Mesopotamian journal of Big Data
Vol. **(2022)** 2022, **pp**. 23–30
DOI: https://doi.org/10.58496/MJBD/2022/003 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

Research Article

Big Data Processing: A review

Taufik Gusman*, [1], Mohammad Naeemullah[2], Adeeb Mansoor Qasim[3],

[1] *Department of Information Technology, Politeknik Negeri Padang, Sumatera Barat, Indonesia*

[2] *Department of Computer Science Maulana Azad College. Rauza Bagh Aurangabad. Maharashtra, India*

[3] *Computer Science Department, AL-Salam University College, Hay AL-khadra'a, Baghdad, 10022, Iraq*

ARTICLE INFO

Article History
Received 19 Jan 2022
Accepted 24 Mar 2022
Published 12 Apr 2022

Keywords
Big data
Parallel Processing
Spark
Machine Learning

ABSTRACT

The processing of "big data," which consists of very vast and complicated datasets, is a fast expanding area. It has been employed in a wide variety of industries and applications, from e-commerce to financial services to transportation, and it has the potential to revolutionise the way organisations function and make decisions. In this work, we discuss the definitions, characteristics, and challenges of large data processing. We also talk about the ethics of using this technology and the prevalent tools and technologies used for processing large amounts of data. Finally, we consider how big data processing is expected to evolve in the years ahead in light of current trends and promising new technologies.

1. INTRODUCTION

The term "big data"[1] is used to describe massive data collections that cannot be efficiently processed by conventional methods. Volume (a lot of data), variety (various kinds of data), and velocity (how quickly data is generated and collected) are three common descriptors of such data sets. As the amount of digital data being generated has grown tremendously in recent years, the importance of big data has risen[2][3]. This has been fueled by the rise of social media and other online platforms, the widespread use of internet-connected devices (such as smartphones, sensors, and smart appliances), and the widespread adoption of data-driven decision making across a wide range of businesses. Both the difficulties and the rewards of processing massive amounts of data are numerous. On the one hand, dealing with and analysing huge data can be a herculean task that necessitates a large investment of time, energy, and other resources. However, the analysis of large amounts of data can yield useful insights and information for corporations, governments, and other institutions. The processing of large amounts of data has several potential advantages.

Better decisions may be made and performance can be optimised when businesses analyse massive amounts of data to acquire a deeper understanding of their customers, markets, and operations. Targeted marketing and product suggestions are only two examples of how big data may be utilised to give clients more individualised service. Big data can aid businesses in streamlining their operations, cutting down on waste, and saving money by revealing previously unseen patterns and trends. New opportunities and ideas for driving innovation and growth can be uncovered with the use of big data. In sum, big data processing has the potential to revolutionise business processes and decision-making, and it will almost certainly continue to be a significant part of the information economy. The term "big data" refers to the massive amounts of data that must be collected, stored, and analysed. Although it can be difficult, the rewards for businesses, governments, and other organizations are substantial. Big data processing presents a number of difficulties, including: The sheer size of some big data sets makes it challenging to keep them organised and run analyses quickly. Structured data (such as databases), unstructured data (such as text, audio, and video), and semi-structured data (such as social media posts) are all examples of what we mean by "big data." It might be difficult to process and analyse data that comes in a variety of formats. The speed at which big data is generated and collected presents challenges for real-time processing and analysis. Data quality: The precision and trustworthiness of an analysis might be compromised

by the presence of errors, inconsistencies, or biases in the underlying big data. Concerns over privacy and security are warranted due to the sensitive nature of much big data. Despite these obstacles, big data processing has the potential to offer many benefits to businesses. The processing of large amounts of data has several potential advantages. Better decisions may be made and performance can be optimised when businesses analyse massive amounts of data to acquire a deeper understanding of their customers, markets, and operations. Targeted marketing and product suggestions are only two examples of how big data may be utilised to give clients more individualised service. Big data can aid businesses in streamlining their operations, cutting down on waste, and saving money by revealing previously unseen patterns and trends. New opportunities and ideas for driving innovation and growth can be uncovered with the use of big data.

While the difficulties of big data processing are real, the opportunities it presents make it a powerful resource for businesses that can master it.

Fig. 1. Big data processing.

2. TYPES OF BIG DATA

Structured data, unstructured data, and semi-structured data are the three primary forms of big data. Data that is stored in a database or other well-defined and predictable format is an example of what is called "structured data"[4]. Because it conforms to a predetermined format or set of rules, structured data is often simple to process and analyse. Data on customers, finances, and inventories are all examples of structured information that may be found in many computerised systems.

Text documents, emails, audio files, and video files are all examples of unstructured data[5] since they lack a standard format or organisation. Due to its lack of uniformity and organisation, unstructured data can be more challenging to process and analyse. There is still value in unstructured data since it may contain useful insights and details. For instance, sentiment analysis of customer reviews from unstructured text data can provide valuable information about product perceptions and customer satisfaction. Moreover, advancements in natural language processing and machine learning techniques have enabled better extraction of meaningful information from unstructured data sources.

