Big Data Analysis

Lecture 01 :
Big Data Overview

Dr. Asmaa Gad Amin
eng_mec_asm@yahoo.com

Grading Policy

- **Quizzes** **10%**
- **Lab +Projects** **20%**
- **Mid-Term Exam** **20%**
- **Final Exam** **50%**

Introduction to Big Data

- An overview of concepts, technologies, and applications.

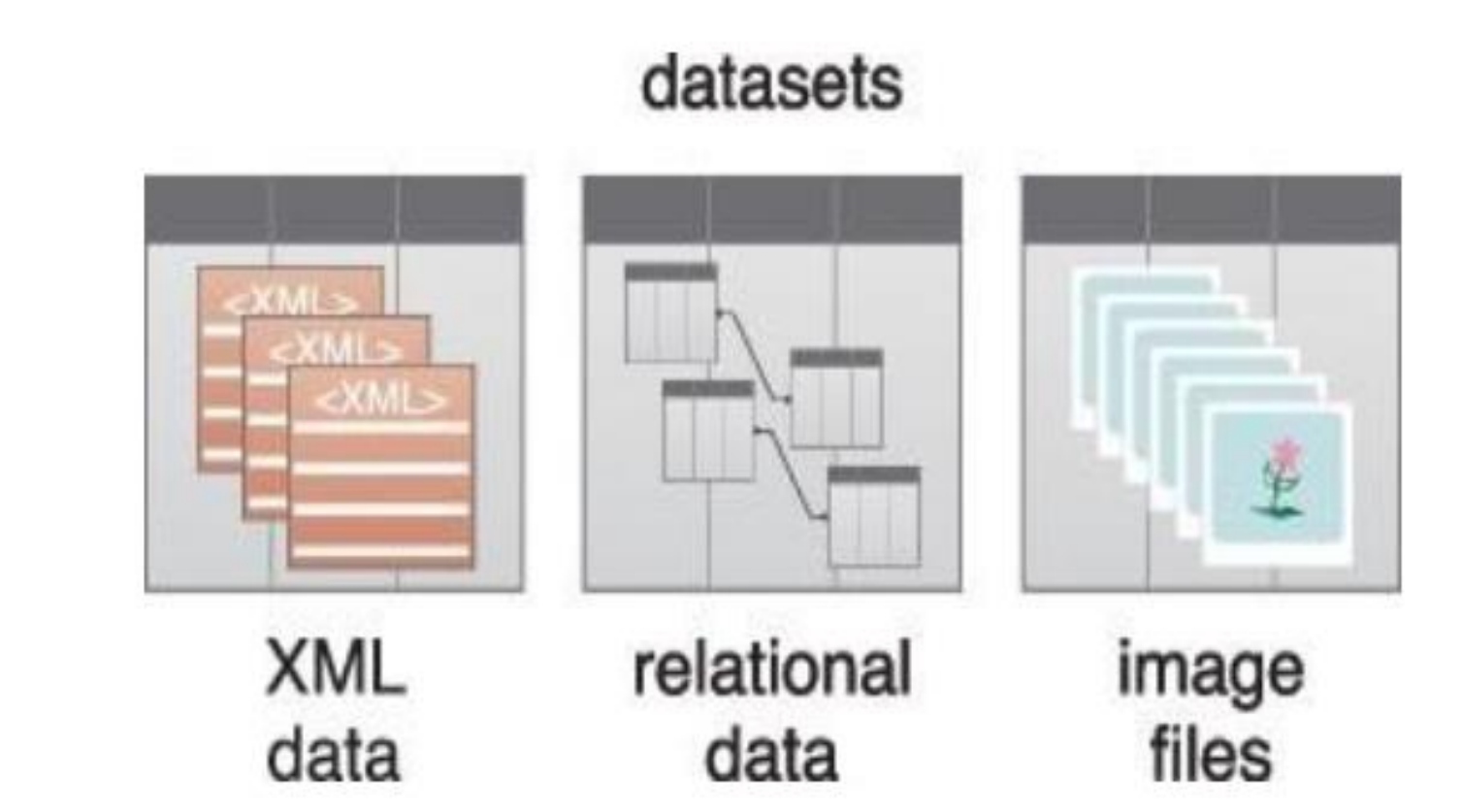
Datasets

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. Some examples of datasets are:

