Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-6, Issue-7, July 2024
DOI: 10.53469/jrse.2024.06(07).09

The Impact of Big Data Analysis on Trends

Gurinder Kaur

Department of Information Technology, St. Joseph's College, Trichy, Tamilnadu, India
gurinder001@gmail.com

Abstract: ***The digitization of modern manufacturing systems has led to a significant increase in data creation, commonly discussed to as Big Data. Despite the existence of various technologies and techniques for collecting such data, its transformation into meaningful information and knowledge is still in its early stages. Advances in sensor networks and Internet of Things (IoT) technology have made it possible to collect enormous amounts of data. However, analyzing such massive quantities of data requires more effective methods that can offer accurate analysis. Artificial Intelligence (AI) techniques, including machine knowledge and evolutionary algorithms, have proven to be capable of delivering precise, fast, and scalable results in big data analytics. Despite the growing awareness in these techniques, there is currently no comprehensive survey available that covers the various artificial intelligence techniques used in big data analytics.***

Keywords: Big data, Artificial intelligence, Machine learning

1. Introduction

The rapid advancements in digital technologies have managed to a significant increase in the amount of digital data being caused [1]. This has resulted in the creation of large volumes of data from several sources such as social networks, smartphones, sensors, and more. The sheer magnitude of this data, which surpasses the capabilities of traditional relational databases and analytical techniques, is referred to as Big Data. To effectively analyze and make sense of such vast datasets, it is imperative to develop new tools and analytical techniques. These tools should be capable of identifying patterns and correlations in rapidly evolving data, enabling us to harness its full potential.

Extraction of important knowledge and insights from large and complex datasets is made possible by the application of modern artificial intelligence algorithms in data analytics. This entails using deep learning models, machine learning algorithms, and natural language processing methods to search through large amounts of data for patterns and relationships. By automating data analysis, the goal of AI-driven big data analytics is to enhance the speed, accuracy, and scalability of the process, empowering organizations to fully leverage their data and gain a competitive edge.

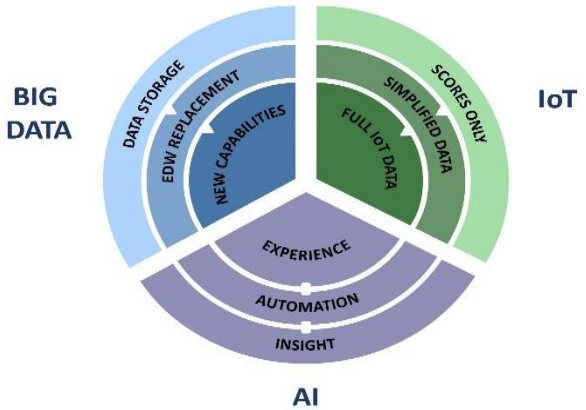

Figure 1: AI in big data

Big data open source technologies have grown in popularity as a result of their undeniable ability to handle vast amounts of data in parallel. This has enabled the quick processing of large datasets using parallel processing and distributed computing methods. These essential features, combined with the capability to manage massive amounts of data, played a significant role in the development of Apache Hadoop, the leading framework for big data processing. Nevertheless, it is worth mentioning that the storage and analysis of such extensive datasets have been extensively examined and compared in the context of relational database management systems (RDBMS) versus the Hadoop environment, offering valuable insights into this widely debated technology.

2. Literature Review

Enormous amounts of data are collected from diverse sources such as sensors, transactional applications, and social media in different formats [2]. Multiple definitions have been proposed for big data. In general, the term Big Data refers to a growing collection of data that encompasses various formats: structured, unstructured, and semi-structured data. Traditional Database Management Systems (DBMSs) are incapable of handling such a massive volume of various data. Consequently, robust Modern technology and sophisticated algorithms are required to process big data.

The concept of big data can be characterized by various V's, including Volume, Velocity, Variety, and Veracity [3].

- Volume denotes to the vast amounts of data produced on a continuous basis. These extensive data sets can be efficiently handled within big data frameworks.
- Velocity represents the rate at which data is produced and processed in order to derive meaningful insights.
- Variety encompasses the diverse range of data formats, including documents, videos, and logs.
- Veracity highlights the factors that determine the quality of data, including biases, noise, and anomalies.

The effective management of large volumes of data is crucial for the successful creation of high-quality data analytics. This involves the efficient collection of data from diverse sources, the utilization of various mechanisms and tools for storage, the elimination of errors through data cleansing, and the transformation of data hooked on a

standardized format. Additionally, data encoding is employed to enhance security and confidentiality. The ultimate objective of this process is to guarantee the accessibility, management, effectual storage, and security of dependable data.

