BIG DATA MANAGEMENT

2025 CGNPH SUMMER SCHOOL

PRESENTED BY DR. COMFORT FOLORUNSO
DEPARTMENT OF SYSTEMS ENGINEERING,
UNIVERSITY OF LAGOS

OUTLINE

- What is Big Data?
- What is Big Data?
- Characteristics of Big Data
- Characteristics of Big Data
- Sources of Big Data
- Big Data Lifecycle
- Big Data Storage and Management
- Tools for Big Data Analytics
- Benefit of Big Data
- Applications of Big Data
- Challenges of Big data

What is Big Data?

- Big data refers to extremely large and diverse collections of structured, unstructured, and semi-structured data that continue to grow exponentially over time.

- These datasets are so huge and complex in volume, velocity, and variety, that traditional data management systems cannot store, process, and analyze them.

Examples of Big Data

- Tracking consumer behavior and shopping habits to deliver hyper-personalized retail product recommendations tailored to individual customers
- Monitoring payment patterns and analyzing them against historical customer activity to detect fraud in real time.
- Analyzing public datasets of satellite imagery and geospatial datasets to visualize, monitor, measure, and predict the social and environmental impacts of supply chain operations.
- Sequencing, Surveillance, Genomics and Phenotypes Data directly from field/clinics or database.

- Combining data and information from every stage of an order's shipment journey with hyperlocal traffic insights to help fleet operators optimize last-mile delivery

- Using AI-powered technologies like natural language processing (NLP) to analyze unstructured medical data (such as research reports, clinical notes, and lab results) to gain new insights for improved treatment development and enhanced patient care

- Using image data from cameras and sensors, as well as GPS data, to detect potholes and improve road maintenance in cities.

- Using pathologic image data from H and E, Immunohistochemistry microscopic scanning cameras for detection of diseases.

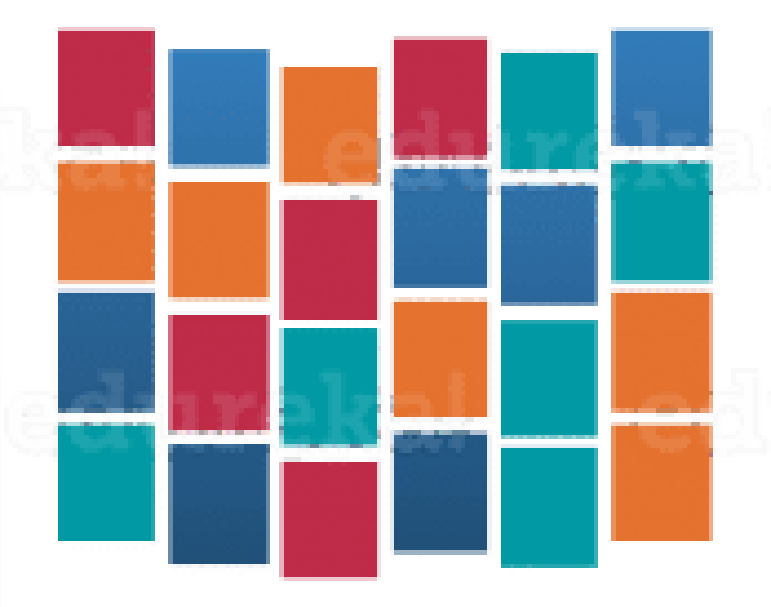
Types of Big Data

1. **Structured Data:** It has a well-defined structure, it follows a consistent order and it is designed in such a way that it can be **easily accessed** and used by a person or a computer. Structured data is usually stored in well-defined columns and rows or in Databases.

- Example: Database Management Systems (**DBMS**), excel file.

2. Semi-Structured Data: It inherits a few properties of Structured Data, but the major part of this kind of data fails to have a definite structure and also, it does not obey the formal structure of data models such as Relational Database Management System (RDBMS).