- tweets stored in a flat file
- a collection of image files in a directory
- an extract of rows from a database table stored in a CSV formatted file

Datasets

- historical weather observations that are stored as XML files

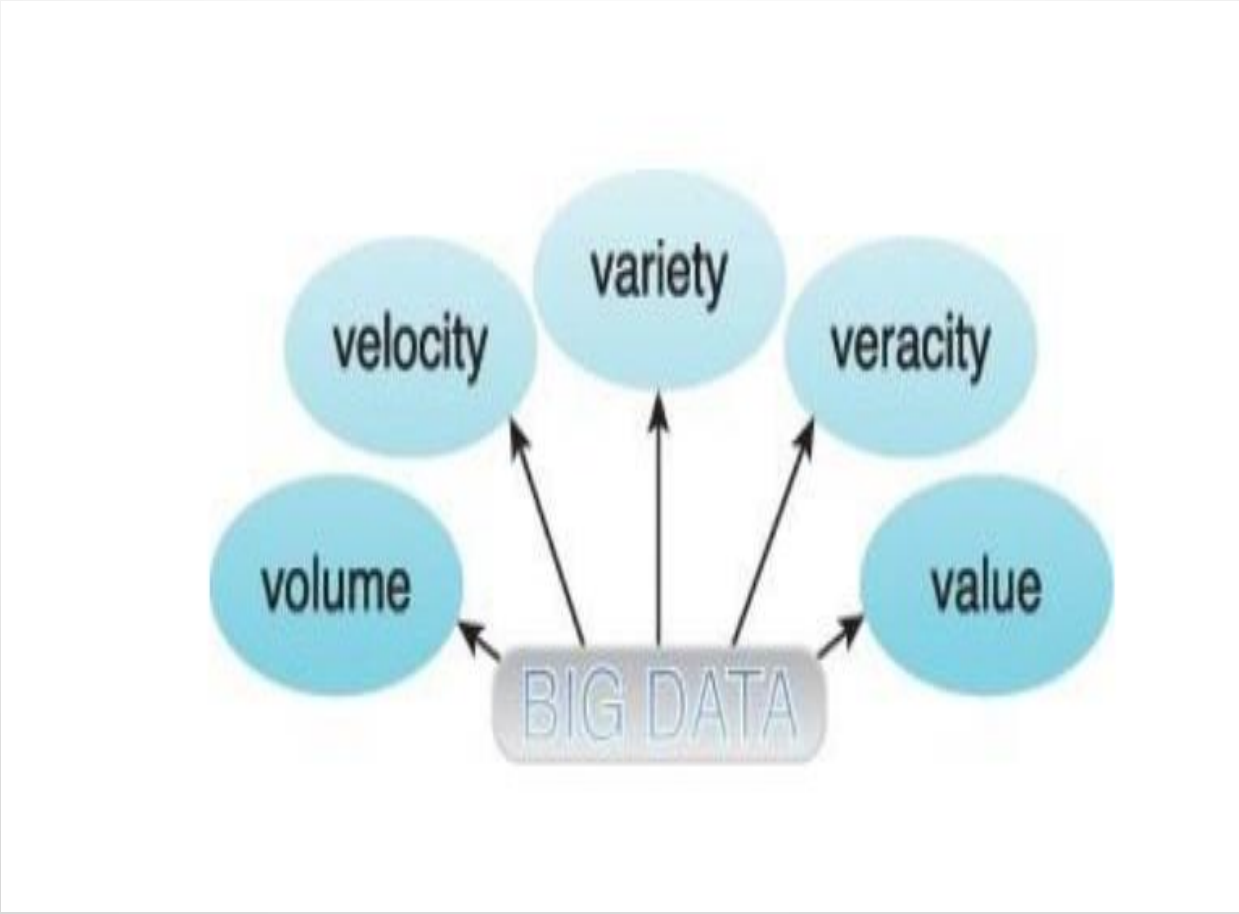


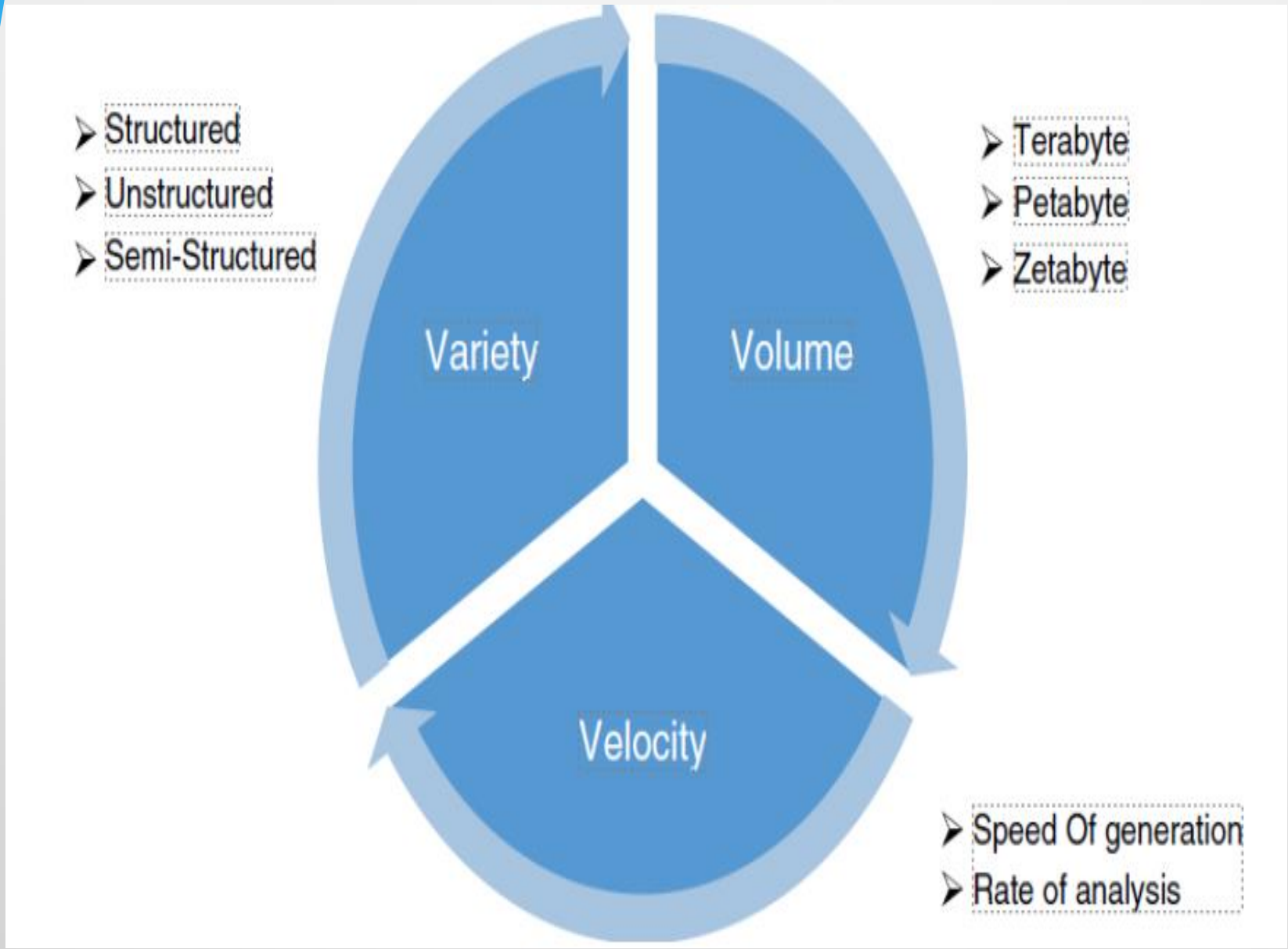
What is Big Data?

- Big Data refers to extremely large datasets that are too complex to be handled by traditional data-processing systems.

