**BYTEWISE   
WEEK 1  
  
TASK 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **DATABASE** | **DATA WAREHOUSE** | **DATA LAKE** | **BIG DATA** |
| **Definition** | Record of data | Analytical processing of data | Record of different kind of data used for ml dl or analytical | Large number of data 5V’s(volume, velocity, variety, veracity, value) |
| **Data detail** | fresh,detailed | refreshed,Summarize form(ETL) | Fresh data |  |
| **scheme** | flexible | rigid |  |  |
|  | OLTP | OLAP |  | hadoop |

**TASK 2**

|  |  |
| --- | --- |
| **Concept** | **Description** |
| **Data Mart** | A subset of a larger data warehouse that is designed to serve a particular business unit or department. It contains a smaller amount of data and is optimized for specific use cases. |
| **Data Lakehouse** | A unified data platform that combines the benefits of a data warehouse and a data lake. It allows for both structured and unstructured data to be stored, managed, and analyzed in a single platform. |
| **Data Mesh** | A distributed architecture that emphasizes domain-driven design and decentralized data ownership. It aims to improve scalability and agility by allowing individual teams to manage their own data products. |
| **DWH vs DATA LAKE** | A data warehouse is a structured repository for data that has been cleansed, transformed, and organized for efficient querying and analysis. A data lake, on the other hand, is a centralized repository for storing raw and unstructured data. The main difference is that a data warehouse is optimized for querying and analysis, while a data lake is optimized for storing and processing large amounts of data. |
| **OLTP vs OLAP** | OLTP stands for Online Transaction Processing, which is a system that manages transactions and queries for a specific business application. OLAP stands for Online Analytical Processing, which is a system that supports complex analysis of data from multiple sources. The main difference is that OLTP is optimized for transaction processing and real-time data access, while OLAP is optimized for data analysis and reporting. |

**Q/A**

**- Can a database be used as DWH?**

Yes, a database can be used as a data warehouse. However, a data warehouse typically involves more than just a database. It also involves a data integration process, data transformation, and data modeling to create a schema that is optimized for reporting and analysis. So, while a database can be a component of a data warehouse, it is not sufficient on its own to constitute a complete data warehouse solution.

**- Major differences between structured and Un-structured data.**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Structured Data** | **Unstructured Data** |
| **Definition** | Data that has a defined schema and format. | Data that has no predefined structure or format. |
| **Examples** | Relational databases, spreadsheets | Emails, social media posts, images, videos, audio recordings |
| **Data organization** | Organized in tables and follows a specific schema | Not organized and has no specific schema |
| **Querying** | Easy to query and analyze using SQL or other query languages | Difficult to query and analyze without preprocessing and natural language processing |
| **Storage** | Typically stored in a structured format in databases or spreadsheets | Can be stored in various formats such as documents, images, or video files |
| **Data processing** | Processing is straightforward and can be automated using ETL tools | Processing requires advanced techniques such as natural language processing and machine learning |
| **Analysis** | Easy to analyze using statistical and analytical tools | Requires advanced analytics tools and techniques such as machine learning and deep learning |

**- What are the duties of a data engineer? (high-level)**

1. Design, build, and maintain data pipelines and data infrastructure for collecting, storing, and processing large volumes of data.
2. Develop and implement ETL (extract, transform, load) processes to move data from various sources into data storage systems.
3. Ensure data quality, accuracy, consistency, and integrity by performing data cleaning, validation, and testing.
4. Optimize database and data processing systems for performance, scalability, and reliability.
5. Collaborate with data scientists, data analysts, and other stakeholders to understand their data requirements and design data solutions that meet their needs.

**TASK 4**

[**Q1,2,4**](https://youtu.be/8MfyfW9CeKA)

**What is ETL? in detail**

ETL stands for Extract, Transform, Load. It is a process used to extract data from various sources, transform it into a desired format, and load it into a target database or data warehouse. The ETL process involves three main steps:

**Extraction**: In this step, data is extracted from various sources such as databases, flat files, or APIs. The data is usually extracted into a staging area where it can be processed and transformed.

