North American Journal of Engineering and Research

Est. 2020

SECURITY AND PRIVACY TECHNIQUE IN BIG DATA: A REVIEW

Kartheek Pamarthi

Email: Kartheek.pamarthi@gmail.com

ABSTRACT

The importance of Big Data as a foundational component of the AI and ML landscape is not going away anytime soon. As a result, the past fifteen years have seen a tremendous investment in Big Data research. The purpose of this literature review is to compile the most recent results from Big Data studies conducted over the past fifteen years. The study will address questions about the main applications of Big Data analytics, the main challenges and limitations encountered by researchers, and the present and future state of Big Data studies. The review follows a predetermined procedure that automatically examines five major digital libraries. Among the more recent branches of computer science is the study of large amounts of data. Social media, online shopping, blogs, financial institutions, healthcare providers, transactions, websites, applications, opinion forums, and a host of other sources all contribute to the accumulation of data. Various businesses, notably healthcare, make excellent use of it after processing. For data processing and analysis in the industrial sector, these massively generated datasets are indispensable. In order to investigate the value and potential of big data in healthcare and industrial processing applications, this article surveys the published works of numerous writers who have helped with data collecting, analysis, processing, and viewing. Data cleansing, missing value analysis, and outlier analysis are some of the opportunities and problems highlighted, in addition to the benefits and applications of big data. Outlier detection models, as well as methods for detecting and cleaning unclean data, were also suggested in this comprehensive review.

Keywords: Big data, Security, Privacy.

INTRODUCTION

During the course of the last fifteen years, Big Data has developed into a fundamental pillar that offers assistance to a wide variety of scientific fields. To name just a few examples, these fields include medicine and healthcare, engineering, and telecommunications. A seemingly endless supply of workable solutions has been the product of an explosion in research efforts over the last fifteen years aimed at overcoming the most intractable problems with Big Data. Countless scholarly studies have been generated as a result, ultimately revealing the nature's duality and contradiction. On one hand, we have evidence that this scientific area has been crucial in shaping the technological accomplishments of our day. However, researchers often make the mistake of confusing the theory (of Big Data) with its practice or use when they base their understanding of Big Data on this seemingly endless universe of tens of thousands of technical papers, each focusing on a specific sector. This approach is no longer sustainable. There have been several attempts to conduct survey studies in an effort to characterize the Big Data ecosystem as a whole, and this fact cannot be ignored. But most of them fell into the trap of using pre-existing models and tried to provide as detailed an answer as they could to the specific needs of a single subfield or a small number of perspectives. Reason being, there is a vast amount of material covered by the subject. We have chosen to undertake a new type of comprehensive investigation that was not skewed by the particularities of any particular industry or muddled by any particular technical viewpoint in order to advance our understanding of the current state of Big Data research during the aforementioned time frame. Given the intricate circumstances mentioned before, it was decided to carry out this investigation. In software engineering, Kitchenham and Charters[1] suggested a method called a systematic literature review (SLR) [2, 3].

In order to accomplish this step, we have followed the methodology. In spite of the fact that SLR is carried out in accordance with a series of clearly delineated procedures, it is necessary to make an initial decision concerning the paramount parameters that should be utilized in order to investigate the subject of the inquiry. When it comes to Big Data, our main goals have been to understand its main uses, identify the obstacles and limitations that researchers face when analyzing the massive datasets they handle, and reveal where the field is heading in terms of future research.

These are all things that we have been working on.
Consequently, we began with three research questions that were a fit for the points that were brought up before. These questions were generally as follows: i) the most popular application areas; ii) the current research problems and constraints; and iii) emerging future trends and directions. The organized methodology offered by SLR served as a framework for these inquiries. From this point on, we proceeded in accordance with the instructions specified in the SLR.

In essence, we began by converting the three research questions that were presented earlier into particular search phrases. Scopus, IEEE Explore, ACM Digital Library, SpringerLink, and Google Scholar were the five separate digital libraries that were subsequently examined using these search phrases. After doing the search, which will be detailed more extensively in the next section of this study, 189 primary papers that met our basic search parameters were located. Out of the 189 studies, only 32 were true reviews of the literature. Focusing on these 32 survey studies allowed us to achieve our research goals of providing a thorough overview of Big Data research over the past 15 years by (a) illuminating the most common application domains, (b) drawing attention to the challenges faced by researchers, and (c) predicting possible future research directions. In this last analysis, we found that while our understanding of Big Data has grown overall, there is a certain amount of stability in terms of research interests in this field. This stability is achieved when the distribution of challenges, future trends, and application domains is balanced. An argument has been advanced that suggests Big Data solutions are quickly finding their way into many facets of daily life, such as the energy sector, smart cities, and healthcare. Conversely, researchers have invested much in data quality management, creating and executing new frameworks to process Big Data in real-time, and ensuring the security and privacy of this data. Future data-driven sophisticated software solutions will incorporate Big Data, but there are several challenges that must be overcome before. Some of these challenges include making systems use less energy, making algorithms more efficient, making frameworks more secure without sacrificing privacy or ethics, integrating AI and ML technologies, making data sets public, improving interoperability among different stakeholders, and considering societal and business changes.