Semi-structured data[6] consists of information that has some form of organization, but not as much as structured data. Semi-structured data can be found in a variety of formats, including social network posts, XML files, and JSON files. When compared to processing and analyzing structured data, semi-structured data can be more challenging, although it is still easier than unstructured data. Because of its advantageous combination of adaptability and structure, semi-structured data is useful in many contexts. For instance, NoSQL databases' schema-less nature and adaptability to many data formats make them ideal for managing semi-structured data.

Each sort of big data provides its own unique characteristics and challenges, necessitating a different set of tools and technologies to handle and analyze the data. Relational databases and SQL-based queries are standard for structured data. Big data technologies like Apache Hadoop and Spark, in conjunction with machine learning techniques, are frequently used to process and generate insights from unstructured and semi-structured data. Taking full advantage of the possibilities of structured, unstructured, and semi-structured data sources will require novel ideas and solutions as big data continues to expand and develop.

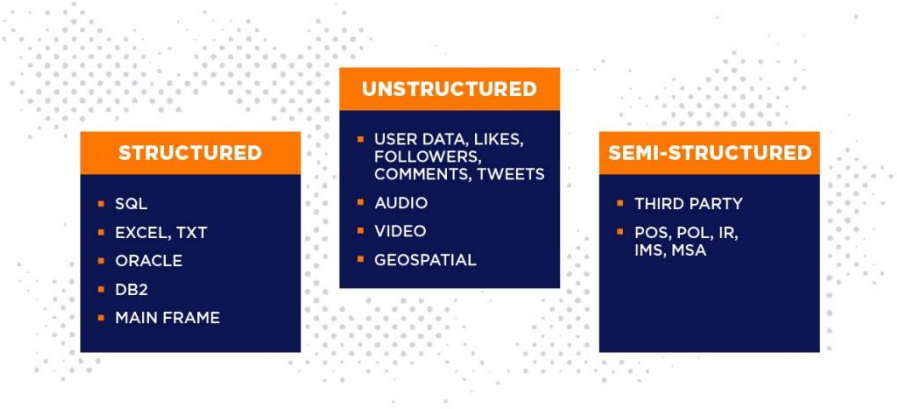

Fig. 2. Types of big data.

3. TOOLS AND TECHNOLOGIES FOR BIG DATA PROCESSING

There are several tools and technologies that are commonly used for big data processing, including:

- Hadoop and MapReduce[7]: Hadoop is an open-source software framework that is commonly used for storing and processing large amounts of data. It is based on the MapReduce programming model, which allows developers to write programs that can process and analyze large data sets in parallel across a distributed computing cluster.
- Spark[8]: Spark is a fast and flexible in-memory data processing engine that is used for a wide range of big data processing tasks, such as ETL (extract, transform, load), machine learning, and stream processing.
- NoSQL databases[9]: NoSQL databases are database management systems that are designed to handle large amounts of unstructured or semi-structured data. Examples of NoSQL databases include MongoDB, Cassandra, and DynamoDB.
- Machine learning and AI[10]: Machine learning and artificial intelligence (AI) algorithms can be used to analyze big data sets and extract insights and patterns that would be difficult or impossible to uncover manually.

- Cloud computing[11]: Cloud-based platforms, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform, can provide scalable and flexible infrastructure for storing and processing big data.

The choice of tools and technologies for big data processing will depend on the specific needs and goals of an organization, as well as the types of data being processed and the required processing speed and scale.

4. CASE STUDIES OF BIG DATA PROCESSING

4.1 Personalization in e-commerce

When talking about online stores, "personalization" refers to how they use information and tools to cater to each customer's preferences. This can be done in a number of ways, including as making suggestions based on a user's browsing or purchasing habits, sending ads that are more relevant to a user's interests, or displaying personalised content.

E-commerce businesses can offer customization in a number of ways by utilising big data processing:

E-commerce businesses can learn more about their customers' likes and dislikes by analysing data like their shopping and browsing habits as well as their demographics. The information gained from this can be utilised to tailor a customer's buying experience in several ways. Artificial intelligence and machine learning Algorithms trained on large amounts of client data can improve the quality of recommendations greatly. Using a customer's prior purchases, search history, and other data points, an online retailer, for instance, could apply machine learning to anticipate which products the client is most likely to be interested in.

Big data processing enables e-commerce businesses to use information about their clients to categorise them into distinct subsets based on their shared traits and patterns of behaviour. This can help them target certain groups of customers with their marketing and customization efforts. E-commerce businesses stand to gain from increased sales and revenue because to the positive effects personalisation has on consumer happiness and loyalty. Online retailers owe it to their customers to be forthright about the information they collect and how it will be used, as well as to protect their security and privacy.

