Disha Mukherjee
Credits—InterviewBit
Year- 2023

# DATA WAREHOUSE

ONE STOP CHEATSHEET FOR EXPERIENCED AND FRESHERS

DISHA MUKHERJEE

Credits
InterviewBit

Disha Mukherjee
Credits—InterviewBit
Year- 2023

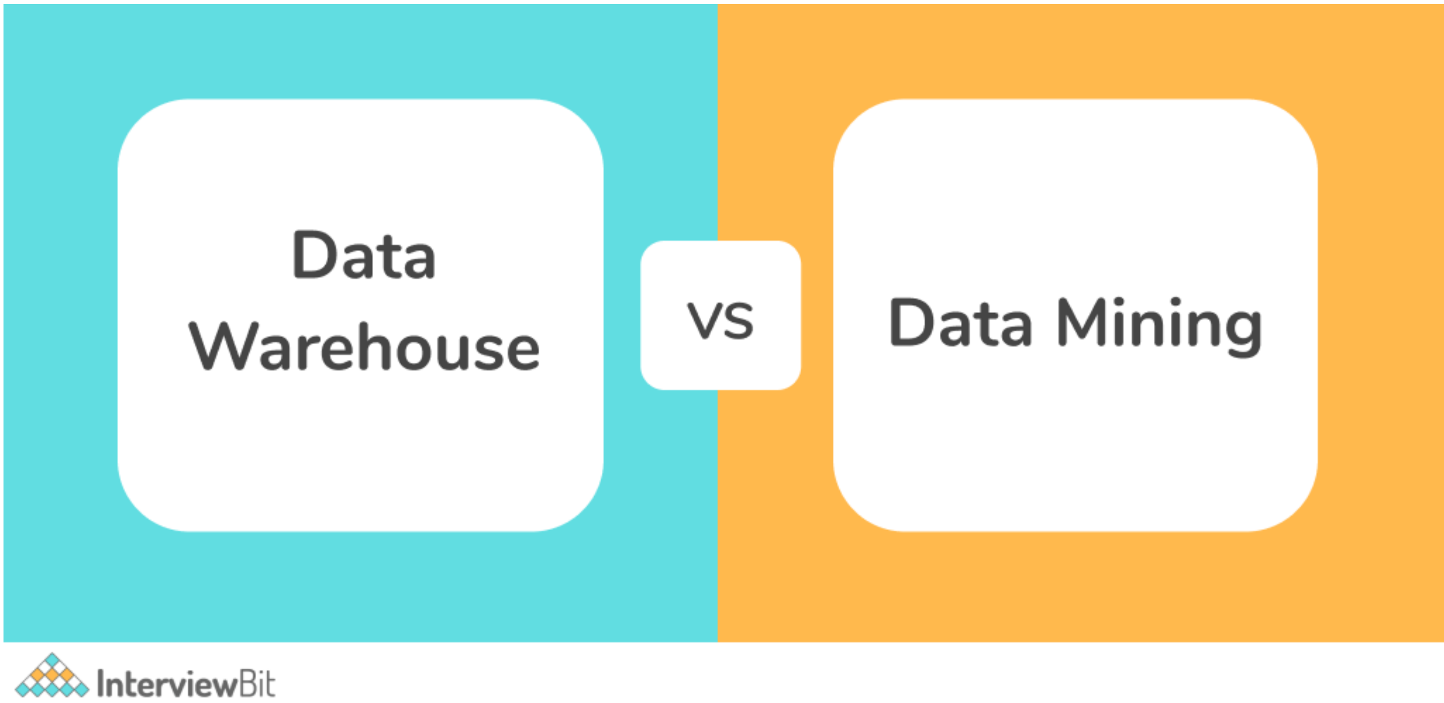
# Data Warehouse Interview Questions for Freshers

## 1. What do you mean by data mining? Differentiate between data mining and data warehousing.

Data mining is the process of collecting information in order to find patterns, trends, and usable data that will help a company to make data-driven decisions from large amounts of data. In other words, Data Mining is the method of analysing hidden patterns of data from various perspectives for categorization into useful data, which is gathered and assembled in specific areas such as data warehouses, efficient analysis, data mining algorithm, assisting decision making, and other data requirements, ultimately resulting in cost-cutting and revenue generation. Data mining is the process of automatically examining enormous amounts of data for patterns and trends that go beyond simple analysis. Data mining estimates the probability of future events by utilising advanced mathematical algorithms for data segments.

Following are the differences between data warehousing and data mining:-

Disha Mukherjee
Credits—InterviewBit
Year- 2023

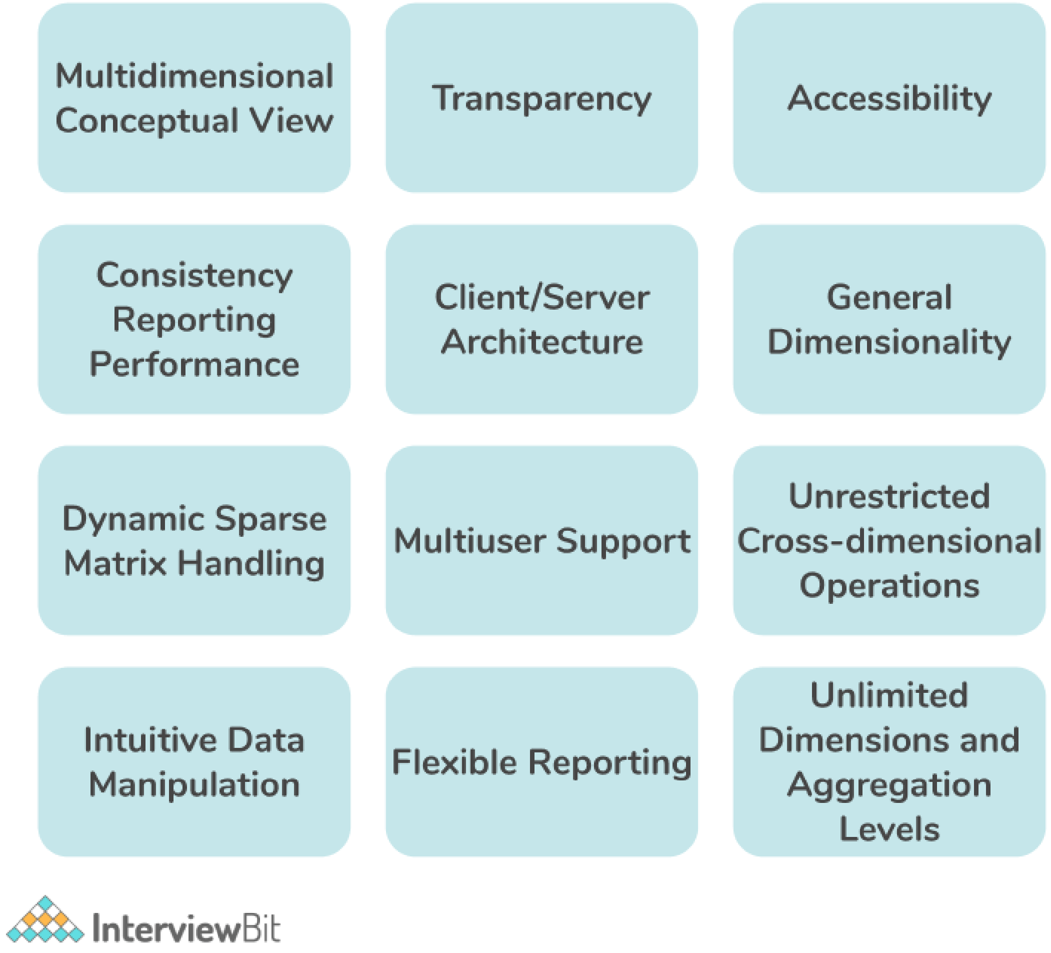


| Data Warehousing | Data Mining |
|---|---|
| A data warehouse is a database system that is intended for analytical rather than transactional purposes. | The technique of examining data patterns is known as data mining. |
| In data warehousing, data is saved on a regular basis. | In data mining, data is evaluated on a regular basis. |
| Engineers are the only ones that do data warehousing. | With the assistance of technologists, business users conduct data mining. |
| Data warehousing is the process of bringing all relevant data together. | Data mining is the process of extracting information from big datasets. |
| Data warehousing can be referred to as a subset of data mining. | Data Mining can be referred to as a super set of data warehousing. |

## 2. What do you mean by OLAP in the context of data warehousing? What guidelines should be followed while selecting an OLAP system?

OLAP is an acronym for **On-Line Analytical Processing**. OLAP is a software technology classification that allows analysts, managers, and executives to get insight into information through quick, reliable, interactive access to data that has been converted from raw data to reflect the true dimensionality of the company as perceived by the clients. OLAP allows for multidimensional examination of corporate data while also allowing for complex estimations, trend analysis, and advanced data modelling. It's rapidly improving the foundation for Intelligent Solutions, which includes Business Performance Management, Strategy, Budgeting, Predicting, Financial Documentation, Analysis, Modeling, Knowledge Discovery, and Data Warehouses Reporting. End-clients can use OLAP to perform ad hoc record analysis in several dimensions, giving them the information and understanding they need to make better choices.

Following guidelines must be followed while selecting an OLAP system:-



Multidimensional Conceptual View

Transparency

Accessibility

Consistency Reporting Performance

Client/Server Architecture

General Dimensionality

Dynamic Sparse Matrix Handling

Multiuser Support

Unrestricted Cross-dimensional Operations

Intuitive Data Manipulation

Flexible Reporting

Unlimited Dimensions and Aggregation Levels

- **Multidimensional Conceptual View:** This is one of an OLAP system's most important capabilities. It is feasible to use methods like slice and dice that require a multidimensional view.

