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

DOI: 10.5958/2249-7137.2021.02117.0

AN OVERVIEW OF BIG DATA

Dr. Ajay Rana*; Mridul**

*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,
INDIA
Email id: mridul@shobhituniversity.ac.in

ABSTRACT

Many businesses and government agencies may now use Big Data to get important information. Such information can aid decision-makers in improving their strategies and plans. It gives the company a competitive advantage and adds value to a variety of economic and social sectors. In fact, a number of governments have launched programs to boost Big Data research and development, with significant funding. In order to maximize profits and optimize resources, the private sector has made numerous investments. This article discusses a variety of Big Data projects, opportunities, examples, and models from a variety of industries, including healthcare, commerce, tourism, and politics. It also includes examples of technologies and solutions that have been developed to address Big Data issues. Data with a lot of fields (columns) has better statistical power, while data with a lot of characteristics or columns has a higher false discovery rate.

KEYWORDS: *Big Data, Big Data Opportunities, Big Data Applications, Big Data Technologies.*

1. INTRODUCTION

Big data is a discipline that deals with methods for analyzing, methodically extracting information from, or otherwise dealing with data volumes that are too big or complicated for conventional data-processing application software to handle[1]. Data capture, storage, analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source are all difficulties in big data analysis. 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[2]. 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. The phrase "big data" is now mostly used to refer to the use of predictive analytics, user behavior analytics, or other sophisticated data analytics techniques to extract value from large amounts of data, rather than a specific data set size.

"There's no denying that the amounts of data currently accessible are massive, but that's not the most important feature of this new data environment." Data collections may be analyzed to discover new connections that can be used to "identify economic trends, prevent illnesses, fight crime, and so on." In fields such as Internet searches, fintech, healthcare analytics, geographic information systems, urban informatics, and business informatics, scientists, corporate executives, medical practitioners, advertising, and governments all face challenges with big data sets. In fields like as meteorology, genomicsconnectomics, complicated physics simulations, biology, and environmental studies, scientists face constraints. As data is collected by devices such as mobile devices, cheap and numerous information-sensing Internet of things devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks, the size and number of available data sets has grown rapidly. Since the 1980s, the world's technological per-capita capacity to store information has approximately doubled every 40 months; in 2012, 2.5 exabytes (2.5260 bytes) of data were produced per day. Between 2013 and 2020, the worldwide data volume is expected to increase rapidly from 4.4 zettabytes to 44 zettabytes, according to an IDC study.

According to IDC, there will be 163 zettabytes of data in 2025[3]. One issue that huge companies have is deciding who should be in charge of big-data projects that impact the whole company. Big data is challenging to handle and analyze for relational database management systems and desktop statistical software packages used to display data[4]. "Massively parallel software operating on tens, hundreds, or even thousands of servers" may be required for big data processing and analysis. What constitutes "big data" varies according to the skills of people who analyze it and the technologies they use. Furthermore, as technology advances, large data becomes a shifting target. When confronted with hundreds of terabytes of data for the first time, some companies may need to rethink their data management choices. For others, tens or hundreds of terabytes may be required before data size becomes a major concern. Many businesses and governments are increasingly seeing the benefits of Big Data.

In reality, effective Big Data mining allows various sectors (economic, social, medical, scientific, and so on) to gain a competitive edge and create value. The 3Vs basic features of Big Data are what characterize it the most. Velocity (data grows and changes quickly), Variety (data comes in a variety of forms), and Volume (a large quantity of data is produced every second) are the three Vs. According to these three qualities must all be present in order for a source to be classified as a Big Data source[5]. We can't talk about Big Data if one of these three Vs doesn't apply. Certain actors have added additional Vs and other features to better define Big Data: Value (relevant information can be extracted for the benefit of many sectors), Complexity (it is difficult to organize and analyze Big data because of evolving data relationships), and Immutability (collected and stored Big Data can be permanent), Vision (the defined purpose of Big Data mining), Verification (processed data comply to some specifications), Validation (the

purpose is fulfilled), Value (pertinent information can be extracted for the benefit of many sectors), Complexity (it is difficult to organize and analyze Big data because of evolving data relationships).

Furthermore, others claim that any massive quantity of digital data sets that we can no longer gather and analyze properly using current infrastructures and technology is by definition Big Data. This paper discusses a variety of Big Data initiatives, possibilities, examples, and models in a variety of industries, including health, research, commerce, transportation, tourism, and politics. It also discusses some of the technologies that are utilized to build Big Data applications[6].

1.1 Big data opportunities:

Many academics have discussed the benefits of Big Data in their respective fields. The major possibilities and advantages of Big Data in the following areas are summarized in this section: health, research, commerce, transportation, tourism, and politics. The aim is to provide a broad overview of the topic[7].

