ACADEMICIA
An International Multidisciplinary Research Journal
(Double Blind Refereed & Peer Reviewed Journal)

DOI: 10.5958/2249-7137.2021.02110.8

A BRIEF DESCRIPTION ON BIG DATA

Dr. Ajay Rana*; Vijay Maheshwari**

*Shobhit Institute of Engineering and Technology,
(Deemed to be University), Meerut, INDIA
Email id: ajay.rana@shobhituniversity.ac.in,

**School of Computer Science and Engineering,
Faculty of Engineering and Technology,
Shobhit Institute of Engineering and Technology,
(Deemed to be University), Meerut, INDIA
Email id: vijay@shobhituniversity.ac.in

ABSTRACT

Big data refers to data or data sets that are so big or complicated that conventional data processing application are insufficient, necessitating the use of distributed databases. Big data has always been at the heart of companies like Google, eBay, LinkedIn, and Facebook. It's a collection of large and complex data sets that includes massive amounts of data, social media analytics, data management capabilities, real-time data, and so on. Sensor design, data curation, sharing, storage, analysis, visualization, and information privacy are all challenges. Big data refers to datasets with a lot of diversity and velocity, making conventional tools and methods challenging to manage. Big data analytics is the study of large amounts of data in order to uncover hidden correlations. Big Data is a kind of data whose complexity necessitates the development of new methods, algorithms, and analytics to manage it and extract value and hidden information. We need a new platform known as Hadoop as the fundamental platform for organizing Big Data and solving the issue of making it usable for analytics.

KEYWORDS: *Big Data, Challenges, Parallel Programming, Map Reduce Technique.*

1. INTRODUCTION

Big data is generated by every digital activity and social media interaction. Systems, sensors, and mobile devices all send data. Big data is pouring in from a number of places at an alarming rate, volume, and diversity. To extract real value from big data, we need the best processing power,

analytical capabilities, and expertise. Accurate big data may help you make more confident decisions. Good choices result in increased operational efficiency, cost savings, and risk reduction. Data collections may be analyzed to discover new connections, which can be used to "identify economic trends, prevent illnesses, and fight crime, among other things." Scientists, corporate executives, media and advertising practitioners, and governments all face challenges with big data sets in fields such as Internet search, banking, and business informatics. Because inexpensive and many information-gathering mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks are rapidly being used, data sets are growing in size[1]–[3]. The "scale" of big data varies greatly, ranging from a few hundred gigabytes to several petabytes. Real life examples are stated below:

- Consumer goods businesses and retailers are using social media platforms like Facebook and Twitter to get unparalleled insight into consumer behavior, preferences, and perceptions.
- Manufacturers may track minute vibration data from their equipment, which varies somewhat as it ages, to determine when it's time to replace or repair it. Changing it too early loses money, while replacing it too late results in a costly work halt. The idea of Big Data is shown in Figure 1.

Figure 1:Illustrates the concept of Big Data[4]**.**

1.1 Three Vs of big data:

a. Volume:

Gigabytes to petabytes of data have been kept in corporate repositories. Many reasons contribute to the growth in data volume, including transaction-based data that has been kept over time, unstructured data from social media, and so on. Sensor and machine-to-machine data is being gathered in large quantities. Excessive data volume was a storage problem in the past. However, when storage prices fall, other challenges arise, such as determining relevance in huge data quantities and using analytics to extract value from relevant data. Amount of data is referred to as volume[5], [6].

b. Velocity:

Data is arriving at breakneck speed and must be processed as quickly as possible. The requirement to deal with fast-moving data in near-real time is being driven by RFID sensors and smart meters. Most companies struggle to respond fast enough to cope with data velocity. The term "velocity" refers to the rate at which data was processed. Big data must be utilized for time-sensitive operations like detecting fraud. It pours into your company in order to enhance its worth.

c. Variety:

