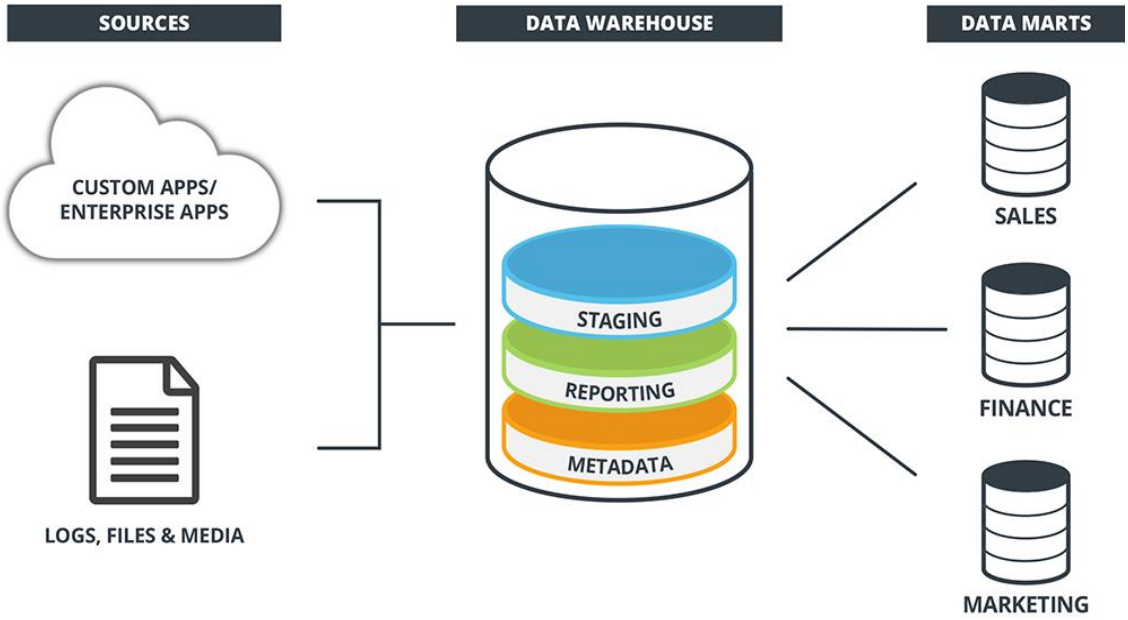
Session 8 & 9 :

Topic 1:Data Ware Housing Concepts and Introduction to Tools

A data warehouse is a data management system that stores current and historical data from multiple sources in a business-friendly manner for easier insights and reporting. Data warehouses are typically used for business intelligence (BI), reporting and data analysis.

Data warehouses make it possible to quickly and easily analyse business data uploaded from operational systems such as point-of-sale systems, inventory management systems, or marketing or sales databases. Data may pass through an operational data store and require data cleansing to ensure data quality before it can be used in the data warehouse for reporting. popular data warehouse tools are **Xplenty, Amazon Redshift, Teradata, Oracle 12c, Informatica, IBM Infosphere, Cloudera, and Panoply**.



The components of data warehouse are as follows:

- **Load manager -** The front component is the load manager. It handles data extraction and warehouse loading. These processes convert data for the data warehouse.
- **Warehouse Manager -** Analyzes data to guarantee consistency, creates indexes and views, generates denormalization and aggregations, and transforms and merges source data.
- **Query Manager -** The backend component is the query manager. It manages user inquiries. These data warehouse components schedule query execution by querying the right tables.

- **End-user access tools -** There are five main categories to end-user access tools: Data query and reporting tools.
    - Application development tools.
    - Executive information systems (EIS) tools.
    - Online analytical processing (OLAP) tools.
    - Data mining tools.

