International Journal of Management, IT & Engineering
Vol. 7 Issue 8, August 2017,
ISSN: 2249-0558
Impact Factor: 7.119Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Rising of Big Data: An Overview

***Kanika Thakur*[1]**
Assistant Professor
L. N. Mishra College of Business Mgt, Bihar

ABSTRACT- The amount of corporate data created annually has increased rapidly, and an increasing quantity of information is being kept in digital formats. One of the difficulties is figuring out how to manage all of these new data kinds and evaluating which information may have value for your organization. Not only is access to new data sources, chosen events or transactions, or blog entries of importance, but also the patterns and interrelationships between these pieces. Collecting a large volume of varied sorts of data in a short period of time does not provide value. We require analytical tools to unearth business-enhancing insights. That is the theme of this paper. Not only does big data introduce new data kinds and storage technologies, but it also introduces new sorts of analysis.

Keywords - Big data, Data mining, Decision making, data Analytics

I. INTRODUCTION

The phrase 'big data' was purportedly developed during a lunch-table conversion at Silicon Graphics Inc. (SGI) in the early 1990s. The usage of big data has risen in popularity as a consequence of the "commercial excitement" produced by technology firms in developing big data analytics markets. Due to the recent fast expansion in the availability of big data as a result of Internet-based technologies such as social media platforms and mobile devices, many industry leaders are unable to manage extraordinarily wide, unpredictable, and high-velocity data. Big data is omnipresent; for example, books in libraries are tagged and tracked, and smart phones are filled with applications that gather vast quantities of data. Other technologies, such as healthcare devices, continually collect data on pulse, blood pressure, hemoglobin, and sleep habits. All of these instances create enormous volumes of data as a result of firms utilizing user preferences for commercial advantage, which may threaten users' privacy. Traditionally, ideas are produced and evaluated in laboratories and then made public via press releases and marketing.

The general populace then embraces these technology. The quick growth and widespread acceptance of big data by the general public left little time for the academic field to mature. Although various books and pieces in electronic media have been published by experts and writers for their work on big data, essential work in academic journals continues to be lacking [1].This poll evaluated problems and issues in numerous fields of big data, with a specific

emphasis on the privacy issue connected with unstructured massive data. Additionally, this survey exposes various privacy attacks and vulnerabilities in existing privacy protection systems.

II. CHARACTERISTICS OF BIG DATA

Doug Laney, a Gartner analyst, characterized Big Data in 2001 as having three "V's" - variety, velocity, and volume, as seen in Figure 1. "Big Data" refers to structured, unstructured, and semistructured data collected from a range of sources. In its most basic form, velocity refers to the rate at which data is generated in real time. It comprises the pace of change, the variable rate of connection of incoming data sets, and activity bursts in a broader sense. The amount of data is one of the characteristics of huge data. The amount and diversity of big data are dominated by valuable information that generated excitement for the vast marketing efforts of software and hardware companies striving to offer their own 'big data solutions.' Commercial businesses are more interested with developing big data solutions that focus on structured data in particular. This essentially overlooks a sizable portion of big data, such as text messages, videos, and audio files acquired from mobile devices – this mostly ignored unstructured data is far more difficult to analyze, making it more difficult for enterprises to deliver commercial big data solutions. According to a recent research, the vast majority of large data is unstructured, with structured data making up a very small portion [2].