- *The healthcare industry*: Big Data analysis provides important information to the health industry. Several Big Data applications have been tried to enhance the commercial and public medical services, as well as to better assist patients and doctors. As described below, Big Data analytics may change the health domain by assisting in the optimization of operational services, providing decision-support tools, and lowering the high cost of this sector.

1) *Health-care cost-cutting*: Big Data analytics aids health-care organizations in determining which departments need to be restructured. It assists in assessing and monitoring service quality, medical unit performance, and human resource and medical equipment requirements in real time[8].

2) *Better disease evolution knowledge*: Data analytics of vast sources of information on viruses and DNAs may aid in disease evolution understanding. Doctors and researchers may use this information to discover new methods to avoid hereditary and genetic illnesses[9].

3) *Assisting medical decision-makers*: For example, a study of past surgery outcomes based on patient profiles might be coupled with an examination of a patient's current symptoms or medical data. This connection aids in the selection of the most appropriate intervention and treatment options based on the patient's profile.

4) *Improved prevention*: Predictive Big Data models can evaluate healthcare data from both the commercial and public sectors to help prevent diseases from spreading. These models work by identifying worrisome signs in the population's health. Decision-makers can devise an effective preventive strategy and halt the epidemic's progress based on the afflicted areas and population symptoms. Medical services may be customized in a variety of ways. In order to enhance patient happiness, several medical initiatives gather and evaluate patient input in real time. Real-time medical data, for example, may be used to monitor patients' health in order to adjust medication dosages or provide medical suggestions based on the studied symptoms[10].

- *The commercial sector*: If properly utilized and handled, data may be a real asset for businesses. Large and diverse data sources (e.g., internal data produced by internal business

operations, external data gathered from public sources, online sites, and data bought from other organizations) may be effectively integrated and analyzed thanks to Big Data advanced technology. In the commerce industry, big data analytics allows for the extraction of useful information. It allows you to get a deeper knowledge of your consumers' habits and preferences. It's also utilized to see how effective business tactics are. Such information allows businesses to adjust and improve their goods, services, and plans (like targeted advertizing in real-time). As a result, it makes it possible to boost customer happiness, profitability, and competitiveness. According to the Big Data market will grow by 45 percent in 2014, reaching a value of $25 billion. In general, Big Data mining provides for improved macroeconomic and microeconomic monitoring, as well as assisting decisionmakers in spotting commercial possibilities and anticipating recessions . Facebook, Google, and Amazon, for example, gather and sell data on web users' behaviors, feedback, comments, and online purchases. Credit card firms (such as Equifaxe and Transunion) follow suit in order to boost revenues and improve services. Furthermore, the development of various ICTs, as well as high connectivity across many organizations (e.g., corporate subsidiaries, partners, suppliers, and consumers online), has resulted in new business models based on real-time Big Data sharing and monetization. Indeed, as points out, businesses may use Big Data at several phases depending on their maturity level:

1. Rather than simply monitoring internal company operations, businesses may use Big Data analysis to better understand consumer behavior and improve their commercial strategy and goods.
2. Businesses that have reached a particular level of maturity are more ready to improve their procedures and identify new possibilities.
3. The company reaches the next maturity level when it is able to monetize the value of the gathered Big Data in addition to optimizing its business model. This may be accomplished by reselling data and analytical findings for extra profit. Another option is to use Big Data insights to improve goods and the consumer experience in shops and online.

- *Agriculture sector*: The use of Big Data produced in the agricultural industry may provide useful information. Such knowledge allows for the optimization of production methods, the adaptation of plans in response to climatic forecasts, the monitoring of demand by area and customer profiles, and much more. This may be accomplished by evaluating data from a variety of heterogeneous sources (e.g., weather and history, demand forecasts, and smart sensors). A Japanese initiative, for example, seeks to create an enhanced recommendation system based on Big Data analysis, according to [8]. The aim is to suggest the best product combinations, restaurants, and manufacturers that provide goods that are in accordance with the consumers' tastes to internet users (e.g., Bio products or products with with no allergic substances). By just filling out the patient's symptoms, the system will also suggest appropriate goods and their providers. Its goal is to link various entities (users, restaurants, and producers) and provide granular data access based on user profiles.
- *The tourism industry*: To enhance tourism operations and better serve visitors, many Big Data models have been created or are in the works. In reality, businesses may utilize Big Data technology to get important insights, such as a better understanding of visitors'

behaviors, the identification of changing preferences and requirements, and the monitoring of tourists' geo-position, activities, and context. It is feasible, for example, to suggest hotels, restaurants, and activities to visitors in real time based on their interests, online habits, and geolocations. Suggested a Tourist Recommendation System in this regard. This system is built on a foundation of comprehensive Big Data analysis and visualization capabilities, including: I an examination of past visitor activity patterns. ii) Real-time analysis of current tourist activities, preferences, profile, and website visits. iii) the tourist's whereabouts is being tracked. iv) Other factors such as weather and traffic congestion are monitored. The aim is to provide customized real-time recommendations.