Benefits of data warehouses

- improved data consistency
- Better business decisions
- Easier access to enterprise data for end-users
- Better documentation of data
- Reduced computer costs and higher productivity
- Enabling end-users to ask ad-hoc queries or reports without deterring the performance of operational systems
- Collection of related data from various sources into a place

Challenges with data warehouses

- **No support for unstructured data** like images, text, IoT data, or messaging frameworks like HL7, JSON, and XML. Traditional data warehouses are only capable of storing clean and highly structured data, even though Gartner estimates that up to 80% of an organization's data is unstructured. Organizations that want to use their unstructured data to unlock the power of AI have to look elsewhere.
- **No support for AI and machine learning.** Data warehouses are purpose-built and optimized for common DWH workloads including historical reporting, BI, and querying — they were never designed for or intended to support machine learning workloads.
- **SQL-only** — DWHs typically offer no support for Python or R, the languages of choice for app developers, data scientists, and machine learning engineers.
- **Duplicated data** — Many enterprises have data warehouses and subject-area or (departmental) data marts in addition to a data lake, which results in duplicated data, lots of redundant ETL, and no single source of truth.
- **Tough to keep in sync** - keeping two copies of the data synchronized between the lake and the warehouse adds complexity and fragility that is tough to manage. Data drift can cause inconsistent reporting and faulty analysis.

- **Closed, proprietary formats increase vendor lock-in** — most enterprise data warehouses use their own proprietary data format, rather than formats based on open source and open standards. This increases vendor lock-in, makes it difficult or impossible to analyze your data with other tools, and makes it more difficult to migrate your data.
- **Expensive** — commercial data warehouses charge you for storing your data, and also for analysing it. Storage and compute costs are therefore still tightly coupled together. Separation of compute and storage with a Lakehouse means you can independently scale either as needed.

Types of Data Warehouse: There are three main types of data warehouse.

1. **Enterprise Data Warehouse (EDW)**

This type of warehouse serves as a key or central database that facilitates decision-support services throughout the enterprise. The advantage to this type of warehouse is that it provides access to cross-organizational information, offers a unified approach to data representation, and allows running complex queries.

2. **Operational Data Store (ODS)**

This type of data warehouse refreshes in real-time. It is often preferred for routine activities like storing employee records. It is required when data warehouse systems do not support reporting needs of the business.

3. **Data Mart**

A data mart is a subset of a data warehouse built to maintain a particular department, region, or business unit. Every department of a business has a central repository or data mart to store data. The data from the data mart is stored in the ODS periodically. The ODS then sends the data to the EDW, where it is stored and used.

Topic 2: Data Warehousing Tools

Data warehouse tools are software applications or platforms designed to facilitate the process of collecting, storing, managing, and analyzing large volumes of data from various sources, such as databases, spreadsheets, cloud services, and even IoT devices.

These tools help to collect, read, write and transfer data from various sources. They are designed to support operations like data sorting, filtering, merging, etc.

Data Warehouse tools do?

1. **Data extraction:** The first and foremost thing that a data warehousing tool does is that it extracts information from all operational sources of an organization such as customer databases.
2. **Data transformation:** The extracted information is then cleaned and validated, so that it is fit to be sent into a data warehouse. Data warehouse tools offer a range of transformation capabilities to clean, standardize, and enrich data.
3. **Data loading:** Next you can load the data in the destination. You can opt for any loading strategy, such as full loads, incremental loads, and real-time streaming, depending on what best suits your needs.
4. **Data Modelling:** Once your data is in the data warehouse, you can use the tools features to define the relationships in your data. You can either use star schema or snowflake schema, which consists of fact tables (containing measures) and dimension tables (containing attributes). For example, fact can be "Sales Revenue" that represents the quantitative data related to each sale transaction, such as the total amount of money generated by each sale. On the other hand, "Product" can be a dimension that provides details about the products sold. It includes attributes like "Product Name," "Product Category," "Manufacturer," and so on.
5. **Query and Analysis:** These tools provide query and reporting capabilities that allow you to extract insights from the data warehouse. You can write SQL queries or use graphical interfaces to create reports and visualizations for analysis.

- **Tools used in Data Warehousing**

1.Astera Data Warehouse Builder

ADWB is an agile meta driven data warehouse tool that simplifies and automates all data warehousing processes, from design and development all the way to deploying and publishing data, giving you a single platform to build on-premises

2. Snowflake

Snowflake is a cloud-based data warehousing platform that offers a fully managed and scalable solution for data storage, processing, and analysis. It is designed to address the challenges of traditional on-premises data warehousing by providing a modern and cloud-native architecture.