Big data analytics plays a crucial role in organizations by enabling them to uncover valued data and patterns that can impact their business operations [4]. To identify the relationships between different features and make accurate predictions about future observations, advanced data analysis techniques are necessary. The term "big data analytics" refers to the application of various methods and tools to gain insights from large datasets [5]. By leveraging the results of big data analytics, organizations can enhance their decision-making processes and improve overall efficiency. Numerous analytical approaches have been established to abstract knowledge from data, ensuring organizations can make informed decisions based on reliable information.

- The analysis of historical data in a business to provide a description of past events falls under the domain of descriptive analytics, as per Joseph and Johnson [6].
- Waller and Fawcett [7] state that Statistical modelling is used in predictive analytics and machine learning techniques to anticipate imminent outcomes.
- Prescriptive analytics, which involves the use of both descriptive and predictive analytics, recommends the most appropriate actions to improve business operations, according to Joseph and Johnson.

A range of analytical techniques such as Statistical analysis, machine learning, rule-based systems, neural networks, data mining, and other techniques are used to uncover hidden patterns and improve decision-making on massive data sets. Researchers are constantly working on improving these techniques, introducing new methods, and exploring the integration of different algorithms to advance this field of study. Nevertheless, further analytical advancements are necessary to tackle the challenges posed by big data [8].

Batch processing, real-time processing, and interactive analytics are distinct platforms within the realm of big data [9]. Batch processing platforms are designed to handle complex computations that require significant period to process data. The most widely used group processing stage is Apache Hadoop, which offers scalability, cost-effectiveness, tractability, and fault tolerance for big data processing. Hadoop encompasses various modules such as the Hadoop Distributed File System (HDFS), Yet Another Resource Negotiator (YARN), and the MapReduce distributed programming model. These modules work together to support the entire big data value chain, including data aggregation, storage, processing, and management.

In Tsai et al. (2015), the authors conducted a comprehensive review of multiple studies pertaining to both contemporary and conventional large data analysis. The review primarily focused on the Knowledge Discovery in Data mining (KDD) process, which encompasses input, analysis, and output. Within the analysis phase, the authors delved into various data mining techniques, including clustering and classification. Additionally, the authors highlighted several open issues and proposed future research directions to enhance the efficiency of these methods. However, it is worth noting that the survey lacked a systematic approach, as the studies were not thoroughly compared, and recently published articles were not incorporated. Furthermore, the authors solely concentrated focusing just on the machine learning subset of AI approaches, ignoring subsets like computational intelligence.

Siddiqa and her colleagues [10] provided a comprehensive introduction to different techniques for managing big data. They presented a detailed classification system that categorized these techniques according to processing, security, pre-processing, and storage. The authors discussed various articles within each category, describing the structures of the proposed methods and comparing different techniques. Additionally, they addressed future research directions and highlighted the open challenges in the field. However, the paper did not outline a specific method for selecting articles.

3. Hadoop Architecture

Apache's Hadoop is a Java-based open-source framework that allows for distributed processing of massive datasets across computer clusters using simple programming prototypes. The framework is optimized for distributed storage and computation in a scalable environment, capable of expanding from a single server to thousands of machines, each with local computation and storage capabilities.

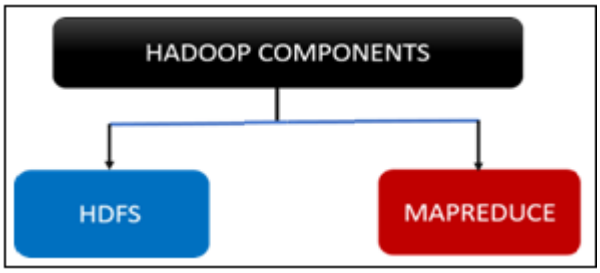

Figure 2: Hadoop Components

Hadoop has been widely implemented across various industries to meet their specific needs. Yahoo was one of the early adopters of Hadoop, and since then, other prominent companies like Facebook, Twitter, and Adobe have integrated it into their systems to enhance their operations.

Big Data may be extremely helpful in the banking and securities industry for tracking fraudulent activity, identifying credit risk, detecting card fraud, preserving audit trails, and managing customer data analytics. This aids in resolving security issues in the banking sector. Big Data is also being used by the Securities Exchange Commission (SEC) to track and monitor activity using network analytics and natural language processing. [11].

