Module 1

Introduction to Big Data

Balasubramanian PG

Table of Contents

01

What is Big Data?

Big data refers to large and complex datasets that are too massive to be processed or analyzed using traditional data processing techniques. It involves capturing, storing, and analyzing vast amounts of information to reveal patterns, trends, and insights

Example of Big Data

Social media: Platforms like Facebook and Twitter generate enormous volumes of data in the form of posts, comments, likes, and shares.

Internet of Things (IoT): Devices such as smart sensors, wearables, and connected appliances produce large streams of data in real-time.

E-commerce: Online retailers collect data on customer behavior, purchase history, and preferences to enhance their marketing strategies and user experience.

Healthcare: Medical records, diagnostic images, and patient data generate substantial data that can be used for research and personalized treatments.

02

Types of Big Data Analytics

Descriptive Analytics

Focus: Describes past events and provides a summary of historical data.

Purpose: Aims to understand what happened and gain insights into past performance.

Techniques and Examples:

Reporting: Generating reports and visualizations to present historical data in a clear and concise manner.

Data Aggregation: Summarizing large datasets to provide an overview of trends and patterns.

Key Performance Indicators (KPIs): Identifying and tracking essential metrics to assess business performance.

Example Scenario: A retail company analyzes its sales data from the past year to create a report with monthly revenue, top-selling products, and sales trends.

Diagnostic Analytics

Focus

Examines historical data to determine why certain events occurred.

Purpose

Seeks to identify the root causes of specific outcomes or trends.

Example

An e-commerce platform investigates a sudden drop in website traffic and finds that a recent website update caused the issue.

Techniques of Diagnostic Analytics

- Data Drill-Down: Navigating through data hierarchies to investigate details and find causes of specific occurrences.
- Data Mining: Applying algorithms to discover hidden patterns and relationships within the data.

Predictive Analytics

Focus

- Uses historical data to make predictions about future events or outcomes.

Purpose

- Anticipates what is likely to happen based on patterns and trends.

Example

- A transportation company uses historical traffic data to predict potential traffic jams and optimize route planning.

Techniques of Predictive Analytics

- Machine Learning: Employing algorithms to build predictive models that can learn from historical data and make future predictions.

- Time Series Analysis: Analyzing data over time to forecast future trends.

Prescriptive Analytics

Method

Focus

- Utilizes predictive models to suggest actions or recommendations.

Purpose

- Recommends the best course of action to achieve a specific goal.

Priority

Example

- A healthcare provider uses a prescriptive analytics model to recommend personalized treatment plans for patients based on their medical history and conditions.

