Day72 Intro. to Big Data

By: Loga Aswin

Big data refers to large volumes of data that are too complex or massive for traditional data processing applications to handle efficiently.

Types Of Big Data

1. Structured Data:

- It is highly organized and formatted, making it easily searchable and accessible.
- It follows a clear data model and schema.
- Stored in databases with defined tables and rows.
- Can be queried using SQL (Structured Query Language).
- Examples:
  - Relational Databases: Data stored in tables with rows and columns, such as customer information, transaction records, etc.
  - Spreadsheets: Organized data in Excel or Google Sheets.
- Use Cases:
  - Financial records in banks.
  - Inventory management systems.
- Tools:
  - SQL databases like MySQL, Oracle, SQL Server.

2. Unstructured Data:

- Unstructured data lacks a specific structure or organization, making it more challenging to process and analyze.
- Characteristics:
- No predefined data model or format.

- Includes text, images, videos, social media posts, emails, etc.
- Difficult to analyze using traditional methods.
- Examples:
  - Social Media Feeds: Tweets, Facebook posts, etc.
  - Multimedia: Images, videos, audio files.
  - Text Documents: Emails, PDFs, Word documents.
- Use Cases:
  - Sentiment analysis of social media data.
  - Image recognition and analysis.
- Tools:
  - Natural Language Processing (NLP) tools like NLTK, spaCy.
  - Image and video processing libraries like OpenCV, TensorFlow.

3. Semi-Structured Data:

- Semi-structured data has some structure but doesn't conform to the strict structure of traditional databases.
- Characteristics:
- Contains tags, markers, or other indicators of structure within the data.
- Flexible and can be easily modified.
- Supports queries and some level of organization.
- Examples:
  - XML Files: Contains structured data with tags defining elements and attributes.
  - JSON (JavaScript Object Notation): Flexible format for data exchange between systems.
  - NoSQL Databases: Document-oriented databases.
- Use Cases:
  - Storing data with varying structures.
  - Web application data.
- Tools:
  - NoSQL databases like MongoDB, Cassandra.

- Parsers for XML and JSON data.

Characteristics of Big Data (4 Vs):

1. **Volume:**
   - Refers to the sheer size of the data.
   - Example: Terabytes, petabytes, or exabytes of data generated from various sources like social media, sensors, etc.
2. **Velocity:**
   - Describes the speed at which data is generated and processed.
   - Example: Real-time data streaming, continuous data flow from IoT devices, social media updates, etc.
3. **Variety:**
   - Indicates the diversity of data types and sources.
   - Example: Structured data (like databases), unstructured data (social media posts, emails), semi-structured data (XML, JSON), multimedia files, etc.
4. **Veracity:**
   - Refers to the reliability and accuracy of the data.
   - Example: Data can be incomplete, inconsistent, or contain errors due to its varied sources and formats.

Example of Big Data:

1. **Social Media:** Platforms like Facebook, Twitter, and Instagram generate massive amounts of unstructured data in the form of posts, comments, images, videos, etc.
2. **Internet of Things (IoT):** Sensors and devices connected to the internet constantly generate data in real-time, such as temperature sensors, GPS devices, etc.
3. **E-commerce:** Online shopping platforms collect huge volumes of structured and unstructured data about customer behavior, purchases, preferences, etc.