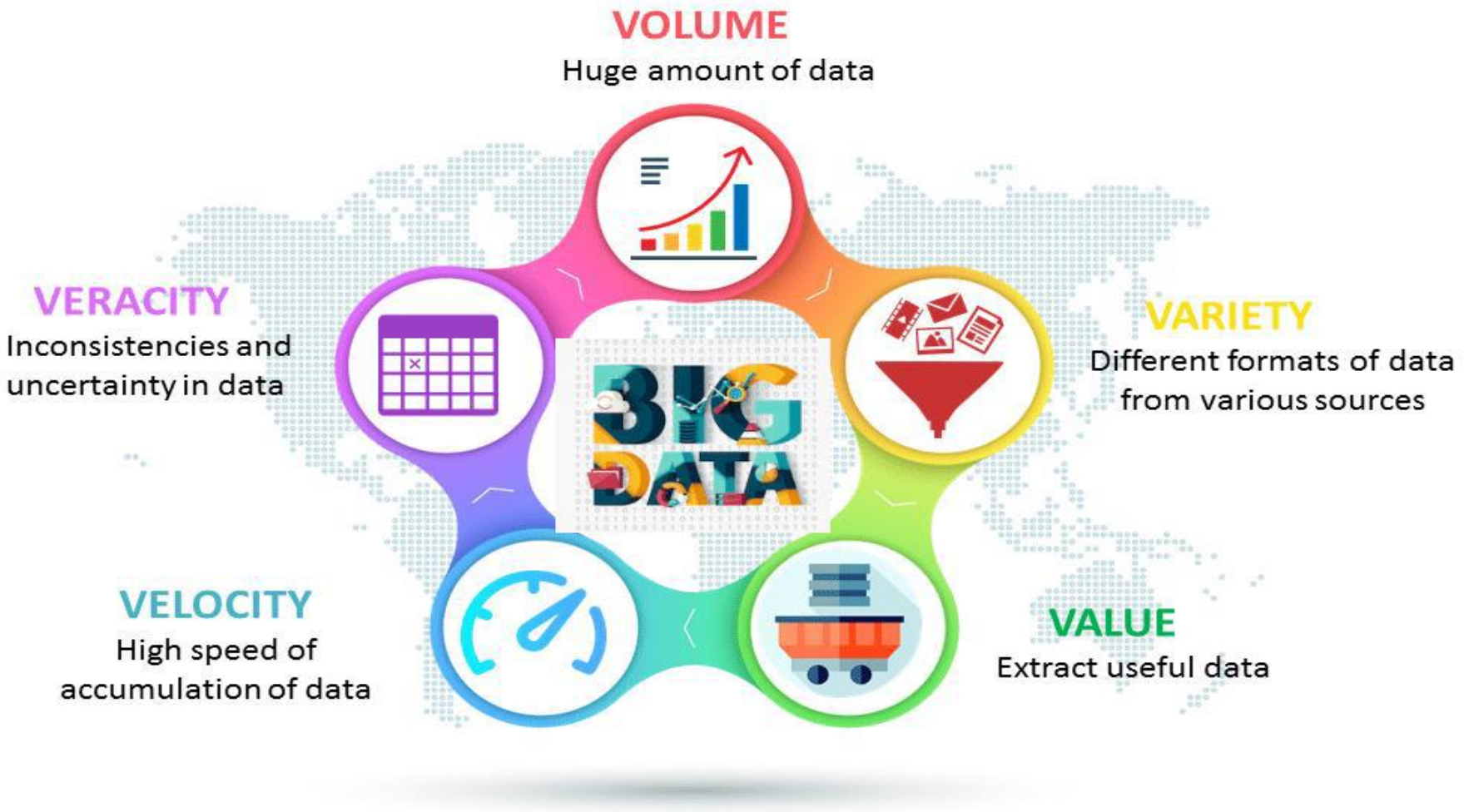
- Example: Comma Separated Values (**CSV**) File.



3. **Unstructured Data:** is completely different in that it neither has a structure nor obeys the formal structural rules of data models. It does not even have a consistent format and it is found to be varying all the time.

