A Survey Paper on Big Data

Gopal Prasad Jaiswal
Department of Artificial Intelligence
SSIPMT Raipur
Raipur,India
gopal.jaiswal@ssipmt.com

Atharv Diwan
Department of Imformation Technology
SSIPMT Raipur
Raipur,India
atharv.diwan@ssipmt.com

Modit Paswan
Department of Imformation Technology
SSIPMT Raipur
Raipur,India
modit.paswan@ssipmt.com

Abstract*— The term "big data" refers to the vast amounts of structured and unstructured data generated by organizations, individuals, and machines. This data is typically characterized by its volume, velocity, and variety, and it can come from a wide range of sources such as social media interactions, customer transaction records, sensor data, and more. Big data is a term used to describe the large, complex and diverse sets of data that are generated at an unprecedented rate from various sources such as social media, transactional systems, and other digital devices. These datasets are often so large and complex that traditional methods of data analysis and processing are no longer sufficient, and new techniques and tools have been developed to deal with them.Big data has transformed the way businesses operate and make decisions, enabling them to leverage insights from data to create competitive advantages. It has also revolutionized the way we understand and approach scientific research, medical analysis, and even social and political phenomena. By collecting and analyzing large volumes of data, researchers and analysts can identify patterns, correlations, and insights that would otherwise be impossible to see.The rise of big data has been made possible by several technological developments, including the widespread use of digital devices, cloud computing, and the development of data storage and processing tools such as Hadoop and Spark. These tools allow organizations to store, process, and analyze massive amounts of data in real-time, enabling them to derive insights and make decisions quickly.However, the use of big data also raises several challenges, including concerns about data privacy, security, and ethics. Collecting and analyzing large amounts of personal data can pose privacy risks, and the misuse of data can lead to security breaches or unethical practices. As a result, the responsible use of big data requires careful consideration and adherence to ethical and legal standards.In conclusion, big data has revolutionized the way we approach decision-making and scientific research, but its use must be approached with caution and responsibility. As the amount of data being generated continues to grow, it is crucial that organizations and individuals adopt ethical and responsible practices for data collection, processing, and analysis. The potential of big data is immense, and it is essential that we continue to explore new ways to harness its power while ensuring that its use is ethical, secure, and beneficial to society as a whole.

I. INTRODUCTION

The term "big data" refers to the vast amounts of structured and unstructured data generated by organizations, individuals, and machines. This data is typically characterized by its volume, velocity, and variety, and it can come from a wide range of sources such as social media interactions, customer transaction records, sensor data, and more.[1]

The growth of big data is driven by several factors, including the increasing use of digital technology, the proliferation of connected devices, and the rise of the Internet of Things (IoT). As a result, the amount of data being generated is growing exponentially, and traditional methods of data storage and management are no longer sufficient.

Big data presents significant challenges and opportunities for organizations that can harness its power. The sheer amount of data can be overwhelming, making it difficult to store, manage, and analyze. However, with the right tools and techniques, organizations can turn raw data into valuable insights that can inform decision-making, drive innovation, and provide a strategic advantage in the marketplace.

One of the most significant benefits of big data is its ability to provide real-time insights into customer behavior and preferences.[2] By analyzing customer transaction data, social media interactions, and other sources of information, organizations can gain a deep understanding of their customers and tailor their products and services accordingly. This can lead to improved customer experiences, increased customer loyalty, and higher revenues.

Big data can also be used to optimize supply chain management, improve manufacturing processes, and enhance product development. By analyzing sensor data and other sources of information, organizations can identify bottlenecks, streamline processes, and reduce costs. This can lead to improved efficiency, higher quality products, and increased competitiveness in the marketplace.

The field of big data has seen rapid growth in recent years, driven by advances in technology such as cloud computing, artificial intelligence, and machine learning. These tools allow organizations to store and process large amounts of data, analyze it in real-time, and derive valuable insights that can inform decision-making[4].

Despite the significant benefits of big data, it also poses significant challenges in terms of privacy, security, and ethical

considerations. The collection and analysis of personal data can raise privacy concerns, and the risk of data breaches and cyber attacks is also a significant concern. Moreover, there is a need for responsible governance and regulation to ensure that big data is used in a manner that is ethical and beneficial for society as a whole.[3]

In conclusion, big data is transforming the way organizations operate, and it has the potential to drive innovation, improve decision-making, and enhance our understanding of the world around us. However, it also requires careful management, governance, and ethical considerations to ensure that it is used in a responsible and beneficial manner[5].

