**Q1: What is data analytics in terms of a data warehouse?**

Data Analytics is the science to check raw data to draw business-driven conclusions of the data. The data warehouse enables data analysis.

**Q2: Define a subject-oriented data warehouse?**

Subject-oriented data warehouses store data around a specific point like sales, client, and product.

**Q3: What does OLAP mean, and what are its types?**

OLAP is a system that processes, manages, and collects multi-dimensional data for management. It stands for Online Analytical Processing.

There are four types of OLAP Servers given below:

* Hybrid OLAP
* Relational OLAP
* Specialized SQL Servers
* Multi-dimensional OLAP

**Q4: What is the difference between OLAP and OLTP?**

OLAP is a software tool used for data analysis that helps in business decisions while OLTP is a transaction-oriented application used in a three-tier architecture. Below are some of the differences between OLAP and OLTP:

|  |  |
| --- | --- |
| **OLAP (Online Analytical Processing)** | **OLTP (Online Transaction Processing)** |
| It contains the historical data collected from different databases. | It contains operational data. |
| It is used in data analytics, [data mining](https://www.upgrad.com/blog/data-mining-techniques/), and decision making. | It is application-oriented and is used for various business-related tasks. |
| It stores a huge amount of data and is in TB. | It stores a small amount of data and is stored in MB, GB, etc. |
| It works slowly because the size of the data is large. | It works very fast and queries take place on 5% of the stored data. |
| It needs the backup of data from time to time only. | Backup and recovery of data occur regularly. |
| It is mainly used for a read operation with write operation occurring rarely. | It is used for both read and write operations. |

**Q5: What functions does OLAP perform?**

A few of the primary functions performed by OLAP are Pivot, Drill-down, Roll-up, Slice, and Dice.

**Q6: What is the ER Diagram?**

ER Diagram stands for Entity-Relationship Diagram that shows the interrelationships between the entities in the database.

**Q7: What is SCD?**

SCD stands for slowly changing dimensions, and it applies to such cases where records change over time.

**Q8: Define the types of SCD.**

There are 3 types of SCD as given below:

SCD 1: The new record replaces the original record.

SCD 2: The new record gets added to the existing client table

SCD 3: The original data gets changes to enter new data.

**Q9: What is a Snowflake Schema?**

Snowflake Schema is a schema having a primary dimension table. One or more dimensions can be joined in the primary dimension table. It is the only table that can join with the fact table.

**Q 10: Define Star Schema.**

Star Schema refers to managing the table in a way that results can readily get recovered in the data warehouse environment.

**Q11: Define BUS Schema.**

The BUS Schema includes the suite of standardized definition and confirmed dimension if a fact table is there.

**Q 12: Define Metadata.**

It refers to data about the data. The Metadata consists of details like ordering of fields, several columns used, data types of the fields, limited width, and fixed width.

**Q13: Define the core dimension.**

Core Dimension is a Dimension Table that is mainly used for data mart or a single fact table.

**Q14: Define the loops in the data warehouse.**

These loops exist between the tables in the [data warehouse](https://www.upgrad.com/blog/an-overview-of-the-key-concepts-of-data-warehousing/). If any loops are between the tables, then the query generation takes more time and creates an enigma. So it is always recommended to avoid any loops between the tables.

**Q15: Explain XMLA.**

XMLA is called XML for Analysis, which offers the standard method to access data from OLAP, data mining, and other data sources available over the internet. It is a simple object access protocol which uses the discover and execute methods. The discovery method retrieves the data from the internet, and the execution method is used to execute applications against different data sources.

**Read:** [Data Science Interview Questions](https://www.upgrad.com/blog/data-science-interview-questions-answers-most-frequently-asked/)

**Q16: Explain the differences between database and data warehouse.**

A database is different from the data warehouse as the database uses the relational model for data storage. In contrast, the data warehouse uses other schemas and start schema is one of them. Below are some of the differences between a database and a data warehouse:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Database** | **Data Warehouse** |
| Data type | Relational data or Object-oriented data | Large volume data |
| Operations | Transaction processing | Data modeling and data analysis |
| Dimensions | Two-dimensional data | Multi-dimensional data |
| Data design | ER based | Star and snowflake schema |
| Size of data | Small | Large |
| Functionality | High performance and availability | High flexibility |

**Q17: Define the Cube in Data warehouse.**

Cubes in a Data warehouse are the representation of multi-dimensional data. The body of the cube consists of data values, and the edge of the cube contains dimension members.

**Q18. Explain the types of a data warehouse?**

The Data warehouse is of the following 3 types:

1. **Enterprise Data Warehouse:** In Enterprise data warehouse, the organizational data from various functional areas get merged into a centralized way. This helps in the extraction and transformation of data, which provides a detailed overview of any object in the data model.
2. **Operational Data Store:** This data warehouse helps to access data directly from the database and also supports transaction processing. It integrates contrast data from different sources, which supports various business operations later.
3. **Data Mart:** This data warehouse stores the data for a specific functional area. Also, it contains the data in the form of subsets, which then gets stored in the data warehouse. It reduces the large volume of data for users to analyze it efficiently and gain insights.

**Q19: Between multidimensional OLAP and relational OLAP, which works faster?**

Multi-dimensional OLAP works faster than Relational OLAP.