A SYSTEMATIC REVIEW ON BIG DATA APPLICATIONS

Big data is defined in a variety of different ways by different writers, however there is no definition that is appropriate. Volatility, diversity, and velocity are the usual ways to characterize big data.

However, it encompasses a wide range of Vs, including veracities, vagueness variability, susceptibility, volatility visualization, and many others. The application areas of big data include business organization, operations, production, marketing, and management of information technology, among other types of applications. All of these sectors rely heavily on multi-level data management and processing to get the most out of the data.

This information must be categorized and filtered specifically for each individual working in the business. The development of data science will be influenced by a variety of algorithms that are utilized for a variety of purposes. The big data management life cycle consists of the following concepts: study, data collection, documentation, integration, preparation, analysis, publication, sharing, storage, and reuse. Many companies in the media, entertainment, and communication sectors use big data to further their own corporate objectives. Business sectors collect and analyze consumer behavior data for the purposes of audience targeting and product suggestion making. Big data analytics, when applied, can solve complex problems [7]. From 2015 to 2019, healthcare big data presentations showcase a range of strategies. Statistical data analysis, hidden Markova model, and bioinformatics-specific machine learning methods are all part of this category of approaches. The healthcare industry is currently one of the industrial domain's victims of data breaches, which happen through a number of medical field gadgets.

It is necessary for healthcare information systems that are utilized effectively by multiple institutions at the same time to concurrently incorporate significant information in an efficient manner. Filtered data contributes to the provision of improved medical services. The profit margin, production, and operations of industrial businesses are all improved by big data [5]. There are numerous applications and technical disciplines that rely on big data, including those that are based on artificial intelligence, data analytics, and other similar fields. By 2022, the Big Data industry in the US will be worth 72.38 billion USD, according to Grand View Research, Inc. Big data management is crucial in healthcare because of the importance of big data in the medical information system, which is required for the assessment of diagnoses using large datasets [8]. The data's origins, uncertainty, and the presence of unclean or noisy data are all potential threats to its integrity. Algorithms like deep learning, machine learning, and natural language generation (NLG) are having an indirect or direct impact on applications.

Some examples of these applications include healthcare, smart cities, industry 4.0, and others. Pattern designing is one application of deep learning, which in turn helps to improve optimization and computations. The employment search, product customisation, cost optimization, and other activities can all benefit from data science. The big data evaluation hierarchy begins in the 1970s with statistical computing and continues through enormous data sets, data mining with statistical learning, business analytics, the 2000s, and big data analytics in 2010. Conversely, companies are ready to deal with some challenges and apply the big data analytics technologies mentioned in [9], as well as to implement them. System health monitoring and management (SHMM) is introduced within the framework of the healthcare system. For the purposes of forecasting, prediction, diagnosis, and related operations, this method integrates both active and passive data. The article discusses healthcare cloud computing and the Internet of Things (IoT) in [10]. Similarly, the healthcare business can benefit greatly from applications that utilize big data. Many different areas and services within healthcare are benefiting from Big Data technologies. This all-inclusive coverage includes, but is not limited to, better medical diagnosis, more accurate treatment, reduced costs, higher operational efficiency, disease detection, and advanced patient care. When it comes to healthcare, data quality, efficiency, and effectiveness are king.

End-users are supported by multimedia approaches, which are used in the healthcare industry as well as other industries to improve efficiency, coordination, and focused system services. The objective of healthcare analytics encompasses a multitude of goals, including but not limited to increased services, real-time monitoring, clinical decision support systems, and patient safety. The industrial sector cannot function without massive data sets and the technical means to access and use these datasets. One may learn about it by looking at the value chain, which goes like this: data gathering, analysis, curation, storage, and usage [11]. Data is retrieved for a variety of industrial operations from a variety of sources, including the internet, industrial incorporation, ERP, CRM, HRM software modules, the social media network, transactions, healthcare, geographical data, remote sensing, audio-video recording, and even more sources.

