organization. E.g., marketing sales, HR, or finance.

**What is OLTP?**

Online transaction processing, shortly known as OLTP, supports transaction-oriented application in 3-tier architecture. OLTP administers the day to day transaction of company or organization.

**What is check constraint?**

Check constraint is used to verify a range of values in a column.

**List out the types of normalization?**

Types of normalizations are: 1) first normal form, 2) second normal form, 3) third normal forms, 4) boyce-codd fourth, and 5) fifth normal forms.

**What is forward data engineering?**

Forward engineering is a technical term used to describe the process of translating a logical model into a physical implement automatically.

**What is PDAP?**

It is a data cube that stores data as a summary. It helps the user to analyse data quickly. The data in PDAP is stored in a way that reporting can be done with ease.

**Explain snowflake schema database design**

A snowflake schema is an arrangement of a dimension table and fact table. Generally, both tables are further broken down into more dimension tables.

**Explain analysis service**

Analysis service gives a combined view of the data that is used in data mining or OLAP.

**What is sequence clustering algorithm?**

Sequence clustering algorithm collects paths which are similar or related to each other and sequences of data having events.

**What is discrete and continuous data?**

Discreet data is a finite data or defined data. E.g., gender, telephone numbers. Continuous data is data that changes in a continuous and ordered manner. E.g., age.

**What is the time series algorithm?**

Time series algorithm is a method to predict continuous values of data in table. E.g., Performance one employee can forecast the profit or influence.

**What is bit mapped index?**

Bitmap indexes are a special type of database index that uses bitmaps (bit arrays) to answer queries by executing bitwise operations

**Explain data warehousing in detail**

Data warehousing is a process for collecting and managing data from varied sources. It provides meaningful business enterprise insights. Data warehousing is typically used to connect and analyse data from heterogeneous sources. It is the core of the BI system, which is built for data analysis and reporting.

**What is junk dimension**

Junk dimension combines two or more related cardinality into one dimension. It is usually Boolean or flag values.

**Explain data scheme**

Data Scheme is a diagrammatic representation that illustrates data relationships and structures.

**Explain data collection frequency**

Data collection frequency is the rate to collect the data. It also passes through various stages. These stages are: 1) extracting from various sources, 3) transforming, 4) cleansing, and 5) storing.

**What is database cardinality?**

Cardinality is a numerical attribute of the relationship between two entities or entity sets.

**What are the different types of cardinal relationships?**

Different types of key cardinal relationships are:

One-to-One Relationships

One-to-Many Relationships

Many-to-One Relationships

Many-to-Many Relationships

|  |  |
| --- | --- |
| **Star Schema** | **Snowflake Schema** |
| Hierarchies for the dimensions are stored in the dimensional table. | Hierarchies are divided into separate tables. |
| It contains a fact table surrounded by dimension tables. | One fact table surrounded by dimension table which is in turn surrounded by dimension table |
| In a star schema, only a single join creates the relationship between the fact table and any dimension tables. | A snowflake schema requires many joins to fetch the data. |
| It has a simple database design | It has a complex database design |
| Denormalized data structure and query also run faster. | Normalized data Structure. |
| High level of data redundancy | Very low-level data redundancy |
|  |  |

**What is index?**

Index is used for a column or group of columns to retrieve data fast.

**What is business or natural keys?**

business or natural keys is a field that uniquely identifies an entity. For example, client ID, employee number, email etc.

**What is a Surrogate Key?**

An artificial key which aims to uniquely identify each record is called a surrogate key. These kinds of key are unique because they are created when you don’t have any natural primary key. They do not lend any meaning to the data in the table. Surrogate key is usually an integer.

**Explain various types of fact tables**

There are three types of fact tables:

Additive: It is a measure that is added to any dimension.

Non-additive: It is a measure that can’t be added to any dimension.

Semi-additive: It is a measure that can be added to a few dimensions.

**What is aggregate table?**

The aggregate table contains aggregated data that can be calculated using functions such as: 1) Average 2) MAX, 3) Count, 4) SUM, 5) SUM, and 6) MIN.

**What is Confirmed dimension?**

A conformed dimension is a dimension which is designed in a way that can be used across many fact tables in various areas of a data warehouse.

**Explain snapshot of data warehouse**

Snapshot is a complete visualization of data at the time when data extraction process begins.

|  |  |
| --- | --- |
| Fact Table | Dimension Table |
| A fact table stores the measurements together with the properties of dimension tables. | The qualities along which the fact table calculates the metric are stored in the dimension tables. |
| There are fewer attributes in a fact table, but there are more records. | The dimension table has fewer records and more attributes. |
| The primary key of a fact table is a concatenation of the primary keys of all dimension tables. | Dimension tables have individual primary keys. |
| You can create fact tables only when a dimension table is complete. | You must create dimension tables first, i.e., before fact tables. |
| There are fewer fact tables in a schema. | There are more dimension tables in a schema |

**Differentiate between the logical data model and physical data model.**

|  |  |
| --- | --- |
| **Logical Data Model** | **Physical Data Model** |
| Entity | Table |
| Attribute | Column |
| Primary Key | Primary Key Constraint |
| Alternate Key | Unique Constraint |
| Inversion Key Entry | Non-Unique Indexes |
| Rule | Check Constraint, Default Value |
| Relationship | Foreign Key |
| Definition | Comment |

**Explain the different types of dimensions with examples.**

There are usually five different categories of dimensions-

a) Conformed dimensions: A conformed dimension is one that is used in multiple domains. You can use it with multiple fact tables in a particular database or even across multiple data marts/warehouses.

b) Role-Playing Dimensions: These are the dimensions in a database that serve various purposes.

c) Slowly Changing Dimension (SCD): SCDs are the most essential dimensions. These are the dimensions where the value of an attribute changes over time. SCDs are classified into five categories: type 0, type 1, type 2, type 3, and type 4.

d) Junk Dimension: This is a dimension table that contains attributes that don't belong in the fact table or any of the existing dimension tables. These dimensions have typical attributes such as flags or indications.

e) Degenerated Dimension: A degenerated dimension is not a fact yet appears as a primary key in the fact table. It doesn't have its own set of dimensions. A single attribute dimension table is another name for it.