IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

The major boom of an Unstructured Big Data on Social Media

Lata V. Upadhyay[1], Dr. Dipti H. Domadiya[2]

[1]Assistant Professor, M.V.M. College of Com., Mgt. & I.T., Rajkot.

[2]Associate Professor, National Computer College, Jamnagar.

ABSTRACT

Big Data refers to the rapidly growing volume, variety, value, veracity, and velocity of data being generated in the modern digital world. It looks at the different kinds and sources of Big Data, such as structured data, semi-structured data, and unstructured data, highlighting the growing significance of the sources and elements of unstructured data from the social media perspective, Time Series Data, Geospatial Data, and Streaming Data. The three main challenges of Big Data, including characteristic challenges, processing challenges, and management challenges, are highlighted in this paper. This paper presents an overview of Big Data's characteristics, types, challenges, and various social media platforms. In conclusion, organizations not at all longer neglect unstructured data today; relatively, they are inventing means of evaluating it to extract information.

KEYWORDS

Big Data, Characteristics, Volume, Variety, Value, Velocity, Veracity, Structured, Semi-structured, Unstructured, Time Series Data, Geospatial Data, Streaming Data, Challenges, and Social Media Platforms.

INTRODUCTION

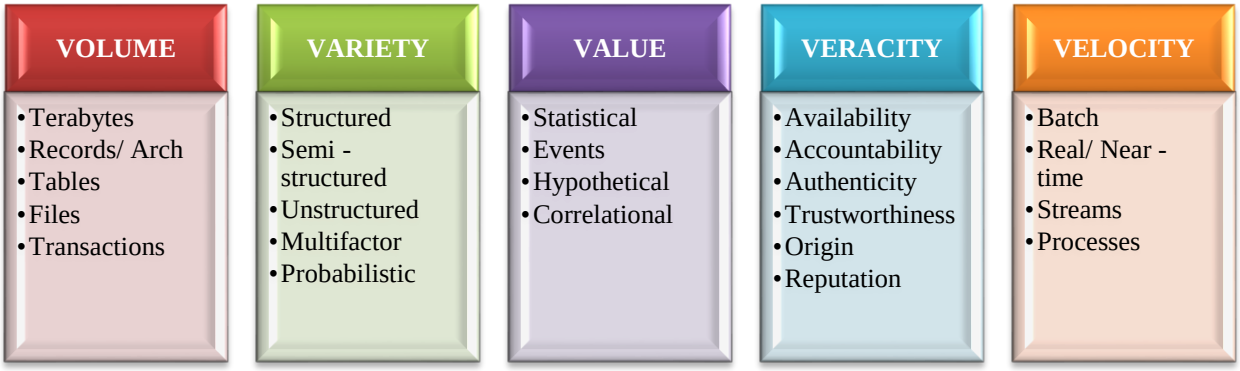
The term Data means a raw material that is limitless and meaningless and available on the earth. Big Data is similar to data, but it is more complex and bigger, which may come from new data sources. So, Big Data can be defined as "the collection, selection, and analysis of data on the large data set." In other words, Big Data refers to extraordinarily massive and intricate data collections that are challenging to manage, process, and analyze using conventional data processing methods.

The term “Big Data” was first introduced to the computing world by Roger Magoulas from O'Reilly Media in 2005 to define a great amount of data that traditional data management techniques cannot manage and process due to the complexity and size of this data [1]. Technological advances in recent years have led to a significant

amount of data that is now generated in everyday life, such as shopping, traveling, banking, manufacturing, trading, public utilities, state, governance, sports, entertainment, science, education, and health [2].

V'S OF BIG DATA CHARACTERISTICS

It includes the five V's of Big Data, which are volume, variety, value, velocity, and veracity of data. Figure - 1 shows five V's of Big Data Characteristics:



[Figure - 1: An overview of the Five V's of Big Data]

Volume

Big data engages the analysis and processing of data at an enormous scale. It refers to the vast amount of data generated from various sources such as sensors, social media, transactions, logs, and more. Data is available at a larger scale in the order of Terabytes or Petabytes and beyond [3]. Data storage devices start from Floppy discs, CD-ROM (Compact Disc Read-Only Memory), DVD (Digital Versatile Disc), USB (Universal Serial Bus), and Hard disks, and at present, on cloud is the biggest example of Big Data Volume.