Data is now available in a variety of forms. Traditional databases store structured and quantitative data. Line-of-business apps generate information. Unstructured text documents, email, video, audio, and financial transactions are all examples of unstructured text documents. Managing, integrating, and regulating various types of data is still a challenge for many businesses. There are many kinds and sources of data. From organized and historical data kept in corporate storage to unstructured, semi structured, audio, video, and other types of data, the diversity of data has expanded. We consider two additional dimensions when thinking about big data:

d. Variability:

Data flows may be extremely erratic with periodic peaks as velocity and types of data increase. It's all over social media. Peak data loads that occur on a daily, seasonal, or event-based basis are impossible to manage. There's much more unstructured data here. The inconsistency of the data, which may stymie the process of correctly processing and maintaining the data. The data's irregularity may sometimes stymie the process of effectively processing and maintaining the data.

e. Complexity:

Today's data is derived from a variety of sources. Linking, matching, and transforming data across systems is still a challenge. Relationships, hierarchies, and various data connections must all be connected and correlated. Otherwise, your data may soon get out of hand. Data management becomes very complicated when huge amounts of data are collected from various sources. Data, in particular, must be linked, integrated, and correlated so that consumers can understand the information or messages that the data is intended to communicate.

f. Veracity"

The significant variability in data quality collected. The accuracy of data analysis is dependent on the accuracy of the original data.

1.2 Parallel Programming & Map reduce:

Data analysis software inherently parallelizes. Many programmers are interested in developing parallel applications. In the area of parallel databases, parallel research has had the greatest success. Parallel databases allow programmers to split up input data tables into parts and execute each piece via the same single-machine program on each processor, rather than having to untangle an algorithm into different threads to run on various cores. Parallel programming is as simple as programming a single computer using this "parallel dataflow" paradigm. It also works

in data centers with “shared-nothing” clusters of computers: the machines may interact using simple data streams rather than costly shared RAM or disk infrastructure[7]–[9].

Data analysis software inherently parallelizes. Many programmers are interested in developing parallel applications. In the area of parallel databases, parallel research has had the greatest success. Parallel databases allow programmers to split up input data tables into parts and execute each piece via the same single-machine program on each processor, rather than having to untangle an algorithm into different threads to run on various cores. Parallel programming is as simple as programming a single computer using this "parallel dataflow" paradigm. It also works in data centers with “shared-nothing” clusters of computers: the machines may interact using simple data streams rather than costly shared RAM or disk infrastructure.

Map Reduce is at the core of Hadoop. This programming paradigm is responsible for Hadoop's enormous scalability over thousands of servers. It can handle petabytes or zetabytes of data stored in Apache Hadoop in batches. If we've worked with clustered scale-out data processing systems before. The Map Reduce idea is therefore easy to grasp. The Map Reduce programming paradigm has turned a new page in the narrative of parallelism. The Map Reduce framework is a parallel dataflow system that divides data across many computers. They all use the same single-node logic. Map Reduce requires programmers to use conventional programming languages such as C, Java, Python, and Perl to create their programs. Map Reduce enables programs to be written to and read from conventional files in a file system rather than needing database schema definitions, in addition to its familiar syntax.

Map Reduce is a term that refers to two different jobs. The first is the task of map, which involves converting one set of data into another. Value pairs are used to break down individual components. Reduce takes the output of a map as input and merges the data values into a smaller set. After the map task, the reduce job is always run. As a result, the name Map Reduce is a series. Figure 2 depicts the Map Reduce idea in Big Data.

