BIG DATA ANALYTICS (CS-431)

Dr. Sriparna Saha
Associate Professor

Website: https://www.iitp.ac.in/~sriparna/
Google Scholar: https://scholar.google.co.in/citations?user=Fj7jA_AAAAAJ&hl=en
Research Lab: SS_Lab
Core Research AREA: NLP, GenAI, LLMs, VLMs, Multimodality, Meta-Learning, Health Care, FinTech, Conversational Agents

TAs: Sarmistha Das, Nitish Kumar, Divyanshu Singh, Aditya Bhagat, Annu Kumari, Harsh Raj

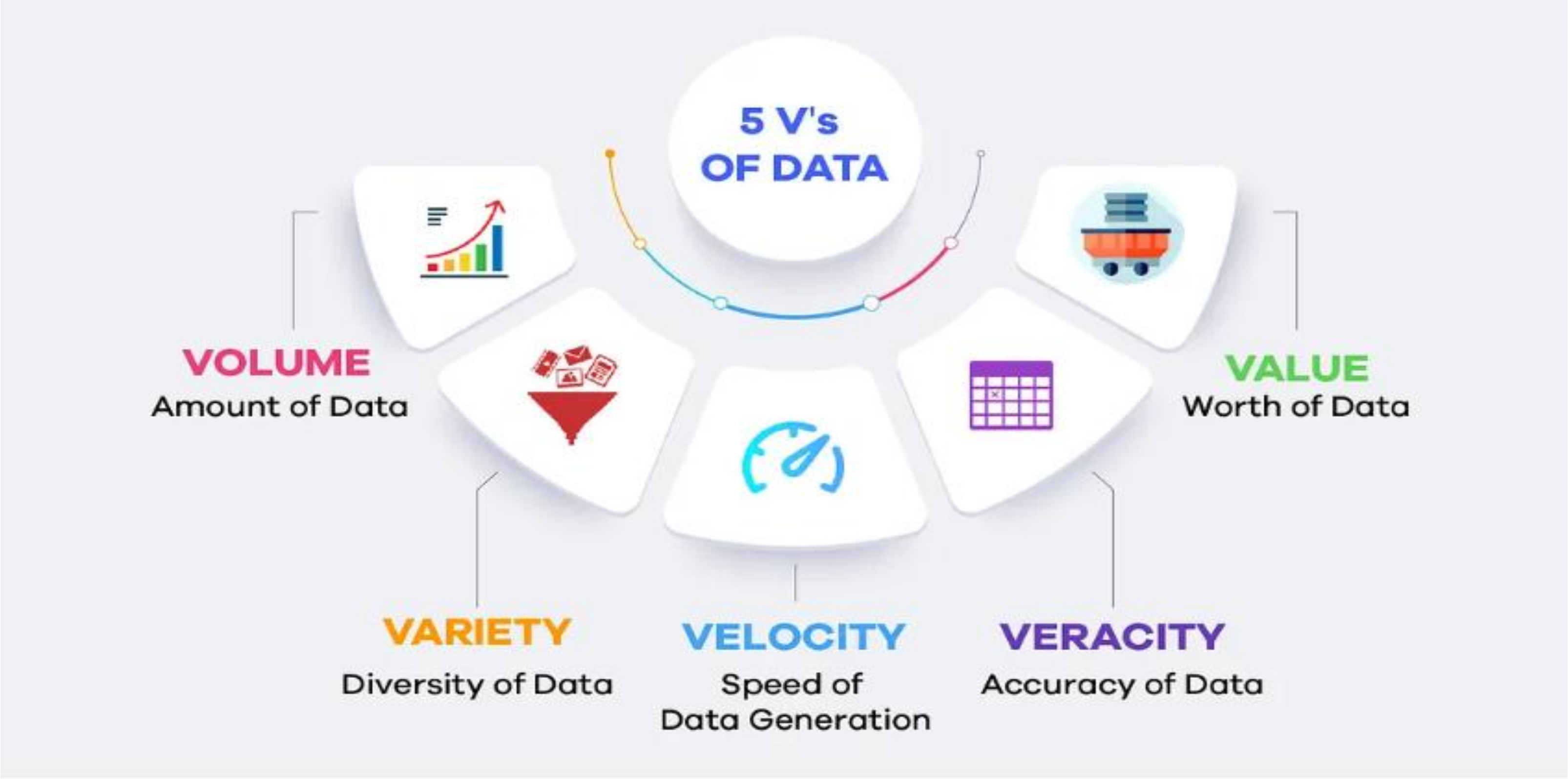
WHAT IS BIG DATA ANALYTICS?