Characteristics of Big Data – The 5 Vs

- 1. Volume
- 2. Velocity
- 3. Variety
- 4. Veracity
- 5. Value

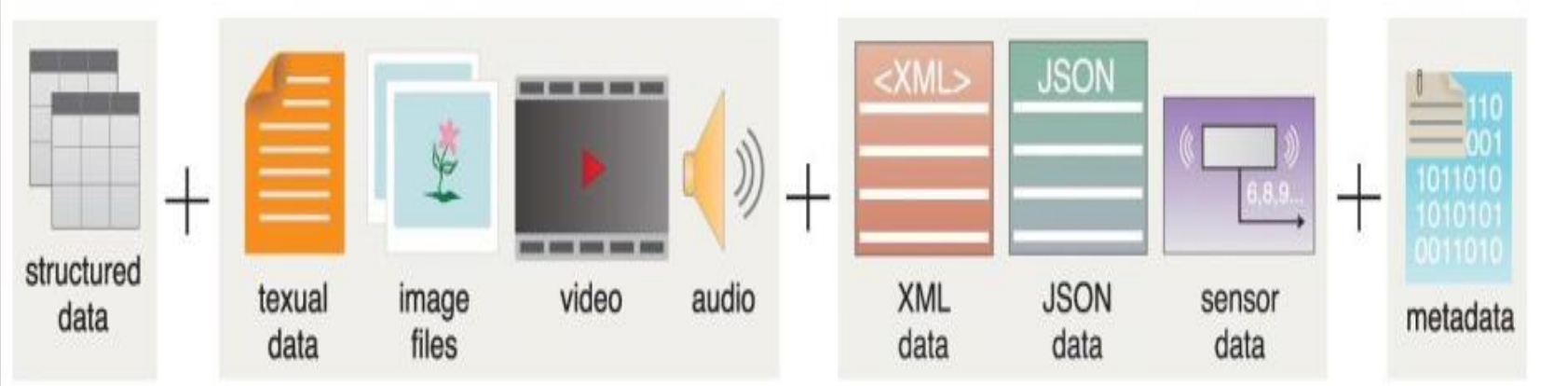



Structured
Unstructured
Semi-Structured
Variety
Volume
Terabyte
Petabyte
Zetabyte
Velocity
Speed Of generation
Rate of analysis

Variety

- Data variety refers to the multiple formats and types of data that need to be supported by Big Data solutions.
- Data variety brings challenges for enterprises in terms of data integration, transformation, processing, and storage.

- Examples of high-variety Big Data datasets include structured, textual, image, video, audio, XML, JSON, sensor data and metadata.