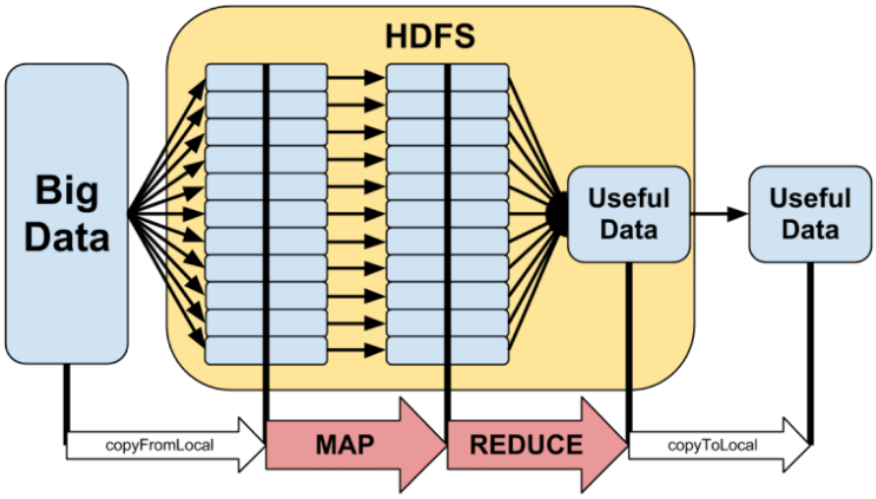

Figure 2: Illustrates the concept of MapReduce in Big Data[10]**.**

1.3 Best Big Data Analytics Use Cases:

a. Sentiment Analysis:

Sentiment research provides valuable corporate information for improving customer experience, revitalizing a brand, and gaining a competitive edge. The capacity to dig for multistructured data collected from many sources into a single database is crucial to effective sentiment analysis.

b. 360-Degree View of Customer:

A 360-degree customer perspective allows you to get a better understanding of your customers' motives and behavior. To get a 360-degree customer evaluation, data from many sources, such as social media, data-gathering sensors, mobile devices, and so on, must be analyzed. As a consequence, more effective micro-segmentation and real-time marketing become possible.

c. Ad Hoc Data Analysis:

Ad-hoc analysis examines just the data that has been requested or required, adding another layer of analysis to data sets that are growing bigger and more diverse. By evaluating relevant data from unstructured sources, both external and internal, big data ad-hoc analytics may aid in the endeavor to acquire deeper insight into consumers.

d. Real-Time Analytics:

Real-time analytics systems rapidly interpret and analyze data sets, delivering findings even as new data is produced and gathered. This fast-paced approach to analytics may result in rapid responses and adjustments. Better sentiment analysis, split testing, and targeted marketing are all possible with it.

e. Multi-Channel Marketing:

Multi-channel marketing integrates various kinds of media, such as business websites, social media, and physical shops, to provide a seamless experience. Multi-channel marketing necessitates an integrated big data strategy at all phases of the purchasing process.

f. Customer Micro-Segmentation:

For smaller groupings, customer micro-segmentation allows for more customized and targeted communications. This customized strategy requires the analysis of large amounts of data gathered from sources such as consumer internet interactions, social media, and so on.

g. Data Warehouse Modernization:

To improve operational efficiency, combine big data and data warehouse capabilities. Optimize your data warehouse so that new kinds of analysis may be performed. Before deciding what data should be transferred to the data warehouse, utilize big data technology to create a staging area or landing zone for your incoming data. Using in sequence integration software and tools, extract rarely used or aged data from warehouse and application databases.

h. Bigdata Challenges:

Big Data's heterogeneity, size, timeliness, complexity, and privacy issues stymie progress at all stages of the value-creation pipeline. The issues begin during data acquisition, when the data tsunami forces us to make ad hoc choices about what data to retain and what to discard, as well as how to save what we keep consistently with the appropriate information. Today, a lot of data isn't in an organized format by default; for example, tweets and blogs are unstructured text, while pictures and video are formatted for storage and presentation. However, this is not the case for

semantic content and search. A key test is converting such information into a structured format for subsequent examination. When data can be linked to other data, its value skyrockets. As a result, data integration is a significant source of value. Today, the bulk of data is produced directly in digital format; we have the potential and the task to influence production in order to ease subsequent linking and to automatically connect data that has not yet been created. Other fundamental problems include data analysis, organization, recovery, and modeling. Data analysis is an obvious bottleneck in many applications, owing to the original methods' limited scalability as well as the complexity of the data to be processed. Finally, non-technical domain specialists must present the findings and explain them in order to extract actionable knowledge.