There are also other sources. It is provided that the definition of big data for remote sensing (RS) and satellites with volume and velocity are provided. In order to handle large applications, scalable data management is essential, and big data management is required for this. Large amounts of data offer a wide variety of new scientific study and value methodologies that can be used to measure economic growth [12]. Within the realm of electronic commerce, the majority of information and data are gathered by electronic means, which can lead to a variety of issues regarding data evaluation and quality.
Big data is perfect for this since it allows us to incorporate alternative healing sources into the data life cycle of accessible health records. There are several steps that make up the big data life cycle, including data generation, acquisition, storage, analytics, and visualization. The data collection system utilized in the agricultural sector is described in detail in [13]. The accuracy, consistency, missing value, data cleaning method, dataset duplication, and usability of big data are all affected by its sources. These factors are relevant in many domains, including healthcare and business analytics. Data analytics, tools for company operations, report preparation, and related tasks can be discovered through industrial processing, depending on the format of the big data. There are a plethora of processes that could benefit from the application of big data in manufacturing. These include, but are not limited to, manufacturing, quality assurance, supply chain risk, tailored product creation, and many more.

The authors have provided an explanation of the data analysis that manufacturers use, which assists in the monitoring and detection of faults during the manufacturing process. Users receive data through the support system, which includes monitoring of processes, detection of anomalies, and analysis of faults. Among the various uses are companies in the chemical industry, complex processes, harvester process control, and other similar applications [14]. One area where big data might assist enhance the efficiency of older manufacturing is in the detection of faults. Failure identification and detection are the two stages that are involved in this process. A number of different industrial domain sectors, including production, quality evaluation, supply chain, and others, are affected by big data, which in turn has an effect on the running costs of an organization.

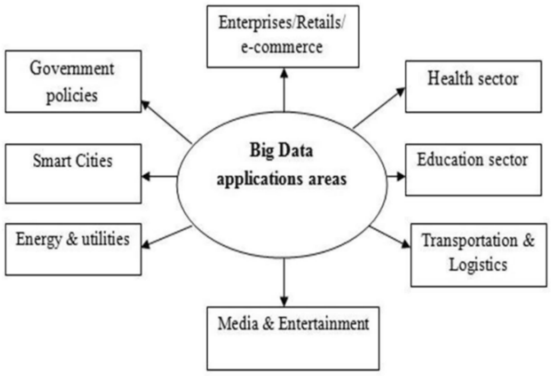

Fig. 1 Big data applications.

Big data technologies have many potential uses; some examples include smart cities, healthcare, education, transportation, social media, business operations, promotions, event planning, and online shopping. Particularly, these sectors generate a great deal of data sources amenable to big data processing and storage. However, privacy and security concerns are growing in importance when considering large data [15]. Cybercriminals, unstructured data, physical risk, and security breaches are some of the many elements that constitute big data, privacy, and security.

These characteristics include volume, variety, velocity, and veracity, among others. An overview of the function that big data plays in medical diagnosis and treatment is provided here. The use of medical instruments like as MRI, CT, FMRI, and others helps to create a significant amount of data in this location. Credit risk assessment, AML, trade analytics, predictive analytics, and KYC are just a few of the numerous banking-related uses for big data. By leveraging big data analytics, video surveillance may be stored, retrieved, processed, accessed, and run more efficiently. The progress of national development, technological breakthroughs in industry, scientific research, interdisciplinary research, and better future forecasting all contribute to the significance of big data. Many other sectors can benefit from big data as well, including the financial sector, insurance, manufacturing, the natural resources sector, and many more (Fig. 1).

Government policies

The use of big data is beneficial and beneficial for a variety of strategies. It is beneficial to the government in that it helps them comprehend the clusters of objects that supply facilities to end-users. Identifying regions and ways that will benefit the population of the nation is made easier for the government by the collection of data sets. It comprises information obtained from each and every transaction as well as other access to information technology resources like e-Mitra. The public's direct beneficiaries of government schemes can be discussed and exchanged on a platform provided by big data. Despite this, screening remains a challenge, as does obtaining the necessary data and information; this is a major problem in surveys of programs for people living below the poverty line (BPL). As of this writing, it is unclear whether the primary beneficiaries would receive any kind of government assistance. Citizens are thus helped in obtaining timely access to government resources by the big data cleaning and clustering approach to data filtration. Applications in agriculture that make effective use of large data, such as crop selection analysis, can also benefit from data cleansing [16].

Smart cities

Intelligent cities, which include "smart health, smart logistics, smart education, smart transport, smart energy, etc.," are the focus of big data applications in recent years. Big data applications have recently become concerned with smart cities. On the other hand, these components call for distinct data sets, acquisition methods, abstraction methods, categorization and grouping of characteristics, and so on. With a living model that is both efficient and practical, people are able to enjoy a high quality of life. Once the system of a city has been built in every aspect, it is accessible to each and every resident.

