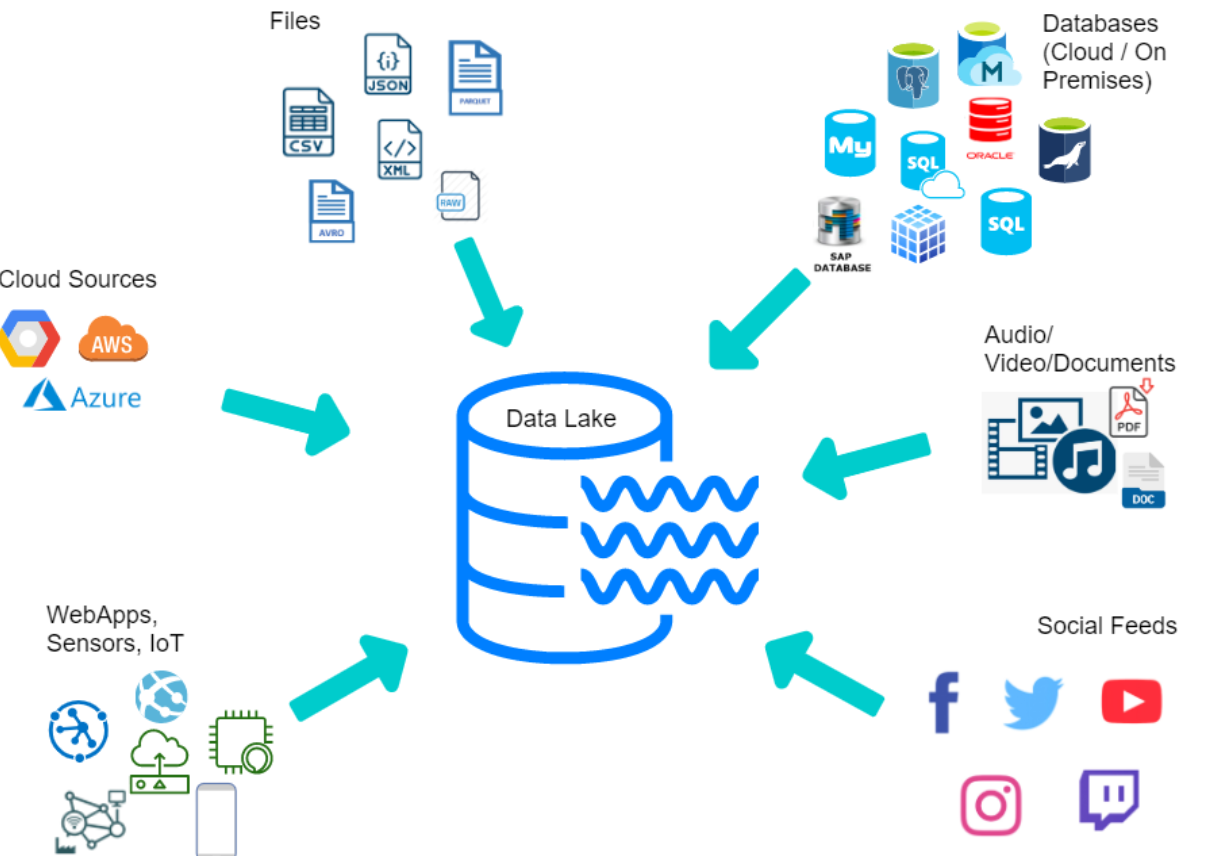
**TASK 1 DATA ENGINEERING BASICS – 1 14-03-23**

**BIG DATA**

It can be defined as data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency. Characteristics of big data include high volume, high velocity and high variety. Sources of data are becoming more complex than those for traditional data because they are being driven by artificial intelligence (AI), mobile devices, social media and the Internet of Things (IoT). For example, the different types of data originate from sensors, devices, video/audio, networks, log files, transactional applications, web and social media — much of it generated in real time and at a very large scale.

**DATA LAKE**

A data lake is a centralized repository for managing extremely large data volumes. It serves as a foundation for collecting and analyzing structured, semi-structured, and unstructured data in its native format to drive new insights, better predictions, and improved optimization. Unlike traditional data warehouses, data lakes can process video, audio, logs, texts, social media, sensor data and documents to power apps, analytics, and AI. Data lakes can be built as part of a data fabric architecture to provide the right data, at the right time, regardless of where it is resides.