i. Volume of data:

The amount of data, particularly machine-generated data, is expanding, as is the pace at which it grows each year, thanks to new data sources that are emerging. In the year 2000, for example, the world's data storage capacity was 800,000 petabytes (PB). It is expected to reach 35 zettabytes (ZB) by 2020, according to IBM. Twitter, for example, produces more than 7 terabytes (TB) of data per day. Facebook has a storage capacity of ten terabytes. Mobile gadgets play a significant role as well.

j. Big data skills are in short supply:

There is already a scarcity of data scientists available. This involves a scarcity of individuals who can work effectively with huge amounts of data and large data sets. Companies need the appropriate mix of people to help them make sense of the data streams that are flooding in. This includes the ability to apply predictive analytics to large data, which is a skill set that even most data scientists lack.

2. DISCUSSION

Big data is a collection of technologies for storing, analyzing, and managing large amounts of data, as well as a macro-tool for seeing patterns in the chaos of this information explosion in order to develop smart solutions. It is now utilized in a wide range of fields, including medical, agriculture, gaming, and environmental protection. The three main ideas of big data were initially linked with three essential concepts: volume, diversity, and velocity. Because large data analysis poses sampling difficulties, only observations and sample were previously allowed. As a result, big data often contains data in quantities that conventional software cannot handle in a reasonable amount of time or for a reasonable price. Big data is used by Amazon, Netflix, and many other businesses to offer services to their consumers.

3. CONCLUSION

The combination of Big Data, low-cost commodity technology, and analytic software has created a watershed moment in data analysis history. Because of the convergence of these developments, we now have the capability to analyze massive data sets rapidly and cost-effectively for the first time in history. All of these skills aren't theoretical or simple. They offer a significant step forward and a clear opportunity to achieve massive improvements in efficiency, production, revenue, and profitability.

When large data systems are accessible, requirements for handing out that may appear impossible now will become commonplace. We learn how to take advantage of them. Systems of

the size of Facebook and Google would have been science fiction not long ago. 100 transactions per second for airline and financial systems was unheard of at the time. A number of new criteria will integrate data from a variety of sources, not all of which will be held by the business. Some will, for example, make advantage of government "open data." There are many opportunities for inventors!

REFERENCES:

1. A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*. 2018, doi: 10.1016/j.jksuci.2017.06.001.

2. A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

3. W. A. Günther, M. H. Rezazade Mehrizi, M. Huysman, and F. Feldberg, "Debating big data: A literature review on realizing value from big data," *J. Strateg. Inf. Syst.*, 2017, doi: 10.1016/j.jsis.2017.07.003.

4. B. Data, B. Data, and B. Data, "Big Data." https://medium.com/@raghav0278/what-is-big-data-and-why-is-it-important-15afe114b8b7 (accessed Aug. 01, 2017).

5. E. Al Nuaimi, H. Al Neyadi, N. Mohamed, and J. Al-Jaroodi, "Applications of big data to smart cities," *J. Internet Serv. Appl.*, 2015, doi: 10.1186/s13174-015-0041-5.

6. R. Kitchin and G. McArdle, "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets," *Big Data Soc.*, 2016, doi: 10.1177/2053951716631130.

7. M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, 2015, doi: 10.1186/s40537-014-0007-7.

8. C. H. Lee and H. J. Yoon, "Medical big data: Promise and challenges," *Kidney Res. Clin. Pract.*, 2017, doi: 10.23876/j.krcp.2017.36.1.3.

9. A. Zwitter, "Big Data ethics," *Big Data and Society*. 2014, doi: 10.1177/2053951714559253.

10. "Big-data-map-reduce-process-22." https://www.researchgate.net/figure/Big-data-map-reduce-process-22_fig4_317825624 (accessed Aug. 01, 2017).