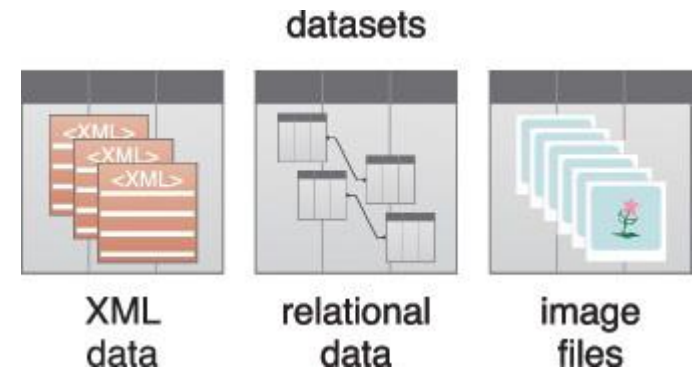# BIG DATA ANALYTICS

Unit - I

Concepts and Terminology

Datasets : Collections or groups of related data are generally referred to as datasets.

Each group or dataset member (datum) shares the same set of attributes or properties as others in the same dataset.

Examples :

✓Tweets stored in a flat file

✓collection of image files in a directory

✓Database table stored in a CSV formatted file

✓Historical weather observations that are stored as XML files

datasets

XML data

relational data

image files

Concepts and Terminology

Data Analysis :Data analysis is the process of examining data to find facts, relationships, patterns, insights and/or trends.
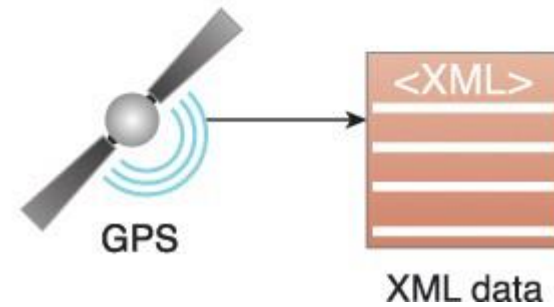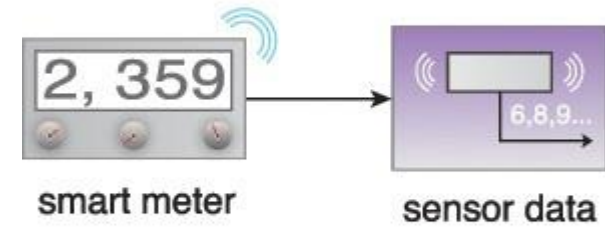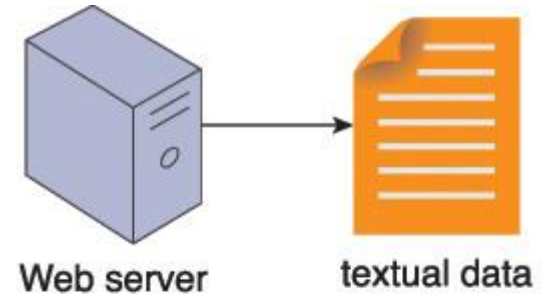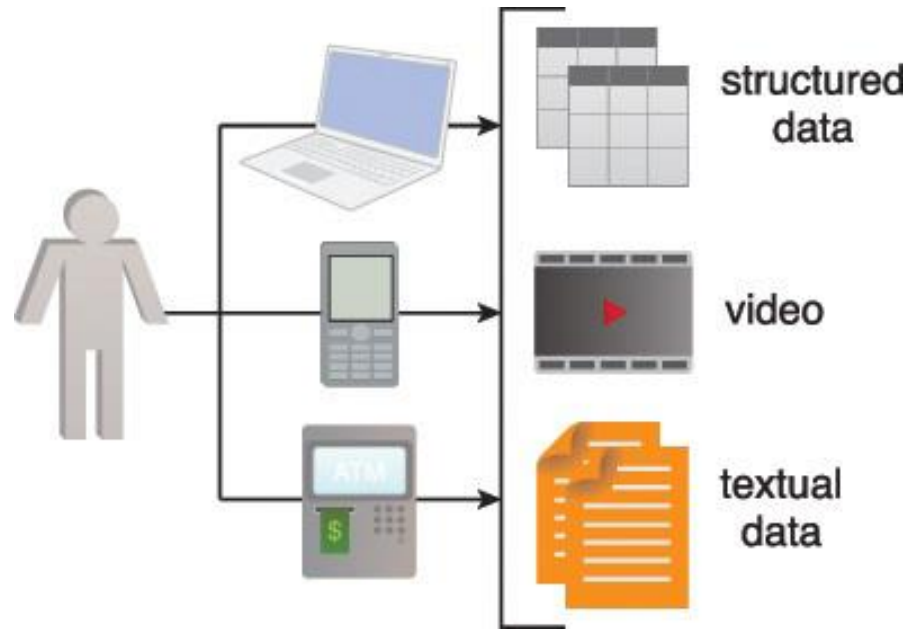
Data Analytics : Data analytics is a broader term that encompasses data analysis. Data analytics is a discipline that includes the management of the complete data lifecycle, which encompasses **collecting, cleansing, organizing, storing, analyzing and governing data**

Business Intelligence (BI) : BI enables an organization to gain insight into the performance of an enterprise by analyzing data generated by its business processes and information systems

Key Performance Indicators (KPI) : A KPI is a metric that can be used to gauge success within a particular business context.

Types of Data

Types of Data

The primary types of data are:

- ✓ **Structured data**
- ✓ **Unstructured data**
- ✓ **Semi-structured data**

Types of Data

Structured Data :

Structured data conforms to a data model or schema and is often stored in tabular form. It is used to capture relationships between different entities and is therefore most often stored in a relational database. Structured data is frequently generated by enterprise applications and information systems like ERP and CRM systems

Unstructured Data :

Data that does not conform to a data model or data schema is known as unstructured data. It is estimated that unstructured data makes up 80% of the data within any given enterprise
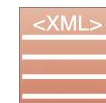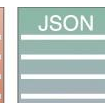

video   image files   audio

Semi-structured data : Type of structured data that does not hold to the tabular structure of data models associated with relational databases or other types of data tables

XML data   JSON data   sensor data