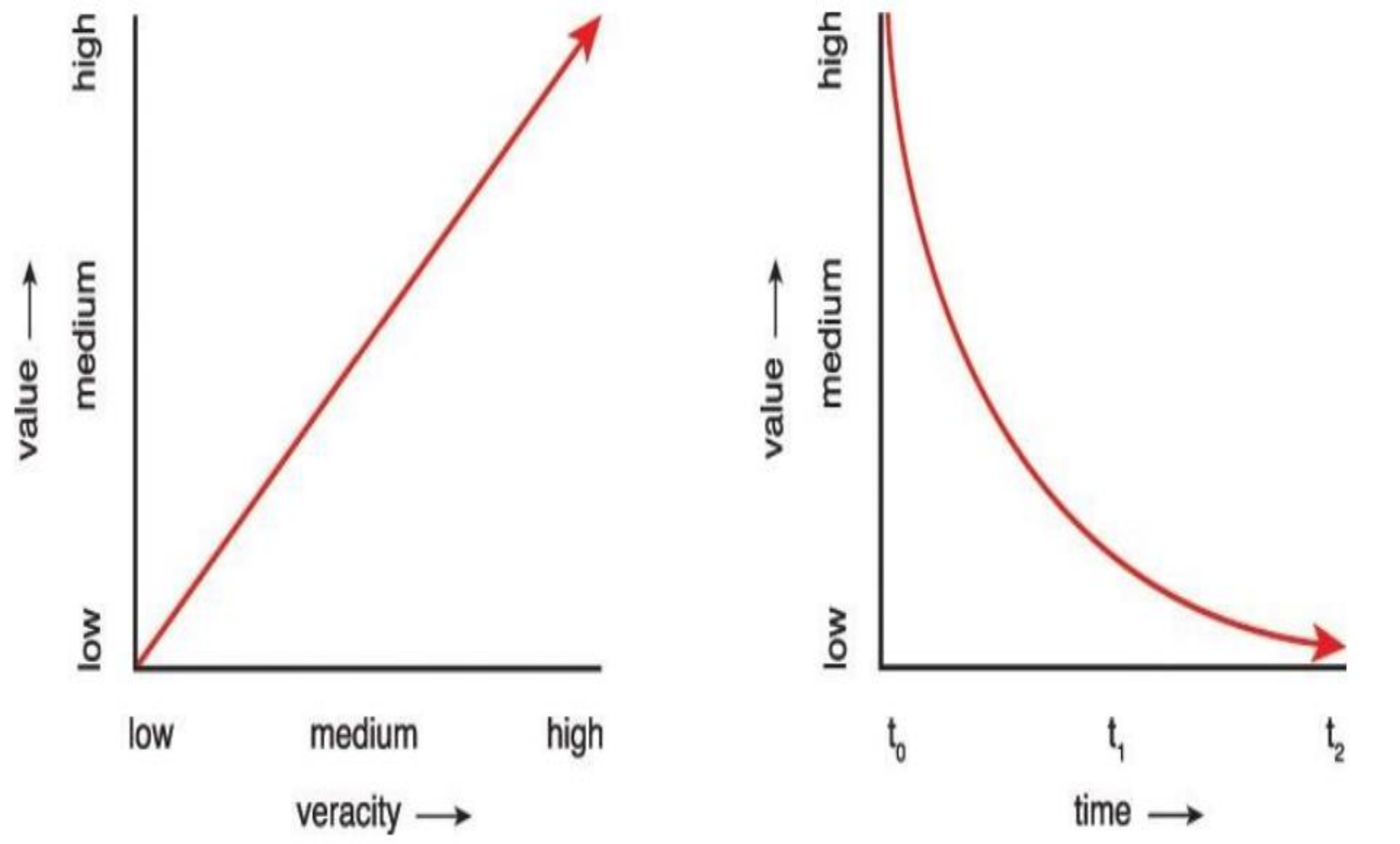
Veracity

- Veracity refers to the quality or fidelity of data. Data that enters Big Data environments needs to be assessed for quality, which can lead to data processing activities to resolve invalid data and remove noise.

Value

- Value is defined as the usefulness of data for an enterprise. The value characteristic is intuitively related to the veracity characteristic in that the higher the data fidelity, the more value it holds for the business.

Value


high
medium
low
value
low
medium
high
veracity
high
medium
low
value
t_0
t_1
t_2
time

Apart from veracity and time, value is also impacted by the following lifecycle-related concerns:

- How well has the data been stored?
- Were valuable attributes of the data removed during data cleaning?
- Are the results of the analysis being accurately communicated to the appropriate decision-makers?

Importance of Big Data

- Enables data-driven decisions
- Enhances customer experiences
- Improves efficiency and innovation
- Supports predictive analytics

Big Data Architecture

Components:

- Data sources
- Data storage
- Data processing
- Data analysis
- Data visualization

Hadoop Ecosystem

- HDFS – Distributed file system
- MapReduce – Processing model
- YARN(Yet Another Resource Negotiator) – Resource manager
- Hive, Pig – Query tools (query language)

Spark Framework

- In-memory data processing
- Faster than MapReduce
- Supports SQL, Streaming, MLlib, GraphX

NoSQL Databases

- Non-relational storage
- Examples: MongoDB, Cassandra
- Scalable and flexible for unstructured data

Big Data Analytics

- Descriptive Analytics
- Predictive Analytics
- Prescriptive Analytics
- Diagnostic Analytics
- Cognitive Analytics

1. Descriptive Analytics

- Answers: What happened?
- Summarizes historical data to identify patterns and trends.
- Tools: Dashboards, Reports, Data Visualization.
- Example: Monthly sales reports.