Variety

Variety is a key aspect of Big Data; data is available in different forms like audio, images, videos, social networking data, text, tables, etc. [5]. Big Data covers varied data types and formats. It comprises Structured Data (traditional relational databases), Unstructured Data (text documents, images, videos, audio files, social media posts, sensor data, etc.). Big data may also consist of Semi-Structured Data which has several organizational properties, although it does not fit precisely into a relational database formation. A common use of Big Data processing is to take unstructured data and extract ordered meaning for consumption either by humans or as a structured input to an application [11].

Value

Value refers to the important feature of Big, Data which is defined by the added value that the collected data can bring to the intended process, activity, or predictive analysis/hypothesis [4]. No matter how quickly or how much data is created, it must be trustworthy and useful. The data would not be suitable for processing or analysis in any other case. According to research, poor-quality data can cause a company to lose nearly 20% of its income.

Velocity

Big Data is often available in real-time and the rate at which it grows is very high [3]. Velocity "means both how fast data is being produced and how fast the data must be processed to meet demand" [11].Velocity describes the speed at which data is created and transferred. For businesses that require their data to flow fast so that it is accessible when needed to make the best business decisions, this is a crucial factor. Data may be emitted by devices, networks, click streams, smart phones, sensor readings, or social media. This information must be swiftly processed and analyzed, often in near real-time. Processing increases effectiveness. The data velocity depends highly on the proliferation of mobile devices and other device sensors connected to the Internet [9].

Veracity

Veracity means the accuracy of collected data; when we are working with a large volume of data, it's not possible to get 100% correct data [5]. Veracity specifies the level of data reliability. In traditional data warehouses, there was always the assumption that the data was certain, clean, and precise [11]. The information gathered can be incomplete, erroneous, or unable to offer meaningful, useful insight. It is crucial to remove the information that is not necessary and use the remaining data for processing because the majority of the data you encounter is unstructured.

Generally, Big Data refers to the huge, diverse, speedily generated, and precious datasets that necessitate advanced analysis and processing methods to extract important insights and gain business value.

CONTENT FORMAT OF BIG DATA

Big data content formats can also be used for the classification of data. Based on the content format, Big Data can be divided into three categories, namely structured, semi-structured, and unstructured. Furthermore, Time Series Data, Geospatial Data, and Streaming Data are also part of Big Data types. Three Basic sources through which these types of data are generated are listed here in below:

[Figure - 2: An overview of types of Big Data]

Structured Data

Structured Data means data that is stored and organized in a predefined arrangement, usually in tabular form using rows and columns. Structured data is understood by computers and is also efficiently organized for human readers [11]. Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource [7]. It pursues a fixed plan and can be easily analyzed and processed using conventional database management systems. Examples are spreadsheets; data stored in relational databases, and structured Extensible Markup Language files.

Unstructured Data

Unstructured Big Data refers to huge volumes of data that do not contain an organized or predefined format. Big Data is only getting bigger 90% of the data in the world today was created within the last two years [11]. It is typically in the form of images, video, audio, text, emails, sensor data, blog posts, geo data, social media posts, and other types of content. A significant portion of the unstructured content collected from social media is text [6]. Various elements of unstructured source types, such as Blog Post, contain the date and time of the post, embedded links, content, comments, author, etc. Geo Data elements include date and time, real-time recording, initial data, number of flows, and location. Opinions, comments, preferences, types of posts, needs and desires, interests, number of posts, date and time of posts, fan views, followers, mentions, likes, check-ins, +1s, pins, etc., are various elements of social media. Unstructured data pretenses are a challenge in terms of processing, storage, and analysis due to their need for organization. Though, it contains valuable insights that can be removed using advanced techniques like image recognition, natural language processing, and machine learning.