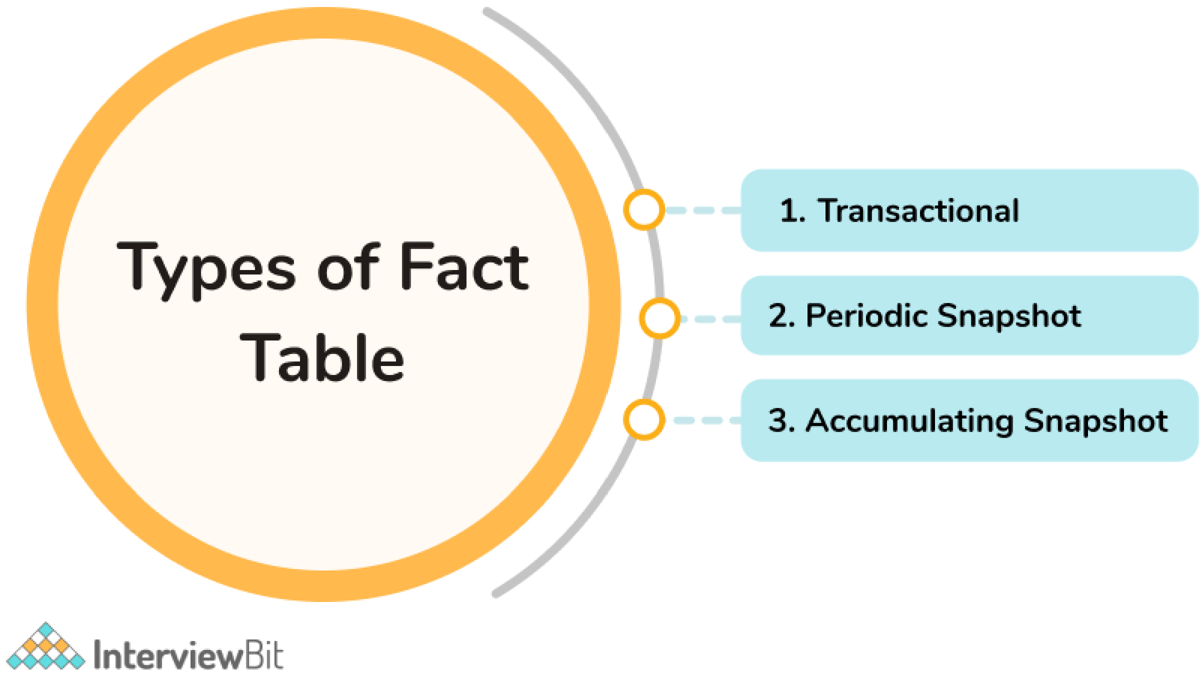
- **Transparency:** Make the technology, the underlying data repository, computing operations, and the disparate nature of source data completely accessible to consumers. Users' efficiency and productivity are improved as a result of this transparency.
- **Accessibility:** OLAP systems must only allow access to the data that is truly needed to do the analysis, giving clients a single, coherent, and consistent picture. The OLAP system must map its own logical schema to the disparate physical data storage, as well as to conduct any required transformations.
- **Consistent Reporting Performance:** As the number of dimensions or the size of the database grows, users should not experience any substantial reduction in documenting performance. That is, as the number of dimensions grows, OLAP performance should not deteriorate.
- **Client/Server Architecture:** Make the OLAP tool's server component clever enough that the various clients can be connected with minimal effort and integration code. The server should be able to map and consolidate data from disparate databases.
- **Generic Dimensionality:** Each dimension in an OLAP method should be seen as equal in terms of structure and operational capabilities. Select dimensions may be granted additional operational capabilities, although such duties should be available to all dimensions.
- **Dynamic Sparse Matrix Handling:** To optimise sparse matrix handling by adapting the physical schema to the unique analytical model being built and loaded. When confronted with a sparse matrix, the system must be able to dynamically assume the information distribution and change storage and access in order to achieve and maintain a constant level of performance.
- **Multiuser Support:** OLAP technologies must allow several users to access data at the same time while maintaining data integrity and security.
- **Unrestricted cross-dimensional Operations:** It gives techniques the ability to determine dimensional order and to perform roll-up and drill-down operations within and across dimensions.
- **Intuitive Data Manipulation:** Reorientation (pivoting), drill-down and roll-up, and other manipulations can be done intuitively and precisely on the cells of the scientific model using point-and-click and drag-and-drop methods. It does away with the need for a menu or several visits to the user interface.
- **Flexible Reporting:** It provides efficiency to corporate clients by allowing them to organize columns, rows, and cells in a way that allows for easy data manipulation, analysis, and synthesis.
- **Infinite Dimensions and Aggregation Levels:** There should be no limit to the number of data dimensions. Within any given consolidation path, each of these

common dimensions must allow for an almost infinite number of customer-defined aggregation levels.

### 3. What do you understand about a fact table in the context of a data warehouse? What are the different types of fact tables?

In a Data Warehouse system, a Fact table is simply a table that holds all of the facts or business information that can be exposed to reporting and analysis when needed. Fields that reflect direct facts, as well as foreign fields that connect the fact table to other dimension tables in the Data Warehouse system, are stored in these tables. Depending on the model type used to construct the Data Warehouse, a Data Warehouse system can have one or more fact tables.

Following are the three types of fact tables:-



- **Transactional Fact Table:** This is a very basic and fundamental view of corporate processes. It can be used to depict the occurrence of an event at any given time. The facts measure are only valid at that specific time and for that specific incident. "One row per line in a transaction," according to the grain associated with the transaction table. It typically comprises data at the detailed level, resulting in a huge number of dimensions linked with it. It captures the smallest or atomic level of dimension measurement. This allows the table to provide users with extensive dimensional grouping, roll-up, and drill-down reporting features. It's packed yet sparse at the same time. It can also be big at the same time, depending on the number of events (transactions) that have occurred.

- **Snapshot Fact Table:** The snapshot depicts the condition of things at a specific point in time, sometimes known as a "picture of the moment." It usually contains a greater number of non-additive and semi-additive information. It aids in the examination of the company's overall performance at regular and predictable times. Unlike the transaction fact table, which adds a new row for each occurrence of an event, this represents the performance of an activity at the end of each day, week, month, or any other time interval. However, to retrieve the detailed data in the transaction fact table, snapshot fact tables or periodic snapshots rely on the transaction fact table. The periodic snapshot tables are typically large and take up a lot of space.
- **Accumulating Fact Table:** These are used to depict the activity of any process with a well-defined beginning and end. Multiple data stamps are commonly found in accumulating snapshots, which reflect the predictable stages or events that occur over the course of a lifespan. There is sometimes an extra column with the date that indicates when the row was last updated.

**You can download a PDF version of Data Warehouse Interview Questions.**

**Download PDF**

---

## 4. What do you mean by dimension table in the context of data warehousing? What are the advantages of using a dimension table?

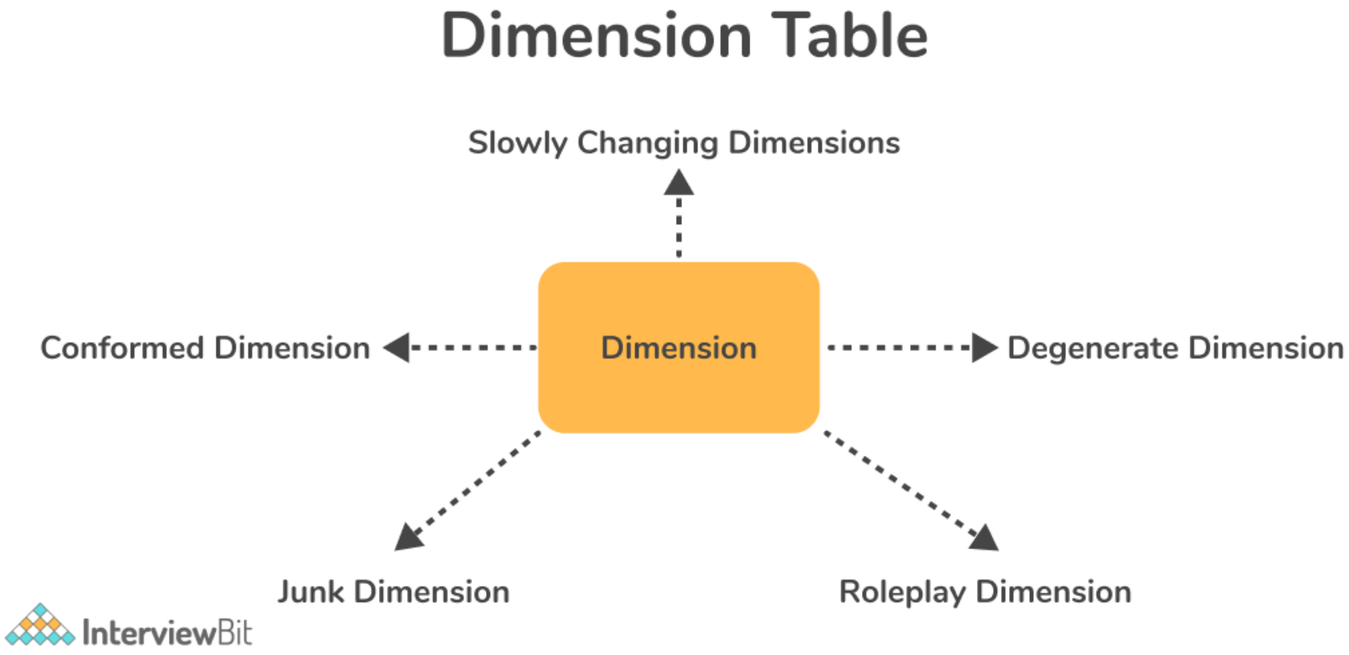
A table in a data warehouse's star schema is referred to as a dimension table. Dimensional data models, which are made up of fact and dimension tables, are used to create data warehouses. Dimension tables contain dimension keys, values, and attributes and are used to describe dimensions. It is usually of a tiny size. The number of rows might range from a few to thousands. It is a description of the objects in the fact table. The term "dimension table" refers to a collection or group of data pertaining to any quantifiable occurrence. They serve as the foundation for dimensional modelling. It includes a column that serves as a primary key, allowing each dimension row or record to be uniquely identified. Through this key, it is linked to the fact tables. When it's constructed, a system-generated key called the surrogate key is used to uniquely identify the rows in the dimension.

Following are the **advantages** of using a dimension table :

- It features a straightforward design.
- It is simple to study and comprehend.
- It stores data that has been de-normalized.
- It aids in the preservation of historical data for any dimension.
- It's simple to get info from it.
- It's simple to build and put into action.
- It provides the context for any business operation.

### 5. What are the different types of dimension tables in the context of data warehousing?

Following are the different types of dimension tables in the context of data warehousing:-



- **Slowly Changing Dimensions (SCD):** Slowly changing dimensions are dimension attributes that tend to vary slowly over time rather than at a regular period of time. For example, the address and phone number may change, but not on a regular basis. Consider the case of a man who travels to several nations and must change his address according to the place he is visiting. This can be accomplished in one of three ways:

  - Type 1: Replaces the value that was previously entered. This strategy is simple to implement and aids in the reduction of costs by saving space. However, in this circumstance, history is lost.
  - Type 2: Insert a new row containing the new value. This method saves the history and allows it to be accessed at any time. However, it takes up a lot of space, which raises the price.
  - Type 3: Add a new column to the table. It is the ideal strategy because history can be easily preserved.

