Unit-1

1)OLTP=Online Transaction Processing

2)OLAP=Online analytical processing

3)DWH=data warehouse

4)EDW=enterprise data warehouse

5)CRM=Customer Relationship Management

6)Explain future trands of datawarehouse.

🡺Data warehousing is very much useful in the future era. For this it use traditional and new improvements because day by day the data will be increased and the market will be grow.

Unstructured data makes a unique challenge for companies wishing to utilize their data for analysis

. • If your company is trying to deal with its unstructured data now, that issue will weaken over time. • In the business world, most unstructured information lies in customer-related content, which is plentiful and accessible.

• Organizations utilize big data tools and software such as Hadoop to process, mine, integrate, store, track, index and report business insights from raw unstructured information.

Data that is in databases is structured; it is been organized in row and columns in data warehouse.

How do you get the information out after store data warehouse? The answer is by searching. To get the information out of structured data.

• Search Engine refers to a huge database of internet resources such as web pages, newsgroups, programs, images etc. It helps to locate information on World Wide Web.

Service-oriented architecture is a software architecture that describes a componentbased approach to allow loosely coupled components to address business requirements.

• This new IT solution will connect the SOA integration layer to near-real-time data warehouse and consider the near-real-time data warehouse as a data stage source for other business and system environments.

More and more data warehousing applications on all fronts are been built using SOA: ETL, reporting, analytics, BI applications, data mining, metadata, data quality and data cleansing.

Real-time data warehousing

• Real-time Data Warehousing describes a system that reflects the state of the warehouse in real time.

If a query run against the real-time data warehouse to understand a particular fact about the business or entity described by the warehouse, the answer reflects the state of that entity at the time the query run. • A real-time data warehouse is a data warehouse that is update (by the ETL) the moment the transaction happens in the source system

7)what is normalization data store?

🡺 Other types of data warehouses put the data not in a dimensional data store b store. A normalized data store is one or more relational databases with little or no data redundancy.

A relational database is a database that consists of entity tables with parent between them. Normalization is a process of removing data redundancy by implementing normalization rules.

There are five degrees of normal forms, from the first normal form to the fifth normal form.

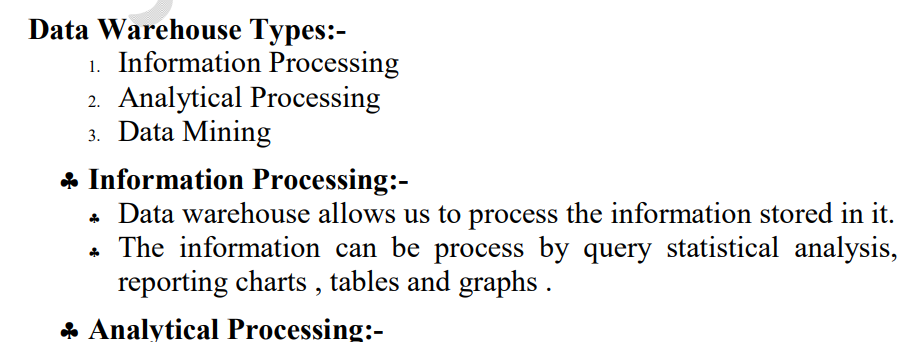
A normalized datastore is usually in third normal form or higher, such as fourth or fifth normal form.

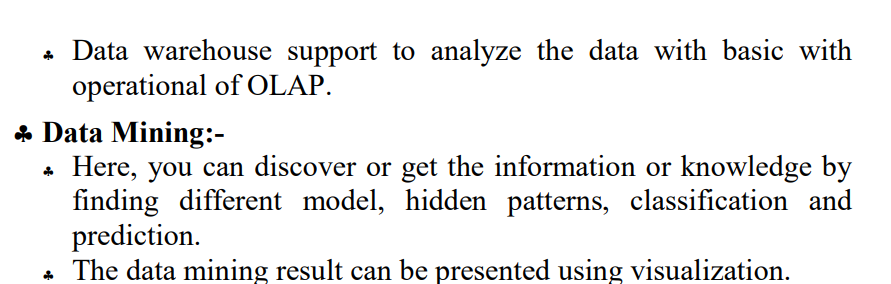
The production of clean data is generally referred to as **Data Normalization**. However, when you dig a little deeper, the meaning or goal of Data Normalization is twofold:

* Data Normalization is the process of organizing data such that it seems**consistent across all records and fields**.
* **It improves the cohesion of entry types,** resulting in better data cleansing, lead creation, and segmentation.

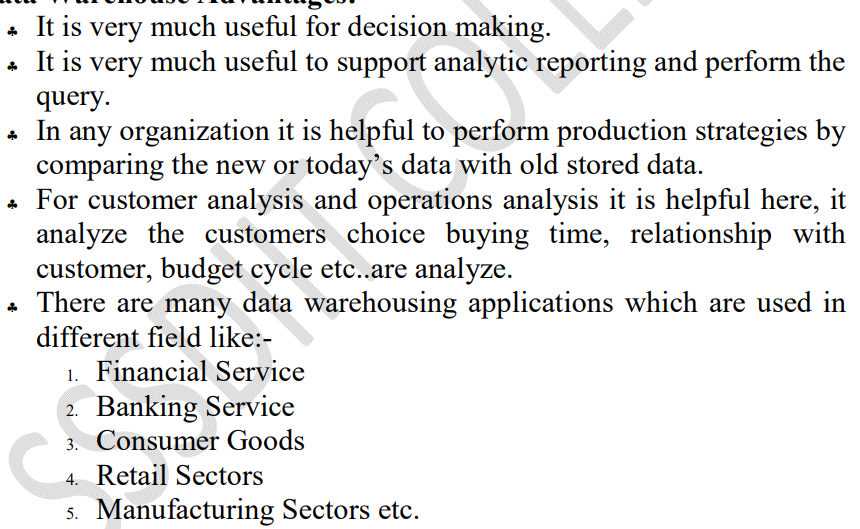
Simply said, this procedure entails**removing Unstructured Material**as well as Redundancy (duplicates) to ensure logical data storage. When Data Normalization is done correctly, standardized data entry is the result. This technique, for example, applies to the recording of URLs, contact names, street locations, phone numbers, and even codes. These standardized data fields can thus be quickly grouped and read.

8)Explain data Warehouse types.