3. SAP Datawarehouse Cloud

SAP Data Warehouse Cloud is a cloud-based data warehousing solution developed by SAP. It is designed to provide organizations with a modern, scalable, and integrated platform for data storage, data modelling, data integration, and analytics. Here are key features and aspects of SAP Data Warehouse Cloud:

4. Oracle Exadata

Oracle Autonomous Data Warehouse (ADW) is a cloud-based data warehousing service offered by Oracle Corporation. It is designed to simplify data management and analytics tasks by automating many of the traditionally complex and time-consuming processes associated with data warehousing. Here are key aspects and features of Oracle Autonomous Data Warehouse:

- It supports data integration and ETL (Extract, Transform, Load) processes with built-in features for data loading and transformation.
- ADW supports various data types and models, including relational, JSON, spatial, and graph data, making it versatile for diverse analytical requirements.

5. Panoply

Panoply is a managed ELT and a cloud data warehouse platform that allows users to set up a data warehouse architecture. The cloud data warehouse eliminates the need for you to set up and maintain your own on-premises data warehouse, saving time and resources.

6. Teradata Vantage

Teradata Vantage is a data warehousing and analytics platform designed to handle large volumes of data and support complex analytical workloads. The platform uses SQL as its primary query language, which means it is mostly meant for users with SQL skills. Here are some key aspects of Teradata Vantage for data warehousing:

- Various sources, including data warehouses, data lakes, on-premises systems, and cloud platforms.
- Built-in analytics functions and supports integration with popular data science and machine learning tools.
- Workload management features to ensure that different types of queries and analytics workloads are appropriately prioritized and allocated resources.

7. Microsoft Azure

Microsoft Azure also offers data warehousing capabilities. If you have data stored in Azure Blob storage or in a data lake, you can introduce analytical capabilities using Azure Synapse, or with Azure HDInsight.

The platform offers built-in analytics capabilities, including integration with Azure Machine Learning and Power BI.

- It comes with MPP architecture, which distributes data and queries across multiple nodes, and allows you to process large data sets quickly and efficiently.

8. Hevo Data

Hevo, is a cloud-based data integration platform designed to streamline the process of collecting, transforming, and loading (ETL) data into data warehouses and other destinations. While it's not a data warehousing tool itself, it facilitates data ingestion and integration. Here are some key features and aspects of Hevo for data warehousing:

- A wide range of pre-built connectors and integrations to collect data from various sources, including databases, cloud applications, file systems, and more.
- Visual data transformation interface that enables you to clean, enrich, and transform data as it flows into the data warehouse.

9.Amazon Redshift:

Amazon Redshift is a cloud-based fully managed petabytes-scale data warehouse By the Amazon Company. It starts with just a few hundred gigabytes of data and scales to petabytes or more. This enables the use of data to accumulate new insights for businesses and customers. It is a relational database management system (RDBMS) therefore it is compatible with other RDBMS applications.

10. Google BigQuery:

BigQuery is a serverless data warehouse that allows scalable analysis over petabytes of data. It's a Platform as a Service that supports querying with the help of ANSI SQL. It additionally has inbuilt machine learning capabilities. BigQuery was declared in 2010 and made available for use there in 2011. Google BigQuery is a cloud-based big data analytics web service to process very huge amount of read-only data sets

11. Amazon DynamoDB:

Amazon DynamoDB is a fully managed proprietary NoSQL data warehouse service that supports key-value and document data structures and is obtainable by Amazon.com as

a part of the Amazon Web Services portfolio. DynamoDB has an identical data model and encompasses a completely different underlying implementation.

12 PostgreSQL:

It is an extremely stable database management system, backed by over twenty years of community development that has contributed to its high levels of resilience, integrity, and correctness. PostgreSQL is employed because the primary data store or data warehouse for several web, mobile, geospatial, and analytics applications. SQL Server is a database management system that is especially used for e-commerce and providing different data warehousing solutions.

13. Teradata:

Teradata is one of the admired Relational Database Management systems. It is appropriate for building big data warehousing applications. Teradata accomplishes this with the help of parallelism. Teradata database system is built on Massively Parallel Processing (MPP) architecture. The Teradata system primarily splits the work among its processes and runs them in parallel to reduce workload and also makes sure that the task is accomplished quickly and successfully. Teradata provides real-time, intelligent answers by processing 100% of the appropriate data, despite the volume of the query.

14.Cloudera:

Cloudera Data Warehousing Platform is that the industry's 1st enterprise data cloud i.e. multi-functional analytics based on a platform that eliminates silos and speeds up the invention of data-driven insights. It applies consistent security, governance, and metadata in shared data cases. Cloudera's trendy Data Warehouse powers superior bismuth and data deposit in each on-premises deployment and as a cloud service.