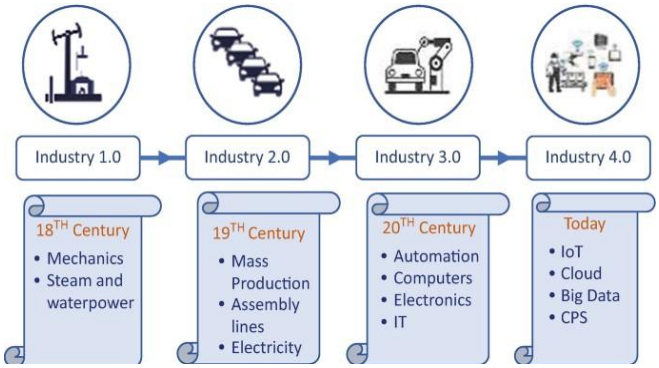

Figure 3: BIG DATA in Industry 5.0

The Big Data framework plays a crucial role in the Healthcare industry by facilitating a thorough examination of data within healthcare facilities. This analysis helps in identifying the availability of resources, tracking the increasing costs, and even monitoring the spread of chronic diseases. Similarly, in the Media and Entertainment sector, Big Data is utilized to gather, analyze, and derive valuable consumer insights. By leveraging social media components, media gratified, and real-time analytics, it identifies patterns that can be used to enhance business operations [12].
Furthermore, the prestigious Grand Slam Wimbledon Championship in Tennis utilizes Big Data to efficiently provide real-time sentiment analysis for television, mobile, and online users. This enables a more engaging and interactive experience for the audience [13].

Big Data has been successfully implemented at The University of Tasmania, an esteemed Australian university, to monitor and oversee the activities of a significant number of individuals, totaling 26,000. This implementation has proven to be highly beneficial in effectively managing and tracking their progress. Furthermore, Big Data has also been utilized to evaluate the effectiveness of teachers in relation to students' learning experiences, academic achievements, behaviour, demographics, and various other factors [14].

In the realm of Developed and Natural Resources, the utilization of Big Data has the potential to greatly enhance the capabilities of the supply chain, leading to improved productivity. Both sectors possess a vast amount of untapped data, which has been accumulating at an accelerated rate. By integrating Big Data technologies into their systems, these industries can significantly enhance efficiency, reliability, overall quality, and ultimately increase profitability [15].

4. AI Methods And Techniques

AI algorithms play a crucial role in enhancing the capabilities of big data analytics. Within the realm of IoT (Internet of Things) Data, there are three distinct types that can be selected: (1) Raw Data, which remains untouched and unstructured, Meta Data, which provides information about the data, Transformed Data, which has been enriched and given added value. The application of AI is instrumental in effectively managing and handling each of these data types by facilitating identification, categorization, and decision-making processes. By combining AI with advanced big data analytics, raw data can be transformed into meaningful and useful information that aids in decision-making.

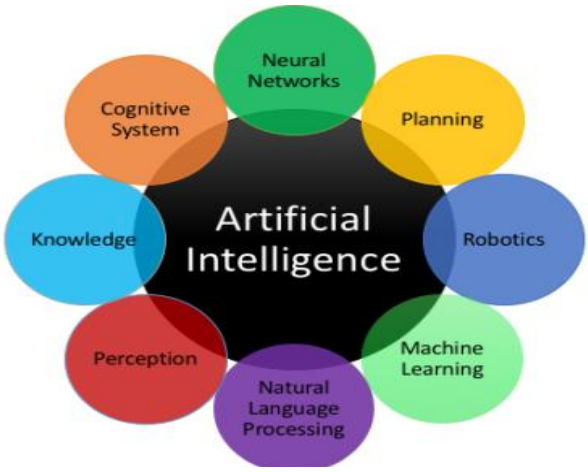

Figure 4: AI Methods

It is critical to use AI to inform decisions in the Internet of Things and data analytics domains, especially when it comes to edge computing networks, streaming data, and real-time analytics. Real-time data is very valuable for a wide range of solutions, segments, and use cases. The capacity to record real-time data, extract valuable attributes, and make judgments in real time will bring a whole new level of logic to services. The core value proposition will often be the data itself, as well as the actionable insights gleaned from it.

5. Big Data And AI Strategies

5.1 Anomalies Detection

By leveraging AI, it becomes possible to detect anomalies or unusual occurrences within a given dataset by analyzing Big Data. This capability can be particularly useful in networks consisting of sensors and predetermined parameters acceptable ranges. Whenever a node within the network falls outside of this range, it is promptly identified as a probable problem that requires kindness.

5.2 Predicting Future Outcomes

Through the application of Bayes theorem, AI can effectively analyze Big Data to determine the probabilities of future events. By considering known conditions that possess a certain probability of influencing the outcome, AI can provide insights into the likelihood of specific events occurring.