Techniques

- Optimization algorithms
- Simulation Models

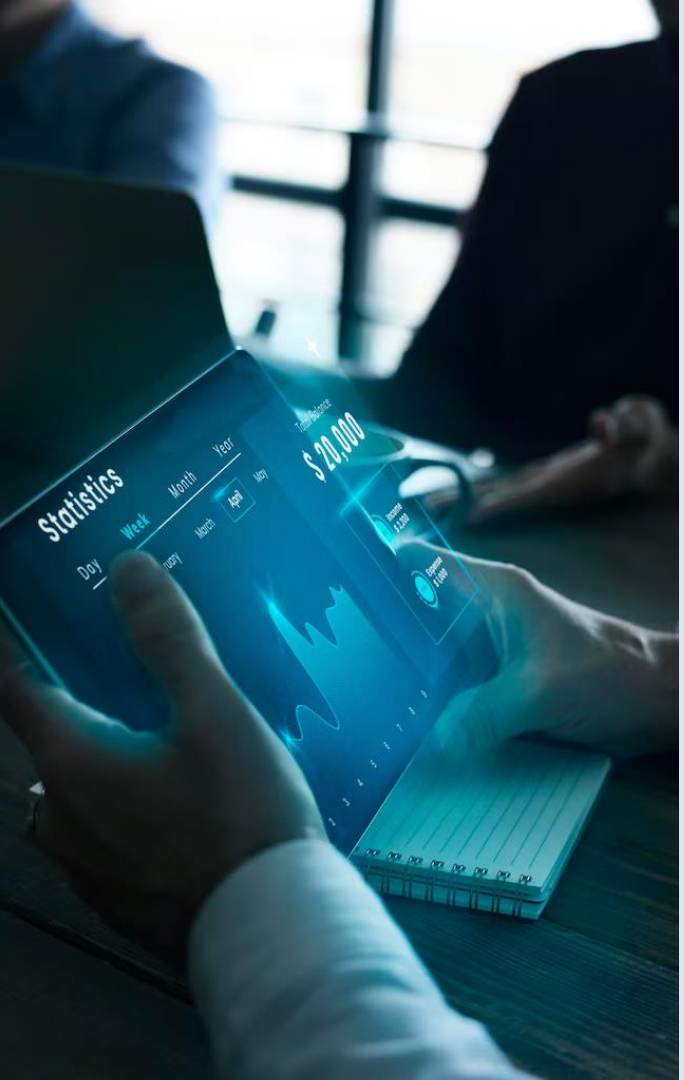
Types of Big Data

Structured Data

Structured data refers to a type of data that has a well-defined and organized format, making it easy to store, access, and analyze. It follows a fixed schema, which means the data is stored in a tabular form with rows and columns, similar to a spreadsheet or relational database. Each column in the structure corresponds to a specific data attribute or field, and each row represents a single data entry or record.

Key Characteristics of Structured Data

1. Tabular Format: Structured data is organized into tables with rows and columns, allowing for efficient storage and retrieval.
2. Fixed Schema: The data follows a predefined and consistent structure, where each field has a specified data type and meaning.
3. Easily Queryable: Due to its organized nature, structured data can be queried using standard database query languages like SQL.
4. High Reliability: The structured format ensures data integrity and consistency, reducing the risk of errors during data analysis.



Use Case

Structured data is commonly used in various business applications and systems, including traditional relational databases and enterprise resource planning (ERP) systems. Its clear organization makes it suitable for straightforward data processing, analysis, and reporting tasks. However, structured data may not be suitable for capturing unstructured or complex information, such as free-form text or multimedia content, which requires different data storage and processing techniques.

Semi Structured Data

Semi-structured data refers to a type of data that does not fit neatly into the rigid structure of traditional relational databases, yet it has some level of organization. Unlike structured data with fixed schemas, semi-structured data allows for flexibility in representing data elements, making it suitable for handling diverse and complex information.

Key Characteristics of Semi Structured Data

1. Flexible Structure: Semi-structured data does not follow a strict, predefined schema like structured data. Instead, it allows for varying structures within the data itself.
2. Self-Describing: Semi-structured data often includes metadata or tags that provide information about the data's structure or meaning.
3. Hierarchical Organization: Semi-structured data is often organized in a hierarchical manner, with nested elements or attributes.
4. Data Elements: Individual data elements within semi-structured data may have different data types and structures.

Types of Semi Structured Data

- XML (eXtensible Markup Language): XML data contains both data and tags that describe the data's structure and meaning.
- JSON (JavaScript Object Notation): JSON data is a lightweight data interchange format with attribute-value pairs
- Some NoSQL databases, like MongoDB, allow for semi-structured data storage, enabling flexibility in the data schema.