1. **Multi-Dimensional OLAP:** In MOLAP, the data gets stored in the multi-dimensional cube. The storage of the data occurs in proprietary formats such as the PowerOLAP.olp file. These products are compatible with excel and make the data interactions easy.
2. **Relational OLAP:** In the Relational OLAP products, the relational database can be accessed with SQL, which is a standard language used to manipulate data in RDBMS. While performing the processing, it accepts the client requests, which are then translated into SQL queries and then get passed into the RDBMS.

**Q20: Explain the differences between divisive hierarchical clustering and agglomerative clustering.**

In the agglomerative hierarchical clustering method, the clusters get to read from the bottom to top, which means that the program first reads the sub-component and then the parent. On the other hand, divisive hierarchical clustering uses the top to bottom approach in which the data at parent level is read first and then at the child level.

In the Agglomerative hierarchical method, objects are present, and each object builds its cluster, and all these clusters together make a large cluster. This method mainly consists of continuous merging that occurs until a single large cluster gets created, while in the divisive clustering method, the division of clusters occurs. The parent cluster gets divided into smaller clusters. This division of clusters continues until each cluster consists of a single object.

**Learn more:**[Data Science Vs Data Mining: Difference Between Data Science & Data Mining](https://www.upgrad.com/blog/data-science-vs-data-mining-difference-between-data-science-data-mining/)

**Q21: What is the chameleon method in a data warehouse?**

Chameleon is the hierarchical clustering method in the data warehouse. This method works on the sparse graph consisting of nodes and edges. These nodes represent the data items, and edges represent the weights. With this representation, the datasets can be created and accessed with ease overcoming the shortcomings of existing methods. The method works in two phases:

* In the first phase, the graph gets partitioned as part of which the data items are divided into many sub-clusters.
* In the second phase, the genuine clusters are searched that can then get combined with other sub-clusters created in the first phase.

**Q22: What is the execution plan, and what approach does the optimizer use during the execution plan?**

The execution plan is the plan used by the optimizer to choose the combination of steps for the execution of SQL queries. The optimizer selects the most efficient combination of steps for executing the SQL queries. The optimizer uses the two approaches in the execution plan, i.e., rule-based and cost-based.

**Q23: What are the different tools used in ETL (Extraction, Transform, and Load)?**

Below is the list of ETL tools:

* Informatica
* Oracle
* Data Stage
* Data Junction
* Ab Initio
* Warehouse builder

**Q24: How are metadata and data dictionaries different?**

Metadata describes the data. It contains all the information about data such as, the source of data, who collected the data, and the data format. It is crucial to understand the information about the data stored in the data warehouses. On the other side, a data dictionary is the basic definition of the database. Data dictionary consists of the files which are present in the database, count of records present in each file, and all the information about the fields in the database.

**Q25: Define Virtual Data warehouse.**

A virtual data warehouse offers a collective view of the complete data. It is like the logical data model of the Metadata, and it has no historical data. A virtual data warehouse is the best way to translate raw data and present it in such a form that it gets used by decision-makers. Data is represented as a semantic map that allows the end-users to view the data in a virtualized form.

**Also Read:** [Data Analyst Interview Questions & Answers](https://www.upgrad.com/blog/data-analyst-interview-questions-and-answer/)

**Q26: What approaches are used to design the data warehouse?**

There are mainly two approaches used for the data warehouse design:

1. **Inmon approach:**It is the top-down approach in which first the data warehouse gets created, and then the data marts are built. In this approach, the data warehouse acts as the center of the Corporate Information Factory, and the data warehouse acts as a logical framework.
2. **Kimball approach:** It is the bottom-up approach in which data mart gets created first. The data mart then integrates to form the complete data warehouse. The integration of different data marts is called the data warehouse bus architecture.

**Q27: What is a real-time data warehouse, and what are its benefits?**

A real-time data warehouse is the data warehouse concept that captures real-time data as soon as it occurs and makes it available in the data warehouse.

**Benefits of a real-time data warehouse:**

* It helps in easy decision-making.
* It removes the batch window.
* It resolves the issue related to the ideal data load.
* It offers an optimized way to run the transformations in the database.
* It offers quick recovery of data.

**Q28: Explain the 3 layer architecture of the ETL cycle.**

The ETL cycle consists of below 3 layers:

1. **Staging layer:** This layer stores the data extracted from multiple data structures.
2. **Data integration layer:** The data from the staging layer transfers into the database with the help of the integration layer. This data then gets organized into the hierarchical groups, also called dimensions, aggregates, and facts. The dimensions and facts together form the schema.
3. **Access layer:** End-users access the data through the access layer and perform the data analysis.

**Q29: What is data purging?**

Data purging is the method of removal of data permanently from the data storage. It is different from data deletion as data deletion only removes the data temporarily while data purging removes the data permanently, and the free space is used for other purposes. Data purging employs different methods. The data purged can be archived if needed.

**Q30: Define the testing phases in a project.**

The ETL test consists five stages as mentioned below:

1. Requirements and data sources identification
2. Data acquisition
3. Implementation of business logic
4. Data building and publishing
5. Reporting

Datawarehousing

Business Intelligence

Dimension Table

Fact Table

Data Mining

OLTP

OLAP

difference between OLTP and OLAP

ODS

difference between View and Materialized View

ETL

VLDB

Aggregate tables

factless fact tables

Datamart

difference between Datawarehouse and OLAP

key columns in Fact and dimension tables

SCD and types of SCD

Dimensional Modelling

surrogate key

ER Modeling vs Dimensional Modeling

different types of datawarehosuing

goal of Optimizer - to find the most efficient way to execute the SQL statements.

Execution Plan

metadata vs data dictionary