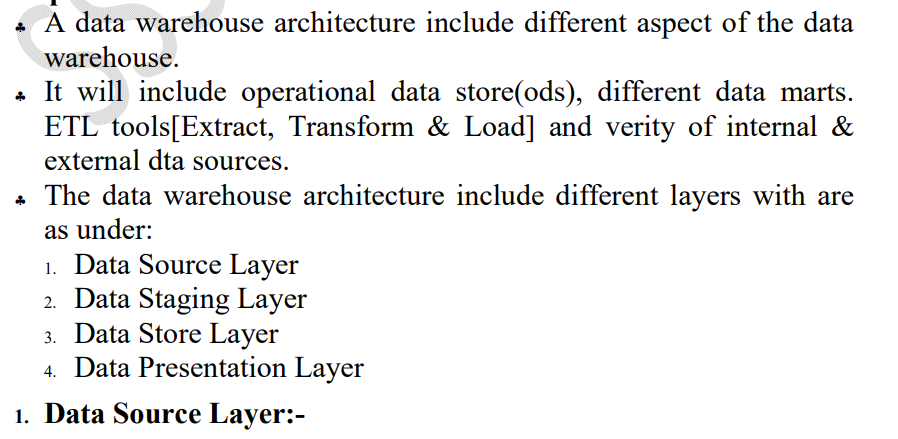
🡪

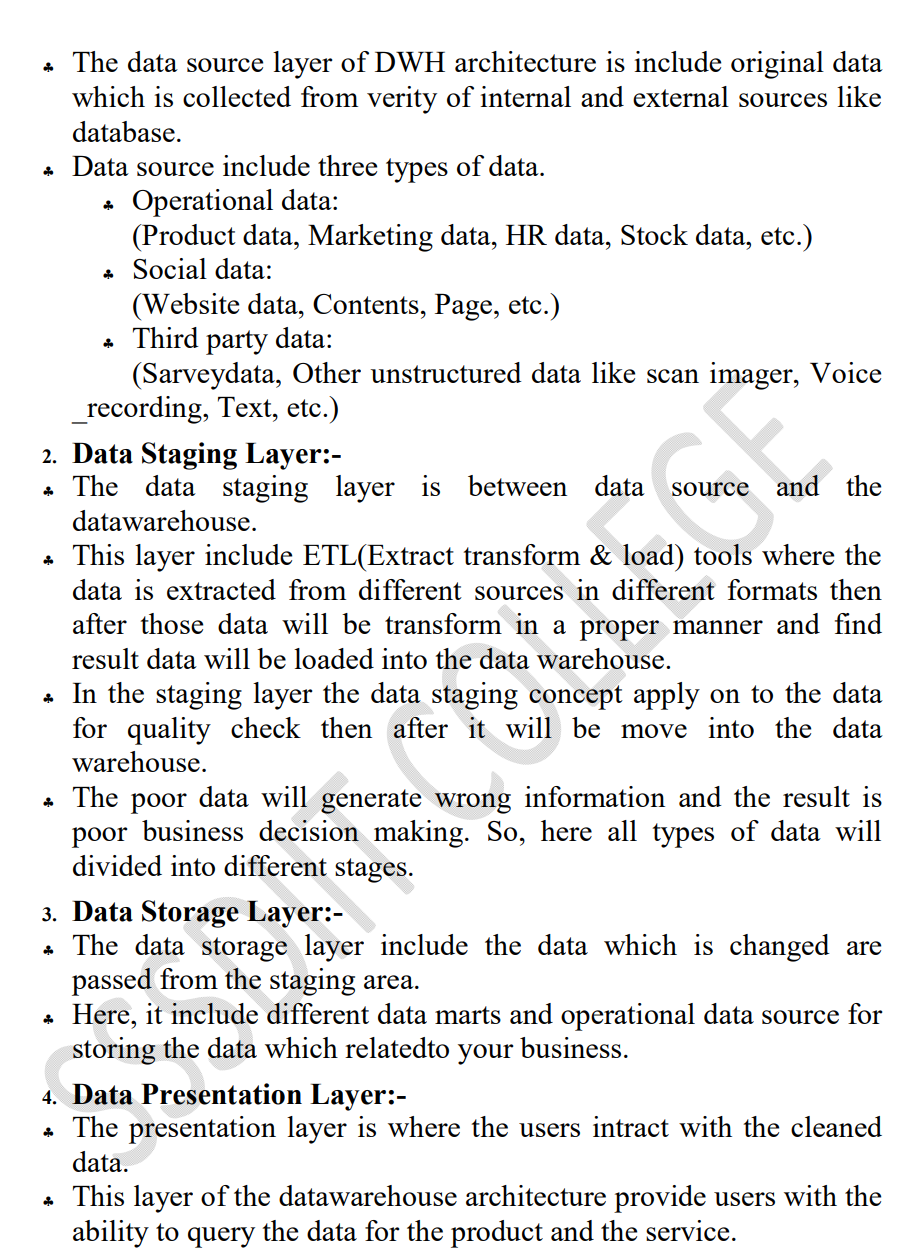


9)Explain Advantages of Datawarehouse.

🡪

10)Explain datawarehouse architecture.

🡪



11)what is datawarehouse ? explain its characteristics.

🡪A **data warehouse** is a type of [data management](https://www.oracle.com/in/database/what-is-data-management/) system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. The data within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications.

**Characteristics of data wre house**

1. Subject Oriented: Focuses on a specific area or subject such as sales, customers, or inventory.
2. Integrated: Integrates data from multiple sources into a single, consistent format.
3. Read-Optimized: Designed for fast querying and analysis, with indexing and aggregations to support reporting.
4. Summary Data: Data is summarized and aggregated for faster querying and analysis.
5. Historical Data: Stores large amounts of historical data, making it possible to analyze trends and patterns over time.
6. Schema-on-Write: Data is transformed and structured according to a predefined schema before it is loaded into the data warehouse.
7. Query-Driven: Supports ad-hoc querying and reporting by business users, without the need for technical support.

12)write advantages and disadvantages of datawarehouse.

🡪**advantages:**

* Data warehouses facilitate end users' access to a variety of data.
* Using numerous data warehouses can increase the operational value of business systems, especially customer relationship management.
* Makes selections with higher quality.
* For the medium and long term, it is especially helpful.
* Storage of analyses and historical search queries is quite beneficial.
* It has a strong capacity for digesting information.
* Access to information is made more flexible and quick because of it.
* Allows for easier corporate decision-making.
* The productivity of businesses rises.
* Gives the company's many departments reliable communication.
* Strengthen connections with customers and suppliers.
* It makes it possible to keep up with business activity and be constantly informed of successful and unsuccessful outcomes.
* transforms information into knowledge and data into information
* You can plan more successfully, thanks to it.
* Cut back on operating expenses and response times.
* The Data Warehouse assists in fusing many data sources, lessening the production system's workload.
* The data warehouse aids in reducing the overall turnaround time for reporting and research.

**Disadvantages:**

* The data warehouses may project substantial expenditures throughout his life. The data warehouse is typically not stationary. Costs for maintenance are considerable.
* Data warehouses could soon become outdated.
* They occasionally need to provide complete information before a request for information, which also costs the organization money.
* Regarding the various objectives a company seeks to achieve, challenges may arise during implementation.
* It might be challenging to include new data sources once a system has been implemented.
* They necessitate an examination of the data model, objects, transactions, and storage.
* They were designed in a sophisticated, multidisciplinary manner.
* The operating systems must be reorganized to accommodate them.
* The data warehouse may seem simple, but it is too complex for the typical person to comprehend.
* The scope of the data storage project will start to expand, despite the best efforts of project management.
* At this point, various business regulations may already be in place for warehouse clients.

13)Explain dataflow architecture of data warehouse.

🡪• The data flow architecture is about how the data stores are arranged within a data warehouse and how the data flows from the source systems to the users through these data stores.

• In data warehousing, the data flow architecture is a configuration of data stores within a data warehouse system, along with the arrangement of how the data flows from the source systems through these data stores to the applications used by the end users.

• This include show the data flows are controlled, logged, and monitored, as well as the mechanism to ensure the quality of the data in the data stores.

• Data stores are important components of data flow architecture. A data store is one or more databases or files containing data warehouse data, arranged in a particular format and involved in data warehouse processes.

• Based on the user accessibility, you can classify data warehouse data stores into three types:

♣ A user-facing data store is a data store that is available to end users and is queried by the end users and end-user applications.

♣ An Internal data store is a data store that is used internally by data warehouse components for the purpose of integrate, cleansing, logging, and preparing data, and it is not open for query by the end users and end-user applications.

♣ A hybrid data store is used for both internal data warehouse mechanisms and for query by the end users and end-user applications.

• A master data store is a user-facing or hybrid data store containing a complete set of data in a data warehouse, including all versions and all historical data. Based on the data format,

• You can classify data warehouse data stores into four types:

♣ A stage is an internal data store used for transforming and preparing the data obtained from the source systems, before the data is loaded to other data stores in a data warehouse.

14)Discuss CRM in details.

🡪Customer Relationship Management (CRM) is the activity of establishing contact and managing communications with customers, analyzing information about customers, campaigning to attract new customers, performing business transactions with customers, servicing customers, and providing support to customers.

There are 3 main types of CRM software: operational CRM systems, collaborative CRM systems, and analytical CRM systems.

Customer Relationship Management (CRM) is the activity of establishing contact and managing communications with customers, analyzing information about customers, campaigning to attract new customers, performing business transactions with customers, servicing customers, and providing support to customers.

**CRM Advantages**

CRM is an integral part of any business’s success. When performing well, it results in higher customer satisfaction, improved customer retention rates, and enhanced customer loyalty.

 It allows the business to nurture relationships with customers, giving them a better experience, and shows that the business is really concerned about their needs.

 CRM also helps in improving the overall sales performance by assisting decision-making processes at all levels of the organization.

 It can also result in better [understanding](https://crm.walkme.com/7-small-business-best-practices-in-microsoft-dynamics-crm-infographic/) customers’ needs, wants, and expectations. This enables you to tailor your products and services according to their demands.

 The importance of CRM has been recognized by most organizations. Implementation of CRM is on the rise and many companies are looking at new ways to utilize this technology.

**CRM Disadvantages**

Customer relationship management (CRM) also has certain disadvantages.

 Although CRM is a great tool for improving customer relationships, it is not without its faults and limitations. There are some aspects of the system that you should be aware of before implementing it in your organization.

 For one thing, they can be quite expensive and [time-consuming](https://www.activecampaign.com/blog/hours-needed-for-crm) to implement. It can take weeks, months, or even several months just to come up with a system that is workable for your organization.

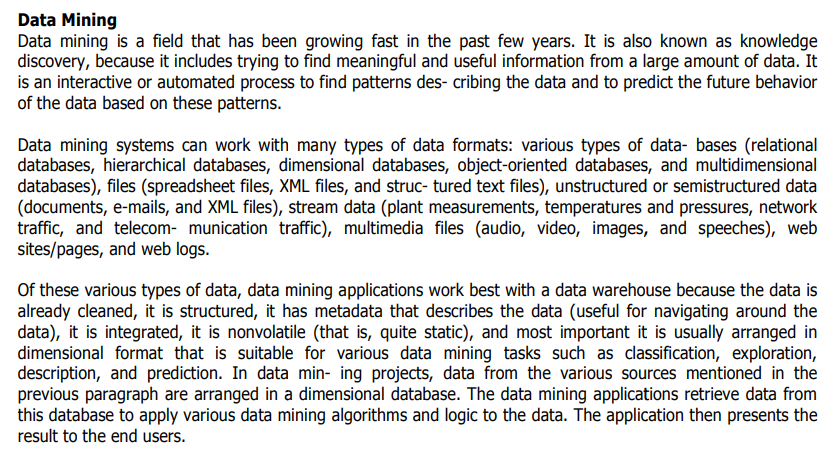
 There are also limitations on data, as CRM requires all your customer data to be in one place.

This means that you will have to go through all your information and decide which information you wish to be stored in the CRM system.

 Furthermore, there are many different CRM systems on the market today and it can be quite challenging to choose the right one for your business.



15)Explain datamining .

🡪

16)

UNIT-2

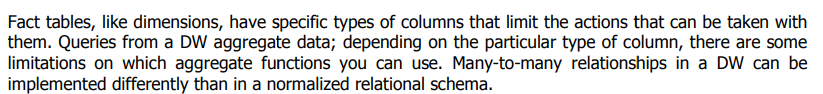
1)List out types of Entity relationship .

🡪

2)what is role playing dimension ?

🡪 A role-playing dimension is a dimension that can be associated with multiple facts in different roles. For example, a date dimension can play the role of order date, ship date, or delivery date for a sales fact.

3)What is fact table ?

