Journal of Computer Science

Review

Mitigating Privacy and Security Risks in the Era of Big Data: A Comprehensive Framework of Best Practices and Protocols

[1]Uma Narayanan, [2]Nithin Puthiya Veettil, [3]Ratheesh Thottungal Krishnankutty, [3]Leya Elizabeth Sunny and [4]Varghese Paul

[1]Department of Computer Science, CVV Institute of Science and Technology, Kerala, India
[2]Tata Consultancies Services, Kerala, India
[3]Division of Information Technology, Cochin University of Science and Technology, Kerala, India
[4]Department of Computer Science, Rajagiri School of Engineering and Technology, Kerala, India

Article history
Received: 21-11-2023
Revised: 25-04-2024
Accepted: 29-04-2024

Corresponding Author:
Uma Narayanan
Department of Computer Science, CVV Institute of Science and Technology, Kerala, India
Email: uma.narayanan@cvv.ac.in

Abstract: Addressing the intricate challenges of security and privacy in big data necessitates a focused research approach toward the development of innovative frameworks, protocols and algorithms. This study advocates for a heightened focus on advancing encryption technologies and methodologies to bolster data anonymization, thereby fortifying the protection of user information. Additionally, emphasis is placed on enhancing data access control mechanisms to create more robust systems, ensuring only authorized users can access and utilize data, thus strengthening defenses against unauthorized breaches. Furthermore, there is a critical need to establish clear ethical guidelines governing the responsible utilization of data resources within societal and legal frameworks. Proactive research initiatives are recommended to devise novel methods for the early detection and prevention of malicious activities within the big data landscape, thereby preserving data integrity and privacy effectively. This survey provides a thorough analysis, consolidating various strategies to address security concerns in big data under a unified framework. It comprehensively covers applications, factors affecting security, challenges faced and potential fixes. By highlighting the importance of securing networks from external influences or attacks, this study contributes to the advancement of a more secure and resilient big data ecosystem.

Keywords: Big Data, Access Method, Security and Privacy, Cryptography, IoT

Introduction

Each day witness the creation of a substantial volume of data. This data deluge as depicted in Fig. 1 is continuously augmented by information originating from sensors, mobile devices, transactions, social media, enterprises and the general populace. The astounding upsurge in data can be attributed to the fact that we have generated more data in the past five years than in any previous period. Consequently, big data, characterized by the abundance of massive data sets, has emerged as one of the most fervently discussed research trends of our time.

Large datasets, which can be either structured or unstructured, can be created by compiling massive amounts of information. Its ability to store a limitless amount of structured and unstructured, social and non-social data has made it widely renowned in a few industries. It is an opportunity to improve business for large associations and corporate developments. Figure 2 illustrates the prediction for big data revenue in America and the upcoming significance of big data.

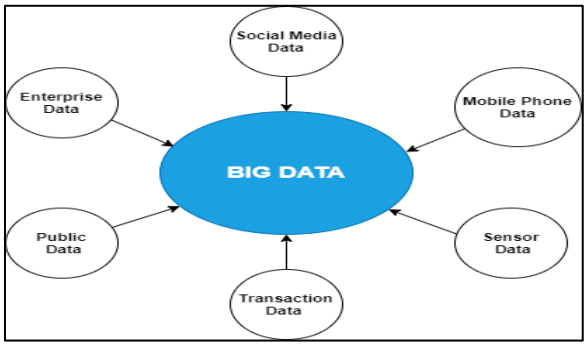

Fig. 1: The Big Data World

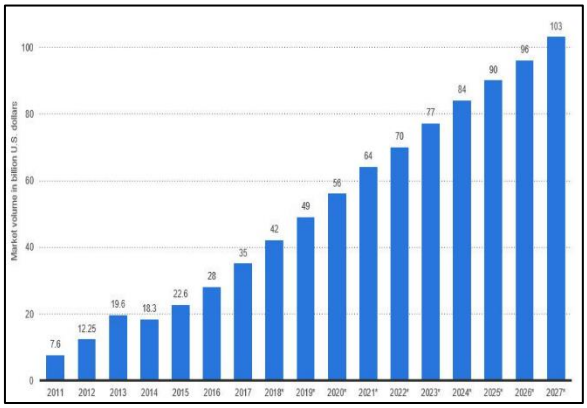

Fig. 2: Forecast of big data market size, based on revenue from 2011-2027 (in billion U.S. dollars) (Holst, 2020)