**DATA BASE**

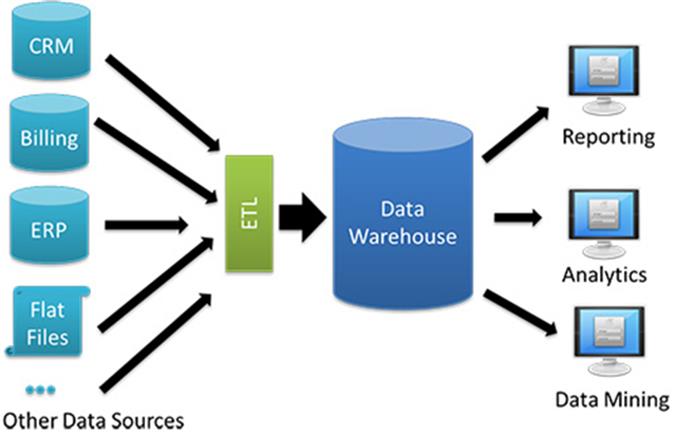
A database is an organized collection of structured information, or data, typically stored electronically in a computer system. A database is usually controlled by a database management system (DBMS). Together, the data and the DBMS, along with the applications that are associated with them, are referred to as a database system, often shortened to just database.

Data within the most common types of databases in operation today is typically modeled in rows and columns in a series of tables to make processing and data querying efficient. The data can then be easily accessed, managed, modified, updated, controlled, and organized. Most databases use structured query language (SQL) for writing and querying data.

**DATA WAREHOUSE**

A data warehouse, or enterprise data warehouse (EDW), is a system that aggregates data from different sources into a single, central, consistent data store to support data analysis, data mining, artificial intelligence (AI), and machine learning. A data warehouse system enables an organization to run powerful analytics on huge volumes (petabytes and petabytes) of historical data in ways that a standard database cannot.

Data warehousing systems have been a part of business intelligence (BI) solutions for over three decades, but they have evolved recently with the emergence of new data types and data hosting methods. Traditionally, a data warehouse was hosted on-premises—often on a mainframe computer—and its functionality was focused on extracting data from other sources, cleansing and preparing the data, and loading and maintaining the data in a relational database. More recently, a data warehouse might be hosted on a dedicated appliance or in the cloud, and most data warehouses have added analytics capabilities and data visualization and presentation tools.



**DATA WARE HOUSE ARCHITECTURE**

Generally speaking, data warehouses have a three-tier architecture, which consists of a:

Bottom tier: The bottom tier consists of a data warehouse server, usually a relational database system, which collects, cleanses, and transforms data from multiple data sources through a process known as Extract, Transform, and Load (ETL) or a process known as Extract, Load, and Transform (ELT).

Middle tier: The middle tier consists of an OLAP (i.e. online analytical processing) server which enables fast query speeds. Three types of OLAP models can be used in this tier, which are known as ROLAP, MOLAP and HOLAP. The type of OLAP model used is dependent on the type of database system that exists.

Top tier: The top tier is represented by some kind of front-end user interface or reporting tool, which enables end users to conduct ad-hoc data analysis on their business data.

**OLAP AND OLTP IN DATA WARE HOUSE**

OLAP (for online analytical processing) is software for performing multidimensional analysis at high speeds on large volumes of data from unified, centralized data store, like a data warehouse. OLTP, or online transactional processing, enables the real-time execution of large numbers of database transactions by large numbers of people, typically over the internet. The main difference between OLAP and OLTP is in the name: OLAP is analytical in nature, and OLTP is transactional.

OLAP tools are designed for multidimensional analysis of data in a data warehouse, which contains both historical and transactional data. Common uses of OLAP include data mining and other business intelligence applications, complex analytical calculations, and predictive scenarios, as well as business reporting functions like financial analysis, budgeting, and forecast planning.

OLTP is designed to support transaction-oriented applications by processing recent transactions as quickly and accurately as possible. Common uses of OLTP include ATMs, e-commerce software, credit card payment processing, online bookings, reservation systems, and record-keeping tools.