🡺

4) SDC=Slowly changing Dimension

5)what is data wrehoue schema?

🡪 Data warehouse schema is a description, represented by objects such as tables and indexes, of how data relates logically within a data warehouse. Star, galaxy, and snowflake schema are types of warehouse schema that describe different logical arrangements of data.

6)Give steps for logica design of datawarehouse.

🡪

7)Give steps for physical design of data warehouse.

🡺

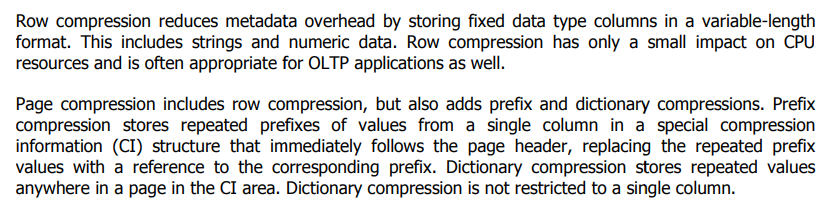
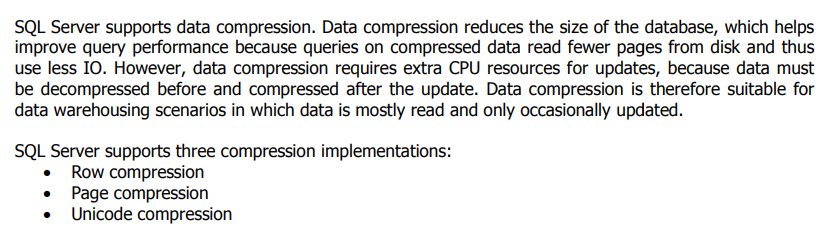
1. dentify business requirements.
2. Create a conceptual model.
3. Develop a logical model.
4. Define the physical model.
5. Extract, transform, and load (ETL) processes.
6. Develop reporting and analysis tools.
7. Implement data quality and data governance processes.

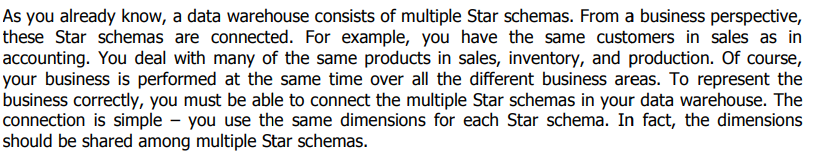
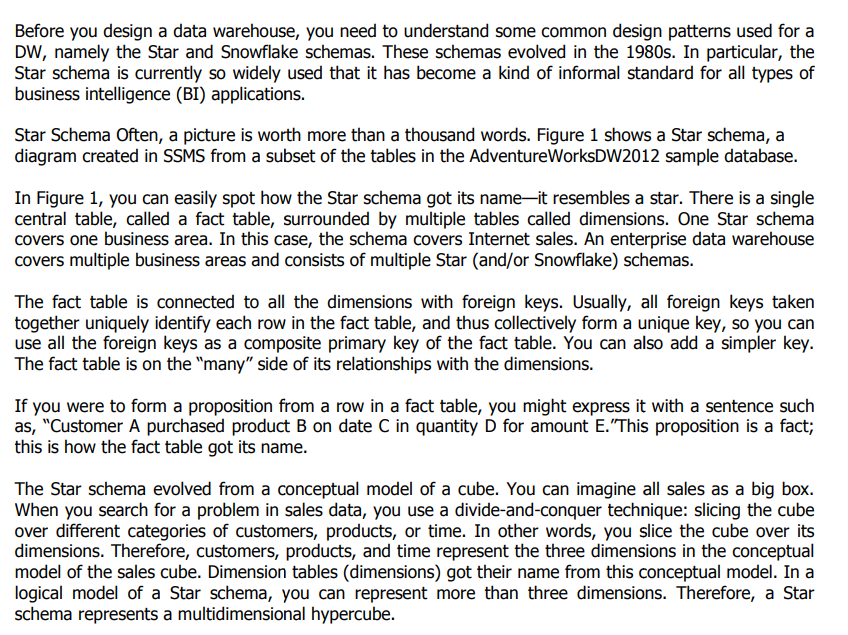
**8)Explain types of dimension.**

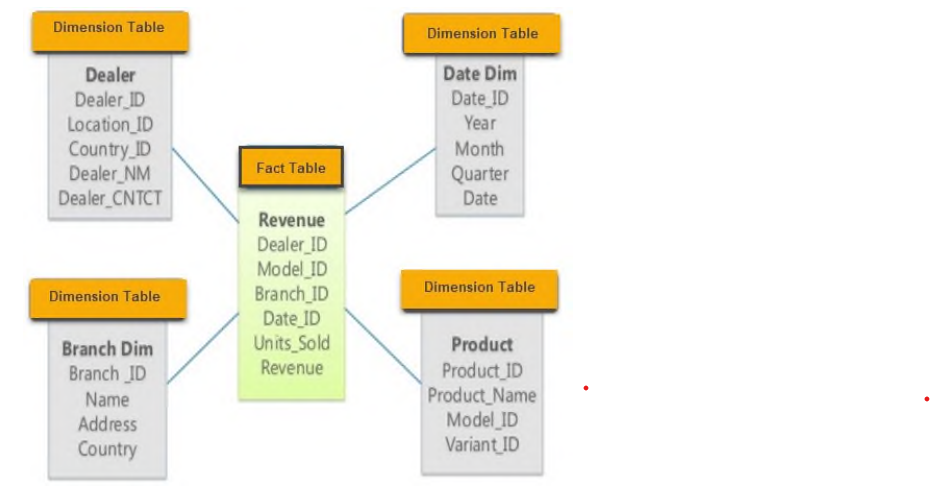
## 🡪

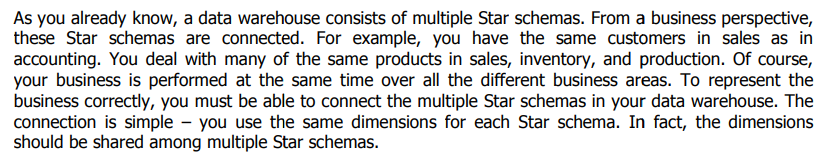
## 

## 

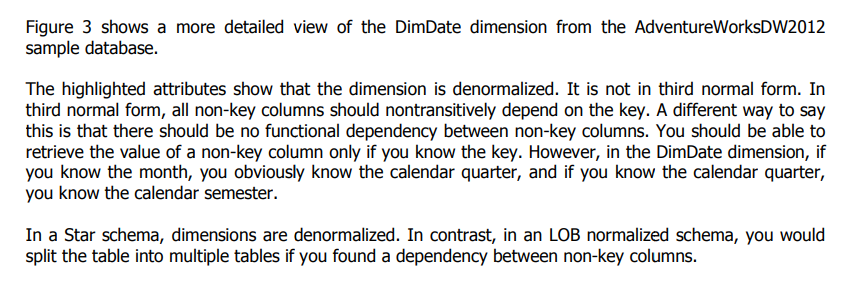
**9)What is data compression?🡪**

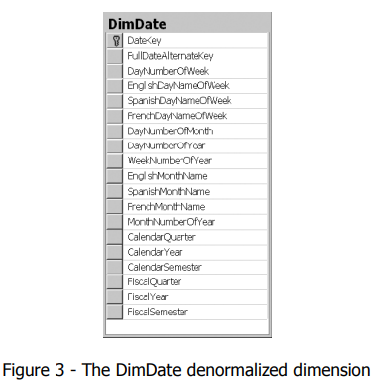
**10)Explain star schema with example.🡪**

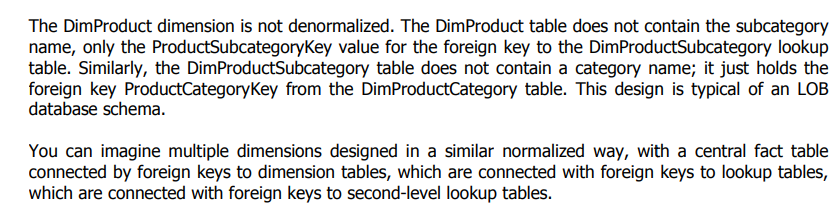
****

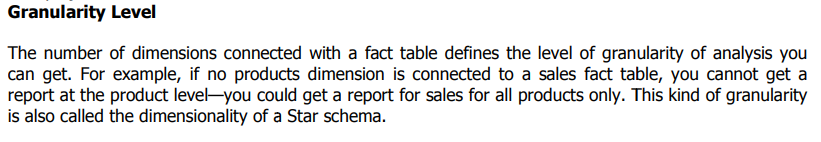
****

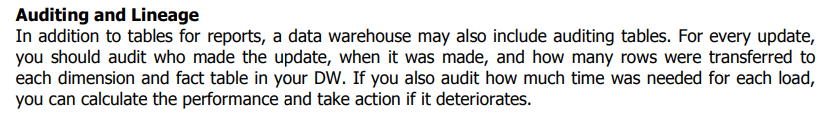
**11)Explain snowflake schema with example.**

****

****

****

****

****

**13)Discuss view in details..**

**🡪** the Data Warehouse Views feature is a method of creating new warehoused tables by modifying an existing table, or joining or consolidating multiple tables together by using SQL.

A database view is a subset of a database and is based on a query that runs on one or more database tables. Database views are saved in the database as named queries and can be used to save frequently used, complex queries. There are two types of database views: dynamic views and static views.

Complex View: A view based on multiple tables, which contain GROUP BY clause and functions. Inline View: A view based on a subquery in FROM Clause, that subquery creates a temporary table and simplifies the complex query. Materialized View: A view that stores the definition as well as data.