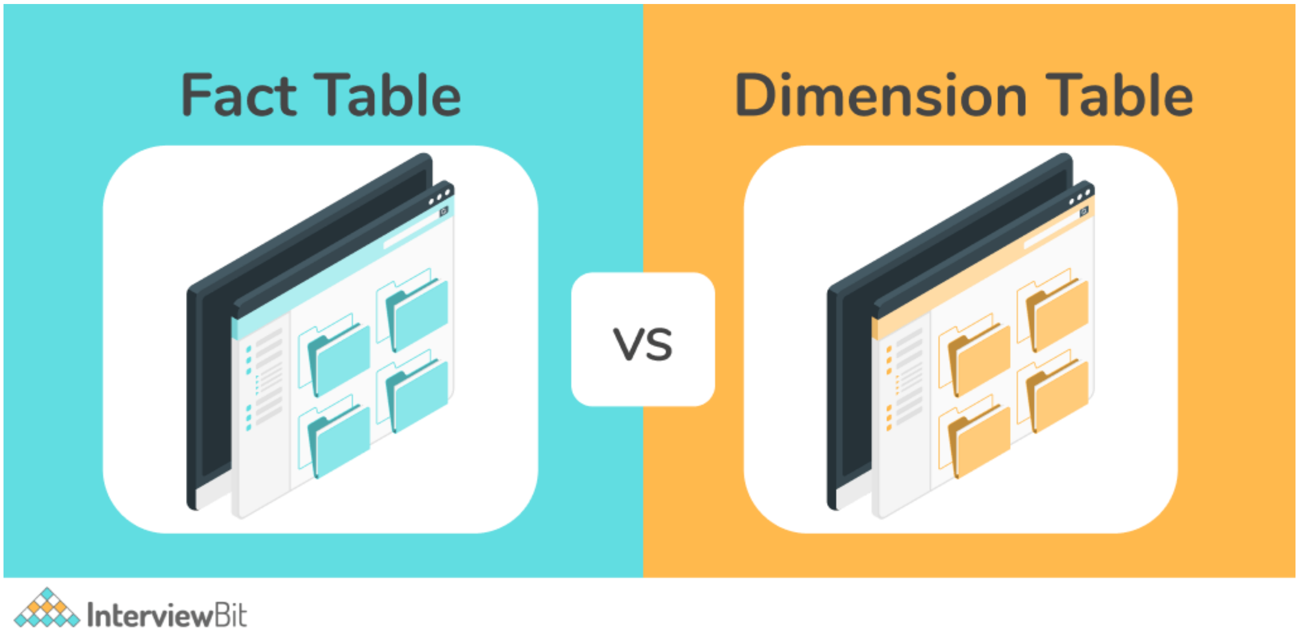
- **Junk Dimension:** A trash dimension is a collection of low-cardinality attributes. It contains a number of varied or disparate features that are unrelated to one another. These can be used to implement RCD (rapidly changing dimension) features like flags and weights, among other things.
- **Conformed Dimension:** Multiple subject areas or data marts share this dimension. It can be utilised in a variety of projects without requiring any changes. This is used to keep things in order. Dimensions that are exactly the same as or a proper subset of any other dimension are known as conformed dimensions.
- **Roleplay Dimension:** Role-play dimension refers to the dimension table that has many relationships with the fact table. In other words, it occurs when the same dimension key and all of its associated attributes are linked to a large number of foreign keys in the fact table. Within the same database, it might serve several roles.
- **Degenerate Dimension:** Degenerate dimension attributes are those that are contained in the fact table itself rather than in a separate dimension table. For instance, a ticket number, an invoice number, a transaction number, and so on.

## 6. Differentiate between fact table and dimension table.

The record of a reality or fact table could be made up of attributes from various dimension tables. The Fact Table, also known as the Reality Table, assists the user in investigating the business aspects that aid him in call taking in order to improve his firm. Dimension Tables, on the other hand, make it easier for the reality table or fact table to collect dimensions from which measurements must be taken.

The following table enlists the difference between a fact table and a dimension table:-

Disha Mukherjee
Credits—InterviewBit
Year- 2023



| Fact Table | Dimension Table |
|---|---|
| It contains the attributes' measurements, facts, or metrics. | It is the companion table that has the attributes that the fact table uses to derive the facts. |
| Data grain (the most atomic level by which facts may be defined) is what defines it. | It is detailed, comprehensive, and lengthy. |
| It is used for analysis and decision-making and contains measures. | It contains information regarding a company's operations and procedures. |
| It contains information in both numeric and textual formats. | It only contains textual information. |
| It has a primary key that works as a foreign key in the dimension table. | It has a foreign key that is linked to the fact table's primary key. |
| It stores the filter domain and reports labels in dimension tables. | It organizes the atomic data into dimensional structures. |
| It does not have a hierarchy. | It has a hierarchy. |
| It has lesser attributes than a dimension table. | It has more attributes than a fact table. |
| It has more records as compared to a dimension table. | It has fewer records than a fact table. |

| Fact Table | Dimension Table |
|---|---|
| Here, the table grows vertically. | Here, the table grows horizontally. |
| It is created after the corresponding dimension table has been created. | It is created prior to the creation of the fact table. |

## 7. What are the advantages of a data warehouse?

Following are the advantages of using a data warehouse:

- **Helps you save time:**
  - To stay ahead of your competitors in today's fast-paced world of cutthroat competition, your company's ability to make smart judgments quickly is critical.
  - A Data warehouse gives you instant access to all of your essential data, so you and your staff don't have to worry about missing a deadline. All you have to do now is deploy your data model to start collecting data in a matter of seconds. You can do this with most warehousing solutions without utilising a sophisticated query or machine learning.
  - With data warehousing, your company won't have to rely on a technical professional to troubleshoot data retrieval issues 24 hours a day, seven days a week. You will save a lot of time this way.
- **Enhances the quality of data:**
  - The high-quality data ensures that your company's policies are founded on accurate information about your operations.
  - You can turn data from numerous sources into a shared structure using data warehousing. You can assure the consistency and integrity of your company's data this way. This allows you to spot and eliminate duplicate data, inaccurately reported data and disinformation.
  - For your firm, implementing a data quality management program may be both costly and time-consuming. You can easily use a data warehouse to reduce the number of these annoyances while saving money and increasing the general productivity of your company.
- **Enhances Business Intelligence (BI):**
  - Throughout your commercial endeavours, you can use a data warehouse to gather, absorb, and derive data from any source. As a result of the capacity to easily consolidate data from several sources, your BI will improve by leaps and bounds.
- **Data standardization and Consistency are achieved:**
  - The uniformity of huge data is another key benefit of having central data repositories. In a similar manner, a data storage or data mart might benefit your company. Because data warehousing stores data from various sources in a consistent manner, such as a

transactional system, each source will produce results that are synchronized with other sources. This ensures that data is of higher quality and homogeneous. As a result, you and your team can rest assured that your data is accurate, resulting in more informed corporate decisions.

- **Enhances Data Security:**
  - A data warehouse improves security by incorporating cutting-edge security features into its design. For any business, consumer data is a vital resource. You can keep all of your data sources integrated and properly protected by adopting a warehousing solution. The risk of a data breach will be greatly reduced as a result of this.
- **Ability to store historical data:**
  - Because a data warehouse can hold enormous amounts of historical data from operational systems, you can readily study different time periods and inclinations that could be game-changing for your business. You can make better corporate judgments about your business plans if you have the correct facts in your hands.

## 8. What are the disadvantages of using a data warehouse?

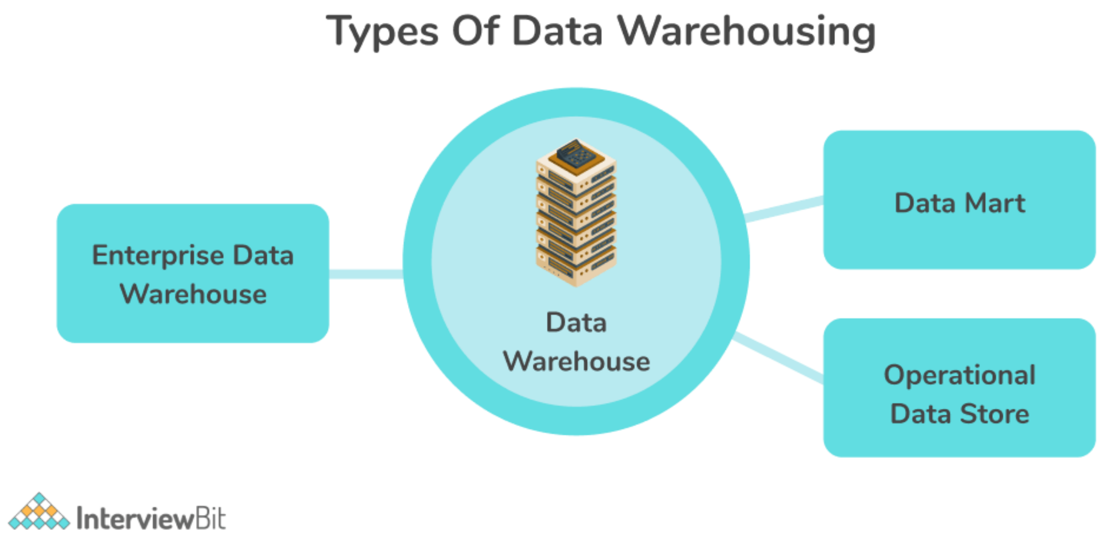
Following are the disadvantages of using a data warehouse:-

- **Loading time of data resources is undervalued:** We frequently underestimate the time it will take to gather, sanitize, and post data to the warehouse. Although some resources are in place to minimize the time and effort spent on the process, it may require a significant amount of the overall production time.
- **Source system flaws that go unnoticed:** After years of non-discovery, hidden flaws linked with the source networks that provide the data warehouse may be discovered. Some fields, for example, may accept nulls when entering new property information, resulting in workers inputting incomplete property data, even if it was available and relevant.
- **Homogenization of data:** Data warehousing also deals with data formats that are comparable across diverse data sources. It's possible that some important data will be lost as a result.