Examples: Audio files, Images, videos, social media post (such as tweets) etc

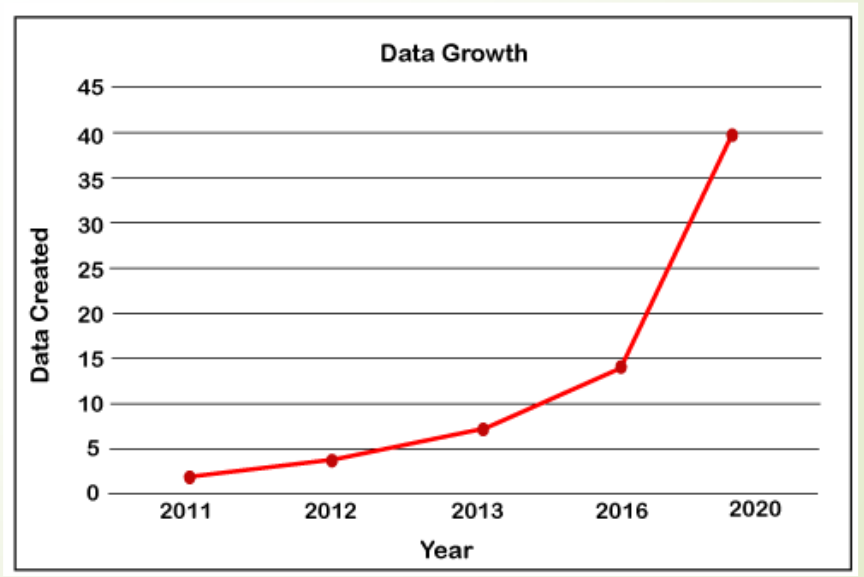
Characteristics of Big Data



1. Volume

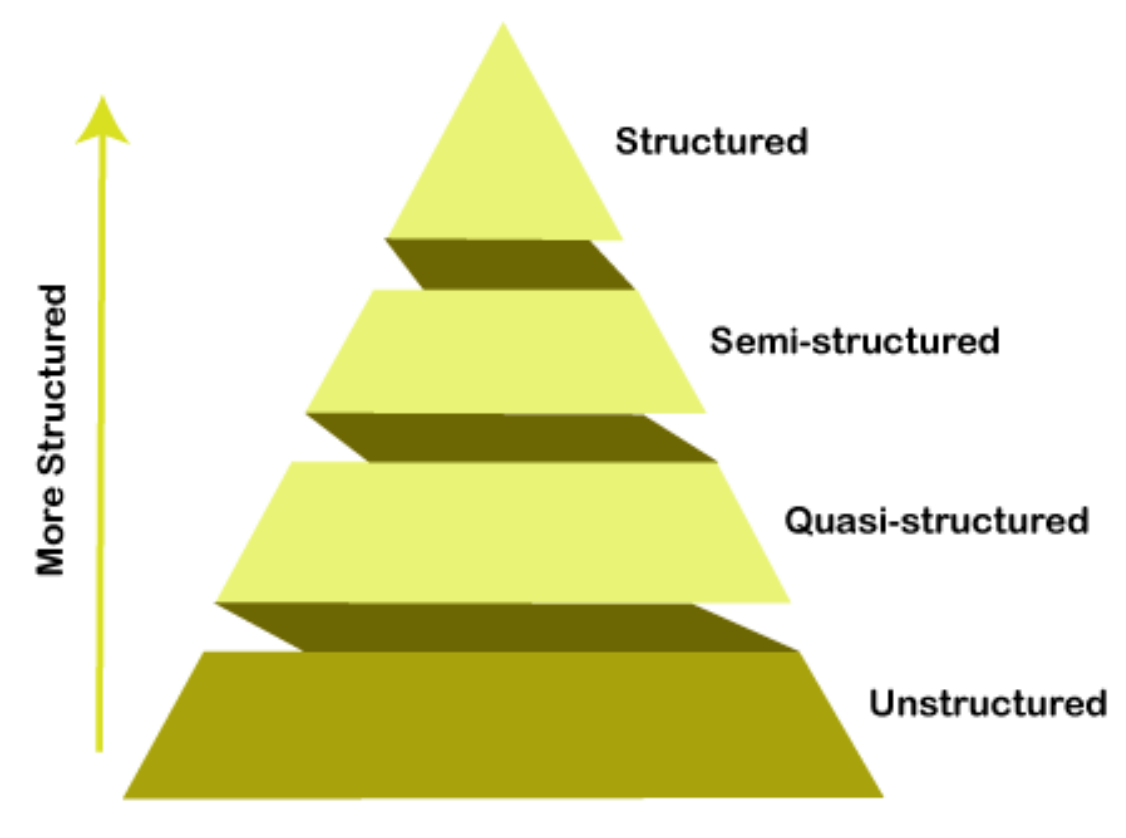
Volume refers to the unimaginable amounts of information generated every second from social media, cellphones, credit cards, machine-to-machine (M2M) sensors, images, video, etc. The **distributed systems**, uses software Framework like **Hadoop** to store data in several locations.

- Facebook alone can generate approximately **4 petabytes (PB)**, or **4 million gigabytes**, of user-generated content on a daily basis (Osman, 2024).



