IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252

Big Data Analytics

Lalit Sehrawat
Dronacharya College of Engineering, Gurgaon, Haryana, India

I. WHAT IS BIG DATA ANALYTICS?

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency. Big data has one or more of the following characteristics: high volume, high velocity or high variety. Artificial intelligence (AI), mobile, social and the Internet of Things (IoT) are driving data complexity through new forms and sources of data. For example, big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media — much of it generated in real time and at a very large scale.

Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable. Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously untapped data sources independently or together with existing enterprise data.

1.1 What is Big Data

Data which are very large in size is called Big Data. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data. It is stated that almost 90% of today's data has been generated in the past 3 years.

Sources of Big Data

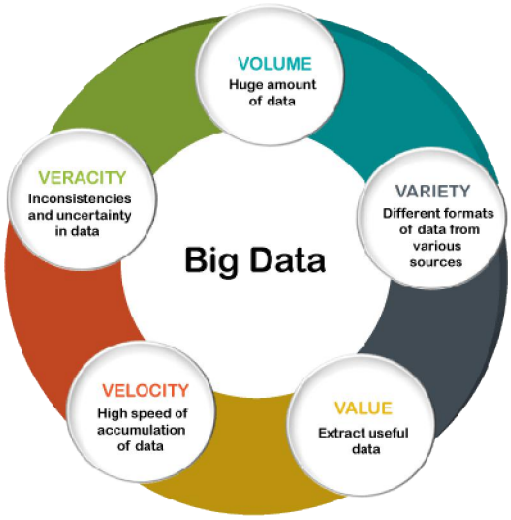
These data come from many sources like

- **Social Networking Sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
- **E-Commerce Site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.
- **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- **Telecom Company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
- **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

1.2 5 V's of Big Data

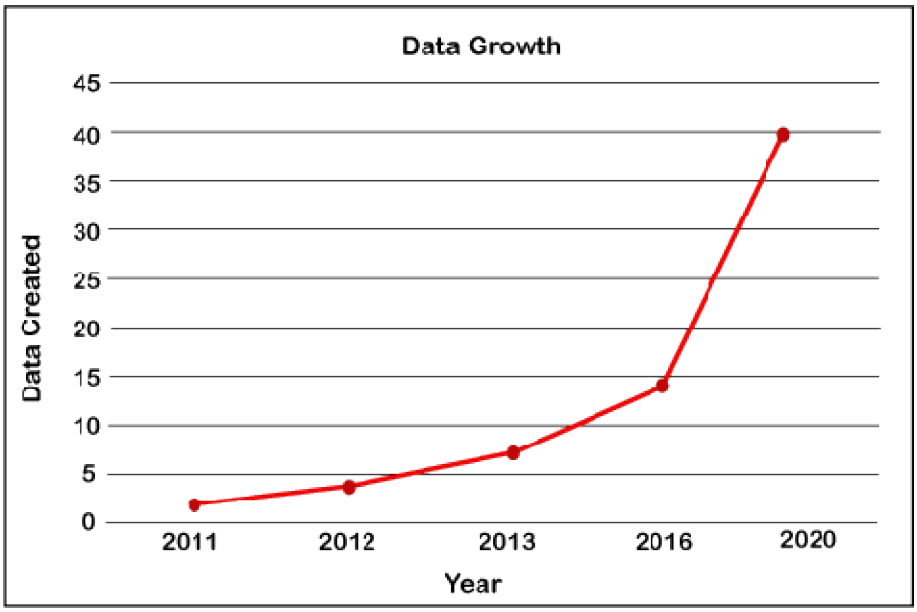
- **Volume**
- **Veracity**
- **Variety**
- **Value**
- **Velocity**

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252



A. Volume

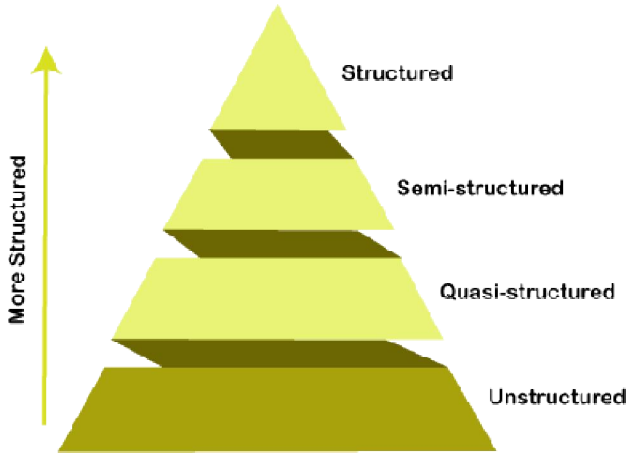
The name Big Data itself is related to an enormous size. Big Data is a vast 'volumes' of data generated from many sources daily, such as **business processes, machines, social media platforms, networks, human interactions,** and many more. **Facebook** can generate approximately a **billion** messages, **4.5 billion** times that the "**Like**" button is recorded, and more than **350 million** new posts are uploaded each day. Big data technologies can handle large amounts of data.