Semi-Structured Data

Semi-Structured Data is not raw data or strictly typed; instead, it has an evolving data model [8]. It lies among Structured and Unstructured Data. It has several organizational structures; however, it does not conform to an unbending schema similar to structured data. Semi-Structured Data frequently includes tags or metadata that offer some level of organization and can be used for analysis. Examples of Semi-Structured Data include JSON data, log files, XML files, and web page data.

Furthermore, the below listed also represents types of Big Data:

Time Series Data

Time series data is recovered and collected over a series of timestamps or intervals. It signifies observations or measurements taken at standard intervals like sensor data, weather data, stock market data, or website traffic data. Time series data is generally used for trend analysis, forecasting, anomaly detection, and other temporal analysis tasks.

Geospatial Data

Geospatial data is related to specific coordinates on the Earth's surface or geographic location. It contains Global Positioning System (GPS) data, maps, satellite imagery, location-based social media data, and other geographic information. Geospatial data is extensively used in applications like logistics, urban planning, Location Based Services (LBS), and environmental monitoring.

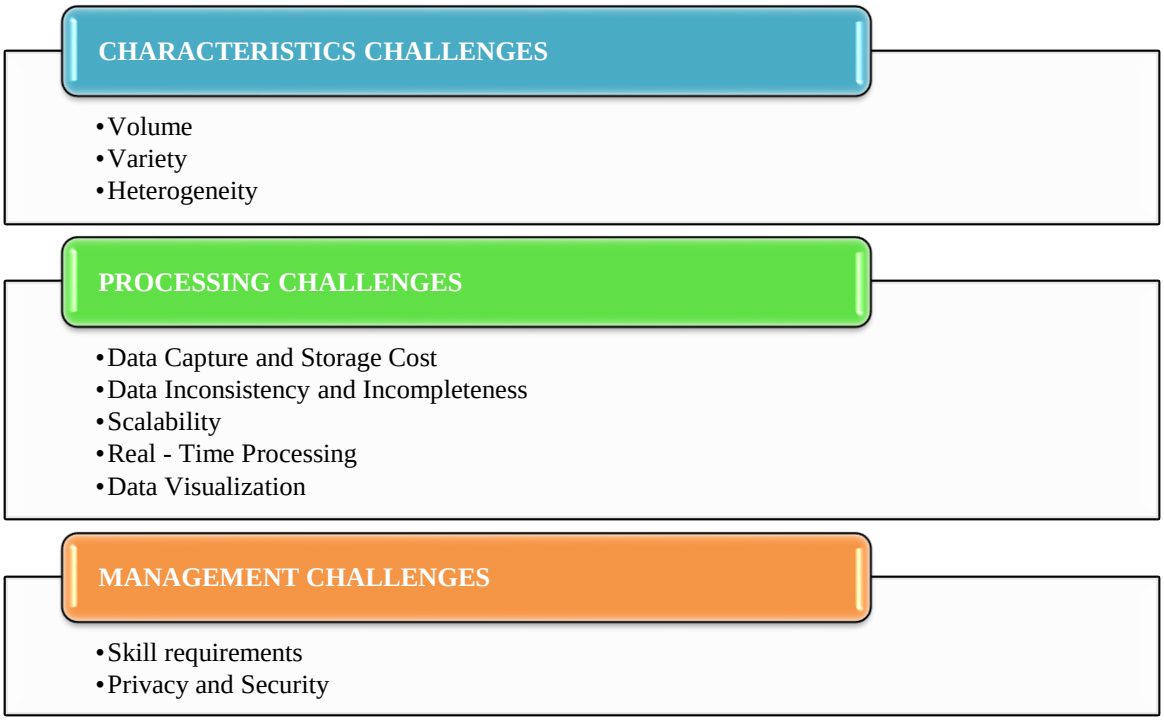
Streaming Data

Streaming data is related to data that is produced continuously in real-time. The speed of marking the generated data is evaluated in terms of the scale of the batch, near real-time, and real-time to reach streaming [10]. It is often generated by devices, social media feeds, sensors, or other sources that produce a constant flow of data. Streaming data requires analysis and processing techniques that can handle the high velocity and real-time nature of the data stream. One of the current main challenges in data mining related to Big Data problems is to find adequate approaches to analyzing massive amounts of online data (or data streams) [6].