Various classification algorithms, such as multi-class classification, Naïve-Bayes classification, and others, have been presented as efficient methods for classification [17]. The provision of support for innovative health services is one way that intelligent cities raise the living conditions of their residents.Among other things, cutting-edge smart health services require a number of infrastructure components, including sensor facilities, cloud services, real-time data processing and sharing, mobile-to-cloud transition, and sensors systemcovercloud. Virtual machine (VM) migration technology transfers heterogeneous data to the cloud to improve storage efficiency and ease of access. A virtual machine (VM) model including an ant colony optimization (ACO) method for heterogeneous cloud computing systems (CCS) has been found by the authors to enhance the quality of services delivered by the new health system. For the decision-making system that will build a smart city, MapReduce and HADOOP are the way to go for huge data based on the Internet of Things. When it comes to building smart city frameworks, the Internet of Things is quickly becoming an essential tool for providing ICT assistance. The Smart City Data Analytics Panel (SCDAP) was set up with the intention of integrating data model management, ICT tools, and functionality into the construction of urban amenities for residential people.

Energy & Utilities

Traditional systems, such as geographic information systems (GIS), have begun to be replaced by cloud-based and big data systems in the energy and utilities industries. The proliferation of open-source technologies that facilitate data cleansing and filtering from massive datasets has paved the way for this shift. The energy and utility sectors benefit from the ubiquitous sensor data generated by industrial activities. Market predictive analysis models and geographical analysis models (including inquiry, building, transportation, and distribution) may find use for this sensor data. When a significant amount of spatial, sensor, and geographic information system (GIS) data is converted using HADOOP, big data open-source solutions have the potential to save both time and money [18].

Education sector

ICT, which stands for information and communication technology, has become an indispensable instrument in the field of education in recent times, and every government has been actively pushing it ever since it was first introduced. It contributes to an increase in the effectiveness and efficiency of education at both the teaching and learning levels. In addition to this, it improves the outcomes and productivity on a variety of levels, and the government has been providing support for the implementation and improvement of assessments on a consistent basis in order to benefit citizen stakeholders. Through the implementation of operational learning stages, the society as a whole is educated and comes to comprehend the significance of growth in every aspect. Innovative education policies create an environment that is conducive to active learning for stakeholders. Citizens are able to access resources, participate in implementation initiatives, and get data and knowledge through the use of information and communication technology technologies. A knowledge-sharing pool was created with the assistance of information and communications technology (ICT) and big data technologies. Institutions interacted with one another, and companies and academic institutions collaborated, all of which contributed to the development of healthy societies and nations [19].

E-commerce

Since this application is now in the commercialization stage, its values are going up, and it's one of the most important uses of big data. By keeping track of clients who buy supported products, e-commerce businesses can utilize big data to identify them based on a health context. Many distinct ideas can give rise to the notion of huge data. Profit maximization through providing goods and services in a competitive market while also meeting the needs of stakeholders and consumers is the fundamental objective of any industry. When it comes to the e-commerce industry, big data analytics are crucial for a number of reasons, such as decision making, market segmentation, infrastructure, and transparency. Many different types of big data, such as company operations, audio, video, and transaction records, have several potential applications in the e-commerce space. But business value defines all the roles of e-commerce in one place [20]. According to [21], a CPG firm produces 4 trillion data points annually, as a consequence of the 1,52,000 samples generated every second. There needs to be efficient and effective management of this data throughout all the data processing phases.
The authors described big data analytics as the study of human interactions with data analytics, integration of processes, and automation; and the process of discovering data values for industries and the economic system. The objective is to enhance the business process system in a way that makes it suitable for organizationally related functions, such as those involving transparency, description, and prediction. This will be achieved by providing business personnel with the necessary knowledge about drivers and how to deploy advanced analytical methods. Concerning supply chain companies' use of big data, twenty-four articles were published in 2016. The authors have mentioned a number of supply chain industries that make use of big data analytics, including healthcare, manufacturing, and finance [22]. The application of IT vision and big data architecture principles can lead to the realization of value for the business.

Transportation and logistics

Another area where big data is making an impact is in the realm of e-commerce and logistics, which together promote many businesses for financial gain. But simultaneously, a lot of problems emerge due to the massive quantities of inaccurate data generated by different kinds of machinery. Data analytics technologies require component integration in the transportation and logistics sector, however problems still can't be solved to an adequate degree. To find useful findings, data analytics technologies like OpenRefine, Knime, R-programming, Orange, and others are used to evaluate, filter, and convert massive amounts of data. Processing is becoming more important to provide features like optimal infrastructure, improved customer service, and the prediction of components like the wrong road, traffic, passenger availability, etc., due to the increasing amount of transportation-related data, which includes passengers on aircraft worldwide [23]. This system for processing data involves refining and collecting many data sets from a variety of sources, as well as aggregating these data sets for processing purposes. The tourism industry is another industry that makes use of big data to locate potential tourists based on the data collected by the industry.