- *Politics Sectors*: Many governments are evaluating various sources of moving or static data in real time (for example, logs, historical events, public and private surveillance cameras, citizen comments on social media, online transactions, GPS data, and mobile communications) . They also make use of data produced by a variety of government ICTs. The aim is to uncover useful information, trends, and correlations, or to develop prediction models that will allow the government to improve its policies and services to people. Another key aim is to maintain constant surveillance and monitoring in order to safeguard people and reduce the effect of crimes. For example, the government may use sophisticated Big Data algorithms to anticipate events that might jeopardize the country's security or to identify suspects, criminal organizations, and terrorists. However, monitoring individuals' communications and activities raises a slew of privacy concerns that are difficult to address .delves into these problems. Furthermore, political scientists and professionals may utilize Big Data analytics to extract useful information. This kind of knowledge allows people to get a better grasp of political problems. For example, offers a geopolitical analysis-based Big Data application. This program assesses President Barack Obama's political beliefs over a certain period of time. The program downloads Obama's speeches from the White House website, cleans them for consistency, and extracts the data sets that are relevant to the use case. The program use data mining methods to measure the president's attention on political topics, examine his emotions, and determine his mode in the face of significant political events. The suggested model may be used to identify political trends, predict the effect of elections on the country's development, verify political views, track political goals, and assess people' confidence in the present political environment, among other things.

1.2 Big Data Technologies:

- *Hadoop Ecosystem*: Hadoop Framework was created to store and process data in a distributed data processing environment using a simple programming paradigm. Data from a variety of high-speed and low-cost devices may be saved and examined. In the last year, businesses have embraced Hadoop as a Big Data Technology for their data warehouse needs. The trend seems to be continuing and accelerating in the next year. Companies that haven't looked into Hadoop yet are likely to see its benefits and applications.
- *Artificial Intelligence*: Artificial intelligence (AI) is a broad field of computer science concerned with the creation of intelligent machines capable of performing tasks that would normally require human intelligence. From Apple's Siri to self-driving cars, AI is rapidly evolving. As an interdisciplinary branch of science, it considers a variety of approaches, such

as increased Machine Learning and Deep Learning, to make a significant change in the majority of tech industries. Existing Big Data Technologies are being revolutionized by AI.

- *NoSQL*: Database NoSQL Database NoSQL Database NoSQL In the database, NoSQL includes a variety of different Big Data Technologies that were created to design modern applications. It depicts a non-SQL or non-relational database with a data acquisition and recovery method. They are used in real-time Web and Big Data Analytics. It saves unstructured data and provides quicker performance and flexibility for a variety of data formats, including MongoDB, Redis, and Cassandra. In a variety of devices, it offers design integrity, simpler horizontal scalability, and control over possibilities. By default, it utilizes data structures that aren't related to databases, which speeds up NoSQL computations. Every day, Facebook, Google, Twitter, and other comparable businesses keep gigabytes of consumer data.
- *Programming in R*: R is an open-source Big Data programming language and technology. The free program is extensively used for statistical computation, visualization, and help communication in unified development environments like Eclipse and Visual Studio. According to experts, it was the most widely spoken language on the planet. Data miners and statisticians utilize the system to create statistical software and, in particular, data analysis.
- *Data Lakes*: In terms of structural and unstructured data, Data Lakes refers to a centralized repository for storing all data types at all levels. Data may be stored in its raw form without being converted into structured data during data accumulation. It allows for real-time data analysis ranging from dashboards and data visualization to Big Data transformation for improved business intelligence. Businesses that utilize Data Lakes remain ahead of the competition by doing new analytics, such as Machine Learning, using new log file sources, social media data, and click-streaming.
- *The Beam*: Apache Beam provides a simple API for building complex Parallel Data Processing pipelines using a variety of Execution Engines or Runners. In 2016, the Apache Software Foundation created these Big Data technologies.
- *Using Docker*: Docker is a Big Data technology that simplifies the creation, deployment, and operation of container applications. Containers assist developers in loading a program with all of the necessary components, such as libraries and other dependencies.
- *Flow of air*: Apache For the administration of data pipelines, Airflow is a Process Management and Scheduling System. DAGs (Directed Acyclic Graphs) tasks are used in Airflow job processes. The process code description makes it simple to handle, verify, and version huge amounts of data.
- *Block chain technology*: Block chain is a Big Data technology that has a unique data secure feature in the digital Bitcoin money that prevents data from being erased or changed once it has been published. It's a highly secure environment that's a great fit for a variety of Big Data applications in sectors like manufacturing, banking, insurance, medical care, and retail, to mention a few.