2. Variety

Big Data is generated in multiple varieties. Compared to the traditional data like phone numbers and addresses, the latest trend of data is in the form of photos, videos, and audios and many more, making about 80% of the data to be completely unstructured.

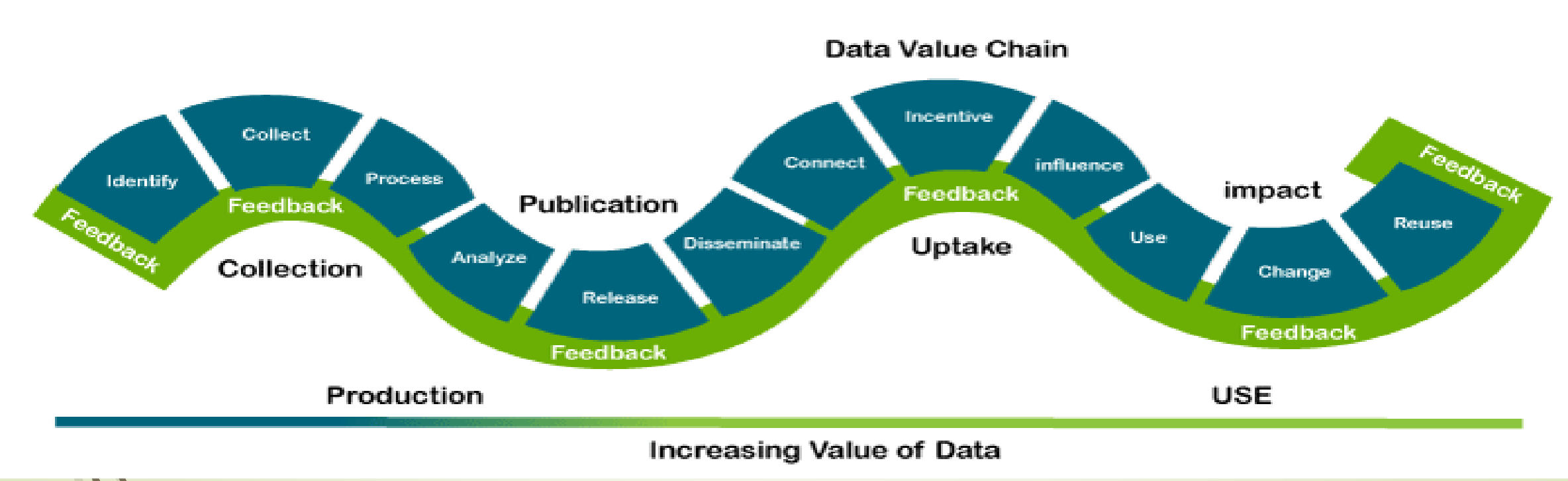


3. Veracity

Veracity basically means the degree of reliability that the data has to offer. Since a major part of the data is unstructured and irrelevant, Big Data needs to find an alternate way to filter them or to translate them out as the data is crucial in business developments. The higher the veracity of the data, the more trustworthy it is.