**Transformation**: Once the data is extracted, it needs to be transformed into a format that is compatible with the target system. This step involves cleaning, filtering, aggregating, and transforming the data to ensure that it is accurate and consistent.

**Loading**: In the final step, the transformed data is loaded into the target database or data warehouse. This step may involve mapping the data to the target schema and performing any necessary data validation or error handling.

The ETL process is an essential component of data integration and plays a critical role in ensuring that data is accurate, consistent, and available for analysis.

**What is ELT? in detail.**

ELT stands for Extract, Load, Transform. It is a process similar to ETL, but with a different sequence of steps. In ELT, data is extracted from various sources and loaded directly into the target database or data warehouse. The transformation step is then performed on the loaded data using the processing power of the target system.

The ELT process involves three main steps:

**Extraction**: In this step, data is extracted from various sources such as databases, flat files, or APIs.

**Loading**: Once the data is extracted, it is loaded directly into the target database or data warehouse.

**Transformation**: In the final step, the data is transformed using the processing power of the target system. This may involve running complex queries, joining tables, or performing other data processing tasks.

The main advantage of ELT over ETL is that it can be faster and more efficient since the processing power of the target system is utilized for the transformation step. ELT is often used in situations where the target system has a large amount of processing power and can handle complex data transformations.

**3 Tier Architecture in DE**

The three-tier architecture is a software architecture pattern used in data engineering to separate the user interface, application logic, and database layers of an application. It consists of three layers:

**Presentation layer**: This layer is responsible for presenting data to the user and receiving input from the user. It includes user interface components such as forms, reports, and dashboards.

**Application layer**: This layer contains the application logic and business rules that process the data. It includes components such as application servers, APIs, and web services.

**Data layer**: This layer contains the data storage and management components such as databases, data warehouses, and data lakes.

The three-tier architecture provides several benefits, including:

Scalability: Each layer can be scaled independently to handle increased load or demand.

Separation of concerns: Each layer has a specific role and responsibility, which makes it easier to develop, test, and maintain the application.

Security: The separation of layers can help to enforce security policies and prevent unauthorized access to sensitive data.

**ETL Tools (any 3)**

There are many ETL tools available, each with their own strengths and weaknesses. Here are three popular ETL tools:

* **Apache NiFi**: A powerful, open-source tool that provides a web-based interface for designing and managing data flows. It supports a wide range of data sources and destinations, including Hadoop, Kafka, and databases.
* **Talend**: A comprehensive ETL tool that provides a wide range of data integration
* **Hevo Data**
* **Luigi**
* **Airflow**
* **Blendo**

**TASK 5**

**What is Historical Load**

Historical load is a process of extracting and loading all historical data from a source system to a target system or data warehouse. It is typically performed as a one-time event to capture the entire history of the data. Historical load is usually necessary when a new data warehouse is created or when existing data is migrated to a new system. The process can be time-consuming and resource-intensive, but it ensures that all historical data is available for analysis.

**What is Full Load**

Full load is a process of extracting and loading all data from a source system to a target system or data warehouse. It is similar to historical load, but it is performed on a regular basis to ensure that the target system is up-to-date with the latest data. Full load typically involves truncating the target tables and reloading all data from scratch. This process can be time-consuming and resource-intensive, but it ensures that the target system is completely refreshed with the latest data.

**What is Incremental Load**

Incremental load is a process of extracting and loading only the new or updated data from a source system to a target system or data warehouse. It is typically performed on a regular basis, such as daily or hourly, to ensure that the target system is up-to-date with the latest data. Incremental load involves comparing the source and target systems to identify new or updated data, and then loading only that data into the target system. This process is more efficient than full load since it only processes the changed data, which can significantly reduce the time and resources required for data integration. Incremental load is often used in conjunction with full load to ensure that the target system is periodically refreshed with all data.