The above are some of the frequent types of Big Data based on their characteristics. It's crucial to remember that these categories might overlap is mixed in different ways and that as technology develops and data sources change, new categories of data may also appear.

CHALLENGES OF UNSTRUCTURED DATA

Big Data is typically defined in contemporary surroundings by its primary properties, such as the huge volume, variety, and heterogeneity of data from which all processes demand new technologies and approaches to derive useful information. In this stage, the important challenges discussed in this study were gathered under three main categories:



[Figure - 3: An overview of the Big Data challenges]

Volume, Variety, and Heterogeneity

Unstructured data can encompass all types of data in great volume and diverse formats. This information may come through various information-sharing channels, including forums, social media, chat, email, online commerce, online communities, and more. Additionally, the data may combine a variety of file types, including pictures, text, audio, and video. Due to the complex nature of the data, the analytical processes are more demanding and expensive. This heterogeneous and complex structure of Big Data is a serious challenge that must be overcome in the Big Data analytics process [15].

Data Capture and Storage Cost

In today's digital environments, it is generated daily 2.5 quintillion bytes of data, and this amount is ever-increasing day by day [15]. Big Data's greatest obstacle is data storage. The quantity of data stored in IT systems nowadays doubles every two years. Most of the data is in unstructured form; it is not possible to store it in databases [13]. Unstructured data frequently requires substantial storage capability due to its huge complexity and volume. Processing and storing unstructured data can be expensive, computational resources, and require scalable storage systems. The technologies and infrastructure needed to analyze, store, and process unstructured data may include higher costs related to traditional structured data.

Data Inconsistency and Incompleteness

Unstructured data frequently lacks homogenous formats and may hold errors, noise, or inconsistencies. It can be challenging to guarantee the accuracy, reliability, and quality of unstructured data, particularly when it is produced from various sources or gathered through User Generated Content (UGC). Data preprocessing and cleaning become complex tasks, necessitating substantial effort and innovative techniques. Considering the heterogeneous nature of Big Data, at the start, the reliability of the data collected from different sources should be increased by verifying as temporal and spatial. After that, a common homogeneous data structure should be determined, and the data obtained from different sources should be combined on a monotype basis [15].

Scalability

Recently, the optimization of speedily rising data appears to be a challenging method for providing scalability; a notable challenge should be resolved in advanced database systems. If the system's performance and capacity comparatively increase following the installation of hardware or an application, it is considered to be scalable. Big Data scalability is based on the essentials of scaling databases from a single node to large clusters. In the scaling of data, the usage of cloud computing processes is ever-increasing each passing day. As a result, providing synchronization between cloud computing processes is one of the main tasks in Big Data processing [16].

Real-Time Processing

In Real-time processing, it is very important that provide the necessary time to perform the needed operations on large-volume data streams [15]. Furthermore, Big Data's timeliness may aid in the processing of event streams to enable real-time decision-making. Data cleaning is related to cleaning unstructured textual data (e.g., normalizing data), especially high-frequency streamed real-time data [14].

Data Visualization

Data visualization uses various presentation techniques like tables, charts, animations, and graphs to convey information to users efficiently and clearly. Visualizing Big Data is more difficult than doing so with traditional small data sets. The visualization process calls for a variety of innovative components and techniques when dealing with complicated and large-scale data sets. In large-scale data visualization, many generative techniques, such as feature extraction and geometric modeling, are suggested for data visualization and presentation [15]. The visualization process should be accepted to construct the huge-scale data more convenient functional, understandable, and manageable.

Skill Requirements

The Big Data sector is an innovative and emerging job market in which work capital is engaged efficiently. In this competitive upbringing, competencies, demand skills, and requirements are ever-progressing and changing. A broad range of knowledge and abilities are demanded of Big Data specialists in the Big Data industry market. Big data specialists should constantly update their knowledge and abilities because of this. To meet this industry's growing demand, the lack of Big Data talent will be a formidable obstacle. In this developing field, the training of Big Data specialists with a strong technical background is a remarkable challenge that should be achieved [16]. Investigating unstructured data involves specialized tools and techniques. Natural Language Processing (NLP), Text mining, machine learning algorithms, and image recognition are engaged to extract insights from unstructured data. Applying and implementing these techniques can be computationally and complex exhaustive, requiring skilled data analysts or scientists with expertise in handling unstructured data.