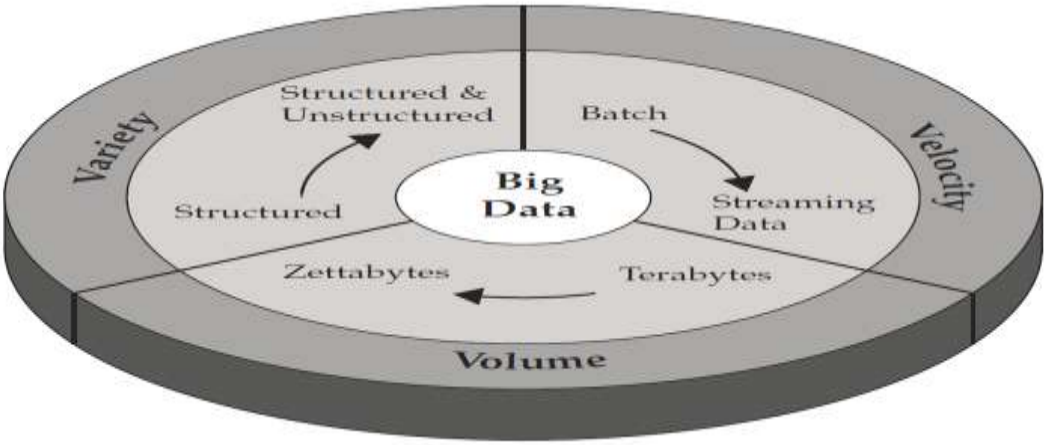

Figure 1

III. CLASSIFICATION OF BIG DATA

Large amounts of data are stored in a number of various formats and have a number of distinct characteristics. Classification of large volumes of data is crucial for understanding the strengths and weaknesses of programs that analyze enormous volumes of data. Big data may be classified

by its storage location, content format, and staging [2]. Each of these groups has its own set of characteristics and interconnections. Data sources are also referred to as data production points. Among these, social media is one of the most relevant and representative sources of big data. Big data is generated through social media networks such as Facebook, Twitter, Instagram, Flickr, YouTube, Google, and Word Press [2]. These websites enable users to connect and form a virtual community around a range of topics. Due to the widespread distribution of personal and inter-personal information among the society, its abuse can have major repercussions and effect [3]. Thus, safeguarding sensitive data from a variety of risks is vital. Social media platforms Social media is a source of information created via URLs for the purpose of sharing and exchanging information and ideas in virtual communities and networks such as collaborative projects, blogs and microblogs, Facebook, and Twitter. Thus, safeguarding sensitive data from a variety of risks is vital. Another source of big data is the Internet of Things (IoT), which is made up of an enormous number of sensors that work together to create tremendous amounts of data. Data is generated by sensing equipment such as mobile phones, satellites, and other sensors used in healthcare and weather stations [4]. Members of this object class include newly designed smart phones, tablets, cameras, and other sensing devices. The fact that these devices are Internet-connected enables sophisticated processing and service provision in a range of industries, including healthcare, banking, and finance. The connectivity of a large number of heterogeneous devices creates huge data [6] that is diverse, variable, unstructured, noisy, and has a high degree of redundancy[7].

To have a better understanding of the properties of big data, it is divided into various categories. The classification is based on the following five criteria: (i) data sources; (ii) content format; (iii) data stores; (iv) data staging; and (v) data processing.

(i) Data Sources - Machine data is data that is generated automatically by hardware or software, such as computers, medical equipment, or other devices, without human intervention. Numerous sensing devices exist for the aim of measuring and converting physical quantities to signals. Financial and employment data, for example, are described as a temporal event. The Internet of Things (IoT) is a collection of objects that are uniquely identifiable as Internet-connected. These products include smartphones, digital cameras, and tablets. When connected to the Internet, these gadgets enable the creation of more intelligent processes and services that satisfy basic, economic, environmental, and health concerns. Numerous connected devices provide a range of services and create enormous amounts of data and information [8].

(ii) Content format - Structured data are typically kept in relational database management systems utilizing SQL, a computer language specialized for data administration and querying. It's simple to save, query, and analyze structured data.

Structured data includes numbers, text, and dates. The term "semi-structured data" refers to data that does not follow a standard database structure. Semi-structured data is structured data that is not organized according to relational database models, such as tables. Capturing semi-structured data for analysis is separate from capturing data in a predefined file format. As a result, acquiring semi-structured data requires the use of complex algorithms that select the next action to do dynamically following data collection [9]. Unstructured data includes text messages, GPS data, videos, and social media data. With the volume of this type of data continuing to expand as a result of Smartphone use, the requirement to evaluate and interpret it has become a challenge.

(iii) Data Stores- Document-oriented data stores are typically used to store and retrieve collections of documents or information, and they support complex data formats such as JSON, XML, and binary (e.g., PDF and MS Word). A document-oriented data storage is comparable to a record or row in a relational database, but it is more versatile and capable of accessing documents based on their contents (e.g., MongoDB, SimpleDB, and CouchDB). Apart from rows, a column-oriented database stores data in columns, with simultaneously recorded attribute values for the same column. Column-oriented databases differ from standard database systems, which store whole rows sequentially [10],. A graph database, such as Neo4j, is used to store and describe data in the form of a graph composed of nodes, edges, and attributes related by relations [12]. Key-value is a relational database system alternative that stores and retrieves data in such a way that it may grow to enormous quantities [13]. Dynamo [14] is an outstanding example of a highly available key-value storage system; it is used by Amazon.com for a number of their services. Similarly, [15] described a scalable key-value store suitable for use in G-store designs that enables transactional multi-key access using a single key enabled by key-value. [16] shown how to do massive tasks in datasets by utilizing a scalable clustering technique. Furthermore, Apache Hbase [17], Apache Cassandra [18], and Voldemort all use key-value stores. Hbase takes use of HDFS, a Cassandra-based open-source version of Google's BigTable. Hbase is a relational database management system (RDBMS) that uses tables, rows, and cells to organize data. Each row is recognized by a row key, and each cell in a table is identified by a row key, a column key, and a version, with the content kept as an uninterpreted array of bytes.