## 9. What are the different types of data warehouse?

Following are the different types of data warehouse:

Disha Mukherjee
Credits—InterviewBit
Year- 2023

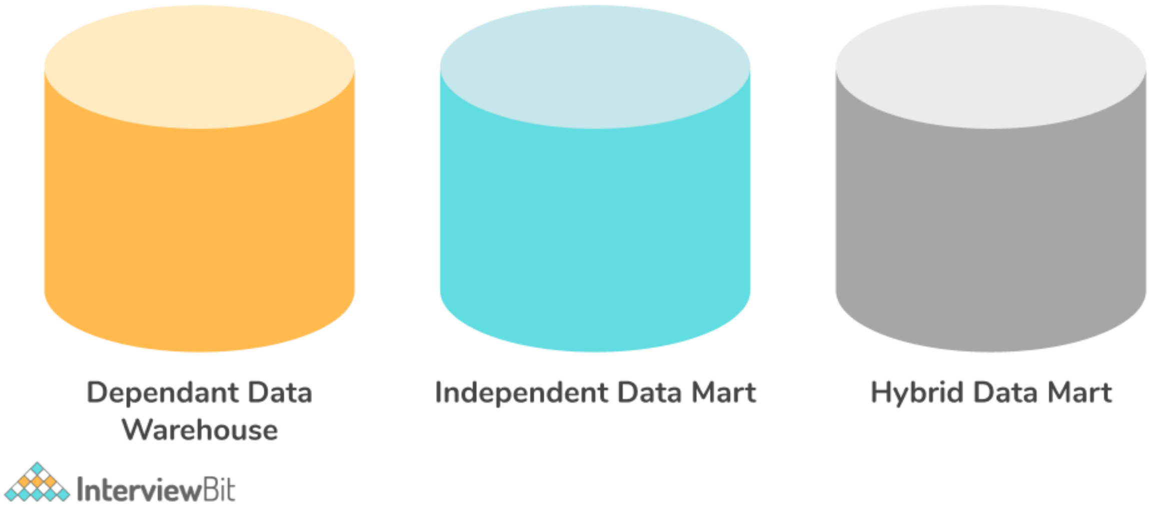


- **Enterprise Data Warehouse:**
  - An enterprise database is a database that brings together the various functional areas of an organisation in a cohesive manner. It's a centralised location where all corporate data from various sources and apps can be accessed. They can be utilised for analytics and by everyone in the organisation once they've been saved. The data can be categorised by subject, and access is granted according to the necessary division. The tasks of extracting, converting, and conforming are taken care of in an Enterprise Datawarehouse.
  - Enterprise Datawarehouse's purpose is to provide a comprehensive overview of any object in the data model. This is performed by finding and wrangling the data from different systems. This is then loaded into a model that is consistent and conformed. The data is acquired by Enterprise Datawarehouse, which can provide access to a single site where various tools can be used to execute analytical functions and generate various predictions. New trends or patterns can be identified by research teams, which can then be focused on to help the company expand.
- **Operational Data Store (ODS):**
  - An operational data store is utilised instead of having an operational decision support system application. It facilitates data access directly from the database, as well as transaction processing. By checking the associated business rules, the data in the Operational Data Store may be cleansed, and any redundancy found can be checked and rectified. It also aids in the integration of disparate data from many sources so that business activities, analysis, and reporting may be carried out quickly and effectively while the process is still ongoing.

  - The majority of current operations are stored here before being migrated to the data warehouse for a longer period of time. It is particularly useful for simple searches and little amounts of data. It functions as short-term or temporary memory, storing recent data. The data warehouse keeps data for a long time and also keeps information that is generally permanent.
- **Data Mart:**
  - Data Mart is referred to as a pattern to get client data in a data warehouse environment. It's a data warehouse-specific structure that's employed by the team's business domain. Every company has its own data mart, which is kept in the data warehouse repository. Dependent, independent, and hybrid data marts are the three types of data marts. Independent data marts collect data from external sources and data warehouses, whereas dependent data marts take data that has already been developed. Data marts can be thought of as logical subsets of a data warehouse.

### 10. What are the different types of data marts in the context of data warehousing?

Following are the different types of data mart in data warehousing:



- **Dependent Data Mart:** A dependent data mart can be developed using data from operational, external, or both sources. It enables the data of the source company to be accessed from a single data warehouse. All data is centralized, which can aid in the development of further data marts.
- **Independent Data Mart:** There is no need for a central data warehouse with this data mart. This is typically established for smaller groups that exist within a company. It has

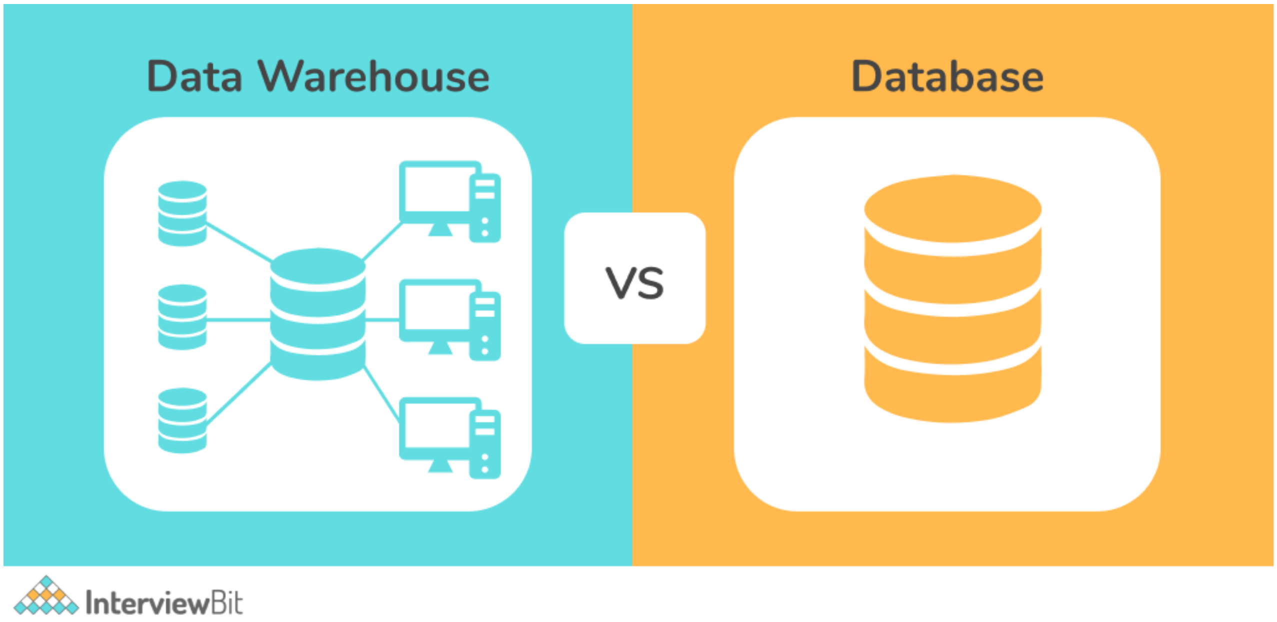
no connection to Enterprise Data Warehouse or any other data warehouse. Each piece of information is self-contained and can be used independently. The analysis can also be carried out independently. It's critical to maintain a consistent and centralized data repository that numerous users can access.

- **Hybrid Data Mart:** A hybrid data mart is utilized when a data warehouse contains inputs from multiple sources, as the name implies. When a user requires an ad hoc integration, this feature comes in handy. This solution can be utilized if an organization requires various database environments and quick implementation. It necessitates the least amount of data purification, and the data mart may accommodate huge storage structures. When smaller data-centric applications are employed, a data mart is most effective.

## 11. Differentiate between data warehouse and database.

**Database:** A database is a logically organized collection of structured data kept electronically in a computer system. A database management system is usually in charge of a database (DBMS). The data, the DBMS, and the applications that go with them are referred to as a database system, which is commonly abbreviated to just a database.

The following table enlists the difference between data warehouse and database:-



| Data Warehouse | Database |
|---|---|
| Data Warehouse uses the OnLine Analytical Processing (OLAP). | Database uses the OnLine Transactional Processing (OLTP). |
| Data Warehouse is mainly used for analyzing the historical data so as to make future decisions based on them. | The database aids in the execution of basic business procedures. |
| Because a data warehouse is denormalized, tables and joins are straightforward. | A database's tables and joins are complicated because they are normalised. |
| It can be referred to as a subject-oriented collection of data. | It can be referred to as an application-oriented collection of data. |
| In this, data modelling techniques are used for designing. | In this, Entity-Relationship (ER) modelling techniques are used for designing. |
| Data may not be up to date in this. | Data is generally up to date in this. |
| The data structure of Data Warehouse is based on a dimensional and normalised approach. For example, a star and snowflake schema is employed. | For data storing, the Flat Relational Approach approach is employed. |
| Generally, highly summarized data is stored in a data warehouse. | Generally, detailed data is stored in a database. |

## 12. What do you mean by a factless fact table in the context of data warehousing?

A fact table with no measures is known as a factless fact table. It's essentially a crossroads of dimensions (it contains nothing but dimensional keys). One form of factless table is used to capture an event, while the other is used to describe conditions.

In the first type of factless fact table, there is no measured value for an event, but it develops the relationship among the dimension members from several dimensions. The existence of the relationship is itself the fact. This type of fact table can be utilised to create valuable reports on its own. Various criteria can be used to count the number of occurrences.

The second type of factless fact table is a tool that's used to back up negative analytical reports. Consider a store that did not sell a product for a period of time. To create such a report, you'll need a factless fact table that captures all of the conceivable product combinations that were on offer. By comparing the factless table to the sales table for the list of things that did sell, you can figure out what's missing.

## 13. What do you mean by Real time data warehousing?

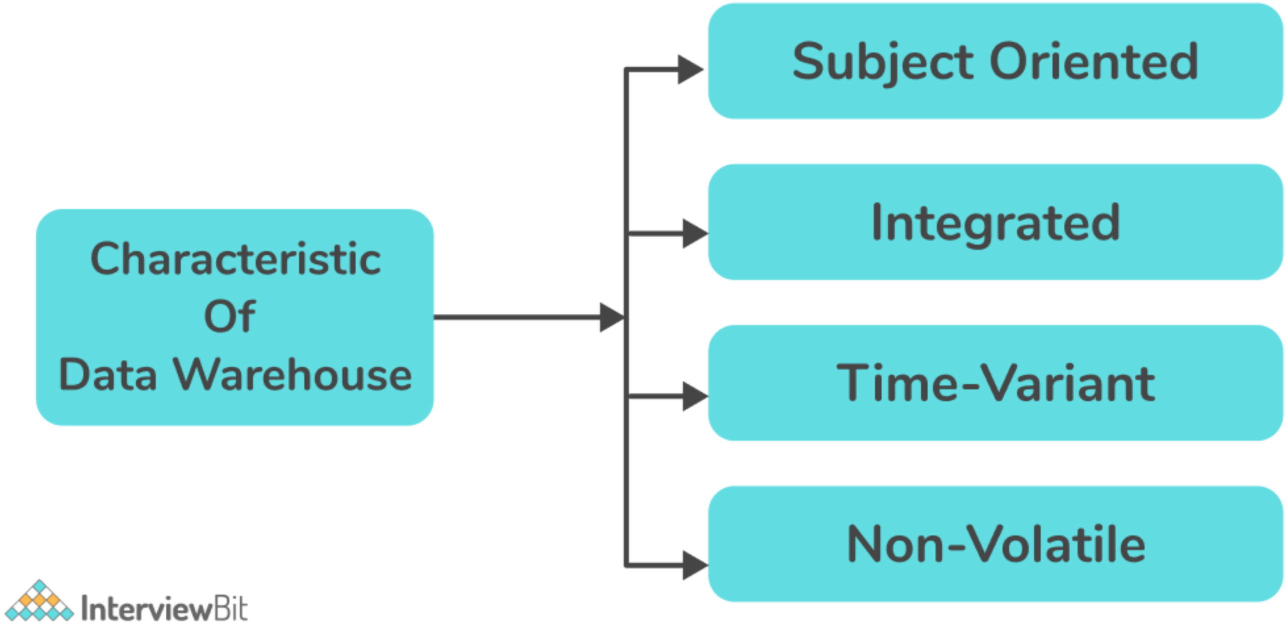
A system that reflects the condition of the warehouse in real time is referred to as real-time data warehousing. If you perform a query on the real-time data warehouse to learn more about a specific aspect of the company or entity described by the warehouse, the result reflects the status of that entity at the time the query was run. Most data warehouses contain data that is highly latent — that is, data that reflects the business at a specific point in time. A real-time data warehouse provides current (or real-time) data with low latency.