Privacy and Security

Unstructured data may comprise personal or sensitive information that grants security and privacy risks. Some specific information, such as individual health information, online shopping memberships, bank account numbers, and social media profiles, require a high level of protection in terms of privacy and security [15]. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of Big Data [12]. Social media emails, posts, and other unstructured data sources can encompass private conversations, confidential data, or personal identifiers. Safeguarding the security and privacy of unstructured data becomes requiring appropriate data governance, critical encryption measures, and access controls. Data protection is related to the security of data. Once you use a large amount of data, you need to make sure it is secure, and Internet Protocol (IP) issues are resolved [14].

UNSTRUCTURED BIG DATA AND SOCIAL MEDIA

Big data from social media, such as Facebook, Google+, LinkedIn, Instagram, YouTube, etc., are considered a natural resource that comes in various types and forms, but unlike any other resource, it grows bigger by the minute. Social media contains useful information about customers, prospects, and competitors, but surfacing

that intelligence requires a focused approach [11]. Unstructured Big Data from social media presents challenges in terms of processing, data storage, maintaining data quality, and privacy concerns.

In short, social media generates gigantic volumes of Unstructured Data, surrounding a variety of data types and involving advanced techniques for analysis and processing. This makes it difficult to extract, refine or analyze as the data is fluctuating and unstructured. Unrefined data cannot be put to good use [3]. The gathering, fusion, processing, and analysis of big social media data from unstructured (or semi-structured) sources to extract value knowledge is an extremely difficult task that has not been completely solved [6]. Organizations have a wonderful opportunity to learn important lessons and make data-driven decisions because of the plethora of data.

Though, it also suggests significant chances for businesses to influence the wealth of information presented on social media platforms for brand monitoring, preferences, market research, reputation management, and customer service improvement, allowing them to build informed decisions and increase a competitive edge.

Social media are computer-mediated technology that facilitates the criteria and sharing of information, ideas, carrier interest, and other expressions via networks. Social media can be defined as a web-based application, a collection of RSS feeds, news, blogs, and unstructured text which is accessible through the web. Media refers to a platform of communication in social media [14].

Here's how Big Data and Social Media are associated:

Generation of Data

Social media platforms such as Twitter, Facebook, LinkedIn, Google+, YouTube, Instagram, etc., produce enormous volumes of Unstructured Big Data in the form of User Generated Content (UGC). This content comprises comments, shares, likes, shares, text-based posts, and multimedia content like images and videos. Unstructured information is 90% of Big Data and is "human information" like emails, videos, tweets, Facebook posts, call center conversations, closed-circuit TV footage, mobile phone calls, and website clicks [11].

Variety of Data

Social media data is extremely varied. It includes a variety of data types, such as images, audio, video, text, and metadata. This variety creates challenges for analyzing, organizing, and extracting insights from the Unstructured Data.

Velocity and Real-Time Data

Social media platforms produce data at a speedy pace, with users constantly sharing, interacting, and posting content. The velocity of social media data is far above the ground requiring real-time or near real-time analysis and processing to capture sentiments, events, and trends as they unfold. Financial Traders, the Internet, Smart phones, and Retailers increase the rate of data inflow again as consumers carry with them a streaming source of reallocated images and audio data [11].

Data Analysis

Evaluating Unstructured Big Data from social media presents valuable openings. Natural Language Processing (NLP) methods can be used to evaluate text data; sentiment analysis can help to realize emotions and user opinions; video and image analysis can remove visual insights; and social network analysis can identify connections and patterns between users.

Customer Insights

Social media data offers organizations a valuable approach to customer preferences, sentiments, and behaviors. By investigating Unstructured Data from social media, businesses can expand a deeper understanding of their target audience, personalize marketing campaigns, improve customer engagement, and make data-driven decisions.