- **Definition:** Big Data Analytics is the complex process of **examining large and varied data sets** (Big Data) to uncover hidden patterns, unknown correlations, market trends, customer preferences, and other useful information.

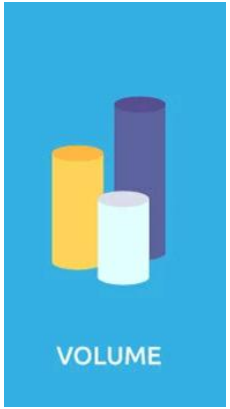
- The goal is to help organizations make **more-informed business decisions**.

- It goes beyond traditional data analysis by using advanced analytic techniques and technologies designed to handle data at a massive scale and speed.

THE 5 VS OF BIG DATA (CHARACTERISTICS)



VOLUME & VELOCITY



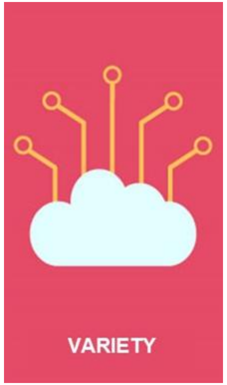
Volume

- Refers to the **immense amount of data** generated and stored.
- Scale has moved from gigabytes to terabytes, petabytes, and even exabytes.
- **Example:** Daily data from social media platforms like Facebook, sensor data from smart devices (IoT), or scientific research data

Velocity

- Refers to the **high speed at which data is generated** and must be processed.
- Data is often streamed in near real-time.
- **Example:** Stock market trading data, live GPS tracking, or real-time monitoring of website traffic.

VARIETY & VERACITY



Variety

- Refers to the **different types of data** available.
- **Structured:** Highly organized data (e.g., SQL databases, spreadsheets).
- **Semi-structured:** Data with some organizational properties (e.g., JSON, XML files).
- **Unstructured:** Unorganized data (e.g., text documents, emails, videos, audio files, images).

Veracity

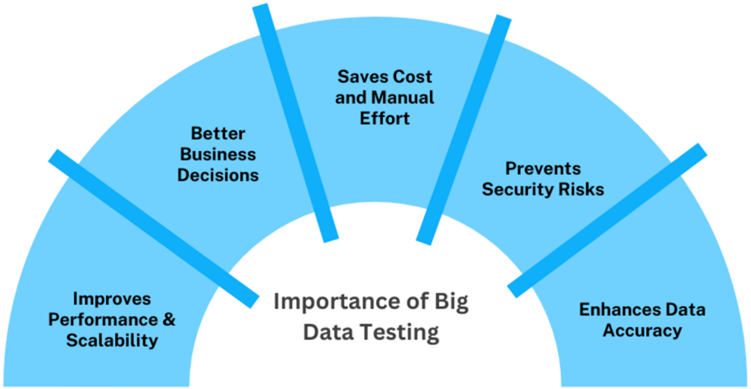
- Refers to the **quality, trustworthiness, and accuracy** of the data.
- Big Data can be messy, inconsistent, and contain biases or abnormalities.
- **Example:** Inaccurate sensor readings due to weather, typos in customer entry forms, or fake social media profiles. Ensuring veracity is crucial for reliable analysis.

VALUE

Value

- This is the most important 'V'. It refers to the **usefulness and tangible benefit** derived from analyzing the data.
- Simply having Big Data is not enough; it must be turned into **actionable insights** that create value.
- **Example:** Analyzing customer purchase history (Volume, Variety) in real-time (Velocity) with high accuracy (Veracity) to provide personalized recommendations that increase sales (Value).

IMPORTANCE OF BIG DATA ANALYTICS



Saves Cost and Manual Effort
Better Business Decisions
Prevents Security Risks
Improves Performance & Scalability
Importance of Big Data Testing
Enhances Data Accuracy

- **Better Decision Making:** Data-driven decisions are more accurate than those based on intuition alone.

- **Increased Efficiency & Cost Reduction:** Identify inefficiencies in operations, supply chains, and marketing to save money.