II. EVOLUTION

The evolution of big data has been a long and continuous process, starting from the early days of computing and culminating in the modern era of data-driven decision making. Here is a more detailed explanation of the evolution of big data:

Early Computing Era:
The first electronic computers were developed in the 1940s and 1950s, and they were primarily used for scientific and military applications. These early computers were capable of processing relatively small amounts of data, and the idea of managing and analyzing large data sets had not yet emerged.

Relational Database Model:
The first significant breakthrough in big data came in the 1970s with the development of the relational database model. This allowed data to be organized and managed in a structured way, making it easier to store and retrieve large volumes of data. This led to the development of early business applications such as financial systems and customer databases.

Data Warehouses and Business Intelligence:
In the 1980s and 1990s, advances in networking and telecommunications technology led to the development of more sophisticated data processing systems, including data warehouses and business intelligence tools. These systems allowed organizations to collect and analyze large amounts of data from multiple sources, providing insights that were previously difficult to obtain.

Rise of the Internet:
The advent of the internet in the 1990s led to a significant increase in the volume and variety of data being generated, as more and more people began to use digital devices to create and share information. This led to the development of new tools and techniques for processing and analyzing large datasets, including data mining and machine learning algorithms.

The Big Data Era:
The 2000s saw the emergence of the big data era, as the volume and velocity of data being generated began to exceed the processing capabilities of traditional data processing systems. This led to the development of new technologies and platforms for storing, processing, and analyzing big data, including Hadoop, Spark, and NoSQL databases.

Cloud Computing:
The rise of cloud computing has further expanded the capabilities of big data, allowing organizations to store and process massive amounts of data without the need for expensive hardware investments.

Internet of Things (IoT):
The Internet of Things has enabled the collection of vast amounts of data from sensors and devices, leading to new insights and opportunities for businesses.

Artificial Intelligence (AI):
AI technologies, such as machine learning, are being used to analyze and extract insights from big data in new and innovative ways, further expanding the potential applications for big data.

In conclusion, the evolution of big data has been a long and continuous process, driven by advances in technology and the growing demand for data-driven insights. The potential applications for big data are vast and continue to expand, making it an exciting and dynamic field for innovation and discovery.

IV ARCHITECTURE

Big data architecture refers to the way in which the various components and layers of a big data system are organized and interconnected to enable the processing, storage, and analysis of large volumes of structured and unstructured data. A typical big data architecture is composed of several layers that work together to extract meaningful insights from the vast amount of data generated by modern business and social environments. These layers include:

Data Sources: This layer includes the various sources from which the data is collected. These sources may include social media platforms, weblogs, sensors, transactional databases, and other structured and unstructured data sources.

Data Ingestion: The data ingestion layer is responsible for extracting data from the various sources and bringing it into the big data system. This layer is responsible for ensuring that data is correctly formatted, cleansed, and transformed before it is stored.

Storage: The storage layer is responsible for storing the massive amounts of data collected from various sources. This layer may use different storage systems such as Hadoop Distributed File System (HDFS), NoSQL databases, and cloud storage systems.

Data Processing: The data processing layer is responsible for processing the stored data and transforming it into usable insights. This layer includes several technologies such as Apache Spark, Hadoop MapReduce, and Apache Flink.

Analysis: The analysis layer involves the use of various analytical tools and techniques to extract meaningful insights from the processed data. This layer may include technologies such as data mining, machine learning, and natural language processing.

Visualization: The visualization layer is responsible for presenting the analyzed data in a human-readable format. This layer includes various visualization tools such as Tableau, QlikView, and Power BI.

Action: The final layer is responsible for taking action based on the insights gained from the analysis. This may include automated actions, such as triggering a marketing campaign or customer service chatbot, or human-led decisions based on the insights gained from the analysis.[8]

In summary, big data architecture is designed to handle the challenges posed by the vast amounts of structured and unstructured data generated by modern business and social environments. The architecture is composed of various layers that work together to collect, store, process, analyze, and visualize data, enabling businesses to make informed decisions based on the insights gained from the analysis. By leveraging big data architecture, businesses can gain a competitive advantage, improve operational efficiency, and better understand customer behavior, among other benefits.[9]