2. Diagnostic Analytics

- Answers: Why did it happen?
- Examines data to discover causes of trends or outcomes.
- Techniques: Drill-down, Correlation analysis, Data mining.
- Example: Analyzing drop in website traffic.

3. Predictive Analytics

- Answers: What will happen?
- Uses statistical models and machine learning to forecast future outcomes.
- Tools: Regression, Decision Trees, Neural Networks.
- Example: Predicting customer churn.

4. Prescriptive Analytics

- Answers: What should we do?
- Suggests optimal actions using optimization and simulation.
- Techniques: AI, Machine Learning, Scenario analysis.
- Example: Recommending best marketing strategy.

5. Cognitive Analytics

- Mimics human thought processes to interpret data.
- Combines AI, NLP, and Machine Learning.
- Example: Chatbots and intelligent assistants for customer support.

Comparison of Analytics Types

- Descriptive → What happened
- Diagnostic → Why it happened
- Predictive → What will happen
- Prescriptive → What should be done
- Cognitive → How systems can learn and adapt

Applications of Big Data

- Healthcare – patient data analysis
- Finance – fraud detection
- Retail – recommendation systems
- IoT – sensor data processing

Challenges in Big Data

- Data privacy and security
- Storage management
- Data integration and quality
- Skilled workforce shortage

Future Trends

- Artificial Intelligence integration
- Edge and Cloud computing
- Real-time analytics
- Data governance and ethics

Summary

- Big Data transforms industries through analytics and innovation.
- Key technologies: Hadoop, Spark, NoSQL.
- Challenges remain in scalability, privacy, and management.

Different Types of Data

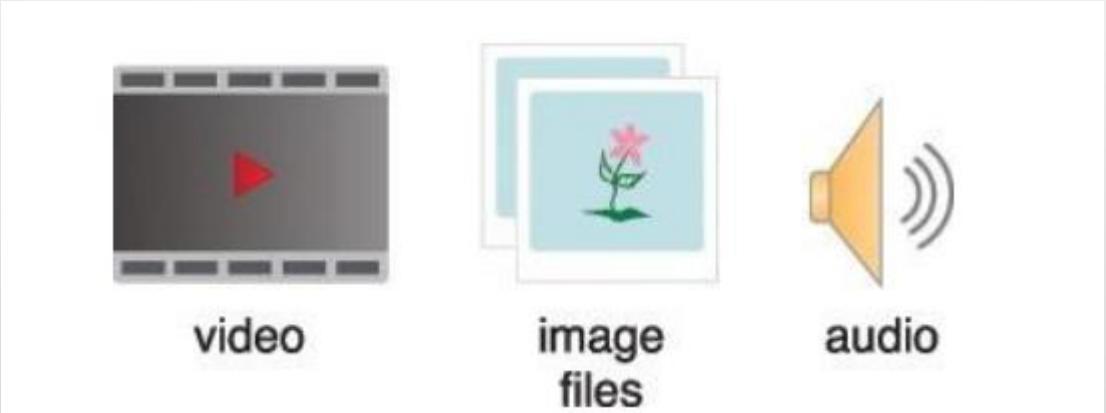
- Understanding how data is classified and used in analytics.

1. Structured Data

- Organized in rows and columns
- Easy to store, query, and analyze
- Examples: Databases, Spreadsheets