- **Product & Service Innovation:** Analyze customer needs and trends to develop new products and enhance existing ones.

- **Enhanced Customer Experience:** Understand customer behavior to deliver personalized services, targeted marketing, and improved support.

- **Risk Management:** Detect fraudulent activities and predict potential risks to the business.

CHALLENGES OF BIG DATA

- **Data Storage and Management:** Storing and managing massive volumes of data is costly and complex.

- **Data Security and Privacy:** Protecting sensitive data from breaches and ensuring compliance with regulations (like GDPR) is a major concern.

- **Data Quality and Veracity:** Ensuring the data is clean, accurate, and reliable for analysis is a constant struggle.

- **Accessibility of Data**:The challenge of making data available to the right people and systems when needed, while still maintaining security.

THE BIG DATA ECOSYSTEM

The Big Data Ecosystem refers to the **entire infrastructure, tools, and applications** used to manage and analyze Big Data. It's a collection of components that work together.

KEY COMPONENTS OF THE BIG DATA ECOSYSTEM

1. Data Sources:

- Where data originates: IoT devices, social media, enterprise applications (ERP, CRM), weblogs, public data, etc.

2. Data Ingestion & Storage:

- Process of moving data from sources to a storage system.
- **Tools:** Apache Kafka, Flume.
- **Storage:** Data Lakes (e.g., Amazon S3), Distributed File Systems (like **HDFS**), and NoSQL databases (e.g., MongoDB, Cassandra).

KEY COMPONENTS OF THE BIG DATA ECOSYSTEM

3. Data Processing & Analytics:

- Where the core analysis happens.
- **Processing Frameworks: Apache Spark** (fast, in-memory processing) and **Hadoop MapReduce** (batch processing).
- **Analytics Tools:** SQL engines (Presto, Hive), Machine Learning libraries (Scikit-learn, TensorFlow), and statistical software (R).

4. Data Visualization & Consumption:

- The final layer where insights are presented to users.
- **Tools:** Business Intelligence (BI) tools like **Tableau**, Power BI, and custom dashboards.

SUMMARY & CONCLUSION

- **Big Data Analytics** unlocks powerful insights from massive, fast-moving, and divers datasets.

- The **5 Vs** (Volume, Velocity, Variety, Veracity, Value) define its scope and challenges.

- While it offers tremendous **importance** for business growth and innovation, it comes with significant **challenges** in storage, security, and skills.

- A robust **Big Data Ecosystem** is essential to effectively ingest, store, process, and analyze data to generate real value.

BOOKS AND REFERENCES

- **Big Data: Principles and Best Practices of Scalable Real-Time Data Systems** By: Marz, N., & Warren, J.
- **Practical Big Data Analytics: Hands-on Techniques to Implement Enterprise Analytics and Machine Learning Using Hadoop, Spark, NoSQL and R** By: Gulla, U., Gupta, S., & Kumar, V.
- **Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data** By: Zikopoulos, P. C., Eaton, C., & deRoos, D.

BIG DATA ANALYTICS (CS-431)

Dr. Sriparna Saha
Associate Professor

Website: https://www.iitp.ac.in/~sriparna/
Google Scholar: https://scholar.google.co.in/citations?user=Fj7jA_AAAAAJ&hl=en
Research Lab: SS_Lab
Core Research AREA: NLP, GenAI, LLMs, VLMs, Multimodality, Meta-Learning, Health Care, FinTech, Conversational Agents

TAs: Sarmistha Das, Nitish Kumar, Divyanshu Singh, Aditya Bhagat, Annu Kumari, Harsh Raj

Big Data Enabling Technologies?

- **Big Data is used for a collection of data sets so large and complex that it is difficult to process using traditional tools.**

A recent survey says that 80% of the data created in the world are unstructured.

One challenge is how we can store and process this big amount of data. In this lecture, we will discuss the top technologies used to store and analyse Big Data.

WHAT IS BIG DATA ANALYTICS?

- **Definition:** Big Data Analytics is the complex process of **examining large and varied data sets** (Big Data) to uncover hidden patterns, unknown correlations, market trends, customer preferences, and other useful information.

- The goal is to help organizations make **more-informed business decisions**.

- It goes beyond traditional data analysis by using advanced analytic techniques and technologies designed to handle data at a massive scale and speed.