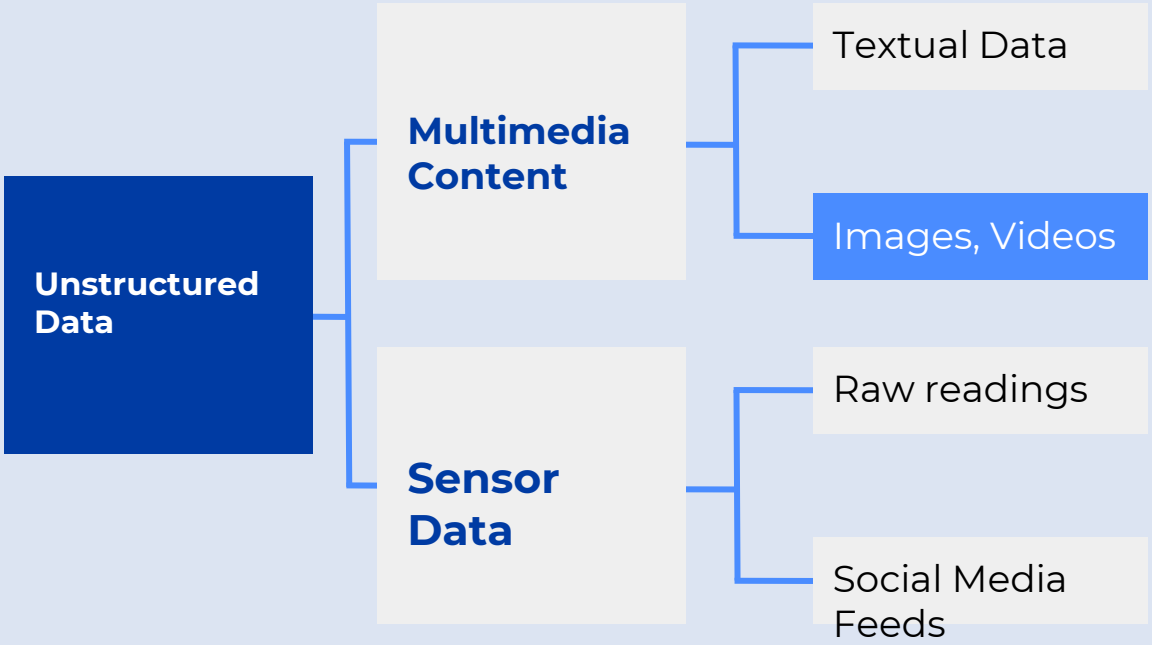
Unstructured Data

Unstructured data refers to a type of data that lacks a predefined or organized structure, making it challenging to store, process, and analyze using traditional methods. Unlike structured data with fixed schemas or semi-structured data with some level of organization, unstructured data does not conform to a specific format or data model.

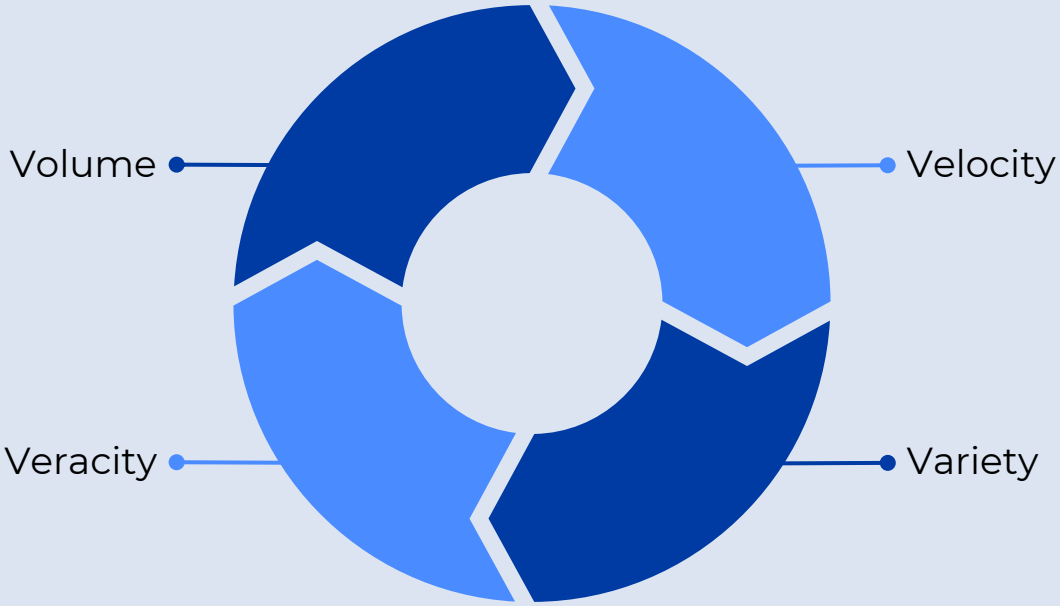
Key characteristics

Lack of Structure	Heterogeneous Format
No Clear Data Mode	Large Volume

Examples


Unstructured Data
Multimedia Content
Textual Data
Images, Videos
Sensor Data
Raw readings
Social Media Feeds

5vs of Big Data



Volume

Definition: Volume refers to the massive scale of data that is generated and collected in big data scenarios.

Significance: Big data involves datasets that are too large to be handled by conventional data processing systems and require specialized tools and techniques to manage and analyze.