Media and entertainment

This is yet another area of big data applications in which media and entertainment companies are looking for ways to gain economic access to the resources that big data has to offer. The media and entertainment businesses are able to make better use of their big data resources when they have access to vast numbers of digital audiences on their platforms. The media and entertainment businesses make use of big data in order to make predictions about what audiences desire, optimize scheduling, boost customer acquisition and retention, and target advertisements more effectively, among other things [24]. One of the issues that big data presents in the media and communications industry is the utilization of mobile material, as well as gathering, analysis, and pattern recognition. There are many different applications of big data in the media and entertainment industry. Some examples of these applications are data journalism, dynamic semantic publishing, social media analysis, crisscross-sealing ducts, product creation, and audio interpretation.

SECURITY MECHANISMS USED IN BIG DATA ANALYTICS

The security procedures that are applicable to big data analytics will be the topic of discussion in this section. In the realm of big data, numerous types of security techniques are utilized. In most cases, security mechanisms can be divided into two distinct kinds. Various mechanisms, both cryptographic and noncryptographic. A classification of Big Data Analytics Security Mechanisms is presented in full below in Figure 2, which can be found below.

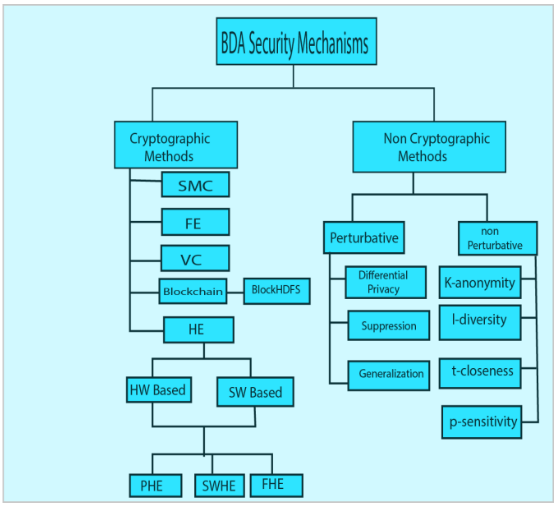

Fig. 2. Big Data Analytics Security Methods.

5.1 Cryptographic Security Mechanisms

A mathematical formula is utilized by cryptographic security systems in order to render the communication unintelligible to any individual who is not allowed to read it. Homomorphic Encryption (HE), Verifiable Computation (VC), and Multiparty Computation (MPC) are a few examples of the cryptographic techniques that have been provided during this discussion. When we outsource our data to cloud service providers, there is a possibility that cloud service providers would use it for their own purpose. As a consequence, we must take precautions to prevent unauthorized access to, alteration of, and sharing of our data with third parties. It is proposed to use probabilistic encryption in conjunction with fully homomorphic encryption. The authors of [25] examine the similarities and differences of HE, VC, and MPC settings in three distinct cloud scenarios: trusted, semi-trusted, and untrusted. Using Fully Homomorphic Encryption (FHE), which is an emerging and powerful cryptosystem that can execute calculations on encrypted data without decrypting it, the authors of [26] offered a privacy-preserving distributed analytics framework for big data in the cloud. This framework was developed by the authors. For the purpose of protecting cloud computing from the significant privacy breaches and leakage threats that were experienced by both the external and internal parties, FHE is utilized. "The existing privacy-preserving data mining approaches have several problems," as stated in [27], including the fact that they are inefficient in protecting data privacy, have poor performance, and rely excessively on a Trusted Third Party (TTP), which is regarded to be a security risk. The authors make use of FHE in order to protect the confidentiality of the data of users while it is being kept and processed in the cloud. A method known as Extremely Distributed Computing (EDC) is utilized by the authors in order to lessen the impact that the processing of encrypted data has on the overall performance of the system. For the purpose of developing a cloud-enabled application that is secure, their testing results on FHE were generated based on two parameters: analysis performance and accuracy. It is possible for FHE to successfully support the operation of addition and multiplication operations at the same time. In spite of this, the applicability of FHE techniques for real-world applications continues to be unattainable due to the large amount of computing overhead they require.