## 14. What do you mean by Active Data Warehousing?

The technical capacity to collect transactions as they change and integrate them into the warehouse, as well as maintaining batch or planned cycle refreshes, is known as active data warehousing. Automating routine processes and choices is possible with an active data warehouse. The active data warehouse sends decisions to the On-Line Transaction Processing (OLTP) systems automatically. An active data warehouse is designed to capture and distribute data in real time. They give you a unified view of your customers across all of your business lines. Business Intelligence Systems are linked to it.

## 15. What are the characteristics of a data warehouse?

Following are the characteristics of a data warehouse:-

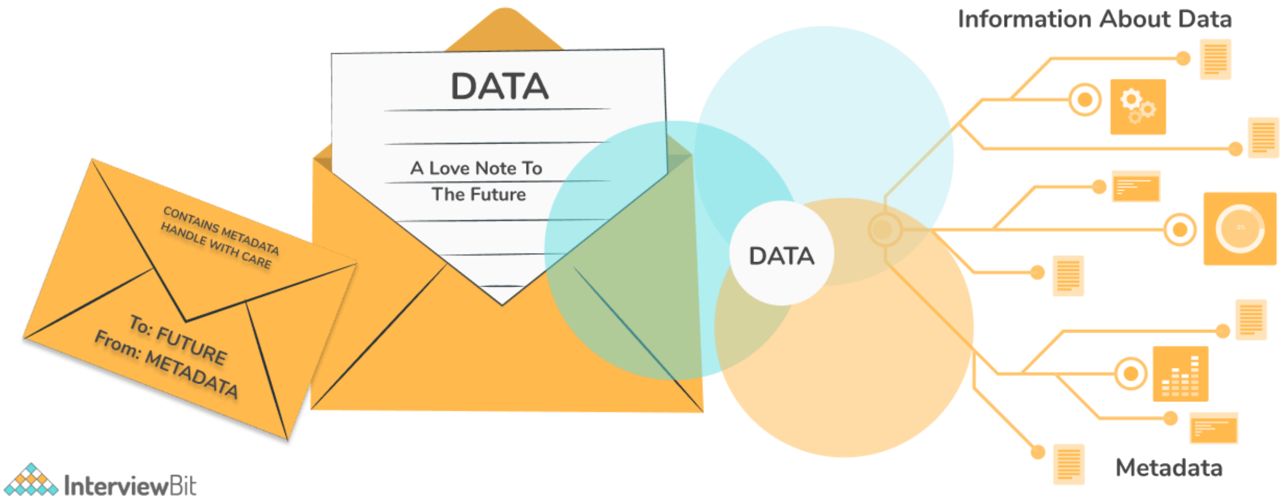


- **Subject-oriented :** Because it distributes information about a theme rather than an organization's actual operations, a data warehouse is always subject-oriented. It is possible to do so with a certain theme. That is to say, the data warehousing procedure is intended to deal with a more defined theme. These themes could include sales, distribution, and marketing, for example. The focus of a data warehouse is never solely on present activities. Instead, it concentrates on demonstrating and analyzing evidence in order to reach diverse conclusions. It also provides a simple and precise demonstration around a specific theme by removing info that isn't needed to make conclusions.
- **Integrated :** It is similar to subject orientation in that it is created in a dependable format. Integration entails the creation of a single entity to scale all related data from several databases. The data has to be stored in several data warehouses in a shared and widely accessible manner. A data warehouse is created by combining information from a variety of sources, such as a mainframe and a relational database. It must also have dependable naming conventions, formats, and codes. The utilization of a data warehouse allows for more effective data analysis. The consistency of name conventions, column scaling, and encoding structure, among other things, should be validated. The data warehouse integration handles a variety of subject-related warehouses.
- **Time-Variant :** Data is kept in this system at various time intervals, such as weekly, monthly, or annually. It discovers a number of time limits that are structured between massive datasets and held in the online transaction process (OLTP). Data warehouse time limitations are more flexible than those of operational systems. The data in the

data warehouse is predictable over a set period of time and provides information from a historical standpoint. It contains explicit or implicit time elements. Another property of time-variance is that data cannot be edited, altered, or updated once it has been placed in the data warehouse.

- **Non-volatile :** The data in a data warehouse is permanent, as the name implies. It also means that when new data is put, it is not erased or removed. It incorporates a massive amount of data that is placed into logical business alteration between the designated quantity. It assesses the analysis in the context of warehousing technologies. Data is read-only and refreshed at scheduled intervals. This is useful for analyzing historical data and understanding how things work. It is not required to have a transaction process, a recapture mechanism, or a concurrency control mechanism. In a data warehouse environment, operations like delete, update, and insert that are performed in an operational application are lost.

## 16. What do you understand about metadata and why is it used for?



Metadata is defined as information about data. Metadata is the context that provides data a more complete identity and serves as the foundation for its interactions with other data. It can also be a useful tool for saving time, staying organised, and getting the most out of the files you're working with. Structural Metadata describes how an object should be classified in order to fit into a wider system of things. Structural Metadata makes a link with other files that allows them to be categorized and used in a variety of ways. Administrative Metadata contains information about an object's history, who owned it previously, and what it can be used for. Rights, licences, and permissions are examples. This information is useful for persons who are in charge of managing and caring for an asset.

When a piece of information is placed in the correct context, it takes on a whole new meaning. Furthermore, better-organized Metadata will considerably reduce search time.

## 17. Enlist a few data warehouse solutions that are currently being used in the industry.

Some of the major data warehouse solutions currently being used in the industry are as follows :

- Snowflakes
- Oracle Exadata
- Apache Hadoop
- SAP BW4HANA
- Microfocus Vertica
- Teradata
- AWS Redshift
- GCP Big Query

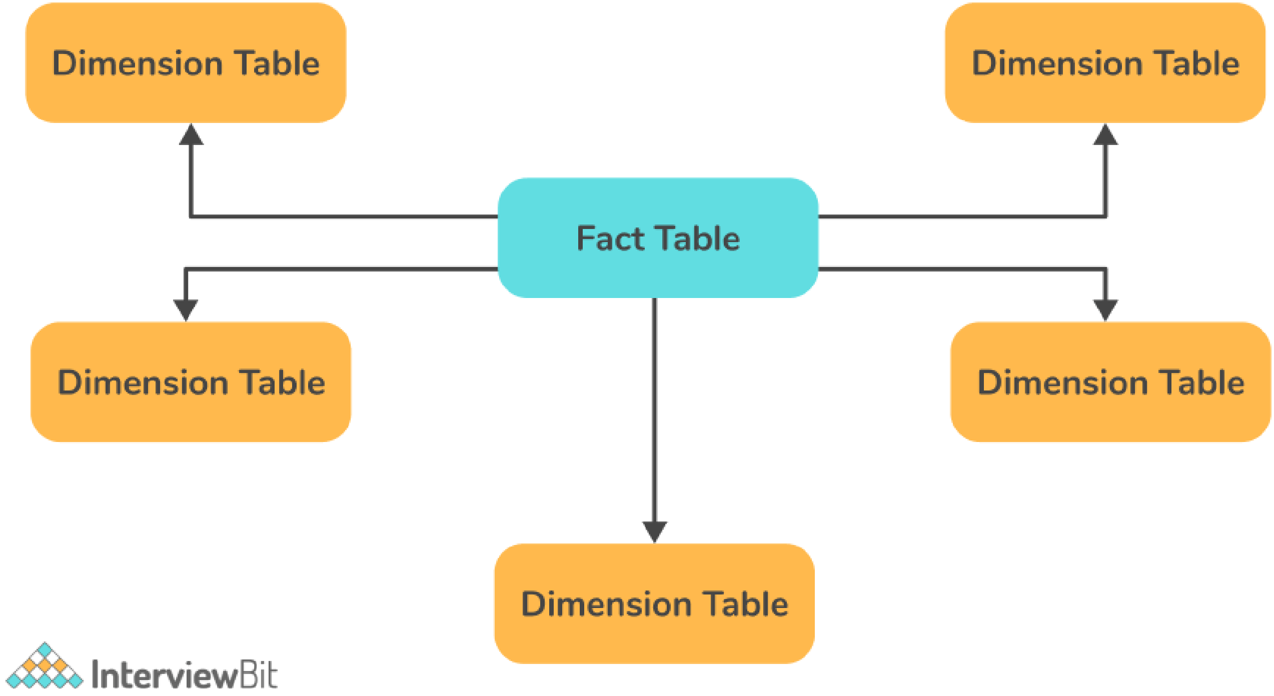
## 18. Enlist some of the renowned ETL tools currently used in the industry.