(iv) Data Staging - Cleaning is the process of locating and correcting erroneous or missing data [19]. The process of changing data to an analyzable format is called

transformation. Consistency with the norm Normalization is a strategy for minimizing duplication in database structures [20].

(v) Data Processing - Numerous businesses have adopted batch map reduce-based systems for long-running batch operations in the last several years [21]. This approach enables the scaling of programs across enormous clusters of servers with thousands of nodes. Time in the present S4 is a well-known and strong real-time process-based big data technology [22]. S4 is a distributed computing platform that enables programmers to quickly construct applications that process an infinite number of real-time data streams. S4 is a general-purpose platform that is pluggable, scalable, and reasonably fault-tolerant. Each of these classes has its own set of characteristics and challenges. The internet, sensors, and all worldwide data collections are used as data sources. The data is stored in a number of formats, ranging from ad hoc to highly structured. The most common database type is the relational database, which comes in a variety of forms [23].

IV. TECHNIQUES FOR ANALYSING BIG DATA

When we use SQL queries to look up financial data or OLAP tools to make sales predictions, we are often aware of the type of data we have and the information it contains. Revenue, geography, and time all have predictable relationships. We may not know the answers, but we do know how the various pieces of the data set connect to one another. Business intelligence users frequently run standard reports from structured databases that have been carefully built to take advantage of these interconnections. Big data analysis is making "meaning" of massive amounts of disparate data that, in their raw form, lack a data model defining what each element means in relation to the others. As we engage on this new type of study, there are numerous new considerations to consider:

(i) Discovery - In many circumstances, we have no idea what we have or how various data sets relate to one another. We must uncover it through an exploring and discovery process.

(ii) Iteration - Because the actual relationships between variables are not always understood in advance, gaining insight is frequently an iterative process as we seek the answers. The essence of repetition is that it occasionally leads us down a route that proves to be fruitless. That is acceptable; experimentation is a necessary part of the process. Numerous analysts and industry experts recommend that we begin with small, well-defined projects, learn from each iteration, and then progress to the next concept or topic of research.

(iii) Adaptable Capacity - Due to the iterative nature of big data research, we must be prepared to spend more time and resources on problem solving.

(iv) Data Mining and Predictive Analytics - Big data analysis is not binary. We are not always aware of the relationships between the various data items. Predictive analytics can provide the insights we need as we mine data for patterns and relationships.

(v) Decision Management - Consider the volume and velocity of transactions. If we are using big data analytics to drive a large number of operational decisions (for example, personalizing a web site or prompting call center operators on consumer habits and activities), we must explore how to automate and improve the execution of all those actions. For instance, we may be unaware of the extent to which social data provides light on sales trends. The difficulty arises in determining which data components are connected to which other data items and in what capacity. The discovery process entails not just exploring the data to determine its utility, but also determining how it ties to our standard enterprise data.Not simply what occurred, but also why. For instance, turnover is a critical measure for many businesses. Churn is quite straightforward to quantify. However, why does this occur? Examining call recording data, customer service enquiries, social media discussion, and other consumer feedback can all assist in explaining why customers defect. Similar techniques can be applied to different sorts of data and circumstances. Why did sales decline in a particular store? Why are some patients more likely to survive than others? The difficulty is to collect the appropriate data, to uncover hidden relationships, and to analyze it properly.

IV. CHALLENGES OF BIG DATA

Over the previous few decades, privacy has been intensively investigated. Previously conducted research has mostly concentrated on cryptography, communication, and information theory. Given the large size of bigdata, it is difficult to successfully apply typical cryptography algorithms. Another limit is imposed by the processing and storage capacities of mobile devices, which make encryption and decryption unfeasible [24]. As a result, present encryption techniques are incompatible with growing demands for massive amounts of data. The failure of fundamental anonymisation techniques exacerbates challenges in the era of big data. There is no widely agreed-upon definition of privacy, as the phrase is very subjective [25]. As a result, a universal definition of privacy is unachievable. Furthermore, the fast use of big data calls into question the reliability of traditional approaches.