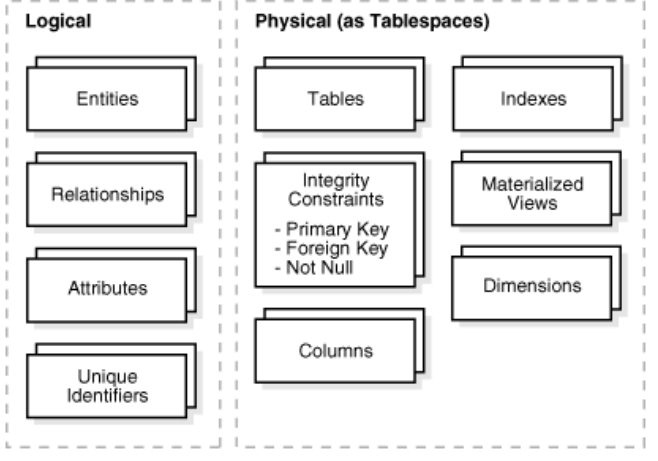
Views are used for security purposes because they provide encapsulation of the name of the table. Data is in the virtual table, not stored permanently. Views display only selected data. We can also use Sql Join s in the Select statement in deriving the data for the view.

**14)Discuss phisycal design for datawarehouse.**

**🡪**During the physical design process, you convert the data gathered during the logical design phase into a description of the physical database structure.

• During the logical design phase, you defined a model for your data warehouse consisting of entities, attributes, and relationships. The entities are linked together using relationships. Attributes are used to describe the entities. The unique identifier (UID) distinguishes between one instance of an entity and another.

• Figure illustrates a graphical way of distinguishing between logical and physical designs.

****

During the physical design process, you translate the expected schemas into actual database structures. At this time, you must map:

• Entities to tables

• Relationships to foreign key constraints

• Attributes to columns

• Primary unique identifiers to primary key constraints

• Unique identifiers to unique key constraints

**Physical Design Structures**

• Once you have converted your logical design to a physical one, you must create some or all of the following structures:

1. Tablespaces

2. Tables and Partitioned Tables

3. Views

4. Integrity Constraints

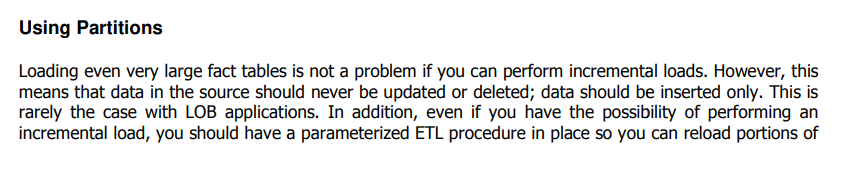
5. Dimensions

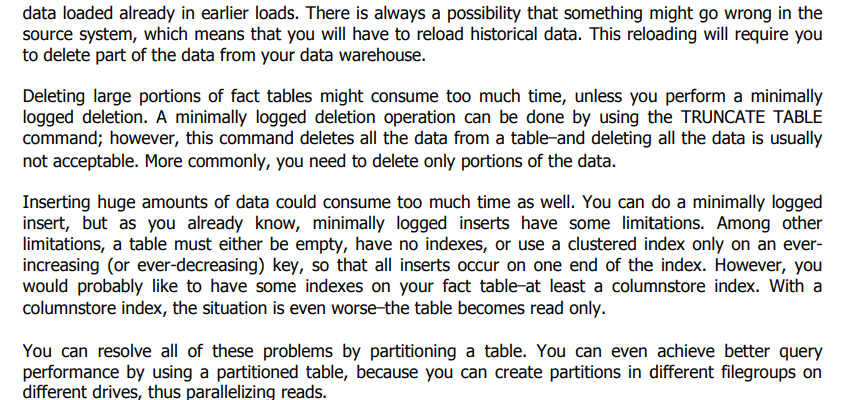
6. Indexes and Partitioned Indexes

7. Materialized Views

**15)Define partitioned table . how its use ful.**

**🡪**A partitioned table is divided into segments, called partitions, that make it easier to manage and query your data. By dividing a large table into smaller partitions, you can improve query performance and control costs by reducing the number of bytes read by a query.

****

****

**16)**

**UNIT-3**

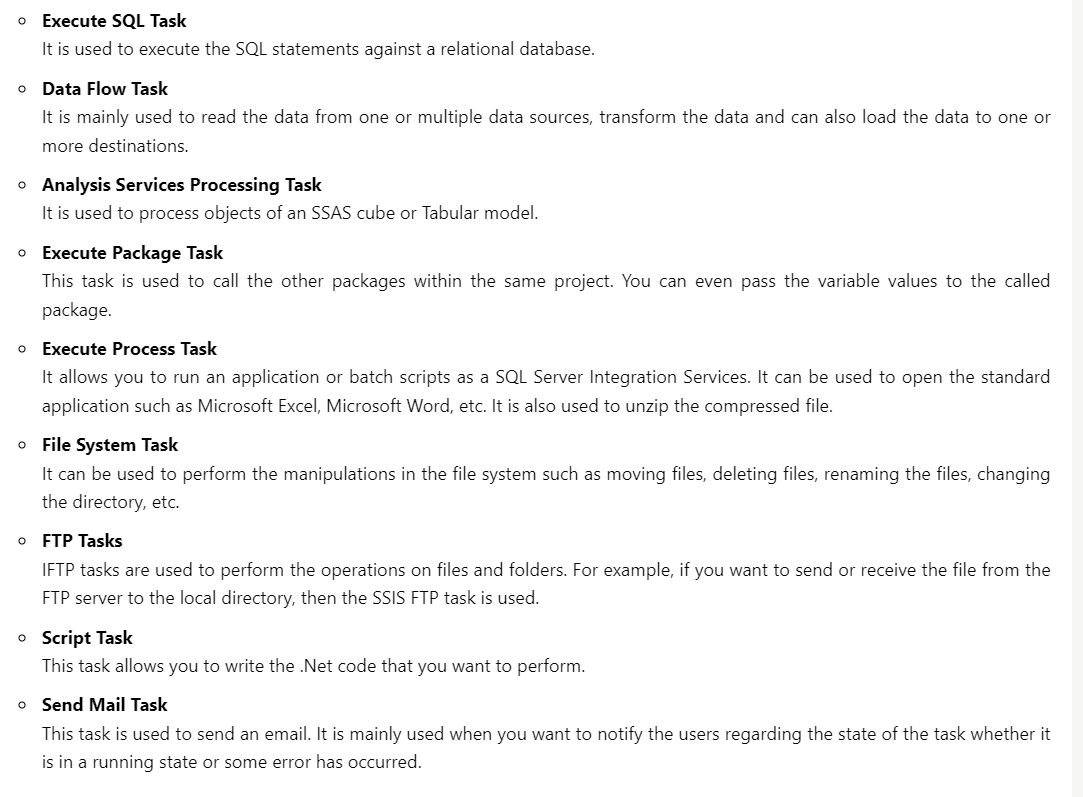
**1)SSIS=** **SQL Server Integration Services**

**2)SSDT=** **SQL Server Data Tools**

**3)ETL=** extract, transform, and load

**4)SMO=** **SQL Server Management Objects**

**5)Explain SSIS task.**

**🡪**

**6)Explain Foreach Container.**

**🡪** Foreach Loop container: You map the enumerated value of the container to a user-defined package variable. The container then uses this variable to dynamically modify the ConnectionString property of the Flat File connection manager and iteratively connect to each flat file in the folder

You can use an SSIS Foreach Loop container to define a control flow task to loop through different types of enumerators, such as files, in a specified folder.

**Add a Foreach Loop container**

1. In **SQL Server Data Tools**, select the **Control Flow** tab.
2. In the **SSIS Toolbox**, expand **Containers**, and then drag a **Foreach Loop Container** onto the design surface of the **Control Flow** tab.
3. Right-click the new **Foreach Loop Container** and select **Edit**.
4. In the **Foreach Loop Editor** dialog, on the **General** page, for **Name**, enter **Foreach File in Folder**. Select **OK**.
5. Right-click the Foreach Loop container, select **Properties**, and in the **Properties** window verify that the **LocaleID** property is set to **English (United States)**.

**Configure the enumerator for the Foreach Loop container**

1Double-click Foreach File in Folder to reopen the Foreach Loop Editor.

2Select Collection.

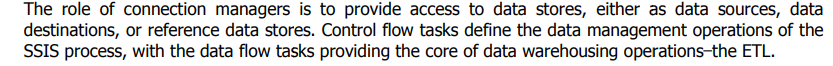
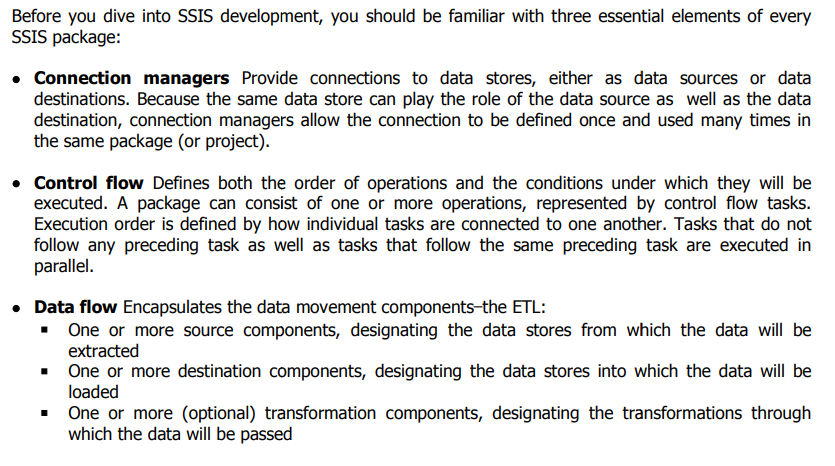
3On the Collection page, select Foreach File Enumerator.

