**HEVO DATA**

**Question 1:** What is data warehouse schema? Explain different types of schema.

Schema is a kind of skeleton for every database. It is used to describe the logical structure for database. Just like databases, data warehousing also has schemas. The schema provides the name and the data type any record will have. It helps simplify the operations and maintenance of any data warehouse.

In data warehousing, we use the following types of schemas:

**1. Star Schema**

The star schema can be easily understood by referring to star topology. Just like star topology has a central node to which all secondary nodes are attached, star schema of data warehouse too has a similar structure. We have a fact table at centre which has keys of all the dimension tables associated with it. Also, the fact table can have attributes of its own.

The dimension tables can be analogized with the secondary nodes of star topology. There can be many dimension tables associated to one fact table. Each dimension table has its own attributes. Each dimension has to be represented using one dimensional table. The dimension tables are not normalized in case of star schema.

**2. Snowflake schema**

This can be viewed as an extension of the star schema. In this schema too, the central fact table is associated to many dimension tables and the fact table stores key to all associated dimension tables.

The change that happens here is that, the dimension tables are normalized. The normalization of dimension tables helps in reduction of redundancy and also helps make the structure storage efficient as compared to star schema. However, normalization of dimension tables may lead to more usage of joins while querying the data warehouse.

In snowflake schema, we have multiple level relationship and a child table might have multiple parents. Usage of low disk space is one of the biggest advantages of snowflake schema.

**3. Fact Constellation schema**

This schema is also known as the Galaxy schema. In this schema, more than one fact table is present. Also, sharing of one dimension table between two or more fact tables is possible.

The Fact constellation schema can be considered as a group of multiple stars, and hence the name. all dimensions that are shared between two or more fact tables are called conformed dimensions. The fact constellation schemas are more complex than other two schemas. They might be more flexible to use but are difficult to implement and preserve.

**Question 2:** What is the difference between OLTP and OLAP? Explain their application with the help of one example each.

The terms OLAP and OLTP stand for On-Line Analytical Processing system and On-Line Transaction Processing system respectively.

**OLAP,** as the name suggests, is used for performing multi-dimensional analysis of large volumes of data. It stores previous data given to it by OLTP and makes it possible for users to view summaries of different kinds of that data. The tables used in OLAP necessarily need not be normalized. OLAP uses more complex and longer queries as compared to OLTP.

In most cases, the OLAP uses OLAP cube, which forms the central part of OLAP. The cube is used as an extension of traditional relational databases by adding in multiple layers of dimensions for data. The data can then be queried, retrieved and analysed multi dimensionally. The OLAP techniques are primarily useful when it comes to applications like data mining, business intelligence, complex calculations, etc. Tools like Microsoft Excel, Oracle OLAP, IBM Cognos use the OLAP technique.

**OLTP** is used when we wish to record the real time updations, transactions or deletions being made by multiple users at a current time. The queries which are used in OLTP are shorter and simpler when compared to OLAP. One of the features of OLTP is that, it makes available the same data to a large number of users at same time without risking the integrity of that data. OLTP is used to support rapid processing of queries such that the response time is only a few milliseconds. Also, the tables which are used in OLTP are highly normalized which sort of makes it easy and fast. If any transaction of OLTP fails, it can affect data integrity and hence special attention is paid to the data integrity of OLTP. Proper backups are made and constant efforts are made to keep systems available 24/7. OLTPs are mainly used when we concern activities like ATM transactions, hotel/railway reservations and online purchases.

**Question 3:** Explain Fact Table and Dimension Table.

Fact tables and dimension tables are part of data warehouse. Fact table, as the name suggests, include any fact about the data present in system and the dimension table explains the properties or attributes of the data in system.

A fact table largely has quantitative data bur may also have textual data. It can usually have just two columns defining the foreign key for its associated dimension tables and the values which are to be analysed. Each fact table contains keys which are used by dimension tables as foreign keys to associate with the fact table. They are designed in deep detail to increase the efficiency. There is no hierarchy present in fact tables. The fact tables can be of 3 types: Additive, Semi-Additive and Non-Additive. A fact table grows vertically and is created after creation of its dimension tables. Fact tables are the centre for star and snowflake schemas.

A dimension table can be looked upon as a companion to fact table. It contains the descriptive attributes or the dimensions whose values are to be stored in fact tables. Each dimension table is associated to a fact table with help of foreign keys. These tables grow horizontally and are more descriptive of data and can have textual data only. A dimension table may have hierarchy and usually has more attributes than fact tables. It has lesser number of records compared to its attributes and is created before the fact table. Multiple dimension tables can be associated to one fact table and also one dimension table maybe associated to multiple fact table depending upon the schema chosen. Dimensions can be of various types like Conformed, Step, Junk, Swappable, Degenerate, etc.

**Question 4:** Compare AWS Redshift and Google BigQuery cloud data warehouses. Which of these is best suited for what purpose?

**AWS Redshift**: we need to define the size while setting it up because it doesn’t demarcate the computation and storage. Also the data flow must match the size of resource selected and this may require optimization of data as well. Redshift is quite difficult to manage if you are not skilled with AWS. Also, it not always possible for the users to independently re-scale their resources as rescaling means hours of reconfiguration and during this time the data available in cluster can’t be modified. But the AWS Redshift provides great security and in addition to database security, we have sign- in security, SSL connection etc.

**Google BigQuery**: Here no sizing is needed because demarcation of storage and computation is present. It needs lower maintenance compared to AWS Redshift and all the backend configuration and syncing is done by Google. There is rapid provision of resources when large amounts of data are needed to be computed and the tool provides full elasticity. Google BigQuery too provides good level of security. Data is encrypted and in transit mode by default.

**Question 5:** What are the benefits of using a data warehouse over a database?

Benefits of data warehousing over database include:

1. We can get historical insights on data which in turn returns enhanced business intelligence.

2. We can easily improve the quality, consistency and conformity of our data.

3. It works greatly when we wish to save time and increase efficiency because data is available all time and decisions can be made quickly on basis of historical data or analysis.

4. As it increases efficiency, it helps companies to generate a higher return on their investments and gain profit in business.

5. It also provides competitive edge over others.

6. It helps get better insights into data and increases security too.

7. By using data warehousing, we can easily scale which again is a big benefit for any organization.

**Question 6:** Explain the components of a Data Analytics Stack. What is its significance?

There are 3 main components of Data Analytics Stack namely:

Data Wrangling: This basically means that the data which is collected from various sources doesn’t always need to be clean, well- organized and recorded. By wrangling our motive is to clean the data and present it in an understandable, more accurate and concise format.

Data Storage: Once the data is cleaned, we need to store that data. Storage can be selected based upon the size, usage and need of organization, customers and users.

Data Analysis: After storage comes the last component ie analysis. Raw facts and figures can’t be of use to individuals or organizations unless conclusions are drawn form them. This is what is done in analysation step ie we draw conclusions and gain insights from data collected.

This is very significant from industry point of view as it would help identify competitor organizations, understand the loopholes, get better insight into customer needs and the market trends.

**Question 7:** What is the difference between SQL and NoSQL databases? Explain with examples.

SQL databases are relational in nature while noSQL database are not relational.

SQL databases have predefined schemas while noSQL database use dynamic database.

SQL is vertically scalable while NoSQL is horizontally scalable.