CONCLUSION

This paper discussed important characteristics, content format, challenges, and different social media platforms such as Twitter, Instagram, Facebook, Google+, LinkedIn, etc. are considered the biggest generator of data, which becomes worthless without Unstructured Big Data. In terms of Volume, Variety, Value, Veracity, and Velocity, this paper rotates around Big Data and its functionality. Executives can acquire appropriate information used for decision-making in less amount of time with unstructured data analysis. Unstructured data has made scope for fraudulent investigation and reliability programs that target and categorize customer separation based on stored behavior analysis. This review will be beneficial for the further development and enhancement of Big Data Analytics from miscellaneous research perceptions.

FUTURE WORK

Future research work can be carried out towards opportunities and applications related to Big Data. There is a need for forthcoming research to understand the critical issues such as security and privacy on data so that various Big Data techniques can be compared to traditional systems. Therefore, there is a significant requirement for using Big Data for observing the data to display the visualization patterns; these outcomes are beneficial for business growth and new trends in the market, etc.

REFERENCES

1. Hadi, H.J., Shnain, A.H., Hadishaheed, S. and Ahmad, A.H., 2015. Big data and five V's characteristics. *International Journal of Advances in Electronics and Computer Science*, *2*(1), pp.16-23.
2. McCaffrey, D.F., Ridgeway, G. and Morral, A.R., 2014. Big data and social media analytics. *Psychological Methods*, *9*(4), pp.403-425.
3. Matilda, S., 2017. Big data in social media environment: A business perspective. In *Decision Management: Concepts, Methodologies, Tools, and Applications* (pp. 1876-1899). IGI Global.
4. Hadi, H.J., Shnain, A.H., Hadishaheed, S. and Ahmad, A.H., 2015. Big data and five V's characteristics. *International Journal of Advances in Electronics and Computer Science*, *2*(1), pp.16-23.

5. Kumar, K.A., 2021. BIG DATA CHARACTERISTICS, CLASSIFICATION, AND CHALLENGES-A REVIEW. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(12), pp.4236-4243.
6. Bello-Orgaz, G., Jung, J.J. and Camacho, D., 2016. Social big data: Recent achievements and new challenges. *Information Fusion*, *28*, pp.45-59.
7. National Information Standards Organization, Understanding Metadata. Bethesda, MD, USA: NISO Press, 2004.
8. Pääkkönen, P. and Pakkala, D., Reference architecture and classification of technologies, products, and services for Big Data systems. Big Data Res. (2015).
9. Yaqoob, I., Hashem, I.A.T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N.B. and Vasilakos, A.V., 2016. Big data: From beginning to future. *International Journal of Information Management*, *36*(6), pp.1231-1247.
10. Sebei, H., HadjTaieb, M.A. and Ben Aouicha, M., 2018. Review of social media analytics process and big data pipeline. *Social Network Analysis and Mining*, *8*(1), p.30.
11. Syed, A., Gillela, K. and Venugopal, C., 2013. The future revolution on big data. *Future*, *2*(6), pp.2446-2451.
12. Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J. and Jagadish, H.V., 2011. Challenges and opportunities with Big Data 2011-1.
13. Kumar, K.A., 2021. BIG DATA CHARACTERISTICS, CLASSIFICATION, AND CHALLENGES-A REVIEW. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(12), pp.4236-4243.
14. Singh, S., Arya, P., Patel, A. and Tiwari, A.K., 2019, February. Social media analysis through big data analytics: a survey. In *Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE)*.
15. Gürcan, F. and Berigel, M., 2018, October. Real-time processing of big data streams: Lifecycle, tools, tasks, and challenges. In *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 1-6). IEEE.
16. Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Mahmoud Ali, W.K., Alam, M., Shiraz, M. and Gani, A., 2014. Big data: survey, technologies, opportunities, and challenges. *The scientific world journal*, *2014*.
17. Eberendu, A.C., 2016. Unstructured Data: an overview of the Data of Big Data. *International Journal of Computer Trends and Technology*, *38*(1), pp.46-50.