4In the Enumerator configuration group, select Browse.

5In the Browse for Folder dialog box, locate the folder on your machine that contains the Currency\_\*.txt files included with the sample data.

6In the Files box, enter Currency\_\*.txt.

**7)Give the list of creating of flow task.**

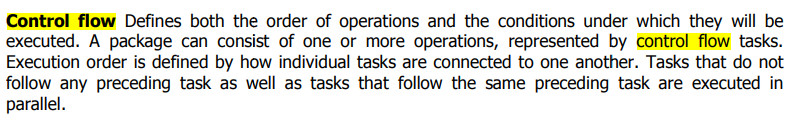
**🡪**

**8)give basic steps for creating package.**

1. **🡪** Defining Business Requirements. ...
2. Setting Up Physical Environments. ...
3. Introducing Data Modelling. ...
4. Choosing Your ETL Solution. ...
5. Online Analytics Processing (OLAP) Cube. ...
6. Creating A Front End. ...
7. Optimising Queries. ...
8. Rolling Out The End Product.

**9)Explain control Flow.**

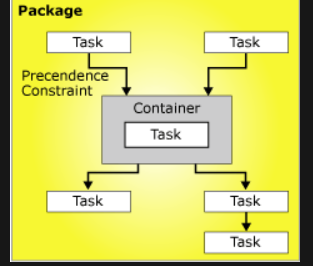
**🡪**

****

A package consists of a control flow and, optionally, one or more data flows. SQL Server Integration Services provides three different types of control flow elements: containers that provide structures in packages, tasks that provide functionality, and precedence constraints that connect the executables, containers, and tasks into an ordered control flow.

For more information, see [Precedence Constraints](https://learn.microsoft.com/en-us/sql/integration-services/control-flow/precedence-constraints?view=sql-server-ver16), [Integration Services Containers](https://learn.microsoft.com/en-us/sql/integration-services/control-flow/integration-services-containers?view=sql-server-ver16), and [Integration Services Tasks](https://learn.microsoft.com/en-us/sql/integration-services/control-flow/integration-services-tasks?view=sql-server-ver16).

The following diagram shows a control flow that has one container and six tasks. Five of the tasks are defined at the package level, and one task is defined at the container level. The task is inside a container.

****

The Integration Services architecture supports the nesting of containers, and a control flow can include multiple levels of nested containers. For example, a package could contain a container such as a Foreach Loop container, which in turn could contain another Foreach Loop container and so on.

Event handlers also have control flows, which are built using the same kinds of control flow elements.

**Control Flow Implementation**

You create the control flow in a package by using the Control Flow tab in SSIS Designer. When the Control Flow tab is active, the Toolbox lists the tasks and containers that you can add to the control flow.

Creating a control flow includes the following tasks:

* Adding containers that implement repeating workflows in a package or divide a control flow into subsets.
* Adding tasks that support data flow, prepare data, perform workflow and business intelligence functions, and implement script.

Integration Services includes a variety of tasks that you can use to create control flow that meets the business requirements of the package. If the package has to work with data, the control flow must include at least one Data Flow task. For example, a package might have to extract data, aggregate data values, and then write the results to a data source. For more information, see [Integration Services Tasks](https://learn.microsoft.com/en-us/sql/integration-services/control-flow/integration-services-tasks?view=sql-server-ver16) and [Add or Delete a Task or a Container in a Control Flow](https://learn.microsoft.com/en-us/sql/integration-services/control-flow/add-or-delete-a-task-or-a-container-in-a-control-flow?view=sql-server-ver16).

* Connecting containers and tasks into an ordered control flow by using precedence constraints.

After you add a task or container to the design surface of the Control Flow tab, SSIS Designer automatically adds a connector to the item. If a package includes two or more items, tasks or containers, you can join them into a control flow by dragging their connectors from one item to another.

The connector between two items represents a precedence constraint. A precedence constraint defines the relationship between the two connected items. It specifies the order in which tasks and containers are executed at run time and the conditions under which tasks and containers run. For example, a precedence constraint can specify that a task must succeed for the next task in the control flow to run. For more information, see [Precedence Constraints](https://learn.microsoft.com/en-us/sql/integration-services/control-flow/precedence-constraints?view=sql-server-ver16).

* Adding connection managers.

Many tasks require a connection to a data source, and you have to add the connection managers that the task requires to the package. Depending on the enumerator type it uses, the Foreach Loop container may also require a connection manager. You can add the connection managers as you construct the control flow item by item or before you start to construct the control flow. For more information, see [Integration Services (SSIS) Connections](https://learn.microsoft.com/en-us/sql/integration-services/connection-manager/integration-services-ssis-connections?view=sql-server-ver16) and [Create Connection Managers](https://learn.microsoft.com/en-us/sql/integration-services/connection-manager/integration-services-ssis-connections?view=sql-server-ver16).

**10)Write a note on ETL Process.**

**🡪**ETL is the most important process in SSIS tool. ETL is used to Extract, Transform, and Load the data into a data warehouse.

ETL is a process responsible for pulling out the data multiple data sources, transforming the data into useful data, and then storing the data into a data warehouse. The data can be in any format xml file, flat file, or any database file.

It also ensures that the data stored in the data warehouse is relevant, accurate, high quality, and useful to the business users.It can be easily accessed so that the data warehouse can be used effectively and efficiently.

It also helps the organization to make data-driven decisions by retrieving the structured and unstructured data from multiple data sources.

**Extract:** In this phase, the original data is checked. It checks the data, whether it consists of any errors or not. It checks for the errors or inconsistency of data by using some artificial intelligence techniques. In short, it verifies whether the quality of the product is met or not.

**Transform:** It is the third phase in ETL. Transformation is the process in which the original format is converted into a required format that you want. Transformation is modelling or changing the data according to the user requirements. The changes can be either change in the number of columns or rows.

**Load :** The fourth phase is Load and index. It loads the data and validates the number of rows that have been processed. Once the loading of data is completed, the indexing is used. Indexing helps you to track the number of rows that are loaded in the data warehouse. Indexing also helps to identify the data, whether it is in the correct format or not.

**11)Explain Sequence container.**

**🡪** In computing, sequence containers refer to a group of container class templates in the standard library of the C++ programming language that implement storage of data elements. Being templates, they can be used to store arbitrary elements, such as integers or custom classes.

there are many benefits of using a Sequence container:

* Disabling groups of tasks to focus package debugging on one subset of the package control flow.
* Managing properties on multiple tasks in one location by setting properties on a Sequence container instead of on the individual tasks.

For example, you can set the Disable property of the Sequence container to True to disable all the tasks and containers in the Sequence container.

* Providing scope for variables that a group of related tasks and containers use.
* Grouping many tasks so you can more easily managed them by collapsing and expanding the Sequence container.

You can also create task groups, which expand and collapse using the Group box. However, the Group box is a design-time feature that has no properties or run-time behavior. For more information, see [Group or Ungroup Components](https://learn.microsoft.com/en-us/sql/integration-services/group-or-ungroup-components?view=sql-server-ver16)

* Set a transaction attribute on the Sequence container to define a transaction for a subset of the package control flow. In this way, you can manage transactions at a more granular level.

For example, if a Sequence container includes two related tasks, one task that deletes data in a table and another task that inserts data into a table, you can configure a transaction to ensure that the delete action is rolled back if the insert action fails.

**12)What is SSIS ? and Discuss SSIS Package.**

**🡪**SSIS stands for SQL Server Integration Services.

It is a component available in the Microsoft SQL Server database software used to perform a wide range of integration tasks.

It is a data warehousing tool used for data extraction, loading the data into another database, transformations such as cleaning, aggregating, merging data, etc.

SSIS tool also contains the graphical tools and window wizards workflow functions such as sending email messages, ftp operations, data sources.

SSIS is used to perform a wide range of transformation and integration tasks. As a whole, the SSIS tool is used in data migration.

SSIS PACKAGE

The Package is a fundamental block where you code in SSIS. Here, code does not mean that you are coding in some programming language; it means the development that you do. The development is done in the SSIS package. SSIS is mainly used for the ETL process, and the ETL process is performed inside the SSIS package.

* **Connections**  
  SSIS package will have some connections, and these connections are used to connect to various data sources.
* **Controlflow elements**  
  SSIS package is composed of two elements, i.e., control flow elements and data flow elements. Control flow elements handle workflows. Workflow means that we are performing some tasks in steps, so the sequence is done through control flow.
* **Data flow elements**  
  The data flow elements perform transformations.

**13)write note on flate file.**

**🡪** A flat file consists of a single table of data. It allows the user to specify data attributes, such as columns and data types table by table, and stores those attributes separate from applications. This type of file is commonly used to import data in data warehousing projects.

**flat file example**

A flat file requiring the employee's identification, address, and expertise is one example. For this project, a flat file database would only contain one table. Each entry in the table describes a record. The columns of the table would include the employee's ID, address, and competence.