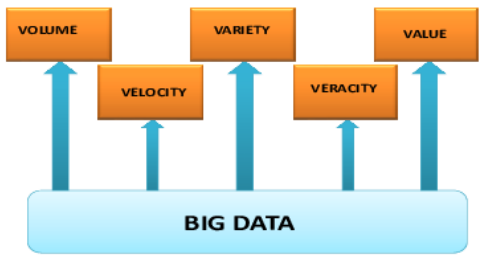

Fig-1 Block Diagram of Big Data

V BACKGROUND

The background of big data can be traced back to the exponential growth of data generated by organizations and individuals with the advancement of technology. The term "big data" was coined to describe the massive amounts of data that were being generated and collected from various sources, including social media, sensors, transactional systems, and mobile devices.

In the early days of computing, data was primarily stored in traditional databases, which were limited in their ability to handle large volumes of data. As technology evolved and data continued to grow in volume, variety, and velocity, it became clear that traditional data processing tools and techniques were no longer sufficient.[10]

To address these challenges, new technologies such as Hadoop, MapReduce, and NoSQL databases were developed. These technologies were designed to handle the large volumes of data generated by modern organizations and provide scalable and cost-effective solutions for storing and processing data.

Today, big data has become a critical aspect of modern business and social environments, enabling organizations to gain valuable insights into customer behavior, market trends, and operational efficiency. Big data is used in a wide range of industries, including healthcare, finance, retail, and manufacturing, among others.[7]

The growth of big data has also led to the emergence of new job roles and professions, such as data scientists, data analysts, and big data architects. These professionals are responsible for developing and implementing big data strategies that enable organizations to extract valuable insights from the vast amount of data generated by their operations.

In conclusion, the background of big data is rooted in the exponential growth of data generated by modern organizations and individuals, and the need for new technologies and tools to handle this data. The emergence of big data has transformed modern business and social environments, and has created new job roles and professions, making big data a critical aspect of modern society.[6]

VI KEY CHARACTERISTICS

There are several key characteristics of big data that distinguish it from traditional data. These characteristics are commonly referred to as the "3Vs": volume, velocity, and variety.

Volume: Big data refers to data sets that are too large to be processed and analyzed using traditional database management tools. The volume of data can range from terabytes to petabytes or even exabytes, and is typically generated from multiple sources.

Velocity: Big data is generated at a high velocity, with data being created and processed in real-time or near real-time. This requires new technologies and tools to capture, store, and process data at high speeds.

Variety: Big data comes in a variety of formats, including structured, semi-structured, and unstructured data. Structured data is data that is organized and can be easily processed, such as data stored in a relational database. Semi-structured data is data that is partially organized, such as data stored in XML or JSON format. Unstructured data is data that has no organizational structure, such as social media data, images, and videos.

In addition to the 3Vs, there are other key characteristics of big data, including:

Veracity: Big data is often characterized by a high degree of uncertainty and inconsistency. This can be due to the large volumes of data, the variety of data sources, and the complexity of data structures. Data quality and accuracy can be difficult to ensure in big data environments.

Value: The ultimate goal of big data is to extract value from the data to drive business insights and decision-making. This requires the ability to effectively analyze and interpret large volumes of data to identify trends, patterns, and correlations.[15]

Variability: Big data is highly variable and can change rapidly over time. New data sources can emerge, and existing data sources can change in structure or format. This requires the ability to adapt quickly to changing data environments.

In summary, the key characteristics of big data include volume, velocity, variety, veracity, value, and variability. These characteristics distinguish big data from traditional data and require new technologies and tools to effectively capture, store, and analyze data.[11]

VII TAXONOMY OF BIG DATA

Taxonomy of big data refers to the categorization of different types of data based on their characteristics, structure, and properties. The following are the three main categories of big data[12]:

Structured Data: This type of data is highly organized and easily searchable using traditional database management systems. Structured data is typically numeric and can be analyzed using mathematical and statistical methods. Examples of structured data include sales transactions, financial data, and customer information.

Unstructured Data: This type of data does not have a predefined structure or format and cannot be easily organized or analyzed using traditional database management systems. Unstructured data includes text documents, social media posts, audio and video files, and images. This type of data requires specialized tools and techniques to analyze and extract insights.