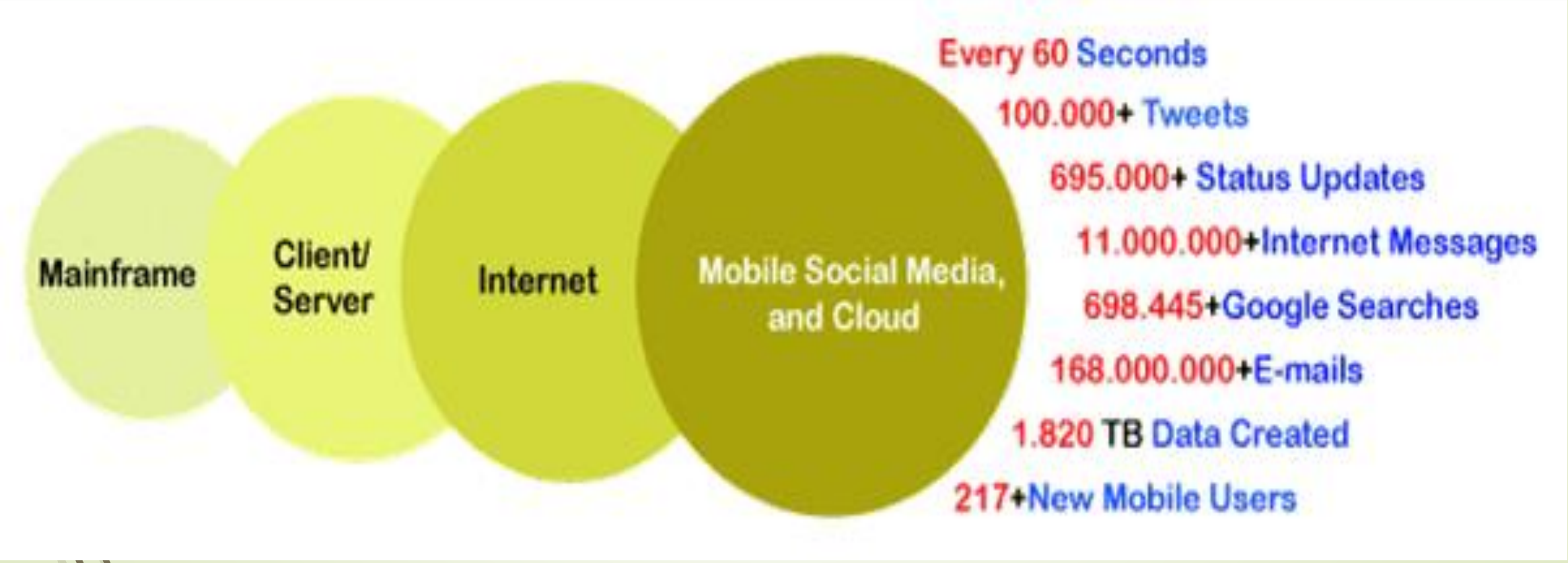
4. Value

It's essential to determine the business value of the data that is collected. Big data must contain the right data and then be effectively analyzed in order to yield insights that can help drive decision-making.



5. Velocity

This describes the enormous amount of data that is available for collection and produced from a variety of sources and devices on a continuous basis.



Other added characteristics include:

- **Variability** – inconsistency of the data flow (e.g., daily sales may spike unpredictably).
- **Visualization** – representing large complex data sets in understandable charts or dashboards.
- **Volatility** – how long data remains relevant and needs to be stored.

Sources of Big Data

- Social networks (Facebook, Twitter)
- IoT devices & sensors
- Business transactions & Customer Relationship Management (CRM)
- Mobile apps & web logs
- Images, audio, videos
- Database – NCBI, Scopus, etc.

Big Data Lifecycle

- **Acquisition:** Collection from various sources
- **Storage:** On-premise / cloud databases, data lakes
- **Processing:** Batch or real-time analytics
- **Visualization:** Dashboards, reports
- **Archiving:** Long-term storage or deletion

Big Data Storage and Management

Data Storage

- **Distributed storage:** Hadoop Distributed File Systems (HDFS), Amazon S3
- **Databases:** NoSQL (MongoDB, Cassandra)
- **Data lakes:** Store raw data
- **Data warehouses:** Structured, optimized for queries

Data Processing

- **Batch processing:** Hadoop MapReduce
- **Real – time processing:** Apache Kafka, Flink
- **Machine learning:** Spark MLlib, TensorFlow on big datasets

Data Management

- **Data Governance & Quality**
  - Define policies for data access & usage
  - Maintain data lineage & metadata
  - Regular data quality checks & cleaning
  - Ensure regulatory compliance (GDPR, HIPAA)
- **Data quality**
- **Data Security**