**Advantages:**

1. **Data Structure:** Relational databases use structured tables with predefined schemas, making them ideal for complex data models and relationships.
2. **Data Integrity:** Relational databases enforce data integrity constraints, ensuring data accuracy and consistency. This includes referential integrity, unique constraints, and data type validation.
3. **Query Capabilities:** Relational databases offer powerful query languages (e.g., SQL) that enable complex data retrieval and analysis, including joins and aggregations.
4. **Scalability:** Modern relational databases support horizontal and vertical scaling, making them suitable for both small and large-scale applications.
5. **Concurrency Control:** Relational databases support concurrent access by multiple users or applications while maintaining data consistency through locking and transaction management.
6. **Security:** Relational databases provide robust security features, including user authentication, authorization, and access control, to protect data from unauthorized access.

**Disadvantages:**

1. **Complexity:** Relational databases can be complex to set up and maintain, especially for users without prior database experience.
2. **Overhead:** Relational databases have higher overhead in terms of storage and processing compared to flat files.
3. **Learning Curve:** Learning SQL and understanding database design concepts can be challenging for beginners.
4. **Cost:** Relational database systems may involve licensing and operational costs, especially for commercial offerings.

**15)Explain foreach container control with example.**

🡪This container does not execute a fixed number of times like for loop, but the number of executions determined by a collection. This can be for example the number of files in a directory or the number of rows in a table. This makes the container more flexible than a for loop container. Foreach Loop containers performs a loop with a collection. The SSIS Foreach Loop Container is more complicated than the For Loop Container since it has many use cases and requires a more complex configuration.

• Foreach File enumerator: - This is the most common type of enumerator .Performs an action for each file in a directory with a given file extension. You will loop through a list of files and simulate some type of action that has occurred inside the container. Foreach File enumerator to enumerate files in a folder. The enumerator can traverse subfolders. For example, you can read all the files that have the \*.log file name extension in the Windows folder and its subfolders. Note that the order in which the files are retrieve cannot be specify.

• Foreach Item enumerator: - Foreach Item enumerator to enumerate items that are collections. For example, you can enumerate the names of executables and working directories that an Execute Process task uses.

• Foreach SMO enumerator: - Foreach SMO enumerator to enumerate SQL Server Management Objects (SMO) objects. For example, you can enumerate and get a list of the views in a SQL Server database.

• Foreach Nodelist enumerator: - Elements and attributes in an XML document. Foreach Nodelist enumerator to enumerate the result set of an XML Path Language (XPath) expression. For example, this expression enumerates and gets a list of all the authors in the classical period: /authors/author[@period='classical'].

• Foreach ADO enumerator: - Foreach ADO enumerator to enumerate rows in tables. For example, you can get the rows in an ADO recordset. The Recordset destination saves data in memory in a recordset that is stored in a package variable of Object data type. You typically use a Foreach Loop container with the Foreach ADO enumerator to process one row of the recordset at a time. • Foreach ADO.NET Schema: - Tables in a dataset or rows in a table. Foreach ADO.NET Schema Rowset enumerator to enumerate the schema information about a data source. For example, you can enumerate and get a list of the tables in the AdventureWorks2012 SQL Server database.

• Foreach from Variable enumerator: - Foreach from Variable enumerator to enumerate the enumerable object that a specified variable contains. The enumerable object can be an array, an ADO.NET DataTable, an Integration Services enumerator, and so on. For example, you can enumerate the values of an array that contains the name of servers.

**UNIT -4**

**1)DQS=Data Quality Services**

**2)DQAF=Data Quality Assessment Framework**

**3)What is data profiling?**

🡪Data profiling is the process of examining, analyzing, and creating useful summaries of data. The process yields a high-level overview which aids in the discovery of [data quality](https://www.talend.com/resources/what-is-data-quality/) issues, risks, and overall trends. Data profiling produces critical insights into data that companies can then leverage to their advantage.

**4)DQS**

**5)DQKB=Data Quality Knowledge Base**

**6)DQM=**Data quality management

**7)Explain Featurs of DQS.**

**🡪**SQL Server Data Quality Services (DQS) is a knowledge-driven data quality product.

DQS enables you to build a knowledge base and use it to perform a variety of critical data quality tasks, including correction, enrichment, standardization, and de-duplication of your data.

DQS enables you to perform data cleansing by using cloud-based reference data services provided by reference data providers.

DQS also provides you with profiling that is integrated into its data-quality tasks, enabling you to analyze the integrity of your data.

DQS consists of Data Quality Server and Data Quality Client, both of which are installed as part of SQL Server.

Data Quality Server is a SQL Server instance feature that consists of three SQL Server catalogs with data-quality functionality and storage.

Data Quality Client is a SQL Server shared feature that business users, information workers, and IT professionals can use to perform computer-assisted data quality analyses and manage their data quality interactively.

You can also perform data quality processes by using the DQS Cleansing component in Integration Services and Master Data Services (MDS) data quality functionality, both of which are based on DQS.

**8)write note on Data Cleansing.**

**🡪** Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Advantages and benefits of data cleaning

Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. Benefits include:

* Removal of errors when multiple sources of data are at play.
* Fewer errors make for happier clients and less-frustrated employees.
* Ability to map the different functions and what your data is intended to do.
* Monitoring errors and better reporting to see where errors are coming from, making it easier to fix incorrect or corrupt data for future applications.
* Using tools for data cleaning will make for more efficient business practices and quicker decision-making

**How to clean data**

1. Step 1: Remove duplicate or irrelevant observations. Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. ...
2. Step 2: Fix structural errors. ...
3. Step 3: Filter unwanted outliers. ...
4. Step 4: Handle missing data. ...
5. Step 5: Validate and QA.

**9)Explain importance of data quality**.

🡪 Data quality is essential for one main reason: You give customers the best experience when you make decisions using accurate data. A great customer experience leads to happy customers, brand loyalty, and higher revenue for your business. If you’re using poor-quality data, you’re mostly guessing at what your customers want. Worse still, you might be actively doing things your customers dislike.

Collecting trustworthy data and updating existing records gives you a better understanding of your customers. It also lets you keep in contact with them using [verified email addresses](https://www.edq.com/email-verification/), mailing information, and phone numbers. This information helps you market effectively and use resources efficiently.

Maintaining data quality can help you stay ahead of your competitors, too. Reliable data keeps your business agile. You can spot trends and industry changes sooner so you can take advantage of new opportunities or tackle challenges before your competitors.

If you want to preserve good data quality, you must constantly manage it to get the best results. Luckily, modern data tools and platforms from companies like Experian help automate and streamline your day-to-day [data validation](https://www.edq.com/data-quality-platform/data-validation/) and management.

Importance of data quality

Decision-making

Better customer experience

Accuracy and precision

Consistency

Generates revenue

Compliance

Data profiling

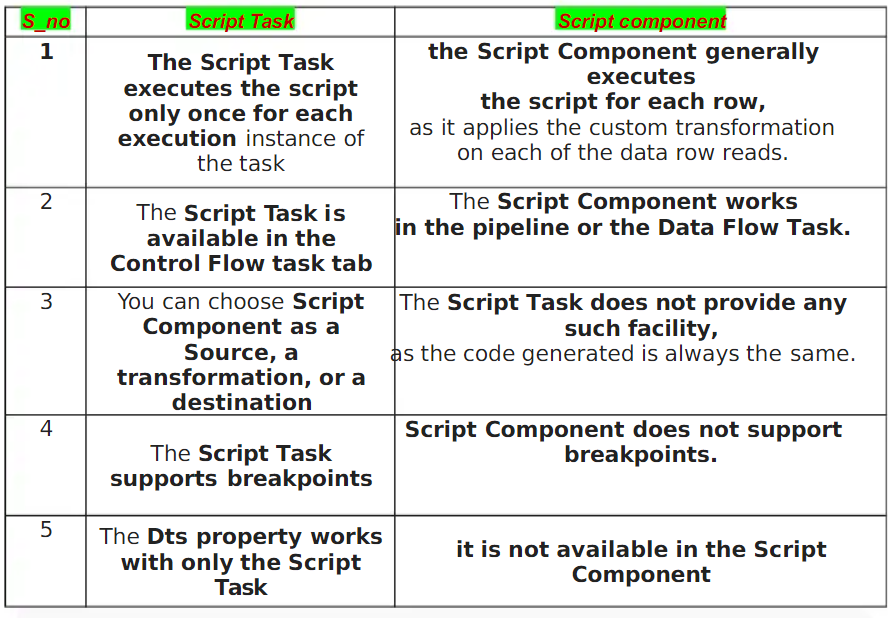
Productivity

Validity

**10)Explain Invalid DQS.**

**🡪**

**11)difference between script task v/s component task.**

**🡪**

**12)Explain data qulity services with match data.**

**🡪**The Data Quality Services (DQS) data matching process enables you to reduce data duplication and improve data accuracy in a data source. Matching analyzes the degree of duplication in all records of a single data source, returning weighted probabilities of a match between each set of records compared. You can then decide which records are matches and take the appropriate action on the source data.

The DQS matching process has the following benefits:

Matching enables you to eliminate differences between data values that should be equal, determining the correct value and reducing the errors that data differences can cause. For example, names and addresses are often the identifying data for a data source, particularly customer data, but the data can become dirty and deteriorate over time. Performing matching to identify and correct these errors can make data use and maintenance much easier.

Matching enables you to ensure that values that are equivalent, but were entered in a different format or style, are rendered uniform.

Matching identifies exact and approximate matches, enabling you to remove duplicate data as you define it. You define the point at which an approximate match is in fact a match. You define which fields are assessed for matching, and which are not.

DQS enables you to create a matching policy using a computer-assisted process, modify it interactively based upon matching results, and add it to a knowledge base that is reusable.

