1. **What is a data block and what is a data file?**

A data block is nothing but a logical space where the RDBMS database data is stored.

A data file is nothing but a file where all the data is available. For every RDBMS database, we will be having one or more data files associated.

1. **Why it is important to enforce compliance data standards?**

The primary idea of keeping high compliance for data standards is because it will help to reduce the data redundancy and helps the team to have a quality data. As this information is actually carried out or used throughout the organization.

1. **Explain the different data models that are available in detail?**

There are three different kinds of data models that are available and they are as follows:

1. **Conceptual**: High-level design of physical data.

2. **Logical**: Includes entity names, ER, attributes, primary keys and foreign keys.

3. **Physical**: Includes primary keys, foreign keys, tables names and column names.

1. **How big data and Hadoop are related to each other?**

Big data and Hadoop are almost synonyms terms. With the rise of big data, Hadoop, a framework that specializes in big data operations also became popular. The framework can be used by professionals to analyze big data and help businesses to make decisions.

1. **Differentiate between dimension and attribute?**

In short, dimensions are nothing but which represents qualitative data. For example, data like a plan, product, class are all considered as dimensions.

The attribute is nothing but a subset of a dimension. Within a dimension table, we will have attributes. The attributes can be textual or descriptive. For example, product name and product category are nothing but an attribute of product dimensions

1. **What are the different types of measures available?**

The three different types of measures are available; they are as follows:

1. Non-additive measures
2. Semi-additive measures
3. Additive measures
4. **What is ODS?**

An operational data store (“ODS”) is a database designed to integrate data from multiple sources for additional operations on the data.

An ODS is designed for relatively simple queries on small amounts of data (such as finding the status of a customer order), rather than the complex queries on large amounts of data typical of the data warehouse.

An ODS is similar to short term memory where it only stores very recent information. On the contrary, the data warehouse is more like long term memory storing relatively permanent information

1. **What is level of Granularity of a fact table?**

The granularity is the lowest level of information stored in the fact table. The depth of data level is known as granularity.

1. **What is** [**junk**](https://www.nuwavesolutions.com/degenerate-junk-dimensions/) **dimension?**

A junk dimension combines two or more related low cardinality flags into a single dimension.

1. **What is degenerate dimension?**

Degenerate dimension attributes exist in the fact table as a part of the primary key but have no corresponding dimension.

1. **What are the different types of SCD's used in data warehousing?**

-- SCD (Slowly changing dimensions), are the dimensions in which the data changes slowly, rather than changing regularly on a time basis.

Three types of SCDs are used in data warehousing, which are defined as:

– SCD1: It is a record that is used to replace the original record even there is only one record existing in the database. The current data will be replaced and the new data will take its place.

– SCD2: It is the new record file that is added to the dimension table. This record exists in the database with the current data and previous data that is stored in the history.

– SCD3: This uses the original data that is modified to the new data. This consists of two records: one record that exist in the database and another record that will replace the old database record with the new information.

1. **Common challenges of data modeling activity.**

1. Domain knowledge (business logic).

2. Building complex and huge model.

3. Inappropriate use of surrogate key.

4. Carrying out unnecessary de-normalization.

1. **What are two types of schemas in data modeling?**

**a. Star Schema**

This schema is divided into two one is fact table and other is dimension table where all the dimension tables are connected to a fact table. The foreign key in fact table refers to primary keys present in dimension tables.

**b. Snowflake Schema**

In this schema the level of normalization is increased, here the fact table will remain the same as of star schema, here dimension tables are normalized. Due to many layers of dimension tables, it looks like a snowflake, thus the name snowflake schema.

1. **What is Data Lake?**

―A data lake is a storage repository that holds a vast amount of raw data in its native format, including structured, semi structured, and unstructured data. The data structure and requirements are not defined until the data is needed.

―If you think of a DataMart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.

James Dixon CTO,

Founder & Chief Geek Pentaho

1. **What is Apache Tez? Is Tez replacement to MapReduce or it is running MapReduce under the hood?**

Apache Tez is a distribured task executor. Framework layer above Tez I.e. hive, pig etc submits a Directed Acyclic Graph of tasks to be executed in this case a map task or reduce task.

So Tez is not replacement of Map reduce it is executing MapR under the hood.