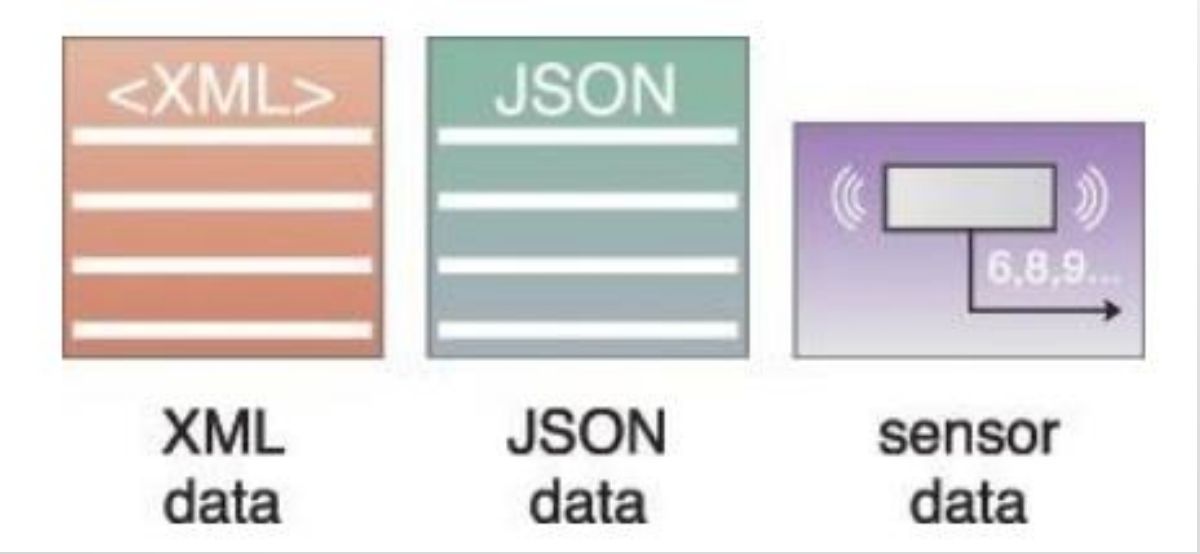
2. Unstructured Data

- No predefined format or organization
- Difficult to search and process
- Examples: Text documents, Images, Videos, Social media posts



3. Semi-Structured Data

- Contains both structured and unstructured elements
- Uses tags or markers to separate data
- Examples: JSON, XML, HTML

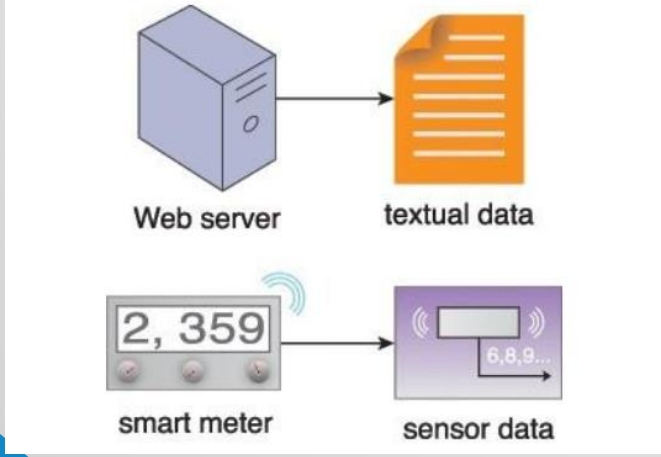


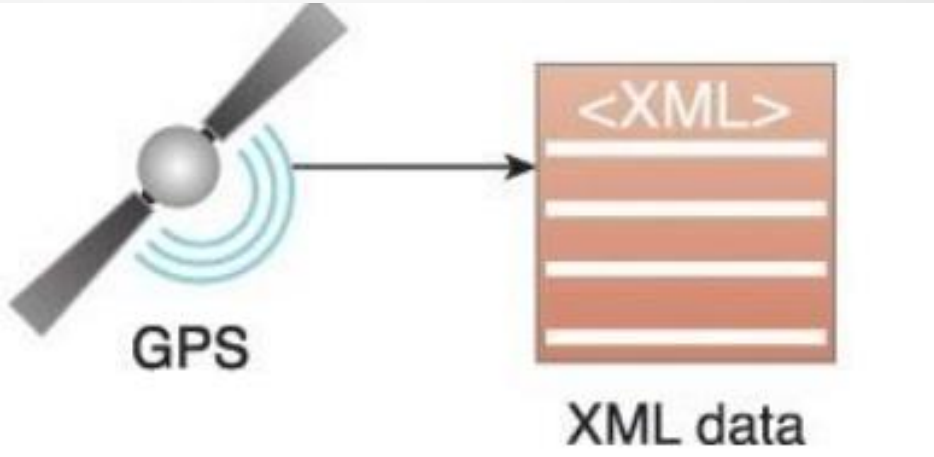
4. Metadata

- Data that describes other data
- Helps organize, find, and understand data
- Example: Author name, file size, creation date

