BASICS OF DATA ENGINEERING [1]

SUBMITTED BY: MUHAMMAD FAHAD

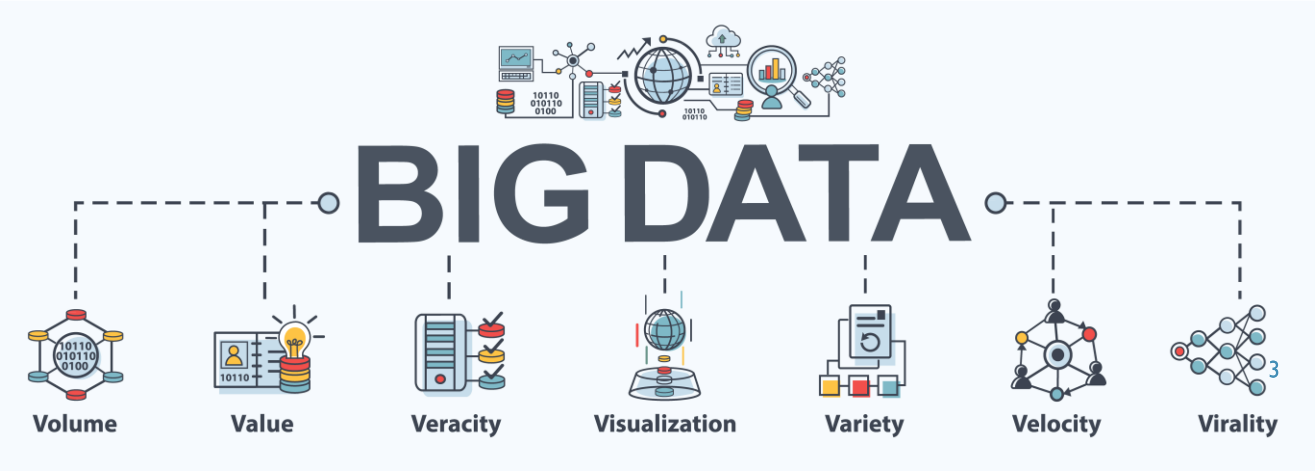
TASKS DAY 1 [14/03/2023]

Task # 1: Please familiarize yourself with the basics of Data Engineering which include the following topics :-

- Big Data
- Data Lake
- Database
- Data Warehouse

BIG DATA

- Big Data is the term used to describe extremely big and complex datasets that require advanced processing techniques in order to successfully analyze and extract knowledge from the data. This data can come from a wide range of sources, including social media, online transactions, sensors and other electronic devices. Big data can be structured, unstructured or semi-structured and ranges from terabytes (TB) to petabytes (PT) or even more.



SUBMITTED BY: MUHAMMAD FAHAD (BW-DE)

TYPES OF DATA

Data can be categorized into 3 types based on their characteristics and properties:

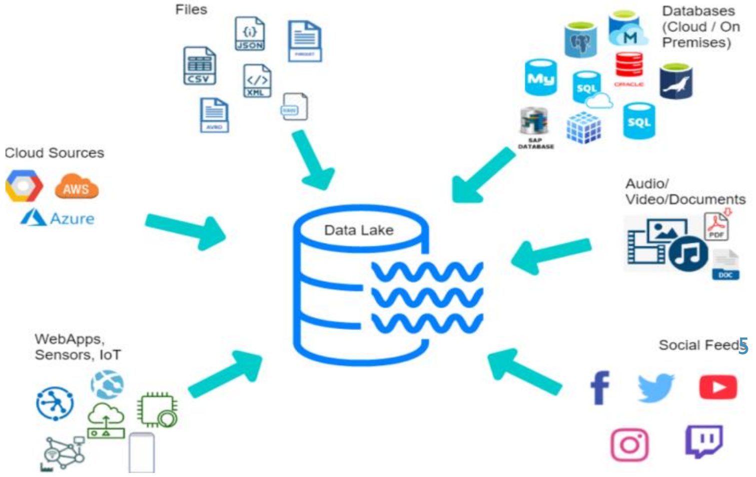
i) Structured Data: Structured data is referred to the data that is highly organized and formatted in a specific way that makes it easy to process and analyze data. Structured data is typical stored in Databases, Spreadsheets or any other formats that allow efficient storing and querying of data. For Example, Customers Data, Financial Transactions, Inventory data etc.

ii) Unstructured Data: Unstructured data is referred to the data that does not have a specific format or structure which makes it difficult to analyze and process the data with traditional data processing tools. A typical example of unstructured data is heterogenous data source containing a combination of simple text files, images, video etc.

iii) Semi-Structured Data: Semi structured data refers to data that have some structure, but not organized as structured data. Semi structured data contain tags, metadata or other markers that help to organize and structure the data. For Example, XML files, JSON files and log files.