Data Analysis

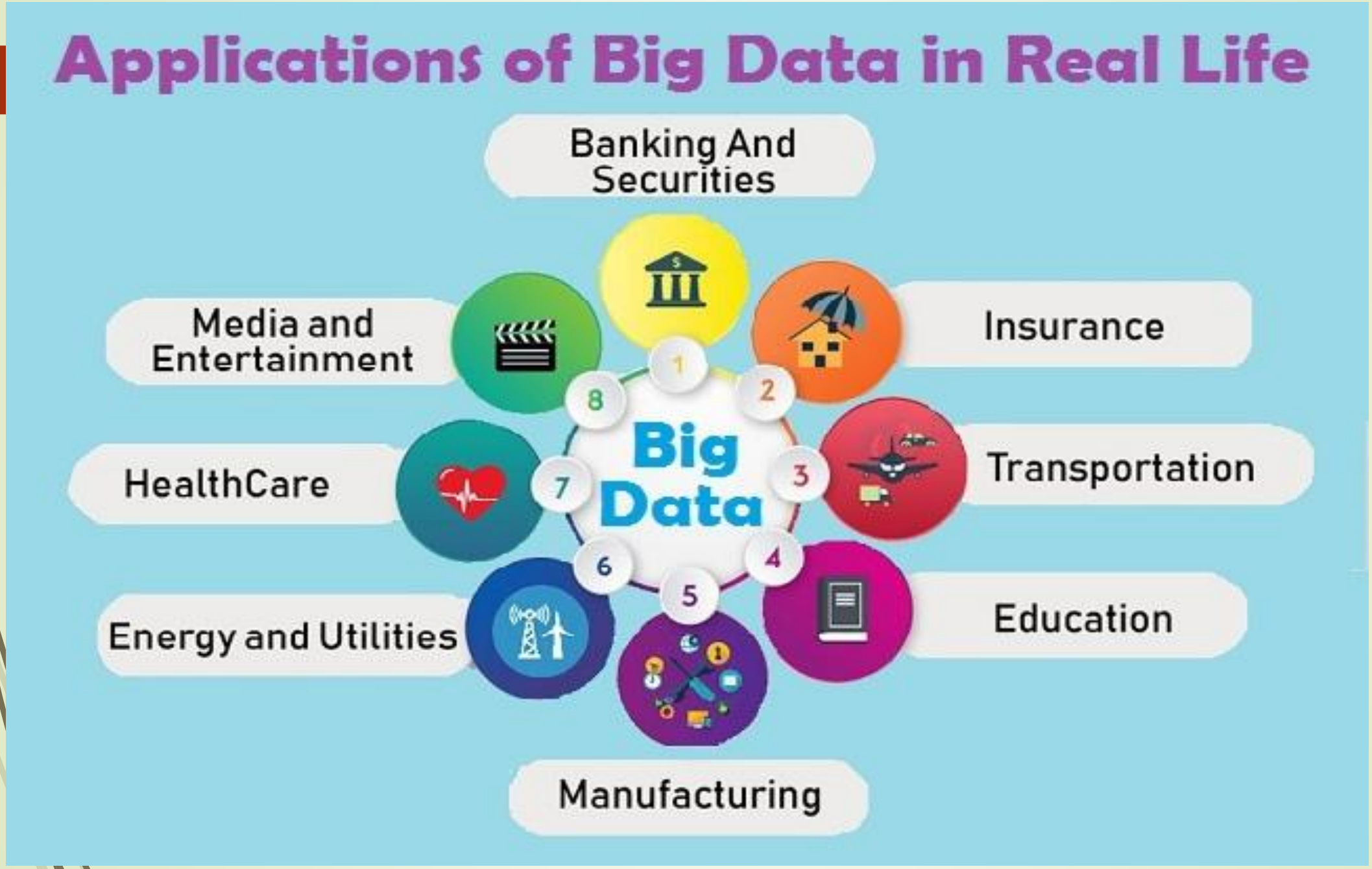
- Data Mining
- Data visualization

Tools for Big Data Analytics

1. **Hadoop-** This tool helps in reserving and studying information.
2. **MongoDB-** It is used on datasets that change constantly.
3. **Talend-** It is helpful in data combining and controlling.
4. **Cassandra**– It is an assigned database use to manage chunks of data.
5. **Spark**– It is helpful in immediate changing and studying a large quantity of information.
6. **STORM**– This is an open-source actual computing arrangement.
7. **Kafka**– This is a distributed running program that keeps fault-tolerant data.

Benefits of Big Data

Gourav, 2021


Applications of Big Data in Real Life
Banking And Securities
Insurance
Transportation
Education
Manufacturing
Energy and Utilities
HealthCare
Media and Entertainment
Big Data
1
2
3
4
5
6
7
8

Gourav, 2021

Challenges of implementing big data analytics

1. Lack of data talent and skills
2. Speed of data growth
3. Problems with data quality
4. Compliance violations
5. Integration Complexity
6. Security concerns

Conclusion

- Big Data is an asset if managed well
- Combines tech, governance & skilled teams
- Key to innovation & competitive advantage

Reference

- Gourav, 2021. An Introductory Guide to Big Data Analytics. https://www.analyticsvidhya.com/blog/2021/09/an-introductory-guide-to-big-data-analytics/

- Maddy Osram, 2024. " Wild and Interesting Facebook Statistics and Facts" _https://kinsta.com/blog/facebook-statistics/?utm_source=chatgpt.com

- Cloud Google. https://cloud.google.com/learn/what-is-big-data