Unit I

Small data:

Small Data refers to datasets that are relatively small in size and can be easily managed, processed, and analyzed using traditional data processing tools and methods. Unlike Big Data, which deals with massive volumes of data that may require specialized architectures and technologies, Small Data can often fit on a single machine or a small cluster of machines.

What is Data:

Data in data analytics refers to the raw information collected from various sources, which serves as the foundation for conducting analytical processes. Data can take various forms, including structured data (organized in a fixed format, like rows and columns in a relational database), semi-structured data (partially organized, like JSON or XML files), and unstructured data (lacking a predefined structure, like text, images, audio, or video files).

- In data analytics, data is processed and analyzed using statistical techniques, algorithms, and machine learning models to derive meaningful insights, patterns, trends, and relationships. The main goal of data analytics is to convert raw data into actionable information that can aid decision-making, improve business processes, and optimize performance in various domains.
- Data analytics involves several key steps, such as data collection, data cleaning, data preparation, data analysis, and data visualization. Throughout these stages, data analysts extract valuable information and create visual representations of the findings to communicate the results effectively to stakeholders or decision-makers.

Feature	Small Data	Big Data
Variety	Data is typically structured and uniform	Data is often unstructured and heterogeneous
Veracity	Data is generally high quality and reliable	Data quality and reliability can vary widely
Processing	Data can often be processed on a single machine or in-memory	Data requires distributed processing frameworks such as MapReduce or Spark
Technology	Traditional	Modern
Analytics	Traditional statistical techniques can be used to analyze data	Advanced analytics techniques such as machine learning are often require
Collection	Generally, it is obtained in an organized manner than is inserted into the database	The Big Data collection is done by using pipelines having queues like AWS Kinesis or Google Pub / Sub to balance high-speed data
Volume	Data in the range of tens or hundreds of Gigabytes	Size of Data is more than Terabytes
Analysis Areas	Data marts(Analysts)	Clusters(Data Scientists), Data marts(Analysts)
Quality	Contains less noise as data is less collected in a controlled manner	Usually, the quality of data is not guaranteed

Feature	Small Data	Big Data
Processing	It requires batch-oriented processing pipelines	It has both batch and stream processing pipelines
Database	SQL	NoSQL
Velocity	A regulated and constant flow of data, data aggregation is slow	Data arrives at extremely high speeds, large volumes of data aggregation in a short time
Structure	Structured data in tabular format with fixed schema(Relational)	Numerous variety of data set including tabular data, text, audio, images, video, logs, JSON etc.(Non Relational)
Scalability	They are usually vertically scaled	They are mostly based on horizontally scaling architectures, which gives more versatility at a lower cost
Query Language	only Sequel	Python, R, Java, Sequel
Hardware	A single server is sufficient	Requires more than one server
Value	Business Intelligence, analysis and reporting	Complex data mining techniques for pattern finding, recommendation, prediction etc.
Optimization	Data can be optimized manually(human powered)	Requires machine learning techniques for data optimization

Feature	Small Data	Big Data
Storage	Storage within enterprises, local servers etc.	Usually requires distributed storage systems on cloud or in external file systems
People	Data Analysts, Database Administrators and Data Engineers	Data Scientists, Data Analysts, Database Administrators and Data Engineers
Security	Security practices for Small Data include user privileges, data encryption, hashing, etc.	Securing Big Data systems are much more complicated. Best security practices include data encryption, cluster network isolation, strong access control protocols etc.
Nomenclature	Database, Data Warehouse, Data Mart	Data Lake
Infrastructure	Predictable resource allocation, mostly vertically scalable hardware.	More agile infrastructure with horizontally scalable hardware
Applications	Small-scale applications, such as personal or small business data management	Large-scale applications, such as enterprise-level data management, internet of things (IoT), and social media analysis