The authors of [28] suggest a homomorphic encryption scheme that uses ECC and SMC to reduce computation and communication costs. They compare this scheme to RSA and Paillier's algorithm and find that it offers better energy efficiency, communication consumption, and privacy protection. The authors also demonstrated the scheme's feasibility, great encryption effect, and high security by applying ECC to the calculation of GPS data of seismic measurements in the seismological Bureau of Fujian Province. This was done to protect national secret data. Existing big data platforms have operational concerns, one of which is the leaking of private data. To make a scalable Big Data platform more secure, the authors of [29] suggested using SHRED algorithms, which stand for "Stretched Homomorphic ReEncryption Decryption," a privacy-preserving approach. Optimization involves the introduction of Laplacian noise. The authors assert that by padding over deterministic cryptosystems, the approach is protected against plaintext attacks. SHRED further guarantees safe recovery of public and private keys. Data breaches expose personal and company information to various privacy concerns, due to the widespread adoption of the Big Data paradigm that does not prioritize security. Companies faced with the difficulty of ensuring the data's security throughout its entire lifecycle: from collection to transmission, storage, and processing. The authors state that HE has performance limitations due to the software library and the type of hardware employed, but that these limitations can be overcome with improvements in order to process massive data sets securely. To address these issues with HE, hardware-based methods are now used to speed up HE encryption and decryption processes. Investigations into hardware implementations including field-programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs), graphics processing units (GPUs), and clustering have yielded improved performance in the execution of operations (such as encryption and decryption). Graphics processing unit (GPU) enabled HE can improve encryption processing capacity by 7.68x and decryption processing capability by 7.4x compared to standard central processing unit (CPU) PCs.

The authors of [30] "suggest a privacy-enhanced federated learning (PEFL) scheme to protect the gradients over an untrusted server and which is a highly promising for big data analytics that can trains a global model across multiple mobile devices." By encrypting the local gradients with a Paillier homomorphic cryptosystem, the scheme safeguards the local model gradients from an untrusted server, which prevents sensitive data from leaking when we upload the vector gradient data to an untrusted cloud server using a federated learning/cooperative learning mechanism. Both the calculation cost and the accuracy of the proposed solution are low. In order to demonstrate that their technique is secure, the authors theoretically analyze and solve multiple cryptographic hard issues. Upon doing extensive experiments, it was found that PEFL (Privacy Enhance Federated Learning) achieves good accuracy in federated learning settings with cheap compute costs; however, the data were not presented numerically.

It is usual practice for enterprises to outsource their data to third-party service providers due to the exponential increase in data created from various sources. Hadoop, a leading Big Data processing platform, restricts access to Big Data solely through the Kerberos authentication mechanism. Nevertheless, privacy and secrecy are under jeopardy. To get around these issues, one solution is to use homomorphic encryption, which enables computation on encrypted data without decrypting it [31]. In addition to Secure Encryption (SE), there are Private Information Retrieval (PIR) and Multiparty Computations (MPC) systems that deal with encrypted data in different ways: searching, retrieving, and cooperative computing. Homomorphic encryption does allow computation on encrypted data, according to the amount of operations supported and homomorphism features. Dependent on the probability features of the encryption algorithms, HE can be classified as either deterministic or probabilistic. Since Probabilistic HE produces unique ciphertexts with the same plaintext and secret key, it outshines deterministic HE. The authors of [32] suggested combining Network Coding and Homomorphic encryption to make wireless devices more secure, faster, and more reliable while still protecting users' personal information. Thanks to the suggested scheme's end-to-end data privacy, public clouds can safely store sensitive data and conduct advanced operations without compromising user privacy. Both internal company data and data from other companies can be combined to form big data. "Intra big data processing" refers to the processing of such data within the same company. Inter large data processing refers to the handling of data from many organizations. It becomes more difficult to process big data when the data originates from diverse organizations. Users' safety and confidentiality is an issue. In order to address these kind of issues, deidentification is employed with the usage of k-anonymity, ldiversity, and t-closeness to improve user privacy. The authors present a privacy-preserving cosine similarity computing protocol that uses lightweight multi-party random masking and polynomial aggregation techniques to efficiently compute the cosine similarity of two vectors without disclosing the vectors to each other. To ensure the privacy of their outsourced data, organizations use a variety of cryptographic techniques. But in real "big data" situations with massive volumes of data, its performance is constrained by the costly computations. The authors present a new randomized encryption scheme called Splayed ASHE (SPLASHE) that prevents frequency attacks based on auxiliary data and a platform called Seabead that enables efficient analytics over large encrypted data sets. The platform relies on symmetric encryption schemes called additively symmetric homomorphic encryption scheme (ASHE), which efficiently performs large-scale aggregations.

Organizations' data security and privacy are at a critical juncture due to the massive amounts of data generated by Internet of Things devices and other Big Data sources. The authors suggested a cryptographic method called Secure Multiparty Computation (Secure MPC) to deal with these kinds of problems. Homomorphic encryption, secret sharing, and Yao's garbled circuits are some of the ways that safe MPC schemes are implemented. The many sources of data in a Big Data setting led to the wide variety of data types received. In [33], the authors provide privacy-preserving procedures for different types of data, including cryptographic approaches (like SMC) and non-cryptographic techniques (like perturbation), and attempt to define the data based on its structure. The authors of [34] provide a method for transmitting data across an unsecured network using homomorphic encryption for secure sum computation. By using secure sum, parties that are working together can add up their private data without disclosing any of that information to each other. By utilizing the additive homomorphic encryption technique, the authors achieve homomorphic encryption. With secure sum, users may add up their personal data without worrying that anybody else would see their secret information. With the use of symmetric key cryptography and homomorphic encryption, the authors present a method for safe sum calculation.

