**Basics of Data Engineering**

**Submitted by:** Afraz Ahmed

* **Big Data:**

Big data is normally data that is huge in volume, and that normal data processing methods wont be sufficient enough to perform actions on such amount of data.

Examples may include data from New York Exchange or data uploaded on Social Media every single day. Even though it is large in volume, Big Data provides you with solutions that you wouldn’t find normally to help your business grow. Big data is characterized by the four Vs: Volume, Velocity, Variety, and Veracity. Volume refers to the huge amount of data generated; velocity refers to the speed at which data is created and processed; variety refers to the different types of data sources, such as text, audio, video, and social media, and veracity refers to the accuracy and trustworthiness of the data.

* **Data lake:**

It is a central repository that stores any kind of unstructured or semi structured data in large amounts. It does not require data to be preprocessed, rather it can store raw data from various sources, including social media, sensors, and mobile devices, among others. The main advantage of a data lake is that it provides a flexible and scalable approach to managing and analyzing data. Data can be stored in a data lake without any predefined structure or schema, and users can access the data easily and quickly for analysis and insights. Data lakes also support a wide range of data types, including text, audio, video, and images.

* **Database:**

The main advantage of a data lake is that it provides a flexible and scalable approach to managing and analyzing data. Data can be stored in a data lake without any predefined structure or schema, and users can access the data easily and quickly for analysis and insights. Data lakes also support a wide range of data types, including text, audio, video, and images. In a database, data is organized into tables, which are made up of rows and columns. Each row represents a single record, and each column represents a specific type of data, such as names, addresses, or phone numbers. Each table is designed to store data related to a specific topic or entity, such as customers, orders, or products.

* **Data Warehouse:**

A data warehouse is a large, centralized repository of data that is used for reporting and analysis. It is designed to support business intelligence activities, such as data mining, trend analysis, and decision-making. A data warehouse is typically made up of multiple databases, which are integrated and organized to provide a single, unified view of the data.