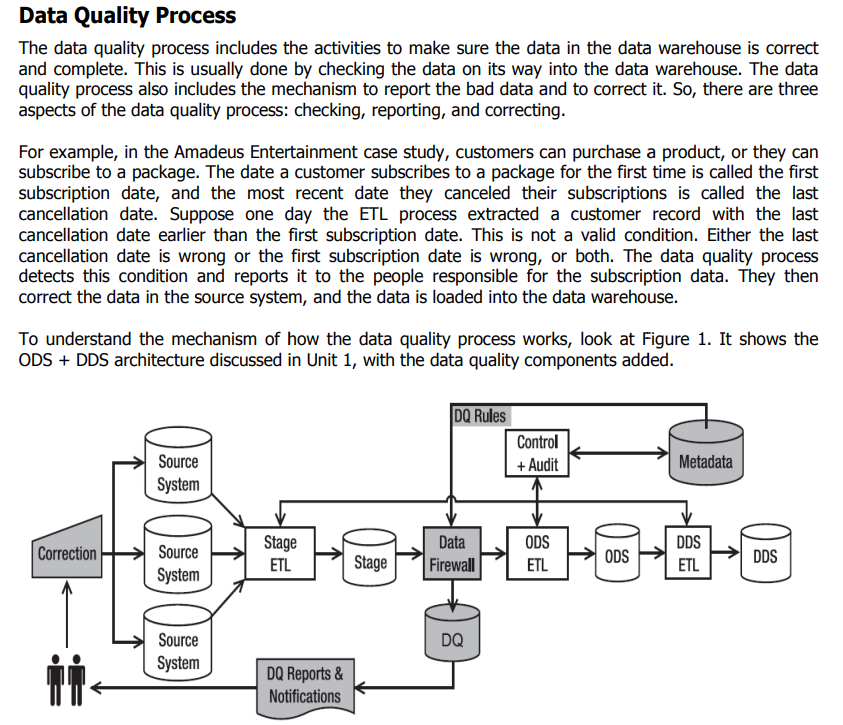
You can re-index data copied from the source to the staging table, or not re-index, depending on the state of the matching policy and the source data. Not re-indexing can improve performance.

HOW tO MATCH DATA:

Create a matching policy in the knowledge base

Perform a de-duplication process in a matching activity that is part of a data quality project.

**13)Explain data qulity process with diagram.**

**🡪**

**14)**

**UNIT-5**

**1)SSAS-**SQL Server Analysis Services

**2)SSRS-**SQL Server Reporting Services

**3)BIML=**Business Intelligence Markup Language

**4)What is data analysis?**

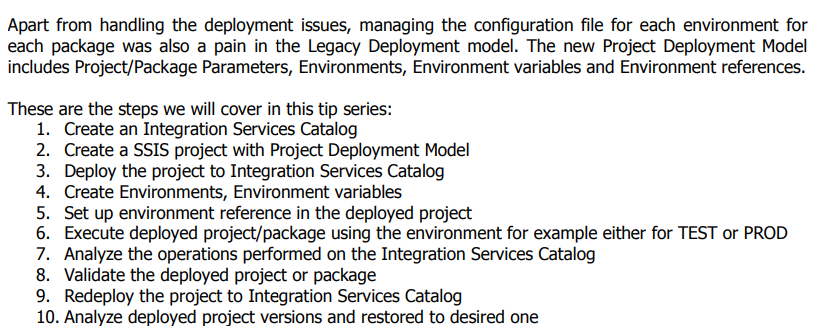
**🡺**a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information; suggesting conclusions; and supporting decision-making

**5)What is deployeement?**

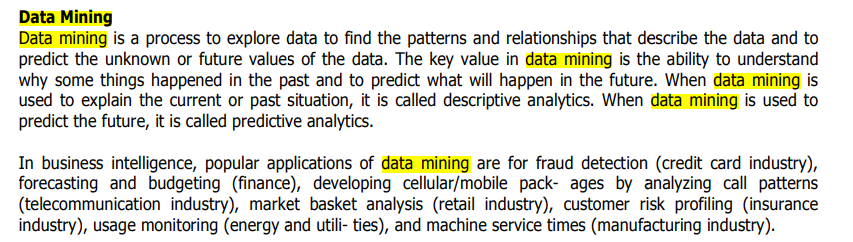
**🡺**  Deployment is the next phase after construction. In the deployment phase, you attend to the last few details, turn the data warehouse on, and let the users reap the benefits. By the time you reach the deployment phase, the majority of the functions are completed. The main concerns in the deployment phase relate to the users getting the training, support, and the hardware and tools they need to get into the warehouse.

To find our place in the whole life cycle of data warehouse development, let us summarize the functions and operations that have been completed up to this point. Here is the list of major activities already completed in the construction phase:

* The infrastructure is in place with the components fully tested.
* The validity of the architecture is already verified.
* The database is defined. Space allocation for the various tables is completed.
* The staging area is fully set up with file allocations.
* The extract, transformation, and all other staging area jobs are tested.



**7)What is data mining?**

**🡪** ****

FEATURES OF DATA MINING:

* It is good with large databases and datasets
* It predicts future results
* It creates actionable insights
* It utilizes the automated discovery of patterns

ADVANTAGES OF DATA MINING:

* Fraud Detection:

It is used to find which insurance claims, phone calls, debit or credit purchases are fraud.

* Trend Analysis:

Existing marketplace trends are analyzed,which provides a strategic benefit as it helps in reduction of costs, as in manufacturing per demand.

* Market Analysis:

It can predict the market and therefore help to make business decisions. For example: it can identify a target market for a retailer, or certain types of products desired by types of customers.

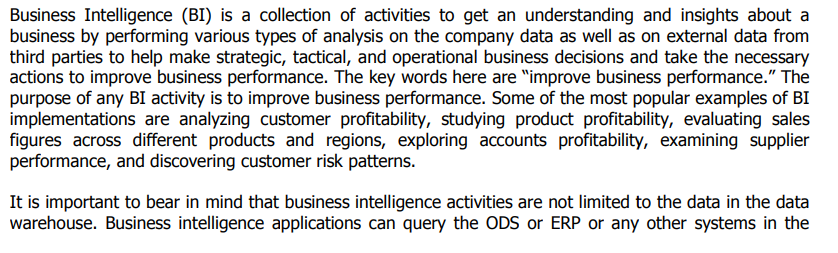
DATA MINING TECHNIQUES:

* Classification:

It is used to fetch the appropriate information from the dataset and to segregate different classes that are present in the dataset. Below are the classification models.

1. K-nearest neighbors
2. Support Vector Machine
3. Gaussian Naïve Bayes, etc.

**8)Explain Bussiness Intaligence**

**🡪**

****

**Adavantages**

* Reduced complexity. ...
* Improved data access. ...
* Enhanced productivity. ...
* Right time, right data. ...
* Increased data quality. ...
* Empowered decision-making. ...
* Lower costs. ...
* Trend insights

**disadvantages**:

* Initial cost
* User resistance
* Data skills gap

benefit:

* Data clarity
* Increased efficiency
* Better customer experience
* Improved employee satisfaction

**methods**

* [**Data mining**](https://www.tableau.com/learn/articles/what-is-data-mining)**:** Using databases, statistics, and [machine learning (ML)](https://www.tableau.com/learn/articles/define-machine-learning) to uncover trends in large datasets
* **Reporting:**Sharing data analysis to stakeholders so they can draw conclusions and make decisions
* **Performance metrics and benchmarking:** Comparing current performance data to historical data to track performance against goals, typically using customized dashboards
* **Descriptive analytics:** Using preliminary data analysis to find out what happened
* **Querying:** Asking the data-specific questions, BI pulling the answers from the data sets
* **Statistical analysis:** Taking the results from descriptive analytics and further exploring the data using statistics such as how this trend happened and why
* [**Data visualization**](https://www.tableau.com/learn/articles/data-visualization)**:** Turning data analysis into visual representations such as charts, graphs, and histograms to more easily consume data
* [**Visual analysis**](https://www.tableau.com/data-insights/reference-library/visual-analytics)**:** Exploring data through visual storytelling to communicate insights on the fly and stay in the flow of analysis
* **Data preparation:** Compiling multiple data sources, identifying the dimensions and measurements, and preparing it for data analysis

**9)Discuus control flow and data flow.**

**🡪**The data flow in the Data Warehouse describes which objects are needed at design time and which objects are needed at runtime to transfer data from a source to BI and cleanse, consolidate and integrate the data so that it can be used for analysis, reporting and possibly for planning.

the metadata description of the source data is modeled with **DataSources**

 A DataSource is a set of fields that are used to extract data of a business unit from a source system and transfer it to the entry layer of the BI system or provide it for direct access.

There is a new object concept available for DataSources in BI. In BI, the DataSource is edited or created independently of 3.x objects on a unified user interface. When the DataSource is activated, the system creates a **PSA** table in the Persistent Staging Area (PSA), the entry layer of BI

**InfoPackage**, you specify the selection parameters for transferring data into the PSA. In the new data flow, InfoPackages are only used to load data into the PSA.

Using the **transformation,** data is copied from a source format to a target format in BI. The transformation process thus allows you to consolidate, cleanse, and integrate data.

**InfoObjects** are the smallest units of BI. You map the information in a structured form that is required for constructing InfoProviders.

**InfoProviders** are persistent data repositories that are used in the layer architecture of the Data Warehouse or in views on data

Use:

Use of the new DataSource permits real-time data acquisition as well as direct access to source systems of type File and DB Connect.