Semi-Structured Data: This type of data has a defined structure, but it is not as rigid as structured data. Semi-structured data includes data that is stored in formats such as XML, JSON, and CSV. This type of data is used in various applications, including data exchange between different systems and web-based data processing.[13]

VIII SECURITY

Security is one of the most critical aspects of big data management, as large volumes of data are often sensitive, containing confidential or personal information. Below are some of the security measures taken in big data:

Access Control: Access control mechanisms are used to manage who has access to what data. This includes authentication of users, authorization of roles and permissions, and encryption of data in transit and at rest.[16]

Data Encryption: Data encryption is a technique that transforms data in a way that it cannot be read or understood without the decryption key. This is especially important for sensitive data like personally identifiable information, financial data, and medical records.

Data Masking: Data masking is a technique used to protect sensitive data by replacing it with fictitious data. This allows organizations to use real data for testing or development purposes without exposing the sensitive information.

Backup and Recovery: Backup and recovery are essential for protecting against data loss. This includes the ability to backup data regularly and store it off-site, as well as having a disaster recovery plan in place to ensure data can be restored quickly in the event of an outage or data loss.

Auditing and Monitoring: Auditing and monitoring systems are used to track and analyze activity on the network, including data access, changes, and transfers. This can help organizations detect and respond to security incidents quickly.[13]

Network Security: Network security measures are implemented to protect data from unauthorized access, such as firewalls, intrusion detection systems, and network segmentation.

Physical Security: Physical security measures protect the infrastructure and equipment that house the data. This includes access control, video surveillance, and environmental controls such as temperature and humidity monitoring.

Compliance and Governance: Compliance and governance frameworks are put in place to ensure that organizations adhere to legal and regulatory requirements, such as HIPAA, GDPR, and PCI-DSS. These frameworks help ensure that sensitive data is protected and that organizations are held accountable for any breaches.[14]

In conclusion, security is a critical aspect of big data management, and there are several measures that organizations

can take to protect their data. Access control, data encryption, data masking, backup and recovery, auditing and monitoring, network security, physical security, and compliance and governance frameworks are all essential components of a robust security strategy for big data.

IX ADVANTAGES

There are several advantages of big data, which include:

Better decision making: Big data helps in analyzing large datasets and finding meaningful patterns and insights. These insights can help organizations make informed decisions and improve their overall performance.

Improved customer service: With big data, organizations can analyze customer behavior and preferences, which can help them personalize their services and improve customer satisfaction.

Cost-effective: Big data technologies have become more affordable over time, making it more accessible to organizations of all sizes. It can help organizations save money by optimizing their operations and identifying areas where they can cut costs.

Competitive advantage: With big data, organizations can gain a competitive advantage by analyzing their data faster and more efficiently than their competitors. They can also use data to identify new market trends and opportunities.

Better risk management: Big data can help organizations identify and mitigate risks more effectively by analyzing data from multiple sources and identifying potential threats before they become significant issues.

Improved productivity: Big data can help organizations automate certain tasks, optimize workflows, and streamline operations, which can lead to improved productivity and efficiency.

Innovation: Big data can help organizations identify new opportunities for innovation and develop new products and services based on customer needs and preferences.

Better marketing: Big data can help organizations personalize their marketing messages and improve the effectiveness of their marketing campaigns by analyzing customer data and identifying the most effective marketing channels and strategies.

Overall, big data has numerous advantages that can help organizations improve their operations, reduce costs, and gain a competitive advantage. However, it is important to note that these advantages can only be realized if organizations invest in the right technology, infrastructure, and talent to manage and analyze their data effectively.

X Disadvantage

Cost: Implementing big data solutions can be expensive, particularly for small and medium-sized businesses. This includes the cost of hardware, software, and skilled personnel required to manage and maintain the system.

Data quality issues: Big data relies on a massive amount of information, and if that information is inaccurate, incomplete, or outdated, it can lead to incorrect insights and decisions. Ensuring data quality is a major challenge in big data[18].

Security risks: Big data is vulnerable to cyber-attacks and hacking, and the consequences of a breach can be catastrophic. As data sets grow larger, it becomes more difficult to secure them against external and internal threats.

Privacy concerns: With so much data being collected and analyzed, privacy concerns are a major issue in big data. This is particularly true with the rise of social media and the collection of personal information by businesses.

Complex technology: Big data requires complex technology to manage and analyze the massive amounts of data involved. This requires a team of skilled personnel to manage and maintain the technology, which can be a challenge for smaller organizations.[17]

Legal and regulatory compliance: As big data becomes more prevalent, there are increasing legal and regulatory compliance issues that must be considered. For example, the European Union's General Data Protection Regulation (GDPR) sets strict guidelines for how personal data is collected, used, and stored.