Some of the renowned ETL tools currently used in the industry are as follows :

- Informatica
- Talend
- Pentaho
- Abnitio
- Oracle Data Integrator
- Xplenty
- Skyvia
- Microsoft – SQL Server Integrated Services (SSIS)

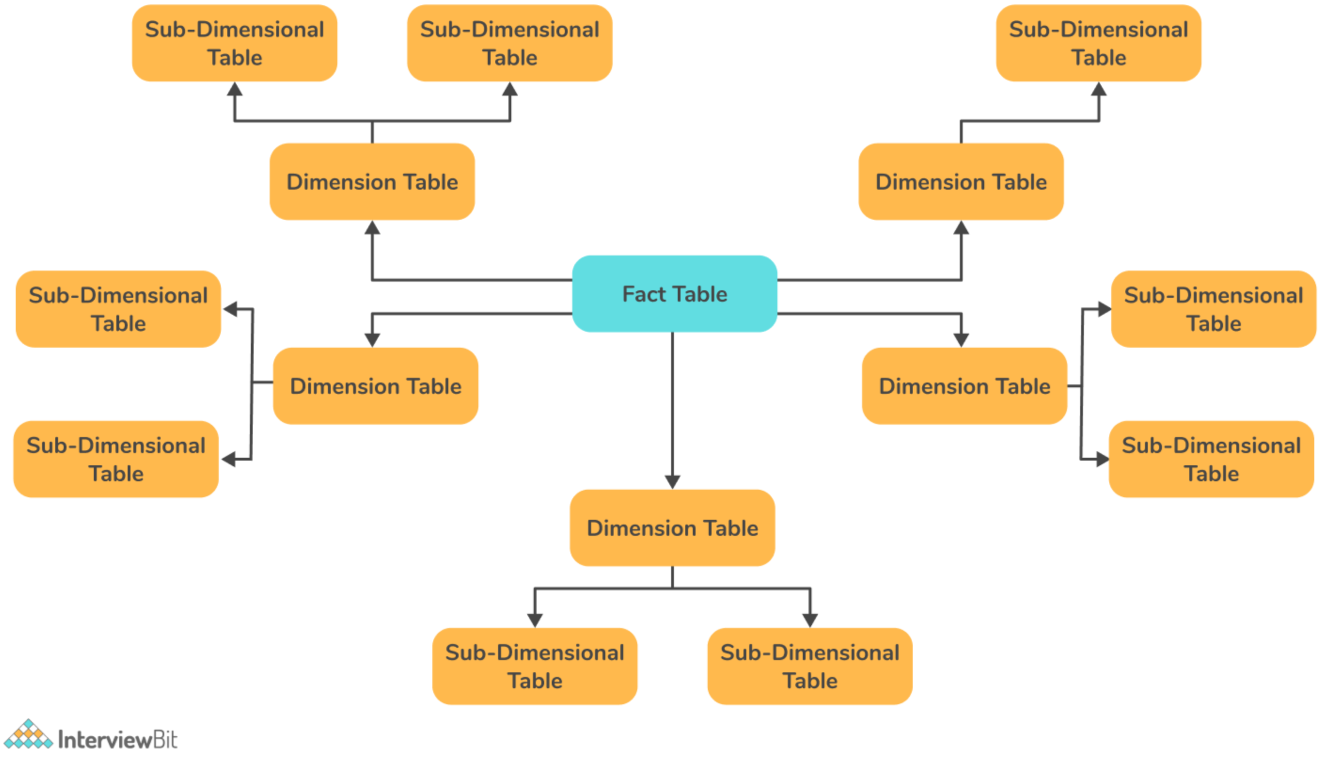
## 19. Explain what you mean by a star schema in the context of data warehousing.

Star schema is a sort of multidimensional model and is used in a data warehouse. The fact tables and dimension tables are both contained in the star schema. There are fewer foreign-key joins in this design. With fact and dimension tables, this schema forms a star.



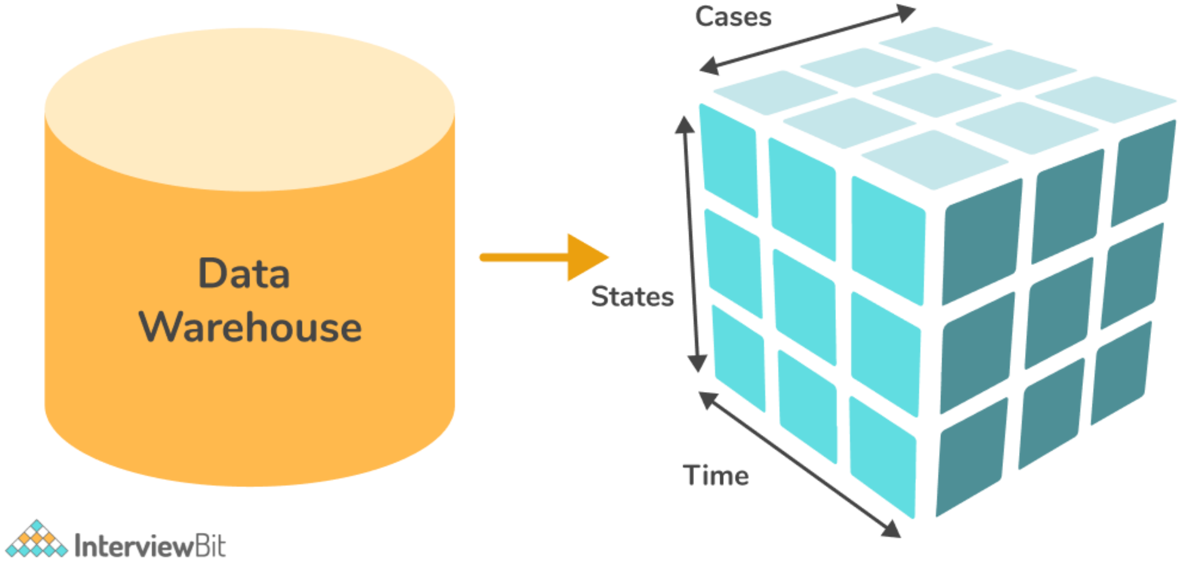
### 20. What do you mean by snowflake schema in the context of data warehousing?

Snowflake Schema is a multidimensional model that is also used in data warehouses. The fact tables, dimension tables, and sub dimension tables are all contained in the snowflake schema. With fact tables, dimension tables, and sub-dimension tables, this schema forms a snowflake.



## 21. What do you understand about a data cube in the context of data warehousing?

A data cube is a multidimensional data model that stores optimized, summarized, or aggregated data for quick and easy analysis using OLAP technologies. The precomputed data is stored in a data cube, which makes online analytical processing easier. We all think of a cube as a three-dimensional structure, however in data warehousing, an n-dimensional data cube can be implemented. A data cube stores information in terms of dimensions and facts.



Data Cubes have two categories. They are as follows :

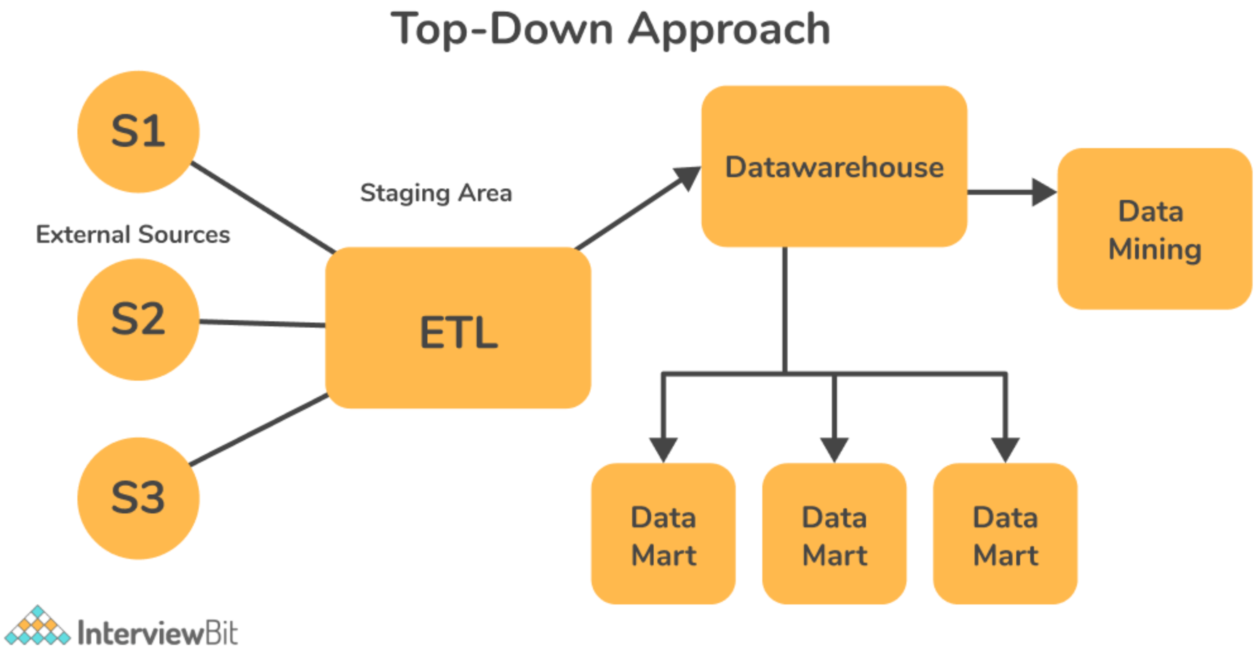
- **Multidimensional Data Cube :** Data is stored in multidimensional arrays, which allows for a multidimensional view of the data. A multidimensional data cube aids in the storage of vast amounts of information. A multidimensional data cube uses indexing to represent each dimension of the data cube, making it easier to access, retrieve, and store data.
- **Relational Data Cube :** The relational data cube can be thought of as an "expanded version of relational DBMS." Data is stored in relational tables, and each relational table represents a data cube's dimension. The relational data cube uses SQL to produce aggregated data, although it is slower than the multidimensional data cube in terms of performance. The relational data cube, on the other hand, is scalable for data that grows over time.

# Data Warehouse Questions for Experienced

## 22. Explain the architecture of a data warehouse.

A data warehouse is a single schema that organizes a heterogeneous collection of multiple data sources. There are two techniques to building a data warehouse. They are as follows:

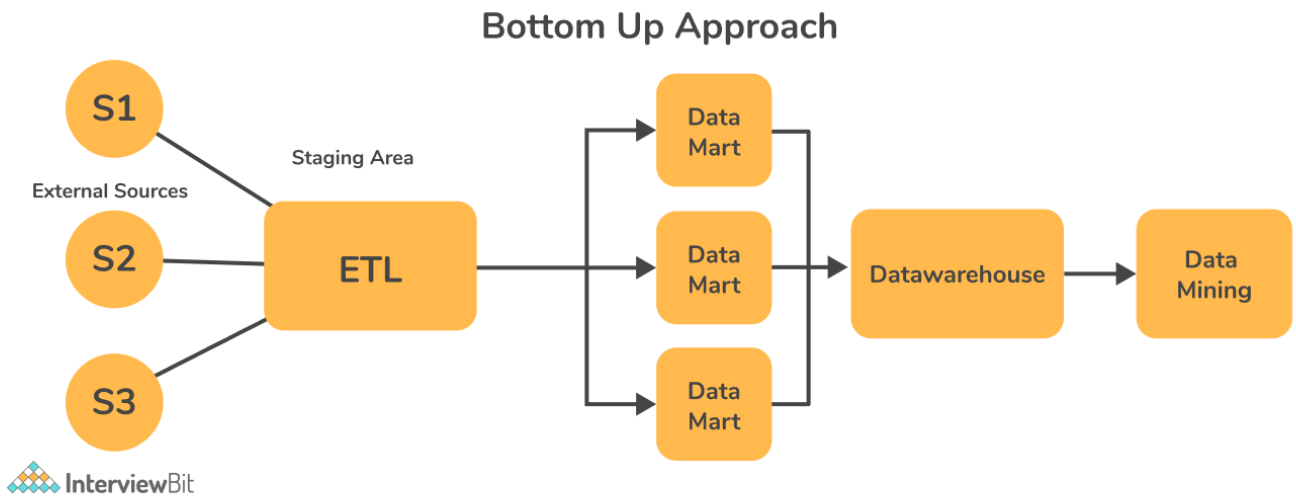
**Top-Down Approach in Data Warehouse:**



Following are the major components :

- **External Sources** - An external source is a location from which data is collected, regardless of the data format. Structured, semi-structured, and unstructured data are all possibilities.
- **Stage Area** - Because the data gathered from external sources does not follow a specific format, it must be validated before being loaded into the data warehouse. ETL tool is used for this purpose in the stage area.
- **Data-warehouse** - After data has been cleansed, it is kept as a central repository in the data warehouse. The meta data is saved here, while the real data is housed in data marts. In this top-down approach, the data warehouse stores the data in its purest form.
- **Data Marts** - A data mart is a storage component as well. It maintains information about a single organization's function that is managed by a single authority. Depending on the functions, an organization can have as many data marts as it wants.
- **Data Mining** - Data mining is the process of analyzing large amounts of data in a data warehouse. With the use of a data mining algorithm, it is used to discover hidden patterns in databases and data warehouses.

