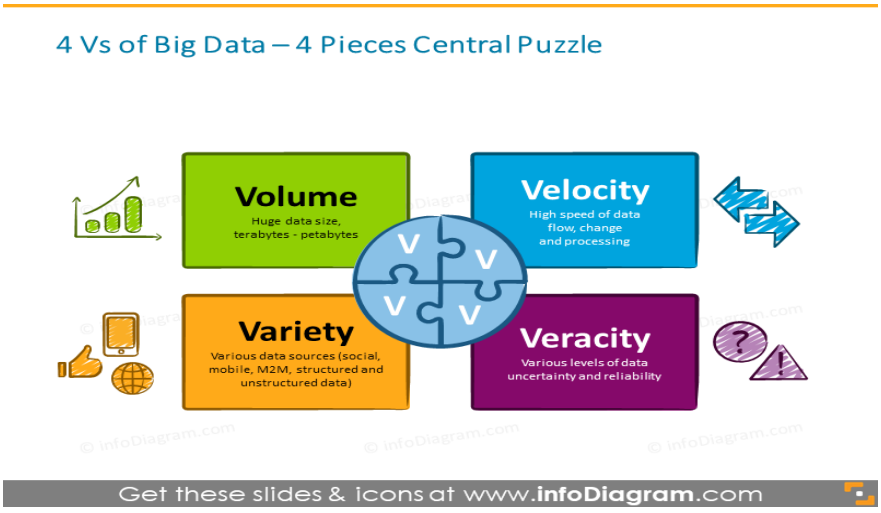
BIG DATA

DHAMINI MACHIREDDY

Big data refers to extremely large and complex datasets that are difficult to manage and process using traditional data processing tools. It's characterized by high volume, velocity, and variety of data, encompassing structured, semi-structured, and unstructured information. Big data analysis provides valuable insights for improved decision-making, process automation, and innovation.

The 4 V's of Big Data are Volume, Velocity, Variety, and Veracity. These characteristics describe the key challenges and opportunities associated with managing and analyzing massive datasets.



- **Volume:**

This refers to the sheer quantity of data being generated and stored. Big data often involves massive datasets, requiring specialized infrastructure and processing techniques.

- **Velocity:**

This describes the speed at which data is generated and processed. Real-time data streams, like those from social media or sensor networks, require efficient processing to extract value quickly.

- **Variety:**

This refers to the different formats and types of data that exist within a big data set. This can include structured data (databases), unstructured data (text, images, videos), and semi-structured data (JSON, XML).

- **Veracity:**

This addresses the quality, accuracy, and reliability of the data. With the vast amount of data available, it's crucial to ensure that the data being used for analysis is trustworthy and representative of the real world

Big Data Architecture

Big Data architecture is a framework of tools and processes used to store, process, and analyze large datasets.

- **Main Components:**

1. **Data Sources** – IoT devices, social media, apps, sensors, web logs, transactions.
2. **Ingestion Layer** – Collects data in real-time or batches
    - Tools: Apache Flume, Kafka, Sqoop
3. **Storage Layer** – Stores structured/unstructured data
    - Tools: HDFS, NoSQL databases, Data Lakes (e.g., Amazon S3, Azure Data Lake)
4. **Processing Layer** – Handles computation/analysis
    - Batch: Hadoop MapReduce
    - Real-time: Spark, Storm, Flink
5. **Analytics Layer** – Extracts insights using ML/AI
    - Tools: Python, R, TensorFlow, Spark MLlib
6. **Visualization Layer** – Converts data into readable formats
    - Tools: Tableau, Power BI, Grafana

Big Data Lifecycle

1. **Data Generation** – From devices, logs, transactions, social media, etc.
2. **Data Collection** – Through APIs, file uploads, web scraping, etc.
3. **Data Storage** – Data is stored in HDFS, cloud storage, or databases.
4. **Data Processing** – Cleaned, formatted, and analyzed using processing engines.
5. **Data Analysis** – Insights are extracted using statistical or machine learning models.
6. **Data Visualization** – Results are shown via dashboards, graphs, or reports.
7. **Decision Making** – Insights help businesses make informed decisions.