These are the top Data Warehousing interview questions and answers that can help you crack your Data Warehousing job interview. You will learn about the difference between a Data Warehouse and a database, cluster analysis, chameleon method, Virtual Data Warehouse, snapshots, ODS for operational reporting, XMLA for accessing data, and types of slowly changing dimensions.

**Top Answers to Data Warehouse Interview Questions**

A Data Warehouse allows you to collect and manage the data that later helps in providing significant business insights. Since it is an important Business Intelligence (BI) field, ‘Data Warehouse Analyst’ is among the most sought-after career options today. This Data Warehouse Interview Questions blog has a compiled list of some of the most important questions that companies generally ask during Data Warehouse job interviews.

The Data Warehouse Interview Questions blog is majorly classified into the parts listed below:

1. Data Warehouse Interview Questions for Freshers
2. Intermediate Data Warehouse Interview Questions
3. Advanced Data Warehouse Interview Questions for Experienced

**Data Warehouse Interview Questions and Answers for Freshers**

**1. Compare a database with Data Warehouse.**

| Criteria | Database | Data Warehouse |
|---|---|---|
| Type of data | Relational or object-oriented data | Large volume with multiple data types |
| Data operations | Transaction processing | Data modeling and analysis |
| Dimensions of data | Two-dimensional data | Multidimensional data |
| Data design | ER-based and application-oriented database design | Star/Snowflake schema and subject-oriented database design |
| Size of the data | Small (in GB) | Large (in TB) |
| Functionality | High availability and performance | High flexibility and user autonomy |

A database uses a relational model to store data, whereas a Data Warehouse uses various schemas such as star schema and others. In star schema, each dimension is represented by only the one-dimensional table. Data Warehouse supports dimensional modeling, which is a design technique to support end-user queries.

**2. What is the purpose of cluster analysis in Data Warehousing?**

Cluster analysis is used to define the object without giving the class label. It analyzes all the data that is present in the Data Warehouse and compares the cluster with the cluster that is already running. It performs the task of assigning some set of objects into groups, also known as clusters. It is used to perform the data mining job using a technique like statistical data analysis. It includes all the information and knowledge around many fields such as Machine Learning, pattern recognition, image analysis, and bio-informatics. Cluster analysis performs the iterative process of knowledge discovery and includes trials and failures. It is used with the pre-processing and other parameters to achieve the properties that are desired to be used.

**Purpose of cluster analysis:**

- Scalability
- Ability to deal with different kinds of attributes
- Discovery of clusters with attribute shape
- High dimensionality
- Ability to deal with noise
- Interpretability

### 3. What is the difference between agglomerative and divisive hierarchical clustering?

- Agglomerative hierarchical clustering method allows clusters to be read from bottom to top so that the program always reads from the sub-component first then moves to the parent; whereas, divisive hierarchical clustering uses top to bottom approach in which the parent is visited first then the child.
- Agglomerative hierarchical method consists of objects in which each object creates its own clusters, and these clusters are grouped together to create a large cluster. It is a process of continuous merging until all the single clusters are merged together into a complete big cluster that will consist of all the objects of child clusters. However, in divisive clustering, the parent cluster is divided into smaller clusters, and it keeps on dividing until each cluster has a single object to represent.

💡 想关注数据工程师的工作机会，请加微信：niuxiaojiang01

### 4. Explain the chameleon method used in Data Warehousing.

Chameleon is a hierarchical clustering algorithm that overcomes the limitations of the existing models and methods present in Data Warehousing. This method operates on the sparse graph having nodes that represent data items and edges which represent the weights of the data items.

This representation allows large datasets to be created and operated successfully. The method finds the clusters that are used in the dataset using the two-phase algorithm.

- The first phase consists of the graph partitioning that allows the clustering of the data items into a large number of sub-clusters.
- The second phase uses an agglomerative hierarchical clustering algorithm to search for the clusters that are genuine and can be combined together with the sub-clusters that are produced.

### 5. What is Virtual Data Warehousing?

- A Virtual Data Warehouse provides a collective view of the completed data. A Virtual Data Warehouse has no historic data. It can be considered as a logical data model of the given metadata.
- Virtual Data Warehousing is a ‘de facto’ information system strategy for supporting analytical decision-making. It is one of the best ways for translating raw data and presenting it in the form that can be used by decision-makers. It provides a semantic map—which allows the end user for viewing as virtualized.