B. Variety

Big Data can be **structured, unstructured, and semi-structured** that are being collected from different sources. Data will only be collected from **databases** and **sheets** in the past, But these days the data will comes in array forms, that are **PDFs, Emails, audios, SM posts, photos, videos,** etc.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252



The data is categorized as below:

1. **Structured data:** In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.
2. **Semi-structured:** In Semi-structured, the schema is not appropriately defined, e.g., **JSON, XML, CSV, TSV**, and **email**. OLTP (**Online Transaction Processing**) systems are built to work with semi-structured data. It is stored in relations, i.e., **tables**.
3. **Unstructured Data**: All the **unstructured files, log files, audio files**, and **image** files are included in the unstructured data. Some organizations have much data available, but they did not know how to **derive** the value of data since the data is raw.
4. **Quasi-structured Data:**The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools.

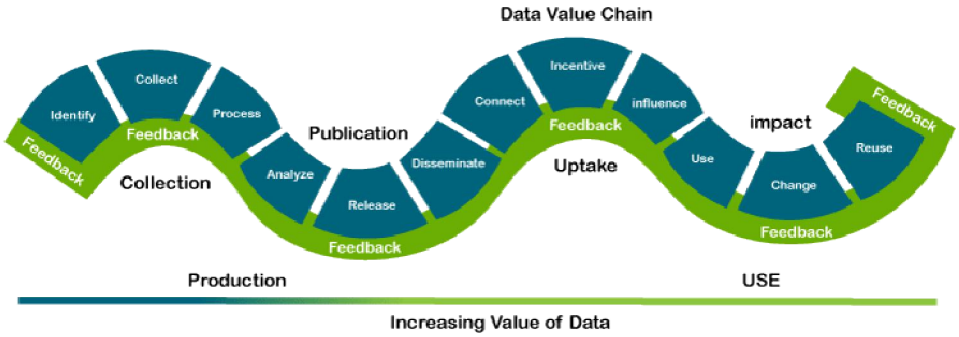
Example: Web server logs, i.e., the log file is created and maintained by some server that contains a list of **activities**.

C. Veracity

Veracity means how much the data is reliable. It has many ways to filter or translate the data. Veracity is the process of being able to handle and manage data efficiently. Big Data is also essential in business development. For example, **Facebook posts** with hashtags.

D. Value

Value is an essential characteristic of big data. It is not the data that we process or store. It is **valuable** and **reliable** data that we **store, process**, and also **analyze**.

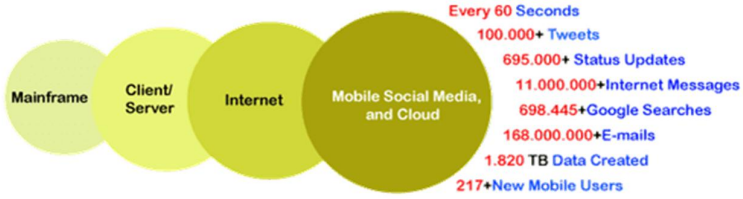


IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, June 2022
Impact Factor: 6.252

E. Velocity

Velocity plays an important role compared to others. Velocity creates the speed by which the data is created in **real-time**. It contains the linking of incoming **data sets speeds, rate of change**, and **activity bursts**. The primary aspect of Big Data is to provide demanding data rapidly.

Big data velocity deals with the speed at the data flows from sources like **application logs, business processes, networks, and social media sites, sensors, mobile devices,** etc.



Applications of Big Data

1. Travel and Tourism
2. Financial and banking sector
3. Healthcare
4. Telecommunication and media
5. Government and Military
6. E-commerce
7. Social Media

Scalable architectures for parallel data processing:

Hadoop or Spark kind of environment is used for offline or online processing of data. The industry is looking for scalable architectures to carry out parallel data processing of big data. There is a lot of progress in recent years, however, there is a huge potential to improve performance.

Handling real-time video analytics in a distributed cloud:

With the increased accessibility to the internet even in developing countries, videos became a common medium of data exchange. There is a role of telecom infrastructure, operators, deployment of the Internet of Things (IoT), and CCTVs in this regard. Can the existing systems be enhanced with low latency and more accuracy? Once the real-time video data is available, the question is how the data can be transferred to the cloud, how it can be processed efficiently both at the edge and in a distributed cloud?

Efficient graph processing at scale:

Social media analytics is one such area that demands efficient graph processing. The role of graph databases in big data analytics is covered extensively in the reference article [4]. Handling efficient graph processing at a large scale is still a fascinating problem to work on. *The research problems to* ***handle noise and uncertainty in the data.***

Identify fake news in near real-time:

This is a very pressing issue to handle the fake news in real-time and at scale as the fake news spread like a virus in a bursty way. The data may come from Twitter or fake URLs or WhatsApp. Sometimes it may look like an authenticated source but still may be fake which makes the problem more interesting to solve

REFERENCES

[1]. https://www.javatpoint.com/what-is-big-data
[2]. https://www.javatpoint.com/big-data-characteristics
[3]. https://towardsdatascience.com/top-20-latest-research-problems-in-big-data-and-data-science-c6fb51e03136