Important Features that Data Warehouse Tools Should Have:

1. **Data Cleansing**

   Many companies use data warehousing to leverage historical data for critical business decisions. Hence, ensuring that only high-quality data is loaded into a data warehouse through data processing is essential. This can be done by making data cleansing a part of the data warehousing process, which can help detect and remove invalid, incomplete, or outdated records from the source datasets.

2. **Data Transformation and Loading**

   Data transformation involves modifying data into a compatible format with the target system, such as a database, to simplify data loading.

3. **Data Governance and Metadata Management**

   Data Governance and Metadata Management play critical roles in a data warehouse tool. Data governance ensures the integrity, compliance, and effective management of data through policies, processes, and controls.

4. **Business Intelligence and Data Analysis**

   Data warehousing and Business Intelligence (BI) are two distinct but closely interlinked technologies that assist an enterprise in making informed decisions. Organizations have much information in raw form in the digital era, generally stored in a data warehouse. It is crucial for data warehouse analytics tools to have BI functionality to aid data retrieval as it helps generate business insights.

<u>Topic: Different algorithms related to Data Warehouse</u>

Data ware house is used to keep data and process it to make the decision. There are following types of algorithms that can be used for the mining form data ware house.

1. C4.5 Algorithm

C4.5 is one of the top data mining algorithms and was developed by Ross Quinlan. C4.5 is used to generate a classifier in the form of a decision tree from a set of data that has already been classified. Classifier here refers to a data mining tool that takes data that we need to classify and tries to predict the class of new data.

2. K-mean Algorithm

One of the most common clustering algorithms, k-means works by creating a k number of groups from a set of objects based on the similarity between objects. It may not be guaranteed that group members will be exactly similar, but group members will be more similar as compared to non-group members. As per standard implementations, k-means is an unsupervised learning algorithm as it learns the cluster on its own without any ex **3. 3.Support Vector Machines**

In terms of tasks, Support vector machine (SVM) works similar to C4.5 algorithm except that SVM doesn't use any decision trees at all. SVM learns the datasets and defines a hyperplane to classify data into two classes.

4. Apriori Algorithm

Apriori algorithm works by learning association rules. Association rules are a data mining technique that is used for learning correlations between variables in a database. Once the association rules are learned, it is applied to a database containing a large number of transactions. Apriori algorithm is used for discovering interesting patterns and mutual relationships and hence is treated as an unsupervised learning approach.

The entire Apriori algorithm is summarized into 3 steps:

- Join: Calculates the frequency of one item set.
- Prune: The itemsets that fulfill the target support and confidence proceed to the next iteration for two item sets.
- Repeat: The above two steps are iterated for each item set level until you sort the scope's required size.

5. Expectation-Maximization Algorithm

EM algorithm work in iterations to optimize the chances of seeing observed data. Next, it estimates the parameters of the statistical model with unobserved variables, thereby generating some observed data. Expectation-Maximization (EM) algorithm is again unsupervised learning since we are using it without providing any labelled class information.

The EM algorithm is unsupervised since it doesn't provide labelled class data. It develops a math model that predicts how the newly collected data will be distributed depending on the given data set.

6. PageRank Algorithm

PageRank is commonly used by search engines like Google. It is a link analysis algorithm that determines the relative importance of an object linked within a network of objects. Link analysis is a type of network analysis that explores the associations among objects. Google search uses this algorithm by understanding the backlinks between web pages.

7. Naive Bayes Algorithm

Naive Bayes is not a single algorithm though it can be seen working efficiently as a single algorithm. Naive Bayes is a bunch of classification algorithms put together. The assumption used by the family of algorithms is that every feature of the data being classified is independent of all other features that are given in the class. Naive Bayes is provided with a labelled training dataset to construct the tables. So it is treated as a supervised learning algorithm.

8.CART Algorithm

CART stands for classification and regression trees. It is a decision tree learning algorithm that gives either regression or classification trees as an output. In CART, the decision tree nodes will have precisely 2 branches. Just like C4.5, CART is also a classifier. The regression or classification tree model is constructed by using labelled training dataset provided by the user. Hence it is treated as a supervised learning technique.