Example: Social media platforms generating billions of posts and interactions daily, resulting in enormous data volumes.

Velocity

Definition: Velocity denotes the high speed at which data is generated, collected, and processed in real-time or near-real-time.

Significance: Big data applications often require quick and immediate processing to respond to rapidly changing conditions and make timely decisions.

Example: Sensor data from IoT devices streaming continuously, requiring real-time analysis for immediate insights.

Variety

Definition: Variety refers to the diverse types and formats of data found in big data environments.

Significance: Big data encompasses structured, semi-structured, and unstructured data from various sources, necessitating flexible data processing techniques.

Example: Data sources including text, images, videos, social media feeds, log files, and sensor readings in different formats.

Veracity

Definition: Veracity addresses the reliability and trustworthiness of the data.

Significance: Big data may contain noisy, incomplete, or inconsistent data, which can impact the accuracy and validity of analysis results.

Example: Social media data with fake accounts, spam, or biased content, affecting sentiment analysis accuracy.

Value

Definition: Value represents the importance and relevance of the insights derived from big data analysis.

Significance: The ultimate goal of big data analysis is to extract valuable and actionable insights that lead to informed decision-making and business improvements.

Example: Analyzing customer data to identify patterns and preferences, enabling personalized marketing strategies that drive sales.

Applications of Big Data

Tracking Customer Spending Habit, Shopping Behavior

Big data analytics is used to collect and analyze vast amounts of customer data from various sources, such as online transactions, loyalty programs, and social media interactions. This data is then used to understand customer preferences, shopping habits, and behavior, enabling businesses to offer personalized product recommendations, targeted promotions, and improved customer experiences.

Recommendation

Recommendation systems, like those used by e-commerce platforms and streaming services, utilize big data to analyze user behavior, historical interactions, and preferences. This data-driven approach helps make personalized recommendations, suggesting products, movies, or content that align with the user's interests, ultimately enhancing user engagement and satisfaction.

Smart Traffic System

In a smart traffic system, big data plays a crucial role in collecting real-time data from various sources such as traffic cameras, GPS devices, and road sensors. Analyzing this data helps optimize traffic flow, detect congestion, and provide dynamic route guidance to drivers, reducing travel time and enhancing overall traffic management.

Secure Air Traffic System

In aviation, big data is used to monitor and analyze flight data, weather conditions, and air traffic patterns. By processing and analyzing this data, the air traffic system can make informed decisions, ensure safety, and manage air traffic flow efficiently.

Auto Driving Car

Self-driving cars rely on big data from multiple sensors, cameras, and LiDAR systems to perceive their environment. The data collected is continuously processed and analyzed to make real-time decisions, enabling the autonomous vehicle to navigate safely and respond to changing road conditions.

Virtual Personal Assistant Tool

Virtual personal assistant tools, like Siri and Google Assistant, leverage big data to understand and respond to user queries effectively. These assistants use natural language processing (NLP) and analyze vast amounts of data to provide accurate and contextually relevant responses.

Internet of Things (IoT)

The Internet of Things (IoT) generates massive amounts of data from connected devices and sensors. Big data technologies process and analyze this data, enabling businesses and individuals to make data-driven decisions, monitor and control devices remotely, and optimize various processes.

Big data analytics tools and technology

R Studio	• Domain-specific programming language for statistical analysis and data visualization. • Developed in 1993 by Ross Ihaka and Robert Gentleman. • Provides better insights through accurate data collection.
Apache Hadoop	• Open-source software framework for storing and processing data on clusters. • Developed in 2005 by Doug Cutting and Mike Cafarella. • Components include Hadoop Distributed File System (HDFS) and MapReduce.

Big data analytics tools and technology

MongoDB	• Document-oriented NoSQL database for high-volume data storage. • Uses collections and documents instead of rows and columns. • Each database contains collections with varying document sizes and content.
Rapid Miner	• Open-source platform for data prep, machine learning, and predictive model deployment. • Offers a powerful graphical user interface for analysis design. • Supports Windows, Macintosh, Linux, and Unix systems.

Apache Spark

Data processing framework for large data sets, with distributed computing capabilities.