5.3 Uncovering Patterns

AI has the ability to examine vast amounts of Big Data in order to recognize patterns that may go unnoticed by human supervision. This capability is particularly valuable as it allows for the discovery of hidden patterns that can provide valuable insights and inform decision-making processes.

5.4 Analyzing Data Bars and Graphs

AI is capable of analyzing Big Data to identify patterns within bars and graphs that are derived from the underlying dataset. This enables the extraction of meaningful information from visual representations, enhancing the

understanding of complex data and facilitating data-driven decision-making.

6. AI/ML Big Data

6.1 Integration of AI in Blockchain-based financial products

The use of distributed ledger technologies (DLT), such as blockchain, has become increasingly widespread in various industries, particularly in finance. The development of blockchain-based applications is attributed to the benefits of speed, efficiency, and transparency that these innovative technologies offer through automation and disintermediation. The adoption of DLTs in finance aims to increase efficiencies by eliminating intermediaries in securities markets, payments, and tokenization of assets, which may lead to changes in the roles and business models of financial operators.

Figure 5: AI in Block chain-based financial products

The industry is promoting the conjunction of AI and DLTs in blockchain-based finance to enhance the efficiency of these systems, but the actual implementation of AI in such projects is still limited.

6.2 Data concentration and competition in AI-enabled financial services/products

Due to high provider concentrations, sensors and preset parameters may not be able to make well-informed decisions as a result of AI advancements having the potential to provide competitive advantages. Smaller financial services providers may find it more difficult to participate in the usage of AI and proprietary models since they lack the financial and human resources to implement in-house AI/ML approaches or leverage big data information sources. Inequitable data access and possible market dominance by a limited number of BigTech corporations may make it harder for smaller competitors to compete in the market for AI-based goods and services.

The possibility of network effects, which could lead to the development of new systemically significant firms, further compounds the risks associated with concentration and dependence on a small number of dominant players. BigTech businesses are an excellent illustration of this kind of possible danger, and the difficulties they face are exacerbated by the fact that they operate outside of the regulatory purview. This is primarily fueled by BigTech's access to and utilization of data, which is made possible by the application of AI algorithms to commercialize that data. A growing number of alternative data suppliers are influencing the economics of database providing, and there is a chance that market concentration will occur.

6.3 Robustness and resilience of AI models: training and testing performance

It is imperative that AI systems operate in a resilient, protected, and reliable manner throughout their lifespan, and any potential hazards must be consistently evaluated and controlled (OECD, 2019. The durability of AI systems can be enhanced by meticulous model training and performance testing based on their intended use.

To capture more complex interactions and non-linear relationships, it is necessary to train models with larger datasets. Higher order effects are more difficult to identify, hence the need for a sufficient amount of data to capture them. However, this poses a challenge as non-linear relationships and tail events are rare and may not be adequately represented in the dataset. Additionally, using excessively large datasets for training runs the risk of making models inflexible, potentially leading to a decrease in performance and learning capabilities.

6.4 Governance of AI systems and accountability

AI models are being used in important decision-making processes, such as determining access to credit or allocating investment portfolios. It is crucial to have strong governance arrangements and accountability mechanisms in place for these AI systems. Organizations and individuals involved those who create, implement, or manage AI systems must to be accountable for guaranteeing their appropriate operation. Human oversight is also necessary throughout the entire AI systems' and products' lifetime to ensure their safety and reliability.

Financial market participants currently rely on the governance and oversight structures in place for the use of AI. These arrangements are similar to those used for conventional algorithms, as AI-based algorithms aren't thought to be essentially different. However, it is important to adapt and develop governance frameworks specifically for AI, as the contemplations and risks accompanying with AI are unique. Explicit governance frameworks that clearly define responsibilities for the development and oversight of AI systems can enhance existing arrangements. Internal governance frameworks can incorporate best practices and minimal requirements for putting policies into action.

6.5 Recent policy activity around AI and finance

AI has become an increasingly important focus in policy-making due to its potential to revolutionize certain markets and the emergence of new risks associated with its use. In May 2019, the OECD took a significant step by adopting its Principles on AI, which are the first internationally recognized standards endorsed by governments to ensure responsible and trustworthy AI. These principles were

developed with input from a diverse group of experts representing various stakeholders. Given the issues addressed by the OECD AI Principles and their connection to sustainable and inclusive growth, they hold great relevance for the application of AI in the global finance sector.