2. DISCUSSION

Big Data refers to all of the data that is being produced at an unprecedented pace across the world. This information may be organized or unstructured. An economy that is strongly knowledge-oriented owes a large portion of today's commercial companies' success. Data is what drives contemporary businesses across the globe, therefore making sense of it and unraveling the different patterns and exposing previously unknown relationships within the enormous sea of data becomes essential and very gratifying. There is a need to turn Big Data into Business Intelligence that businesses can use right now. For businesses of whatever size, region, market share, customer segmentation, or other categorizations, better data leads to better decision making and a better method to plan. Hadoop is the platform of choice for handling massive amounts of data.

3. CONCLUSION

There are many instances of Big Data possibilities and solutions, as we've seen in this article. We can conclude that analyzing Big Data is beneficial for gaining reliable insight. Such knowledge enables decision-makers to make sound decisions, improve policies and strategies, maximize profits, and improve the competitiveness of businesses. Furthermore, by allowing advanced complex analysis across multiple sources, the Big Data revolution contributes to the enrichment of various scientific fields. New business models, such as data value monetization and business metamorphosis, have emerged as a result of Big Data analytics. In fact, organizations use Big Data to varying degrees of maturity. They can not only rely on real-time Big Data analysis to optimize strategies and processes as they grow and mature, but they can also monetize the value of Big Data. They can then focus their efforts not only on improving services and products, but also on developing their ecosystem. A single platform that links all ecosystem participants is needed to support the growth of organizations. To support the needs of various parties (e.g., governments, enterprises, customers, administrations, suppliers, social network communities, and users), this platform should rely on Big Data analysis and modeling. Entities should be able to get a better understanding, immediate feedback, and customized suggestions via such a platform. The goal is to maximize all of the entities' profits. Traditional technologies, on the other hand, are incapable of dealing with Big Data challenges (i.e., velocity, volume, variety and complexity). To ensure performance, results reliability, data availability, and scalability, Big Data modeling and mining necessitate advanced technologies and methods. Another challenge is striking a balance between various security and privacy requirements, fast, reliable processing, and granular role-based access to a number of highly connected Big Data sources. To meet such challenges, a variety of technologies have been developed. However, there are numerous drawbacks. Many areas are still open to research in order to improve the features and capabilities of Big Data applications.

REFERENCES

1. V. Marín Díaz, “Trabajar en la era digital. Tecnologías y competencias para la transformación digital.,” *Pixel-Bit, Rev. Medios y Educ.*, 2016.

2. “Big Data: una herramienta para la administración pública,” *Ciencias la Inf.*, 2016.

3. J. van den Belt and C. Nillesen, “Kritische succesfactoren voor de inzet van Big Data: is groot wel anders?,” *Maandbl. Voor Account. en Bedrijfsecon.*, 2014, doi: 10.5117/mab.88.31273.

4. Y. Taminiau, S. Heusinkveld, and J. Both, "Nevenschade of nieuwe kansen? De impact van accountancywetgeving op de professie van fiscalist," *Maandbl. Voor Account. en Bedrijfsecon.*, 2016, doi: 10.5117/mab.90.31264.

5. B. Majoor, "Kip of ei: de juiste focus in het vraagstuk van kwaliteit van accountantscontrole?," *Maandbl. Voor Account. en Bedrijfsecon.*, 2017, doi: 10.5117/mab.91.24056.

6. Q. Kunyuan, "Innovation and Corporate Debt Financing: Evidence from China's Listed Companies," *South. Financ.*, 2012.

7. W. Snoei and N. van Nieuw Amerongen, "Toepassing van (big) data-analyse in de MKB-jaarrekeningcontrole in een relatief eenvoudige omgeving," *Maandbl. Voor Account. en Bedrijfsecon.*, 2015, doi: 10.5117/mab.89.31178.

8. A. R. Kan, "Big Data in Zicht," *Natl. Denk Tank*, 2014.

9. V.-D. Tran And N. D. T. Huynh, "Exploring The Relationships Among Social Benefits, Online Social Network Dependency, Satisfaction, And Youth?S Habit Formation," *Main Issues Pedagog. Psychol.*, 2017, Doi: 10.24234/Miopap.V15i3.177.

10. H. K. L. Nguyen and B. N. Nguyen, "Mapping biomass and carbon stock of forest by remote sensing and GIS technology at Bach Ma National Park, Thua Thien Hue province," *J. Vietnamese Environ.*, 2017, doi: 10.13141/jve.vol8.no2.pp80-87.