5. Machine-Generated Data

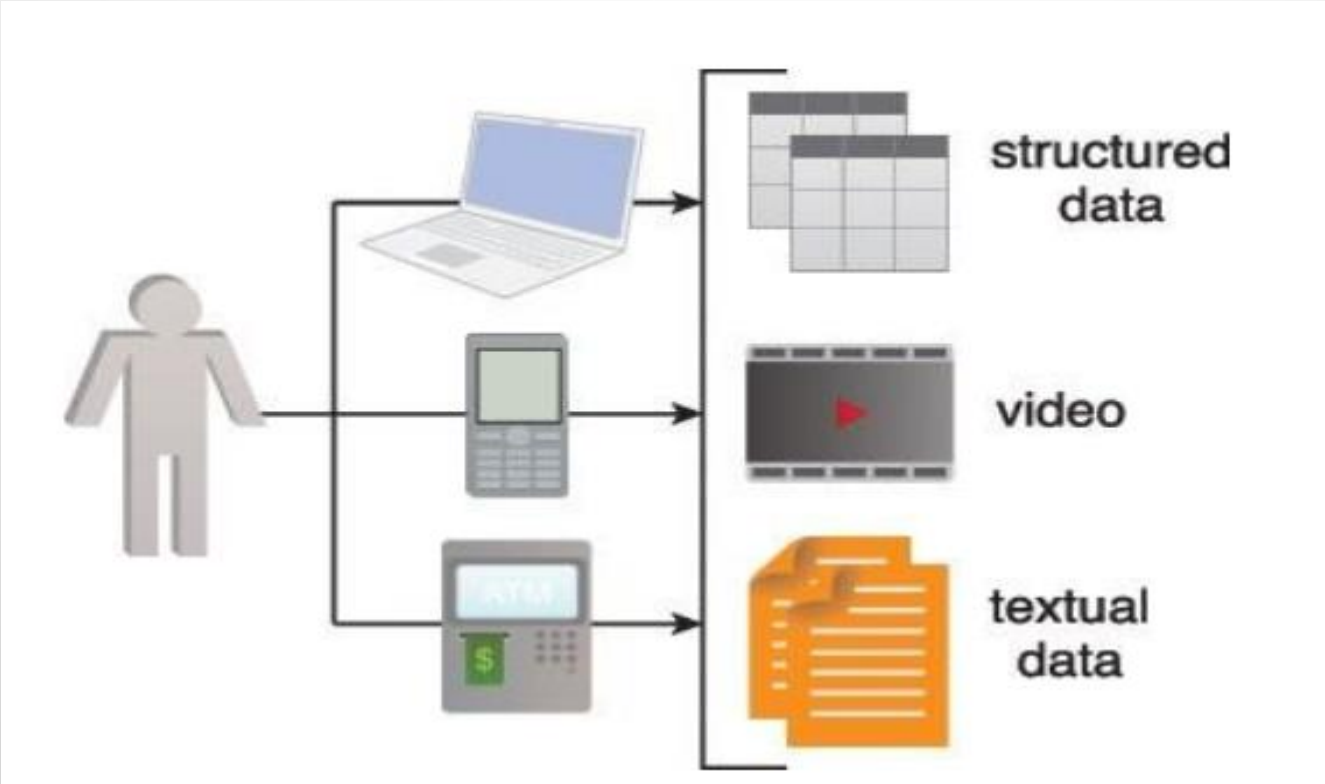
- Produced automatically by devices or systems
- Examples: Sensor data, Server logs, IoT data





6. Human-Generated Data

- Created by people through online interactions
- Examples: Emails, Tweets, Reviews, Posts



7. Real-Time Data

- Data that is created, processed, and analyzed immediately
- Used in systems requiring instant feedback
- Examples: Stock trading, GPS tracking, Smart grids

Summary

- Data can be structured, unstructured, or semi-structured.
- Understanding the type of data is crucial for effective storage, processing, and analysis.

Thank You