**Bottom Up Approach in Data Warehouse:**



Following are the steps involved in the bottom up approach:

- The data is first gathered from external sources (same as happens in top-down approach).
- The data is then imported into data marts rather than data warehouses after passing through the staging area (as stated above). The data marts are built first, and they allow for reporting. It focuses on a specific industry.
- After that, the data marts are incorporated into the data warehouse.

## 23. What are the advantages and disadvantages of the top down approach of data warehouse architecture?

Following are the **advantages** of the top down approach :

- Because data marts are formed from data warehouses, they have a consistent dimensional perspective.
- This methodology is also thought to be the most effective for corporate reforms. As a result, large corporations choose to take this method.
- It is simple to create a data mart from a data warehouse.

The **disadvantage** of the top down approach is that the cost, time, and effort required to design and maintain it are all very expensive.

## 24. What are the advantages and disadvantages of the bottom up approach of data warehouse architecture?

InterviewBit

Credits—InterviewBit
Year- 2023

Following are the **advantages** of the bottom up approach :

- The reports are generated quickly since the data marts are created first.
- We can fit a greater number of data marts here, allowing us to expand our data warehouse.
- In addition, the cost and effort required to build this model are quite minimal.

Because the dimensional view of data marts is not consistent as it is in the top-down approach, this model is not as strong as the top-down approach and this is a **disadvantage** of the bottom up approach.

### 25. Differentiate between a data warehouse and a data mart.

Following table enlists the difference between a data warehouse and a data mart:

| Data Warehouse | Data Mart |
|---|---|
| A data warehouse is a huge collection of data gathered from several departments or groups inside a company. | A data mart is a Data Warehouse's single subtype. It is created to fulfill the requirements of a certain user group. |
| It aids in strategic decision-making. | It aids in the making of tactical business decisions. |
| The process of designing a Data Warehouse is fairly challenging. | The Data Mart design procedure is simple. |
| Data warehousing involves a big portion of the company, which is why it takes so long to process. | Because they can only handle tiny amounts of data, data marts are simple to use, create, and install. |
| The primary goal of a Data Warehouse is to create a unified environment and a consistent view of the business at any given moment in time. | A data mart is primarily used at the department level in a business division. |
| When opposed to data mart, the data kept in the Data Warehouse is always detailed. | Data Marts are designed for certain user groups. As a result, the data is brief and limited. |

| Data Warehouse | Data Mart |
|---|---|
| Data is collected from a variety of sources in a data warehouse. | Data in Data Mart comes from a limited number of sources. |
| The Data Warehouse might be anywhere from 100 GB to 1 TB+ in size. | Data Mart is less than 100 GB in size. |
| The time it takes to implement a Data Warehouse might range from months to years. | The Data Mart implementation process is only a few months long. |
| From the perspective of the end-users, the data stored is read-only. | The transaction data is provided straight from the Data Warehouse. |

## 26. What do you mean by data purging in the context of data warehousing?

Data purging is a term that describes techniques for permanently erasing and removing data from a storage space. Data purging, which is typically contrasted with data deletion, involves a variety of procedures and techniques.

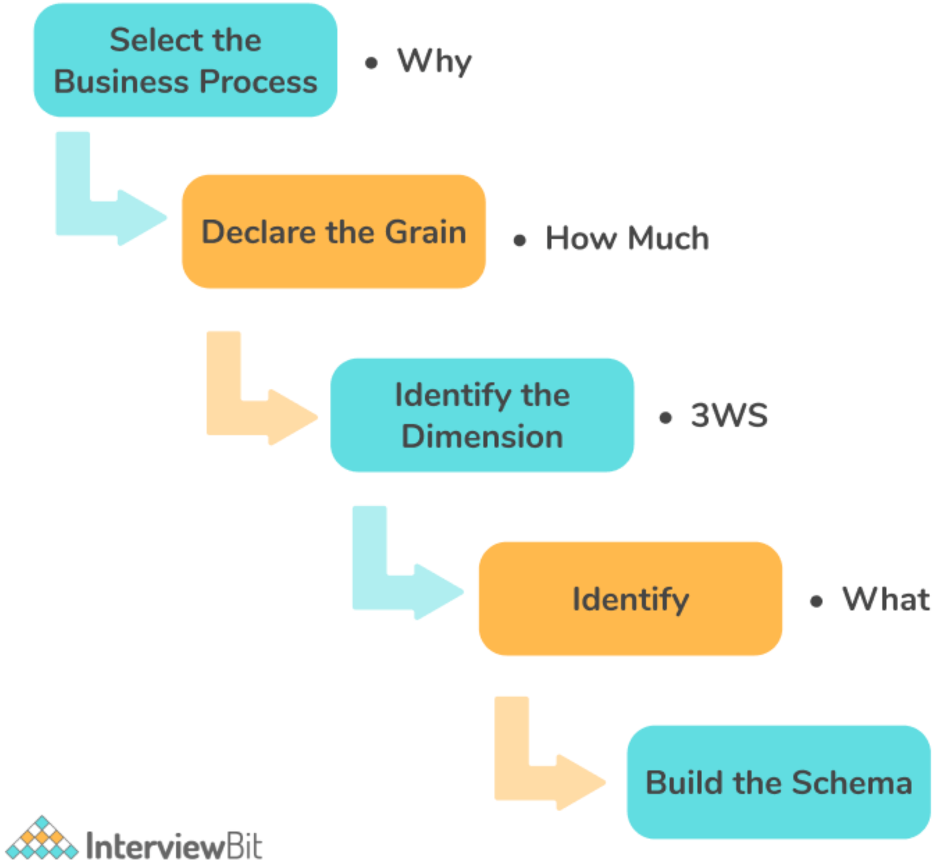
InterviewBit

Purging removes data permanently and frees up memory or storage space for other purposes, whereas deletion is commonly thought of as a temporary preference. Automatic data purging features are one of the methods for data cleansing in database administration. Some Microsoft products, for example, feature an automatic purge strategy that uses a circular buffer mechanism, in which older data is purged to create room for fresh data. Administrators must manually remove data from the database in other circumstances.

## 27. What do you mean by dimensional modelling in the context of data warehousing?

Dimensional Modelling (DM) is a data structure technique that is specifically designed for data storage in a data warehouse. The goal of dimensional modelling is to optimise the database so that data can be retrieved more quickly. In a data warehouse, a dimensional model is used to read, summarise, and analyse numeric data such as values, balances, counts, weights, and so on. Relation models, on the other hand, are designed for adding, modifying, and deleting data in a real-time Online Transaction System.

Following are the steps that should be followed while creating a dimensional model:



- **Identifying the business process :** The first step is to identify the specific business processes that a data warehouse should address. This might be Marketing, Sales, or Human Resources, depending on the organization's data analytic needs. The quality of data available for that process is also a factor in deciding which business process to use. It is the most crucial step in the Data Modeling process, and a failure here would result in a cascade of irreversible flaws.
- **Identifying the grain :** The level of detail for the business problem/solution is described by the grain. It's the procedure for determining the lowest level of data in any table in your data warehouse. If a table contains sales data for each day, the granularity should be daily. Monthly granularity is defined as a table that contains total sales data for each month.
- **Identifying the dimension :** Date, shop, inventory, and other nouns are examples of dimensions. All of the data should be saved in these dimensions. The date dimension, for example, could include information such as the year, month, and weekday.
- **Identifying the fact :** This stage is linked to the system's business users because it is here that they gain access to data housed in the data warehouse. The majority of the rows in the fact table are numerical values such as price or cost per unit.
- **Building the schema :** The Dimension Model is implemented in this step. The database structure is referred to as a schema (arrangement of tables).

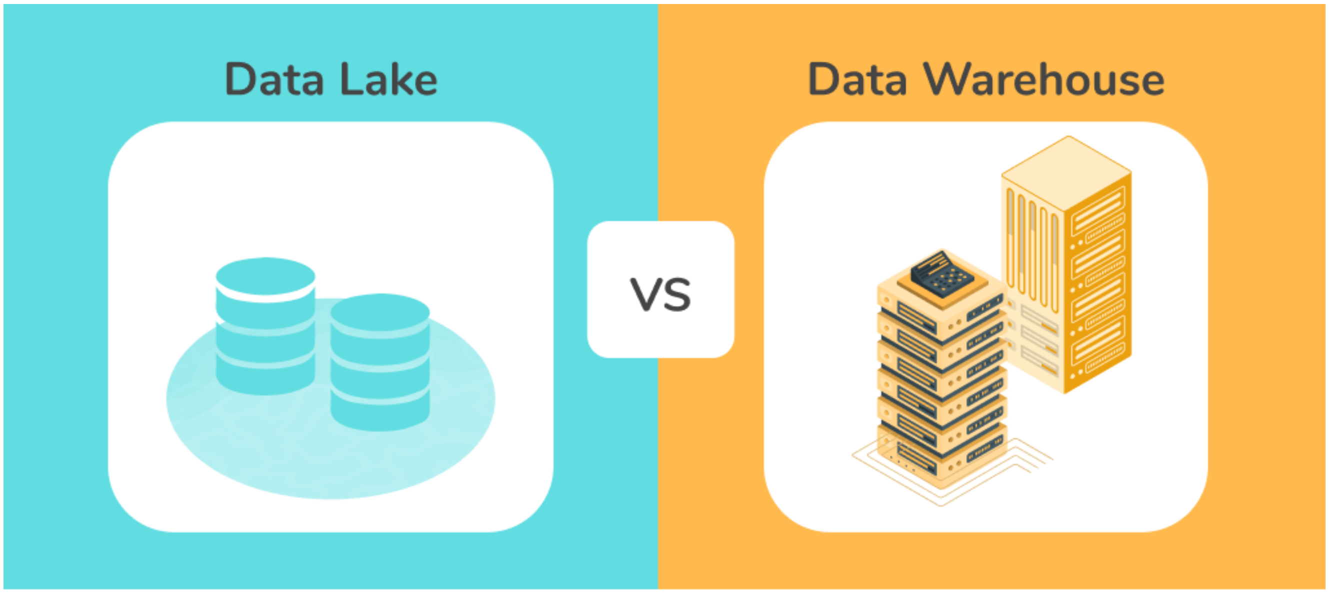
## 28. What do you understand by data lake in the context of data warehousing? Differentiate between data lake and data warehouse.