Topic: Importance of Data ware house and its applications

Data warehouse platforms are different from operational databases because they store historical information, making it easier for business leaders to analyse data over a specific period of time. Data warehouse platforms also sort data based on different subject matter, such as customers, products or business activities.

The Importance of Data Warehousing:

Data warehousing is vital to a business. It helps them store essential data from their past to current activities.

1. Accessible Data to Boost Efficiency

A business's data serves as the foundation of its products and services. Therefore, a business needs to access data right away. A data warehouse allows you to access essential files from different sources simultaneously.

2. Ensures Data Quality and Consistency

A data warehouse collects information from many sources. It converts different data into a uniform format, following the data analytics requirements. Data warehousing ensures production standards and quality information.

3. Retrieving Historical Data

Historical data and files are essential in a business. It allows you and other stakeholders to review past activities of your organization. You can get ideas from this information to improve your business.

4. Helps in Decision-Making

Organizations use business intelligence tools and systems for their data. It's the process of gathering, keeping, and interpreting historical and current data.

5. Generates Revenue

Using a data warehouse is more of an investment for a business. It helps the organization generate more revenue and lessen costs.

A better decision strengthens the company, resulting in producing high revenue. It prevents them from spending more, too. It allows them to see things they must do and not to do in growing the business.

6. Secure Data

Data warehousing allows access to information for all departments. However, you can limit the accessibility of this information for safety reasons. There are several ways you can protect your data from unauthorized personnel.

Applications of Data Warehouse

1. Banking Industry

Bankers can better manage all of their available resources with the right Data Warehousing solution. They can better analyze consumer data, government regulations, and market trends to facilitate better decision-making.

2. Finance Industry

Similar to the applications seen in banking, they mainly revolve around evaluation and trends of customer expenses which aid in maximizing the profits earned by their clients.

3. Consumer Goods Industry

They are used to predict consumer trends, inventory management, and market and advertising research. An in-depth analysis of sales and production is also carried out. Apart from these, information is exchanged between business partners and clientele.

4. Government and Education

The federal government uses the warehouses for compliance research, while the state government uses them for human resources services such as recruitment and accounting services such as payroll management.

5. Healthcare

The healthcare industry is another important application of data warehouses. All clinical, financial, and personnel data is saved in the warehouse, and analysis is performed to provide useful insights into allocating resources effectively.

6. Hospitality Industry

Hotel and restaurant services, automobile rental services, and vacation home services dominate this market. They employ warehousing services to create and assess their ad and promotion programs, which target customers based on their feedback and travel patterns.

7. Insurance

In the insurance industry, there's a saying that goes, "If it ain't broke, don't. "Insurance is not something that can be purchased. It can only be purchased. "Apart from keeping track of existing participants, the warehouses are largely used to evaluate data patterns and customer trends. Warehouses can also be used to create customized consumer offers and promotions.

8. Manufacturing and Distribution Industry

Manufacturing and distribution companies may gather all of their data under one roof with the help of a good data warehousing system, predict market changes, analyze the latest trends, view development areas, and finally make result-driven decisions.

9. The Retailers

Retailers act as go-betweens for producers and customers. In order to ensure their continuous presence on the market, they must keep records of both parties.

10. Services Sector

In the service industry, data warehouses are used to keep track of financial records, revenue trends, consumer profiling, resource management, and human resources.

11. Transportation Industry

Client data is recorded in data warehouses in the transportation industry, allowing traders to experiment with target marketing, where marketing campaigns are created with the needs of the customer in mind.

How data warehousing works?

Data warehousing works by integrating disparate datasets and transforming them into multidimensional schemas for easy analysis and interpretation. A data warehouse typically has three layers:

- A bottom layer that houses a relational database
- A middle layer with online analytical processing (OLAP) servers to analyse the stored data
- A top layer that comprises the front-end client interface

Key Components of a data warehouse

Data warehousing is a complex process involving multiple tools and processes. Most of these tools belong to one of four key components:

Central database: Traditional data warehouses were housed on premise or in the cloud. But due to the need for high-speed performance, in-memory databases are now becoming popular.

ETL tools: ETL (extract, transform, load) tools extract data from source systems, combine and process it, and then load the enriched data into the target database.

Metadata: Metadata describes the source data, adds context, and makes it intelligible.

Access tools: These include tools for query and reporting, application development, analytical processing, and data mining. Data access tools allow authorised users to interact with the data through a user-friendly interface.