In light of connectivity and data transmission, data is provided in vast quantity and big data ought to be established for information mining computations. In order to fully take advantage of this opportunity, competent frameworks must be developed, keeping in mind the current challenges posed by its structure, scalability, security and analysis. Today's engineering of information processing frameworks has seen a shift from a centralized to a distributed architecture. Big data has quite recently taken the center period of the digital world. It's at the front line of everyone's mind. Big data is one of the biggest buzzwords around at the moment and it is taking the world by storm it is overpowering the world and will influence everyone's life. It will change our point of view towards everything.

Key Enablers for the Growth of "Big Data" are Increase of Storage Capacities

A long time ago, organizations used millions of huge boxes to store our information. Presently, exploiting cloud innovation, there is an inconceivable increment in processing power at a small amount of expense. You could now store the worldwide exchange information for the greatest relies upon the highly fundamental hardware equipment. Cost is not an obstacle for any organization. Influence the adaptable, easy cloud innovation that is accessible and that can be set up in only two or three days.

Increase of Processing Power

The processing power of the central processing unit can be calculated using the Floating Operation Per Second (FLOPS). The accompanying examinations are drawn between the most dominant PC processors from 1956-2015. Over that timeframe, the case is that there has been a one-trillion increment in flops of PC handling power. With this increment of processing capability, we can analyze the huge amount of data using single laptops. So if we use cloud computing, it can be quickly processed with less expense, which was unthinkable a few years back.

Availability of Data

According to Forbes magazine, the last two years alone have witnessed the creation of more data than in the entire history preceding them. To put this exponential data growth into perspective, let's examine some of the staggering data sources that contribute to this surge in just a single minute. For instance, within a mere 60 sec, more than 60 new blogs and a staggering 1500 blog posts come into existence, while over 70 new domain names are registered. Simultaneously, YouTube users upload an astonishing 600 new videos. Social media platforms further amplify this data explosion, with Facebook users collectively sharing over 100 terabytes of data each day. The sheer volume of data generation continues as users send 31 million messages and watch 2.7 million videos every minute. In addition, a staggering 694,445 searches are conducted on Google, 320 new accounts are created and more than 98,000 tweets are sent out every minute on Twitter. These astounding statistics underscore the crucial need for big data analytics to make sense of this immense data influx. The insights derived from this data can benefit businesses, academia and society as a whole.

This complicated topic by presenting a thorough overview of the most recent advancements in big data analytics and security. Our numerous contributions include a number of well-designed taxonomies for big data analytics and security that show fascinating connections between different variables and ideas. We also go over significant research-related lessons learned, emphasizing the parallel and non-converged security research operations that might lead to serious security flaws. In order to close the gap and increase overall security, we recommend implementing an enhanced dependability foundation. In addition, we offer insights into open research topics that can direct additional study in this field. In order to advance the state of the art in big data analytics and security, our survey offers a significant resource for academics and industry professionals.

Application of Big Data

There are several applications of big data technology (Agrawal *et al*., 2011; Chen *et al*., 2014; Manyika *et al*., 2011; Warren and Marz., 2015; Schmidt and Hildebrandt., 2017; Stergiou and Psannis., 2017a; Zikopoulos and Eaton, 2011). Some of them are presented here. Figure 3 shows the various applications of big data technology.