DATA LAKE

A Data Lake is a centralized repository that allows to store all structured, unstructured and unstructured data at any scale. It can store data in its native format and process any variety of it, ignoring the size limits. It provides secure and scalable platform that allows enterprises to ingest any data from any system at any speed even if the data comes from cloud or edge-computing systems. It stores any type or volume of data in full fidelity and process data in real time or batch mode.



BENEFITS OF DATA LAKE

- Scalability: Data lakes can scale to handle large volumes of data, making them ideal for organizations that deal with large and diverse datasets.

- Flexibility: Data lakes can store data in its raw form, enabling organizations to use a wide range of tools and technologies to analyze and process the data.

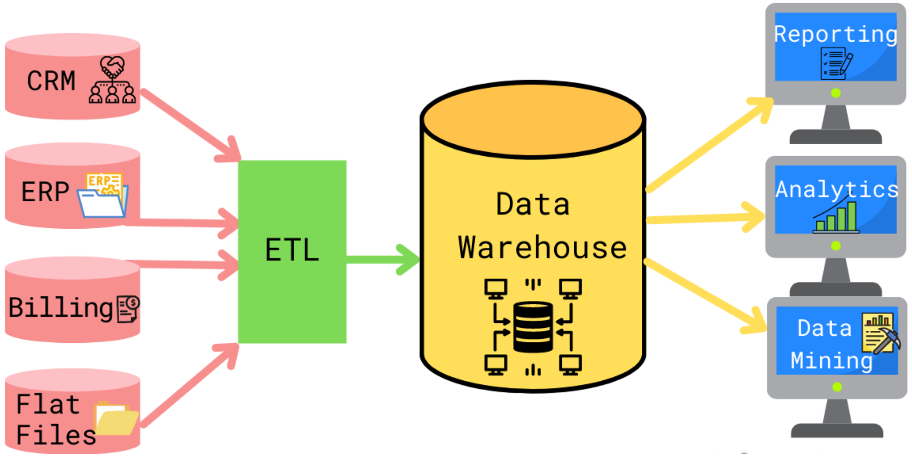
- Cost-Effective: Data lakes can be less expensive to build and maintain than traditional data warehousing solutions, as they do not require upfront data modeling and schema design.

- Speed: Data lakes can provide faster access to data, as data is stored in a single location and does not need to be moved between different systems.

- Integration: Data lakes can integrate with a wide range of data sources and tools, making it easier for organizations to use and analyze data from different sources.

DATA WAREHOUSE

Data warehouse is a large central repository of information/ data that can be used for reporting and analysis to make more informed decisions. It is designed to support business intelligence (BI) activities, such as data mining, reporting, and online analytical processing (OLAP). Data flows into a data warehouse from transactional system, relational databases and other sources. Business analysts, Data engineers, data scientists and decision makers access the data from data warehouse. Data warehouses provide a structured way to store and analyze large volumes of data, making it easier for organizations to make data-driven decisions and gain insights into their operations.



TYPES OF DATA WAREHOUSE

- Enterprise Data Warehouse: An EDW is a centralized repository of all the data collected from various sources across the organization. It is designed to support the decision-making needs of the entire organization and is used by all departments. The data is cleaned, integrated, and transformed into a consistent format for easy access and analysis.

- Data Mart: A data mart is a subset of an EDW that is designed to support the decision-making needs of a specific department or business unit. It contains a subset of data from the EDW and is typically designed for a specific group of users, such as sales or marketing.

- Operational Data Store: Operational Data Store, which is also called ODS, are nothing but data store required when neither Data warehouse nor OLTP systems support organizations reporting needs. In ODS, Data warehouse is refreshed in real time. Hence, it is widely preferred for routine activities like storing records of the Employees.

CHARACTERISTICS OF DATA WAREHOUSE

- Subject-Oriented: A data warehouse uses a theme, and delivers information about a specific subject instead of a company's current operations. In other words, the data warehousing process is more equipped to handle a specific theme. Examples of themes or subjects include sales, distributions, marketing, etc.

- Integrated: Integration is defined as establishing a connection between large amount of data from multiple databases or sources. However, it is also essential for the data to be stored in the data warehouse in a unified manner. The process of data warehousing integrates data from multiple sources, such as a mainframe, relational databases, flat files, etc. Furthermore, it helps maintain consistent codes, attribute measures, naming conventions, and, formats.

- Time-variant: Time-variant in a DW is more extensive as compared to other operating systems. Data stored in a data warehouse is recalled with a specific time period and provides information from a historical perspective.

- Non-volatile: In the non-volatile data warehouse, data is permanent i.e., when new data is inserted, previous data is not replaced, omitted, or deleted. In this data warehouse, data is read-only and only refreshes at certain intervals. The two data operations performed in the data warehouse are data access and data loading.

DATABASE

A database is an organized collection of structured data / information typically stored in a computer system and accessed electronically. It is essentially a structured way of organizing and managing data so that it can be easily accessed, managed and updated. Databases are typically organized into table that contain related data. For example, a customer database might contain tables for customer information, purchase history, and billing information. Each table contains columns that define the data that is stored, such as a customer's name, address, phone number, and email address.