A Data Lake is a large-scale storage repository for structured, semi-structured, and unstructured data. It's a location where you can save any type of data in its original format, with no restrictions on account size or file size. It provides a significant amount of data for improved analytical performance and native integration.

A data lake is a huge container that looks a lot like a lake or a river. Similar to how a lake has various tributaries, a data lake has structured data, unstructured data, machine-to-machine communication, and logs flowing through in real-time.

The following table enlists the differences between data lake and data warehouse:

Disha Mukherjee
Credits—InterviewBit
Year- 2023



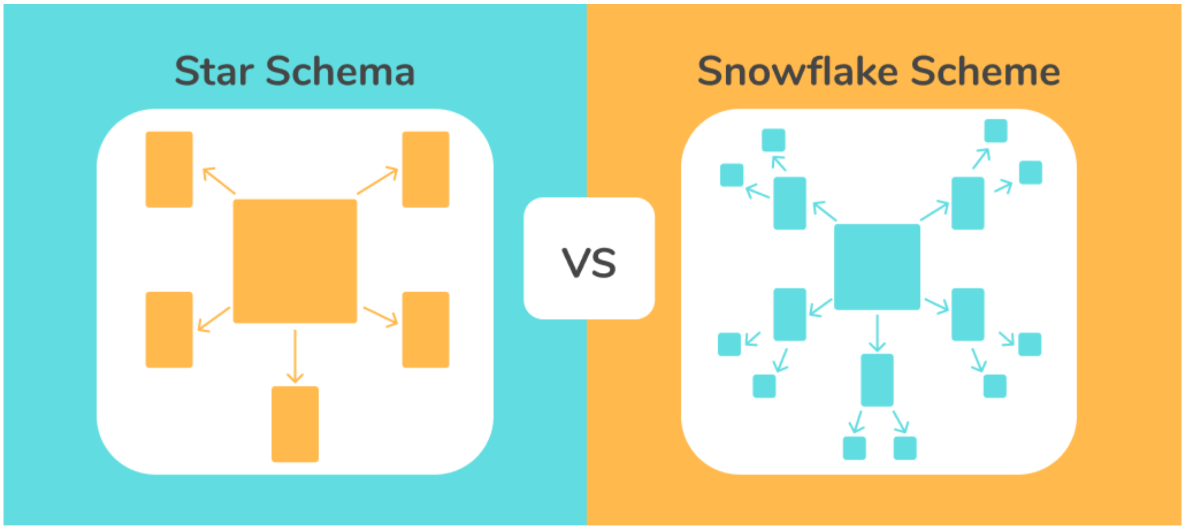
InterviewBit

| Data Lake | Data Warehouse |
|---|---|
| All data is stored in the data lake, regardless of its source or structure. The data is stored in its unprocessed state. When it is ready to be used, it is converted. | Data extracted from transactional systems or data consisting of quantitative measures and their properties will be stored in a data warehouse. The information has been cleansed and changed. |
| Captures semi-structured and unstructured data in their original form from source systems. | Captures structured data and organises it according to defined standards for data warehouse purposes. |
| The data lake is appropriate for those that perform in-depth analysis. Data scientists, for example, require advanced analytical techniques that include predictive modelling and statistical analysis. | Because it is highly structured, easy to use, and understand, the data warehouse is perfect for operational users. |
| The cost of storing data in big data technology is less than that of storing data in a data warehouse. | Data warehouse storage is more expensive and time-consuming. |
| The schema is usually developed after the data has been stored. This provides a great level of flexibility and convenience | Schema is usually defined before data is saved. Work is required at the start of the process, but performance, security, and integration are all advantages. |

| Data Lake | Data Warehouse |
|---|---|
| of data collecting, but it necessitates labour at the end of the process. | |
| Users can access data in data lakes before it has been transformed, cleansed, or structured. In comparison to a traditional data warehouse, it allows consumers to get to their results faster. | Pre-defined inquiries for pre-defined data kinds are answered by data warehouses. As a result, any updates to the data warehouse take longer. |

## 29. Differentiate between star schema and snowflake schema in the context of data warehousing.

Following table enlists the difference between the star schema and the snowflake schema:



InterviewBit

| Star Schema | Snowflake Schema |
|---|---|
| The fact tables and dimension tables are both contained in the star schema. | The fact tables, dimension tables, and sub dimension tables are all contained in the snowflake schema. |
| It is a top-down model. | It is, however, a bottom-up model. |
| The star schema takes up more room. | It takes up less space. |

| **Star Schema** | **Snowflake Schema** |
|---|---|
| Star schema has a low query complexity. | Snowflake schema has a higher query complexity than star schema. |
| It is really simple to comprehend. | It is tough to comprehend. |
| It contains less foreign keys. | It has a greater number of foreign keys. |
| It has a lot of redundancy in its data. | It has a low level of data redundancy. |
| The execution of queries takes less time. | The execution of queries takes longer than star schema. |
| Normalization is not employed in the star schema. | Both normalisation and denormalization are used in this. |
| It has a pretty simple design. | It has a complicated design. |

## 30. Differentiate between Agglomerative hierarchical clustering and Divisive clustering.

**Agglomerative hierarchical clustering :** Flat clustering returns an unstructured set of clusters. On the other hand, this structure is more informative. We don't have to define the number of clusters in advance with this clustering procedure. Bottom-up algorithms start by treating each piece of data as a singleton cluster, then agglomerate pairs of clusters until all of them are merged into a single cluster that contains all of the data.

**Divisive Clustering :** This approach also eliminates the need to define the number of clusters ahead of time. It necessitates a method for breaking a cluster that contains all of the data and then recursively splitting clusters until all of the data has been split into singletons.

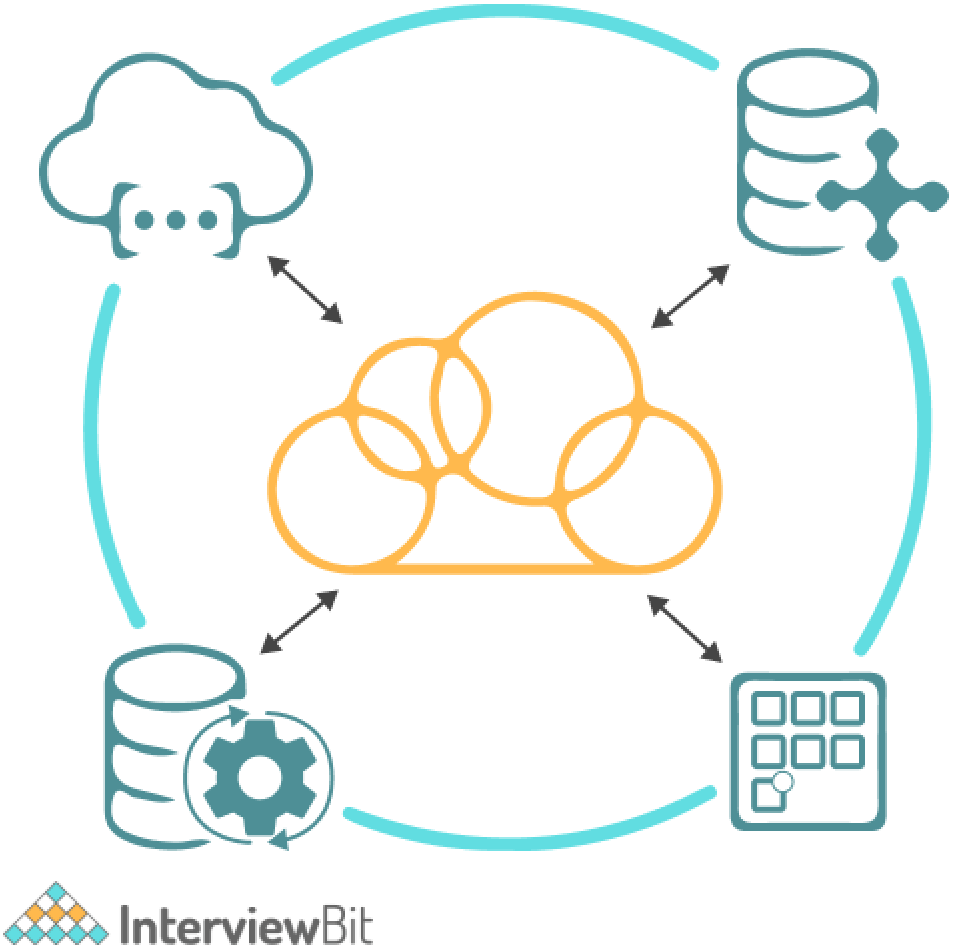
Following are the differences between the two :

- When compared to agglomerative clustering, divisive clustering is more complicated since we require a flat clustering algorithm as a "subroutine" to split each cluster until each data has its own singleton cluster.
- If we don't create a complete hierarchy all the way down to individual data leaves, divisive clustering is more efficient.
- A divisive algorithm is also more precise. Without first examining the global distribution of data, agglomerative clustering makes judgments based on local patterns or neighbour points. These early decisions are irreversible. When generating

top-level dividing decisions, divisive clustering takes into account the global distribution of data.

## 31. What are the advantages of a cloud based data warehouse?

Following are the advantages of a cloud-based data warehouse:



- **Total cost of ownership is low:** The low cost of cloud data warehouses is one of the reasons they are becoming more popular. On-premises data warehouses necessitate high-cost technology, lengthy upgrades, ongoing maintenance, and outage management.
- **Increased performance and speed:** To keep up with the expanding number of data sources, cloud data warehouses are crucial. Cloud data warehouses can easily and quickly integrate with additional data sources as needed, and deploy the updated

solution to production. Cloud data warehouses significantly improve speed and performance, allowing IT to focus on more innovative projects.

- **Enhanced Security:** Cloud security engineers can create and iterate on precise data-protection measures. Furthermore, cloud encryption technologies such as multi-factor authentication make data transfer between regions and resources extremely safe.
- **Improved Disaster Recovery:** Physical assets are not required to prepare cloud data warehouses for disasters. Instead, almost all cloud data warehouses offer asynchronous data duplication and execute automatic snapshots and backups. This data is kept across multiple nodes, allowing duplicate data to be accessed at any time without stopping present activity.

## Conclusion:

In this article, we have covered the most frequently asked interview questions on data warehousing. ETL tools are often required in a data warehouse and so one can expect interview questions on ETL tools as well in a data warehouse interview.