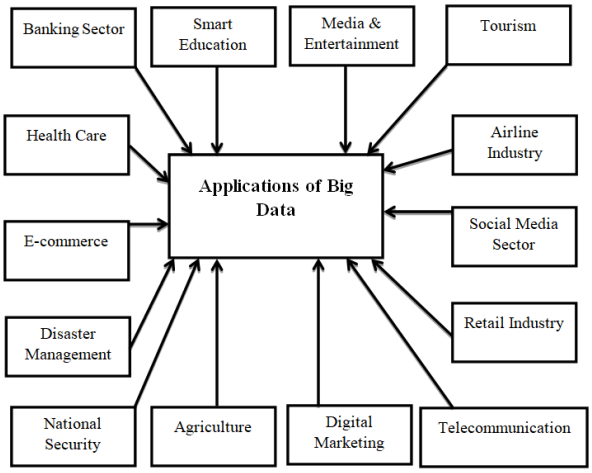

Fig. 3: Applications of big data

Banking Sector

Protection concerns are regularly brought up by the use of client information. A large-scale information inquiry may find sensitive personal information by revealing hidden connections between seemingly unrelated pieces of information. According to research, 62% of investors are cautious when using vast amounts of information due to protection concerns. Additionally, spreading client information throughout offices to provide more opulent experiences increases the risk of security. For instance, the income, investments, mortgages and protection strategies of the clients ended up in the wrong hands. Such incidents exacerbate information security concerns and discourage customers from disclosing personal information in exchange for tailored offers (Srivastava *et al*., 2017; Cerchiello and Giudici, 2016; Das., 2020). The use case is displayed in Fig. 4.

Smart Education

Education is the foundation of any country. In the event that education isn't rendered most appropriately for each learner, he/she won't perform well. Moreover, the obligation to guarantee quality education totally relies upon the legislature and educational institutions. Big data can create exceptional outcomes and present inventive information-driven methodologies for educational institutions as shown in Fig. 5. Due to the COVID-19 pandemic, the mode of education has shifted to Online. The pandemic has upset the education segment, a basic determinant of a nation's economic future. Clearly, the pandemic has transformed the incredibly antiquated, chalk-talk teaching approach into one that is innovation-driven. Policymakers are being forced by this disruption in the delivery of education to figure out how to increase commitment at scale while ensuring full e-learning arrangements and managing the digital environment. In numerous nations, the employment of big data in schools and universities is normal (Olanrewaju *et al*., 2016; West, 2012; Drigas and Leliopoulos, 2014; Sedkaoui and Khelfaoui, 2019; Elia *et al*., 2019; Maldonado-Mahauad *et al*., 2018; Cantabella *et al*., 2019).

Media and Entertainment

The entertainment sectors and media cooperate must drive electronic change to scatter their things and substance in the present market as quickly as possible. The accessibility of looking and getting to any content anyplace with any gadget turns into an across-the-board practice. We now have access to everything at our fingertips and big data has been the backbone of this amazing transformation (Suri and Singh, 2018; Arsenault, 2017; W. E. F., 2016).

Social Media

Social media life is one of the most mainstream advanced media divisions in the present world. The foundation of online life-promoting promotion lies in the use of big data. In spite of the fact that it isn't allowed to utilize all kinds of data in online networking, it is significant for legitimate upkeep and client satisfaction. It examines the inclination, conduct and peak timing of a client to stay relevant and competitive (Yadranjiaghdam *et al*., 2017; Tsou, 2015; Felt, 2016).