The data transfer process (DTP) makes the transfer processes in the data warehousing layers more transparent. The performance of the transfer processes increases when you optimize parallelization. With the DTP, delta processes can be separated for different targets and filtering options can be used for the persistent objects on different levels. Error handling can also be defined for DataStore objects with the DTP.

The use of transformations simplifies the maintenance of rules for cleansing and consolidating data. Instead of two rules (transfer rules and update rules), as in the past, only the transformation rules are still needed.

## Control flow:

A control flow is a graphical model that sequences data flows and mining flows, integrates external commands, programs, and stored procedures, and provides conditional processing logic for a warehousing application. Deployed warehousing applications consist of control flows that you can run and monitor from the Administration Console.

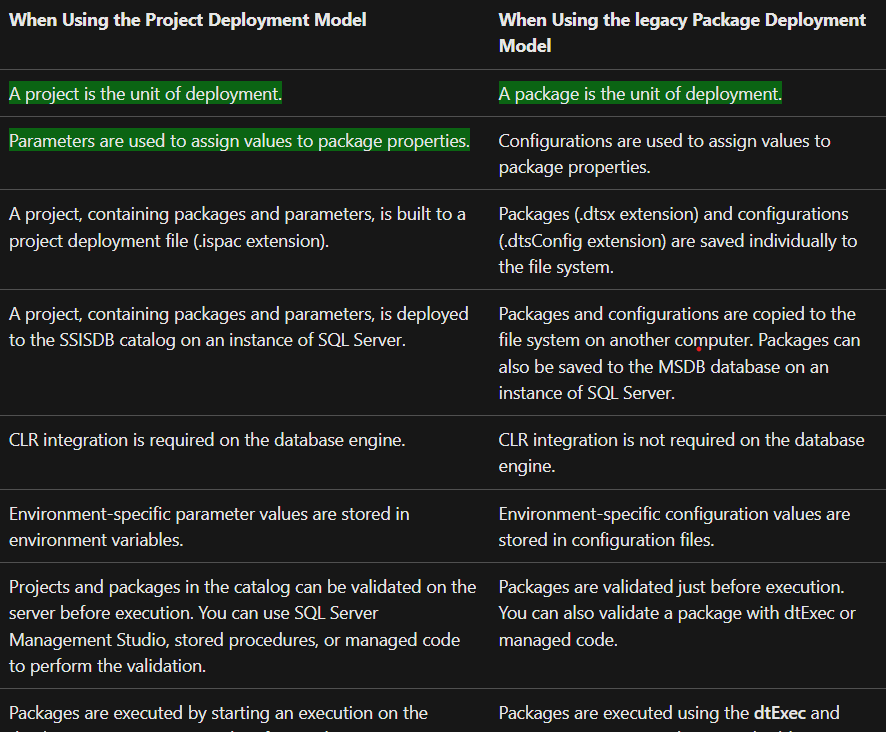
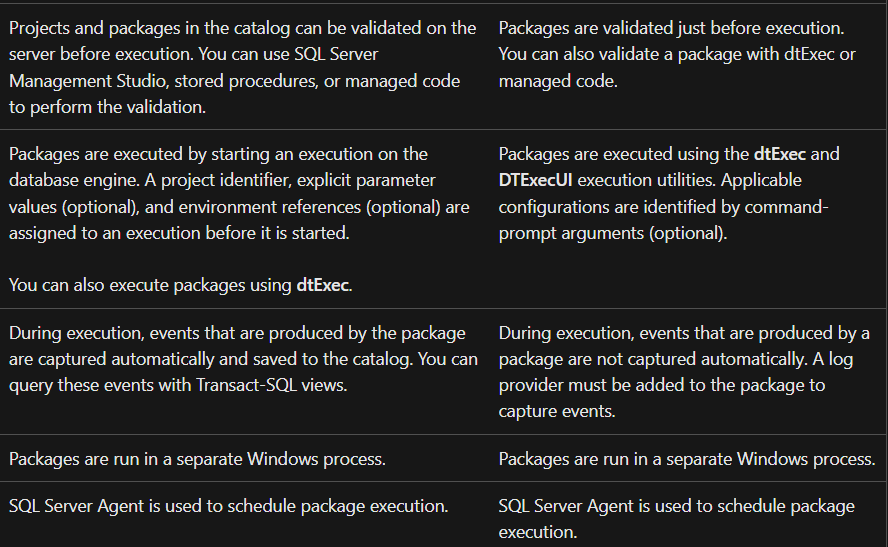
An individual control flow in the runtime environment is equivalent to a control flow in the design-time environment. When you schedule or start a control flow, you are running all of the activities within a particular control flow.

To run or schedule a control flow, you use the Manage Control Flows page of the console.

* [**Control flow instances**](https://www.ibm.com/docs/en/SSEPGG_9.7.0/com.ibm.dwe.admin.doc/cf.instances.html)  
  Control flow instances are specific jobs that are created when you start or schedule control flows from the console. Each instance has a specific name, either user-defined or system-generated, so you can monitor multiple instances of the same control flow when they run.
* [**Instance states**](https://www.ibm.com/docs/en/SSEPGG_9.7.0/com.ibm.dwe.admin.doc/instance.states.html)  
  When you schedule a control flow, each instance goes through a cycle of states: ideally from Scheduled to Running to Successful. To troubleshoot problems, you need to understand these states and what actions you can take for each state.
* [**Missed control flow instance**](https://www.ibm.com/docs/en/SSEPGG_9.7.0/com.ibm.dwe.admin.doc/missed_control_flow_instance.html)  
  You can manage the way in which the Administration Console runs the control flow instances that were missed at the scheduled times.
* [**Control flows and activities**](https://www.ibm.com/docs/en/SSEPGG_9.7.0/com.ibm.dwe.admin.doc/cfs.activities.html)  
  Warehousing applications consist of control flows and activities. Control flows are runtime data warehousing processes.
* [**Activity attributes**](https://www.ibm.com/docs/en/SSEPGG_9.7.0/com.ibm.dwe.admin.doc/activity.attributes.html)  
  Activity attributes are the set of properties that are associated with a particular type of activity. These attributes are treated as system-defined variables with the phase automatically set to run time.
* [**Variables in control flows**](https://www.ibm.com/docs/en/SSEPGG_9.7.0/com.ibm.dwe.admin.doc/variables.cf.html)  
  A variable is a user-defined name that represents data that can be changed within an activity property in a data flow or control flow. By setting variables for certain resources, you can defer the definition of critical properties until a later phase in the life cycle of the warehousing application.
* [**Control flow profiles**](https://www.ibm.com/docs/en/SSEPGG_9.7.0/com.ibm.dwe.admin.doc/copcflprofiles.html)  
  Control flow profiles store variable values that you can use during runtime or execution time of the control flow. You can create multiple control flow profiles in the Administration Console.

**10)Differentiate project deployment model v/s package deployment package.**

**🡺**

**** ****

**11)write note on busssiness intelligence report.**

**🡪** Business Intelligence reporting is broadly defined as the process of using a BI tool to prepare and analyze data to find and share actionable insights. In this way, BI reporting helps users to improve decisions and business performance.

BI reporting—preparing, analyzing, and portraying business metrics—is fundamental to every business. This article will walk you through the basics you need to know. BI Reporting is divided into two categories in business intelligence. Managed reporting occurs when a technical employee such as an IT associate or data analyst prepares the data for non-technical users. Ad-hoc reporting in a BI platform allows non-technical users to create reports from scratch or edit pre-existing reports without having to make requests from IT. Reports allow business users to see data trends over time, slice and dice tables to discover relationships between variables. Smart BI tools have features like [Natural Language Processing (NLP)](https://www.tableau.com/reports/business-intelligence-trends/natural-language) so users can query the data using questions without coding. Reports can take various formats: tables, spreadsheets, and PDFs are all common outputs. Visualizations can include charts, graphs, or histograms. Reports can also be customized dashboards accessed through a web browser. With current technology, BI reports can be automated, and run on a pre-determined schedule.

**How BI reporting works**

BI reporting tools pull and read data from your company’s data sources, on premises and in the cloud. The reporting tool is able to identify measurements such as sales, revenue, inventory counts, etc. and apply dimensions such as date, purchase order, or customer information to create analyses. Users can choose dimensions and measurements using drag-and-drop functionality to create reports. Your team can create tables, histograms or pie charts depending on the data chosen and create forecasting models. They can also create dashboards with these charts and have them update on a schedule or in near real-time. There are a variety of possibilities and capabilities in BI reporting, and Tableau offers [demonstrations and training videos](https://www.tableau.com/learn/get-started) to learn more about how BI reporting works.

**The goal of BI reporting**

The goal of BI reporting is to provide data insights to support decision making. However, analysts and IT developers shouldn’t be the only ones with the ability and responsibility to slice data to uncover insights and trends. Company decision-makers need personalized and organized reports in their preferred formats. They should be able to look at the report, understand the data, draw a conclusion, and make informed decisions. Another goal of BI reporting is to eliminate data silos and create a single source of truth. When teams work with the same datasets and use the same platform for analysis, there are no longer multiple versions of the truth. Data analysis is more accurate, takes less time, and decision makers are able to explore and edit their data in one place. However, these goals cannot be accomplished without a BI reporting tool.

**Features of reporting:**

* **Plug and play functionality.** The right platform fits within your existing data infrastructure and works with your security and governance environment without additional infrastructure.
* **Intuitive user experience.** Non-analysts should be able to easily query and produce the answers they’re looking for.
* **Customizable dashboards.** Platform allows all users to create visually-driven insights at-a-glance.
* **Scalability.** As your business grows, the platform is flexible enough to grow with you.