Lack of skilled personnel: The field of big data requires skilled personnel with expertise in data science, machine learning, and other specialized areas. However, there is currently a shortage of skilled personnel in these areas, making it difficult for organizations to implement and manage big data systems.

Ethical concerns: With so much data being collected and analyzed, there are ethical concerns about how that data is being used. For example, companies may use data to make decisions that affect people's lives, but without proper safeguards in place, those decisions may be biased or unfair[20].

XI CONCLUSION

big data has revolutionized the way businesses operate, by providing insights that were previously unavailable. With the help of big data, companies can now make more informed decisions, develop better products, and provide more personalized services to their customers. However, as with any technology, there are also certain challenges that need to be addressed, such as data privacy and security concerns, the need for skilled professionals to manage and analyze the data, and the potential for biases in the algorithms used for analysis.

Despite these challenges, the benefits of big data cannot be overlooked. Its ability to process and analyze vast amounts of data has enabled businesses to gain new insights and make better decisions, improving efficiency and productivity. Additionally, big data has played a crucial role in scientific research and healthcare, allowing for more accurate diagnoses and personalized treatment plans.

As big data continues to evolve, it is important for businesses and organizations to stay up-to-date with the latest developments and innovations in the field. This includes investing in the necessary infrastructure and talent to manage and analyze the data, as well as implementing strong security measures to protect the privacy of the data. By doing so, companies can continue to reap the benefits of big data and stay ahead of the competition in an ever-changing landscape[19].

REFERENCES

[1] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171-209.

[2] Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward. In Proceedings of the 46th Hawaii International Conference on System Sciences (pp. 995-1004).

[3] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

[4] McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. Harvard Business Review, 90(10), 60-68.

[5] Russell, M., Kusiak, A., & Chen, H. (2013). Big data and analytics in manufacturing: A review. International Journal of Advanced Manufacturing Technology, 68(5-8), 1493-1508

[6] Sun, Y., & Han, J. (2012). Mining heterogeneous information networks: principles and methodologies. Synthesis Lectures on Data Mining and Knowledge Discovery, 3(1), 1-159.

[7] Voulgaris, S., & Datcu, M. (2016). Big data from space: a challenge for image analysis and remote sensing retrieval. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(8), 3577-3585.

[8] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 1(1), 1-26.

[9] Deshpande, A., Jadon, R. S., & Madaan, J. (2018). A review of big data analytics and its applications. International Journal of Advanced Research in Computer Science, 9(3), 76-83.

[10] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

[11] Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5), 662-679.

[12] Furrier J. Big data market $50 billion by 2017—HP vertica comes out #1—according to wikibon research, SiliconANGLE, Tech. Rep. 2012. [Online]. Available: http://siliconangle.com/blog/2012/02/15/big-data-market-15-billion-by-2017-hp-vertica-comes-out-1-according-to-wikibon-research/.

[13] Cannataro M, Congiusta A, Pugliese A, Talia D, Trunfio P. Distributed data mining on grids: services, tools, and applications. IEEE Trans Syst Man Cyber Part B Cyber. 2004;34(6):2451–65.

[14] McQueen JB. Some methods of classification and analysis of multivariate observations. In: Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, 1967. pp 281–297.

[15] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In : Proceedings of the ACM SIGMOD International Conference on Management of Data, 2000. pp. 1–12.

[16] srikant R, Agrawal R. Mining sequential patterns: generalizations and performance improvements. In: Proceedings of the International Conference on Extending Database Technology: Advances in Database Technology, 1996. pp 3–17.

[17] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996. pp 226–231.

[18] Burdick D, Calimlim M, Gehrke J. MAFIA: a maximal frequent itemset algorithm for transactional databases. In: Proceedings of the International Conference on Data Engineering, 2001. pp 443–452.

[19] Chen B, Haas P, Scheuermann P. A new two-phase sampling based algorithm for discovering association rules. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002. pp 462–468.

[20] Pei J, Han J, Asl MB, Pinto H, Chen Q, Dayal U, Hsu MC. PrefixSpan mining sequential patterns efficiently by prefix projected pattern growth. In: Proceedings of the International Conference on Data Engineering, 2001. pp 215–226.