CONCLUSION

Big data analytics is the most important field since it has the ability to bring about a great deal of advantages and innovations, and it also has a bright future for both the academic world and the business world. When handled in the appropriate manner, it is a remarkable domain. Large amounts of data provide a number of challenges, the most significant of which is its scale, which necessitates appropriate storage, administration, integration, processing, and analysis. Despite the fact that big data analytics is helpful for making informed decisions, it will lead to severe security challenges that need to be addressed. Therefore, it became of utmost importance to ensure that the security of big data analytics was maintained. This article is comprised of a discussion of contemporary difficulties and challenges pertaining to the security implications of big data analytics. In the context of big data analytics, this study identified a number of difficulties that need to be addressed from the perspective of privacy and security. There have been two stages in which the results of this investigation have been described. In the first place, it addresses the current research emphasis and themes that have been documented in the selected top journals and conferences in the field through the use of well-established and acknowledged databases. In the second place, it presents an argument concerning the data that was analyzed and derived from the chosen body of literature. With the help of the study, we were able to investigate the existing research on Big Data Analytics Security problems by utilizing the well-known systems learning methodologies.

REFERENCES

1. Cappi, R., Casini, L., Tosi, D., Roccetti, M.: Questioning the seasonality of sars-cov-2: A

fourier spectral analysis. BMJ open 12(4) (2022)

2. Naghib, A., Jafari Navimipour, N., Hosseinzadeh, M., Sharifi, A.: A comprehensive and systematic literature review on the big data management techniques in the internet of things. Wireless Networks 29(3), 1085–1144 (2023) https://doi.org/10.1007/s11276-022-03177-5
3. Sarker, S., Arefin, M.S., Kowsher, M., Bhuiyan, T., Dhar, P.K., Kwon, O.-J.: A Comprehensive Review on Big Data for Industries: Challenges and Opportunities. IEEE Access 11, 744–769 (2023) https://doi.org/10.1109/ACCESS.2022.3232526
4. Bansal, M., Chana, I., Clarke, S.: A Survey on IoT Big Data: Current Status, 13 V's Challenges, and Future Directions. ACM Comput. Surv. 53(6) (2020) https://doi.org/10.1145/3419634
5. Zhong, Y., Chen, L., Dan, C., Rezaeipanah, A.: A systematic survey of data mining and big data analysis in internet of things. Journal of Supercomputing 78(17), 18405–18453 (2022) https://doi.org/10.1007/s11227-022-04594-1
6. Kushwaha, A.K., Kar, A.K., Dwivedi, Y.K.: Applications of big data in emerging management disciplines: A literature review using text mining. International Journal of Information Management Data Insights 1(2) (2021) https://doi.org/10.1016/j.jjimei.2021.100017
7. Amalina F, et al. Blending Big Data Analytics: Review on Challenges and a recent study. IEEE Access. 2020;8:3629–45. https://doi.org/10.1109/ACCESS.2019.2923270
8. Nazir S, et al. A comprehensive analysis of healthcare big data management, analytics and scientific programming. IEEE Access. 2020;8:95714–33. https://doi.org/10.1109/ACCESS.2020.2995572
9. Seh AH, et al. Healthcare Data Breaches: insights and implications. Healthcare. 2020;8(2):133. https://doi.org/10.3390/healthcare8020133
10. 12. Rathee G, Sharma A, Saini H, Kumar R, Iqbal R. A hybrid framework for multimedia data processing in IoT-healthcare using blockchain technology. Multimed Tools Appl. 2020;79:15–6. https://doi.org/10.1007/s11042-019-07835-3
11. Rabhi L, Falih N, Afraites A, Bouikhalene B. “Big Data Approach and its applications in Various Fields: Review,” Procedia Comput. Sci, vol. 155, no. 2018, pp. 599–605, 2019, https://doi.org/10.1016/j.procs.2019.08.084
12. Rahul K, Banyal RK, Goswami P. “Analysis and processing aspects of data in big data applications,” vol. 0529, no. May, 2020, https://doi.org/10.1080/09720529.2020.1721869
13. Shah D, Wang J, He QP. “Feature engineering in big data analytics for IoT-enabled smart manufacturing – comparison between deep learning and statistical learning,” vol. 141, 2020, https://doi.org/10.1016/j.compchemeng.2020.106970
14. Bonde M, Bossen C, Danholt P. Data-work and friction: investigating the practices of repurposing healthcare data. Health Inf J. 2019;25(3):558–66. https://doi.org/10.1177/1460458219856462
15. Bossen C, Pine KH, Cabitza F, Ellingsen G, Piras EM. Data work in healthcare: an introduction. Health Inf J. 2019;25(3):465– 74. https://doi.org/10.1177/1460458219864730
16. C. STAMFORD, “Gartner Forecasts Worldwide IT Spending to Exceed $4 Trillion in 2022,” Gartner. 2021, [Online]. Available: https://www.gartner.com/en/newsroom/ press-releases/2022-04-06-gartner-forecasts-worldwide-it-spending-to-reach-4-point-four-trillion-in-2022
17. Sun Z, Strang KD, Pambel F. Privacy and security in the big data paradigm. J Comput Inf Syst. 2020;60(2):146–55. https://doi.org/10.1080/08874417.2017.1418631
18. Sivaparthipan CB, Karthikeyan N, Karthik S. Designing statistical assessment healthcare information system for diabetics analysis using big data. Multimed Tools Appl. 2020;79:13–4. https://doi.org/10.1007/s11042-018-6648-3
19. “Telehealth _ Telemedicine Market. - Global Opportunity Analysis and Industry Forecast (2018–2023) _ Meticulous Market Research Pvt.”.
20. 142. Lv Z, Qiao L. Analysis of healthcare big data. Futur Gener Comput Syst. 2020;109:103–10. https://doi.org/10.1016/j.future.2020.03.039