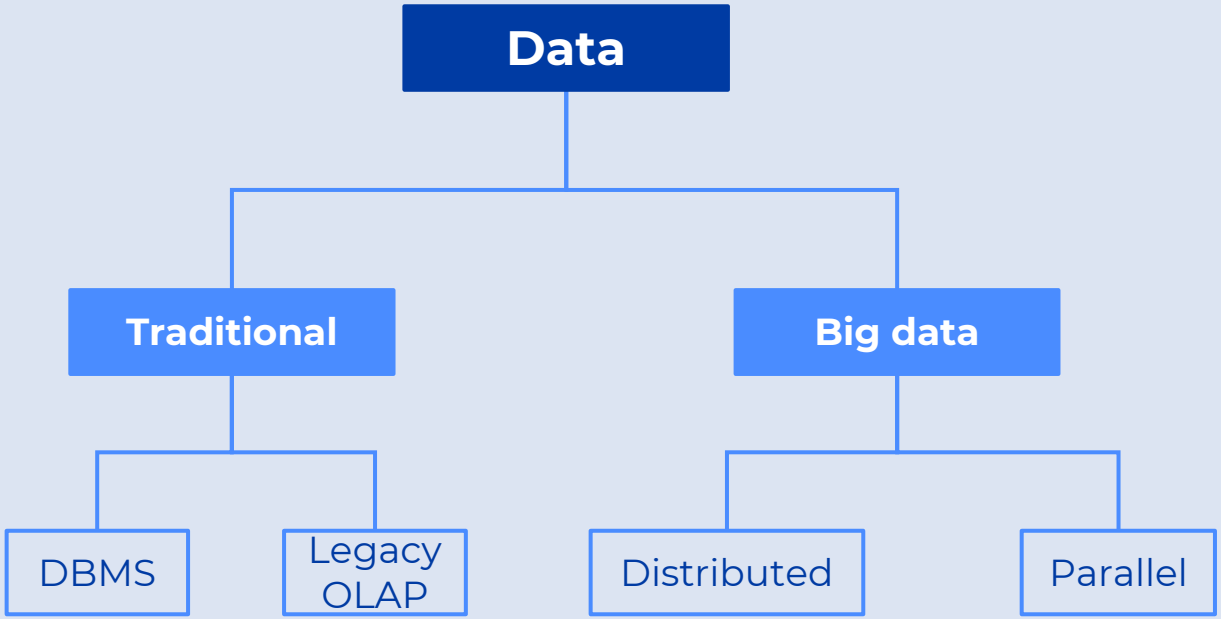
In-built support for streaming, SQL, machine learning, and graph processing.

Known for its speed and efficiency in big data transformations.

Microsoft Azure

- Public cloud computing platform by Microsoft with computing, analytics, and storage services.
- Offers big data cloud offerings in Standard and Premium categories.
- Provides reliable analytics, enterprise-grade security, and high productivity for developers.

Traditional Data v/s Big data



Traditional Data	Big Data
Traditional data is generated in enterprise level.	Big data is generated outside the enterprise level.
Its volume ranges from Gigabytes to Terabytes.	Its volume ranges from Petabytes to Zettabytes or Exabytes.
Traditional database system deals with structured data.	Big data system deals with structured, semi-structured,database, and unstructured data.
Traditional data is generated per hour or per day or more.	But big data is generated more frequently mainly per seconds.
Traditional data source is centralized and it is managed in centralized form.	Big data source is distributed and it is managed in distributed form.
Data integration is very easy.	Data integration is very difficult.
Normal system configuration is capable to process traditional data.	High system configuration is required to process big data.

Traditional Data	Big Data
The size of the data is very small.	The size is more than the traditional data size.
Traditional data base tools are required to perform any data base operation.	Special kind of data base tools are required to perform any databaseschema-based operation.
Normal functions can manipulate data.	Special kind of functions can manipulate data.
Its data model is strict schema based and it is static.	Its data model is a flat schema based and it is dynamic.
Traditional data is stable and inter relationship.	Big data is not stable and unknown relationship.
Traditional data is in manageable volume.	Big data is in huge volume which becomes unmanageable.
It is easy to manage and manipulate the data.	It is difficult to manage and manipulate the data.
Its data sources includes ERP transaction data, CRM transaction data, financial data, organizational data, web transaction data etc.	Its data sources includes social media, device data, sensor data, video, images, audio etc.

Thanks