## 6. What is Active Data Warehousing?

- An Active Data Warehouse represents a single state of a business. Active Data Warehousing considers the analytic perspectives of customers and suppliers. It helps deliver the updated data through reports.
- A form of repository of captured transactional data is known as 'Active Data Warehousing.' Using this concept, trends and patterns are found to be used for future decision-making. Active Data Warehouse has a feature which can integrate the changes of data while scheduled cycles refresh. Enterprises utilize an Active Data Warehouse in drawing the company's image in a statistical manner.

## 7. What is a snapshot with reference to Data Warehouse?

- A snapshot refers to a complete visualization of data at the time of extraction. It occupies less space and can be used to back up and restore data quickly.
- A snapshot is a process of knowing about the activities performed. It is stored in a report format from a specific catalog. The report is generated soon after the catalog is disconnected.

## 8. What is XMLA?

- XMLA is XML for Analysis which can be considered as a standard for accessing data in OLAP, data mining, or data sources on the Internet. It is Simple Object Access Protocol. XMLA uses 'Discover' and 'Execute' methods. Discover fetches information from the Internet, while 'Execute' allows the applications to execute against the data sources.
- XMLA is an industry standard for accessing data in analytical systems, such as OLAP. It is based on XML, SOAP, and HTTP.
- XMLA specifies MDXML as a query language. In the XMLA 1.1 version, the only construct in the MDXML is an MDX statement enclosed in the tag.

#### Intermediate Data Warehouse Interview Questions and Answers

## 9. What is ODS?

- An operational data store (ODS) is a database designed to integrate data from multiple sources for additional operations on the data. Unlike a master data store, the data is not sent back to operational systems. It may be passed for further operations and to the Data Warehouse for reporting.
- In ODS, data can be scrubbed, resolved for redundancy, and checked for compliance with the corresponding business rules. This data store can be used for integrating disparate data from multiple sources so that business operations, analysis, and reporting can be carried out. This is the place where most of the data used in the current operation is housed before it's transferred to the Data Warehouse for longer-term storage or archiving.
- 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 the 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.

## 10. What is the level of granularity of a fact table?

A fact table is usually designed at a low level of granularity. This means that we need to find the lowest level of information that can be stored in a fact table e.g., employee performance is a very high level of granularity. Employee_performance_daily and employee_perfomance_weekly can be considered as lower levels of granularity.

The granularity is the lowest level of information stored in the fact table. The depth of the data level is known as granularity. In date dimension, the level could be year, month, quarter, period, week, and day of granularity.

The process consists of the following two steps:

- Determining the dimensions that are to be included
- Determining the location to find the hierarchy of each dimension of the information

The above factors of determination will be re-sent as per the requirements.

## 11. What is the difference between ‘view’ and ‘materialized view’?

**View:**

- Tail raid data representation is provided with a view to access data from its table.
- It has logical structure that does not occupy space.
- Changes get affected in the corresponding tables.

**Materialized view:**

- Pre-calculated data persists in the materialized view.
- It has physical data space occupation.
- Changes will not get affected in the corresponding tables.

## 12. What is junk dimension?

- In scenarios where certain data may not be appropriate to store in the schema, the data (or attributes) can be stored in a junk dimension. The nature of the data of junk dimension is usually Boolean or flag values.
- A single dimension is formed by lumping a number of small dimensions. This is called a junk dimension. Junk dimension has unrelated attributes. The process of grouping random flags and text attributes in a dimension by transmitting them to a distinguished sub-dimension is related to junk dimension.

## 13. What are the different types of SCDs used in Data Warehousing?

SCDs (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:**

- **SCD1:** It is a record that is used to replace the original record even when 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 the 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 exists in the database and another record that will replace the old database record with the new information.

### 14. Which one is faster, Multidimensional OLAP or Relational OLAP?

Multidimensional OLAP (MOLAP) is faster than Relational OLAP (ROLAP).

- **MOLAP:** Here, data is stored in a multidimensional cube. The storage is not in the relational database but in proprietary formats (one example is PowerOLAP's .olp file). MOLAP products are compatible with Excel, which can make data interactions easy to learn.
- **ROLAP:** ROLAP products access a relational database by using SQL (structured query language), which is the standard language that is used to define and manipulate data in an RDBMS. Subsequent processing may occur in the RDBMS or within a mid-tier server, which accepts requests from clients, translates them into SQL statements, and passes them on to the RDBMS.

### 15. What is Hybrid SCD?

Hybrid SCDs are a combination of both SCD1 and SCD2.