The European Commission released a White Paper in 2020 that presents various policy and governing options for establishing an AI ecosystem that is both excellent and trustworthy. The proposal includes specific measures for supporting, developing, and adopting AI throughout the EU economy and public administration. It also offers potential frameworks for regulating AI in the future and addresses safety and liability concerns. Additionally, the European Union is taking practical steps to implement these measures, including funding pilot projects through the Infinitech consortium to facilitate AI-driven innovation, improve industry investment, and comply with regulations.

7. Conclusion

The demand for AI is expected to remain high in the foreseeable future as data and AI continue to merge into a mutually beneficial relationship. Without data, AI becomes ineffective, and without AI, data becomes overwhelming. By establishing connections between different data sets, a comprehensive understanding of complex problems can be achieved, leading to the discovery of new insights driven by AI. Machine learning-based appliances utilize learning methods to enhance automated decision-making, resulting in improved efficiency and precision. On the other hand, using data that is inconsistent or incomplete could produce unreliable results.

High efficiency and precision are provided by search-based optimization techniques, which use a variety of objective functions to find the best options from a number of alternatives. However, these techniques might not be scalable. By utilizing a knowledge foundation, reasoning and knowledge-based methods improve the quality of analytics.. While they may have a simpler development process, their coverage for different scenarios may be limited. Nevertheless, the mechanisms that do cover specific scenarios provide high accuracy. Companies of smaller scale are increasingly utilizing AI and Big Data in their operations. They make use of the IT hardware resources provided by data centers and leverage cloud-based AI tools to analyze the vast amount of data they gather.

References

[1] Klein S. IoT solutions in Microsoft's azure IoT suite. Berkeley: Apress; 2017. The world of big data and IoT; pp. 3–13. 2017.
[2] Chang WL, Grady N, NBD-PWG NIST Big Data Public Working Group . NIST big data interoperability framework. Gaithersburg: National Institute of Standards and Technology (NIST); 2015.
[3] Feng M, Zheng J, Ren J, Hussain A, Li X, Xi Y, Liu Q. Big data analytics and mining for effective visualization and trends forecasting of crime data. IEEE Access. 2019;7:106111–106123. doi: 10.1109/ACCESS.2019.2930410.
[4] Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. International Journal of Information Management. 2015;35(2):137–144. doi: 10.1016/j.ijinfomgt.2014.10.007. 2015
[5] Ianni M, Masciari E, Mazzeoc GM, Mezzanzanica M, Zaniolo C. Fast and effective Big Data exploration by clustering. Future Generation Computer Systems. 2020;102:84–94. doi: 10.1016/j.future.2019.07.077.2020
[6] Joseph RC, Johnson NA. Big data and transformational government. It Professional. 2013;15(6):43–48.2013
[7] Waller MA, Fawcett SE. Data science, predictive analytics and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics. 2013;34(2):77–84. doi: 10.1111/jbl.12010, 2013
[8] Qiu J, Wu Q, Ding G, Xu Y, Feng S. A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing. 2016: 67. doi: 10.1186/s13634-016-0355-x.
[9] Borodo SM, Shamsuddin SM, Hasan S. Big data platforms and techniques. Indonesian Journal of Electrical Engineering and Computer Science. 2016;1(1):191–200. doi: 10.11591/ijeecs.v1.i1.pp191-200.
[10] Siddiqa A, Abaker I, Hashem T, Yaqoob I, Marjani M, Shamshirband S, Gani A, Nasaruddin F. A survey of big data management: taxonomy and state-of-the-art. Journal of Network and Computer Applications. 2016;71:151–166. doi: 10.1016/j.jnca.2016.04.008.
[11] Antony prakash," Security Process in Hadoop Using Diverse Approach", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, ISSN: 2456-3307 (www.ijsrcseit.com), doi : https://doi.org/10.32628/CSEIT239023.
[12] Azhir E, Navimipour NJ, Hosseinzadeh M, Sharifi A, Darwesh A. An efficient automated incremental density-based algorithm for clustering and classification. Future Generation Computer Systems. 2021;114:665–678. doi: 10.1016/j.future.2020.08.031.
[13] El-bana S, Al-Kabbany A, Sharkas M. A multi-task pipeline with specialized streams for classification and segmentation of infection manifestations in COVID-19 scans. PeerJ Computer Science. 2020;6:e303. doi: 10.7717/peerj-cs.303
[14] Mittal S, Sangwan OP. Big data analytics using machine learning techniques. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence); Piscataway. 2019.
[15] Yun D, Wu CQ, Rao NS, Kettimuthu R. Advising big data transfer over dedicated connections based on profiling optimization. IEEE/ACM Transactions on Networking. 2019; 27(6):2280–2293.