4.2 Fraud detection in financial services

Identification and prevention of financial crimes like identity theft, credit card fraud, and money laundering are the goals of fraud detection in the financial sector. Due to the ability to analyse massive amounts of data from a variety of sources in near real-time, big data processing has the potential to be an effective tool for fraud detection in the financial sector.

Big data processing has many potential applications in the financial services industry, including:

Big data processing enables real-time analysis of financial transactions, allowing banks to spot suspicious patterns and conduct further investigation. If a customer suddenly starts making more or different types of transactions, a bank could employ big data processing to figure out what's going on. Financial institutions can employ big data processing to spot red flags like multiple account openings or large-dollar transactions in a short period of time, both of which could be signs of fraud.

Customers' demographics, account and transaction histories can be analysed by financial organisations using big data processing to spot signs of fraud. Big data processing could help a bank spot a customer who has a penchant for making unusually large transactions or who has created many accounts in different names.

Algorithms developed using machine learning and artificial intelligence may examine large datasets in search of anomalies that might point to fraud. Machine learning could be used by a bank to determine the legitimacy of a transaction by analysing data such the customer's history and the details of the transaction.
Overall, big data processing can be a powerful tool for detecting and preventing fraud in the financial industry, but it is important for financial institutions to ensure that they are using this technology in a responsible and ethical manner.

4.3 Traffic prediction in transportation

The term "traffic prediction" is used in the transportation industry to describe the practise of utilising historical data and current technological advancements to anticipate future traffic conditions on roads, highways, and other forms of transportation infrastructure. Because it enables transportation agencies to analyse enormous amounts of data from numerous sources in real time, big data processing can be a useful tool for traffic prediction.

Some applications of big data processing in transportation traffic prediction include:

Real-time traffic data from sources like traffic sensors, GPS data from automobiles, and social media feeds can be analysed by transportation authorities using big data processing to reveal patterns and trends in traffic. This will allow them to better anticipate short-term traffic circumstances and alter traffic management methods accordingly. Long-term traffic trends can be predicted by analysing data on population growth, land use patterns, and transportation infrastructure with the use of big data processing by transportation agencies. This can aid in anticipating transit demands and deciding where to place necessary infrastructure expenditures.

More accurate traffic predictions can be made with the use of machine learning and artificial intelligence algorithms by analysing large datasets. Machine learning might be used by the transport department to forecast how busy a certain road will be depending on the time of day, the weather, and the day of the week. Congestion may be reduced, and transport networks can function more efficiently and reliably, with the use of big data processing that can improve traffic prediction and management.

5. ETHICAL CONSIDERATIONS IN BIG DATA PROCESSING

Many ethical questions are raised by big data processing, and businesses that exploit this technology should give them serious thought. Ethical issues that must be taken into account while processing massive amounts of data include: Data privacy and security are major concerns in the big data realm because of the prevalence of personally identifiable information, financial details, and health records. Organisations must take precautions to prevent unauthorised access to and use of such information. This involves getting people's permission before collecting or using their data and establishing stringent security measures to prevent data breaches and unauthorised access.

Fairness and bias: Big data processing can be prone to bias, which might have unintended consequences like bias or discrimination. An algorithm that uses machine learning to make predictions or choices may be unjust if it is trained with data that is itself biassed or does not accurately reflect the population. Companies should be alert to the possibility of bias in their data and take appropriate measures to reduce or eliminate it. Data collection and use, as well as the analytical procedures and algorithms, should be open books at companies that employ big data processing. If people know how their data will be used, they'll be better able to decide whether or not to disclose it.

The vast potential of big data processing is undeniable. However, alongside the benefits come significant ethical considerations that need to be addressed. Here's a breakdown of some key areas:

1. Privacy: Big data often involves collecting and analyzing massive amounts of personal information. This raises concerns about individual privacy and the potential for invasion. Questions arise around informed consent, data ownership, and how much control users have over their information.
2. Security: Breaches of big data systems can expose sensitive information, leading to identity theft, fraud, and other harms. Robust security measures like encryption and access controls are crucial to safeguard data integrity and confidentiality.
3. Bias: Algorithms used in big data processing can inherit biases from the data they're trained on. This can lead to discriminatory outcomes in areas like loan approvals, job applications, or criminal justice. Ensuring fairness and transparency in algorithms is essential to mitigate bias and promote ethical decision-making.
4. Transparency: The complex nature of big data analytics can make it difficult for individuals to understand how their data is used and how it impacts them.
   Transparency is key. Individuals should have the right to know how data is collected, used, and how it affects them.
5. Governance: The widespread use of big data necessitates strong governance frameworks and regulations.
   These frameworks should establish guidelines for data collection, storage, usage, and security to ensure responsible big data practices. Organisations that process large amounts of data have a need to act ethically and with consideration for the interests of those affected by their work. This includes being forthright about data collection and use, considering the effects on individuals and communities, and accepting responsibility for unintended outcomes.