It may happen that in a table, some columns are important and we need to track changes for them, i.e., capture the historical data for them, whereas in some columns even if the data changes we do not have to bother. For such tables, we implement Hybrid SCDs, wherein some columns are Type 1 and some are Type 2.

### 16. Why do we override the execute method in Struts?

As part of Struts Framework, we can develop the Action Servlets and the ActionForm Servlets and other servlet classes.

In case of ActionForm class, we can develop the validate() method. This method will return the ActionErrors object. In this method, we can write the validation code.

- If this method returns null or ActionErrors with size = 0, the web container will call **execute()** as part of the Action class.
- If it returns size > 0, it will not call the execute() method. It will rather execute the jsp, servlet, or html file as the value for the input attribute as part of the attribute in the struts-config.xml file.

**Advanced Data Warehouse Interview Questions and Answers for Experienced**

### 17. What is VLDB?

A very large database (VLDB) is a database that contains an extremely large number of tuples (database rows) or occupies an extremely large physical file system storage space. A one terabyte database would normally be considered to be a VLDB.

### 18. How do you load the time dimension?

Time dimensions are usually loaded by a program that loops through all possible dates appearing in the data. It is not unusual for 100 years to be represented in a time dimension, with one row per day.

### 19. What are conformed dimensions?

- Conformed dimensions are the dimensions which can be used across multiple data marts in combination with multiple fact tables accordingly.
- A conformed dimension is a dimension that has exactly the same meaning and content when being referred from different fact tables. It can refer to multiple tables in multiple data marts within the same organization.

### 20. What is the main difference between Inmon and Kimball philosophies of Data Warehousing?

Both differ in the concept of building the Data Warehouse.

- **Kimball** views Data Warehousing as a constituency of data marts. Data marts are focused on delivering business objectives for departments in an organization, and the Data Warehouse is a conformed dimension of the data marts. Hence, a unified view of the enterprise can be obtained from the dimension modeling on a local departmental level.
- **Inmon** explains in creating a Data Warehouse on a subject-by-subject area basis. Hence, the development of the Data Warehouse can start with data from the online store. Other subject areas can be added to the Data Warehouse as their needs arise. Point-of-sale (POS) data can be added later if management decides that it is necessary.
- Hence, the process will be as follows:Kimball > First Data Marts > Combined Ways > Data WarehouseInmon > First Data Warehouse > Data marts

### 21. What is the difference between a data warehouse and a data mart?

A **data warehouse** is a set of data isolated from operational systems. This helps an organization deal with its decision-making process. A **data mart** is a subset of a data warehouse that is geared to a particular business line. Data marts provide the stock of condensed data collected in the organization for research on a particular field or entity.

A data warehouse typically has a size greater than 100 GB, while the size of a data mart is generally less than 100 GB. Due to the disparity in scope, the design and utility of data marts are comparatively simpler.

### 22. Explain the ETL cycle's 3-layer architecture.

The staging layer, the data integration layer, and the access layer are the three layers that are involved in an ETL cycle.

- **Staging layer**: It is used to store the data extracted from various data structures of the source.
- **Data integration layer**: Data from the staging layer is transformed and transferred to the database using the integration layer. The data is arranged into hierarchical groups (often referred to as dimensions), facts, and aggregates. In a DW system, the combination of facts and dimensions tables is called a schema.
- **Access layer**: For analytical reporting, end-users use the access layer to retrieve the data.

### 23. What does data purging mean?

Data purging is a process, involving methods that can erase data permanently from the storage. Several techniques and strategies are used for data purging.

The process of data purging often contrasts with data deletion. Deleting data is more of a temporary process, while data purging permanently removes data. This, in turn, frees up storage and/or memory space, which can be utilized for other purposes.

The purging process allows us to archive data even if it is permanently removed from the main source, giving us an option to retrieve the data from the archive if it is needed. The deleting process also permanently removes the data but does not necessarily involve keeping a backup, and it generally involves insignificant amounts of data.

### 24. Can you define the five main testing phases of a project?

The ETL test is performed in five stages as follows:

- The identification of data sources and requirements
- The acquisition of data
- Implementing business logic and dimensional modeling
- Building and publishing data
- Reports building

### 25. What do you mean by the slice action? How many slice-operated dimensions are used?

A slice operation is the filtration process in a data warehouse. It selects a specific dimension from a given cube and provides a new sub-cube. In the slice operation, only a single dimension is used.