V. CONCLUSION

There is a strong perception of the relevance of data analysis in the context of Big Data, as well as the possible rewards and tactics for success. It is not so much a fad as it is a long-term organizational fixture that will have a demonstrable long-term impact on enterprises and organizations of all sizes. The analysis of Big Data can provide insights that are not immediately evident or would be difficult to find using normal methods. This method focuses on discovering invisible threads, trends, or patterns that are not immediately apparent to the naked eye. That appears to be straightforward, doesn't it? To be sure, it demands the creation of novel technologies and capacities for evaluating data flow and making conclusions.

REFERENCES

1. Worldometers, "Real time world statistics," 2014, http://www.worldometers.info/world-population/.
2. D. Che, M. Safran, and Z. Peng, "From Big Data to Big Data Mining: challenges, issues, and opportunities," in *Database Systems for Advanced Applications*, pp. 1–15, Springer, Berlin, Germany, 2013.
3. M. Chen, S. Mao, and Y. Liu, "Big data: a survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.
4. S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big data: issues and challenges moving forward," in *Proceedings of the IEEE 46th Annual Hawaii International Conference on System Sciences (HICSS '13)*, pp. 995–1004, January 2013.
5. R. Cumbley and P. Church, "Is "Big Data" creepy?" *Computer Law and Security Review*, vol. 29, no. 5, pp. 601–609, 2013.
6. S. Hendrickson, *Getting Started with Hadoop with Amazon's Elastic MapReduce*, EMR, 2010.
7. M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *Science*, vol. 332, no. 6025, pp. 60–65, 2011.
8. B.P. Rao, P. Saluia, N. Sharma, A. Mittal, S.V. Sharma, Cloud computing for Internet of Things & sensing based applications, in: Sensing Technology (ICST), 2012 Sixth International Conference on, IEEE, 2012, pp. 374-380.
9. B. Franks, Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics, Wiley. com, 2012.
10. D.J. Abadi, P.A. Boncz, S. Harizopoulos, Column-oriented database systems, Proc. VLDB Endow., 2 (2009) 1664-1665.
11. F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, R.E. Gruber, Bigtable: A distributed storage system for structured data, ACM Transactions on Computer Systems (TOCS), 26 (2008) 4.
12. P. Neubauer, Graph databases, NOSQL and Neo4j, in, 2010.
13. M. Seeger, S. Ultra-Large-Sites, Key-Value stores: a practical overview, Computer Science and Media, (2009).
14. G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, W. Vogels, Dynamo: amazon's highly available key-value store, in: SOSP, 2007, pp. 205-220.
15. S. Das, D. Agrawal, A. El Abbadi, G-store: a scalable data store for transactional multi key access in the cloud, in: Proceedings of the 1st ACM symposium on Cloud computing, ACM, 2010, pp. 163-174.
16. F. Lin, W.W. Cohen, Power iteration clustering, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 655-662.

17. R.C. Taylor, An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics, BMC bioinformatics, 11 (2010) S1.
18. A. Lakshman, P. Malik, The Apache cassandra project, in, 2011.
19. E. Rahm, H.H. Do, Data cleaning: Problems and current approaches, IEEE Data Eng. Bull., 23 (2000) 3-13.
20. J. Quackenbush, Microarray data normalization and transformation, Nature genetics, 32 (2002) 496-501.
21. Y. Chen, S. Alspaugh, R. Katz, Interactive analytical processing in big data systems: a cross-industry study of MapReduce workloads, Proc. VLDB Endow., 5 (2012) 1802-1813.
22. L. Neumeyer, B. Robbins, A. Nair, A. Kesari, S4: Distributed Stream Computing Platform, in: Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, 2010, pp. 170-177.
23. J. Hurwitz, A. Nugent, F. Halper, M. Kaufman, Big data for dummies, For Dummies, 2013.
24. M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R. Buyya, Big data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, 79 (2015), pp.3-15.
25. I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani and S. Ullah Khan, The rise of big data on cloud computing: Review and open research issues, Information Systems, 47 (2014), pp. 98-115.