6. FUTURE DIRECTIONS IN BIG DATA PROCESSING

Technological developments and shifts in how businesses make use of data are likely to keep big data processing on the cutting edge in the coming years. Future directions in big data processing are anticipated to be influenced by the following:

Algorithms based on machine learning and artificial intelligence (AI) are expected to play an increasingly significant role in big data processing, helping businesses analyse data more quickly and effectively, and gleaning insights and patterns that would be extremely challenging, if not impossible, to discover otherwise.
Data processing at the network's periphery (i.e., near the source of the data) rather than at a centralised data centre is known as edge computing, and its evolution is expected to be spurred by the expansion of the IoT. As a result, data may be processed and analysed more quickly, with less latency and bandwidth needed.

Big data processing is projected to become more intertwined with other cutting-edge methods, such as blockchain, 5G, and quantum computing. Data processing and analysis can be accelerated and made more accurate with the help of technologies like blockchain, 5G networks, and quantum computers. Regulating data collection, use, and sharing is anticipated to expand as big data processing becomes more pervasive and affects more facets of our life. Standards and guidelines for the ethical use of data and algorithms, as well as new legislation, may be necessary.

Big data processing is expected to continue to play an increasingly important role across a wide range of industries and applications, and its future is likely to be influenced by a combination of technology improvements and legislative developments.

7. CONCLUSION AND FURTHER WORK

In conclusion, big data processing is a potent instrument that helps businesses make sense of massive amounts of data. Structured data, unstructured data, and semi-structured data are all examples of big data, which is also characterised by its volume, diversity, and velocity. Hadoop, Spark, NoSQL databases, machine learning, and the cloud are just some of the tools and technologies available for processing massive amounts of data. Personalization in e-commerce, fraud detection in financial services, and traffic prediction in transportation are just a few examples of how big data processing can revolutionise business operations and decision-making. Privacy and security, bias and fairness, and openness are only a few of the ethical concerns brought up by the processing of big data. Future developments in big data processing are anticipated to be influenced by technological breakthroughs and shifts in how businesses make use of information. It's possible that regulations about privacy and security will get more stringent, and it's expected to become more intertwined with other developing technologies like blockchain, 5G, and quantum computing. The processing of large amounts of data will certainly remain an important factor in many fields and functions.

Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors would like to express their gratitude to the Maulana Azad College, and the Al Salam University College for their moral support. Please accept my sincere gratitude for the useful recommendations and constructive remarks provided by the anonymous reviewers.

Funding

The authors receive no funding for this work.

References

[1] N. Khan *et al.*, "Big data: survey, technologies, opportunities, and challenges," *The scientific world journal*, vol. 2014, 2014.

[2] P. A. Sri and M. Anusha, "Big data-survey," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 4, no. 1, pp. 74-80, 2016.

[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] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big Data technologies: A survey," *Journal of King Saud University-Computer and Information Sciences*, vol. 30, no. 4, pp. 431-448, 2018.

[5] N. D. Zaki, N. Y. Hashim, Y. M. Mohialden, M. A. Mohammed, T. Sutikno, and A. H. Ali, "A real-time big data sentiment analysis for iraqi tweets using spark streaming," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 4, pp. 1411-1419, 2020.

[6] A. H. Ali, "A survey on vertical and horizontal scaling platforms for big data analytics," *International Journal of Integrated Engineering*, vol. 11, no. 6, pp. 138-150, 2019.

[7] A. H. Ali and M. Z. Abdullah, "Recent trends in distributed online stream processing platform for big data: Survey," in *2018 1st Annual International Conference on Information and Sciences (AiCIS)*, 2018, pp. 140-145: IEEE.

[8] S. Salloum, R. Dautov, X. Chen, P. X. Peng, and J. Z. Huang, "Big data analytics on Apache Spark," *International Journal of Data Science and Analytics*, vol. 1, no. 3, pp. 145-164, 2016.

[9] M. Diogo, B. Cabral, and J. Bernardino, "Consistency models of NoSQL databases," *Future Internet,* vol. 11, no. 2, p. 43, 2019.
[10] S. Alotaibi, R. Mehmood, I. Katib, O. Rana, and A. Albeshri, "Sehaa: A big data analytics tool for healthcare symptoms and diseases detection using Twitter, Apache Spark, and Machine Learning," *Applied Sciences,* vol. 10, no. 4, p. 1398, 2020.
[11] L. M. Dang, M. J. Piran, D. Han, K. Min, and H. Moon, "A survey on internet of things and cloud computing for healthcare," *Electronics,* vol. 8, no. 7, p. 768, 2019.