21. Alkhalil, A., Abdallah, M.A.E., Alogali, A., Aljaloud, A.: Applying big data analytics in higher education: A systematic mapping study. International Journal of Information and Communication Technology Education 17(3), 29–51 (2021) https://doi.org/10.4018/IJICTE.20210701. oa3
22. Rahmani, A.M., Azhir, E., Ali, S., Mohammadi, M., Ahmed, O.H., Ghafour, M.Y., Ahmed, S.H., Hosseinzadeh, M.: Artificial intelligence approaches and mechanisms for big data analytics: a systematic study. PeerJ Computer Science 7, 1–28 (2021) https://doi.org/10.7717/peerj-cs.488
23. Lundberg, L.: Bibliometric mining of research directions and trends for big data. Journal of Big Data 10(1) (2023) https://doi.org/10.1186/s40537-023-00793-6
24. Ikegwu, A.C., Nweke, H.F., Anikwe, C.V., Alo, U.R., Okonkwo, O.R.: Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions. Cluster Computing 25(5), 3343–3387 (2022) https://doi.org/10.1007/ s10586-022-03568-5
25. Faroukhi AZ, Alaouib IE, Gahia Y, Aminea A. A Multi-Layer Big Data Value Chain Approach for Security Issues, in The 2nd International Workshop on Emerging Networks and Communications, Leuven, Belgiu; 2020.
26. Narayanan U, Paul V, Joseph S. A novel system architecture for secure authentication and data sharing in cloud enabled Big Data Environment, Journal of King Saud University – Computer and Information Sciences; 2020.
27. Asif M, Abbas S, Khan M, Ftima A, Khan MA, Lee SW. MapReduce Based Intelligent Model for Intrusion Detection Using Machine Learning Technique, Journal of King Saud University - Computer and Information Sciences; 2021.
28. Nambiara S, Kalambur S, Sitaram D. Modeling Access Control on Streaming Data in Apache Storm, in Procedia Computer Science, Bengaluru, India; 2020.
29. Tawalbeh LA, Saldamli G. Reconsidering big data security and privacy in cloud and mobile cloud systems, Journal of King Saud University – Computer and Information Sciences. 2021;33:810– 819.
30. Shoba V, Parameswari R. A Pragmatic Approach for Privacy Preserving Healthcare Using Stretched Homomorphic Re-Encryption Decryption Algorithm, International Journal of Advanced Science and Technology. 2020;20(7): 8850-8860.
31. Cunha M, Mendes R, Vilela JP. A survey of privacy-preserving mechanisms for heterogeneous data types, Computer Science Review. 2021;41.
32. Sidhu HJS, Khanna MS. Cloud’s Transformative Involvement in Managing BIG-DATA ANALYTICS For Securing Data in Transit, Storage And Use: A Study, in Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC); 2020.
33. Anwar MJ, Gill AQ, Hussain FK, Imran M. Secure big data ecosystem architecture challenges and solutions, Journal on wirlesscommninication and Networking; 2021.
34. Mothukuri V, Cheerla SS, Parizi RM, Zhang Q, Choo KKR. BlockHDFS: Blockchain-integrated Hadoop distributed file system for secure provenance traceability, Blockchain: Research and Applications. 2021;2.