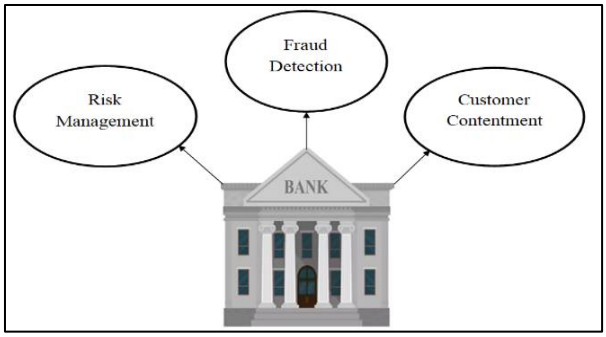

Fig. 4: Use of big data in the banking sector

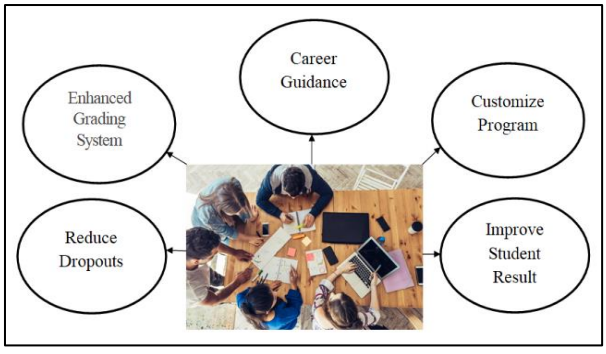

Fig. 5: Big data in the education field

Tourism

The tourism industry is primarily founded on the enthusiasm of a region toward the traveler group and how they present the most attractive package of visits to the traveler upon request. Present-day travelers are bound to utilize a computerized world as opposed to an office setup in a traditional way. Vast information accumulates information on tourists all around the globe about spots and people that can be enormously useful for the country and tourist company in the following ways as shown in Fig. 6 (Fuchs *et al*., 2014; Miah *et al*., 2017).

Airline Industry

The aircraft business makes the best use of big data as it gives them moment-to-minute operational information. It assists with the accumulated data about customer service, ticketing, climate estimates and so forth. A little carrier can likewise respond and make choices for consumer satisfaction and to fulfill the assistance of big data (Kasturi *et al*., 2016; Oh, 2017).

Retail Industry

The retail business is driving from the front in a nation's economy. Big data gives an open door for this division by the investigation of the aggressive commercial center and customer satisfaction as shown in Fig. 7. It decides customer commitment and consumer loyalty by gathering diverse information. It can improve the presentation and proficiency through the basic investigation of the data gathered by big data (Chen *et al*., 2012; Waller and Awcett, 2013; Erevelles *et al*., 2016).

Telecommunication

Telecommunication is one of the most famous client gatherings of big data applications. With the expanding measure of information going through various correspondence channels, it is imperative to gather this data to boost the benefits and compelling techniques for organizations. It visualizes information that improves the board and consumer satisfaction (He *et al*., 2016, Van, 2013; Ahmad *et al*., 2019; Khan *et al*., 2019; Dam, 2013) is as shown in Fig. 8.

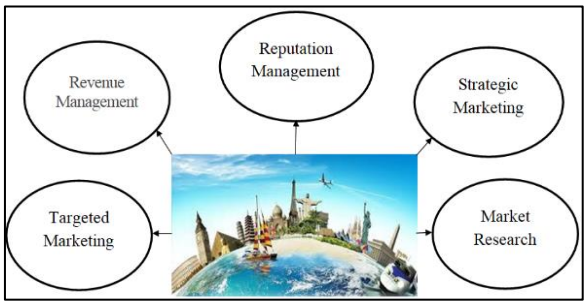

Fig. 6: Big data in tourism

Fig. 7: Big data in retail

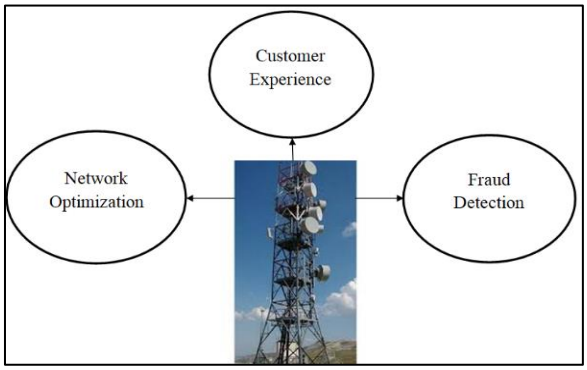

Fig. 8: Big data in telecommunication

Digital Marketing

Marketing trends for the business have totally changed. Computerized promotion is the way to make any business effective. Presently, not just the enormous organizations can run showcasing limited time exercises but additionally, the little business people can run effective publicizing efforts via web-based networking media stages and promote their items. Big data has made computerized advertising extremely amazing and it has become a fundamental piece of any business (Leeflang *et al*., 2014; Brown and Harmon, 2014; Das, 2021).

Agriculture

In agriculture, Big data is doing a compelling job to improve the productivity of the farmers. The objective is to limit the farmers' misfortune and increase the age of important nourishment grains for the residents of the countries. Data science has helped a great deal to acquaint advanced and modern strategies with the current farming customs. Employments of big data make us ready to meet the necessary measure of yearly production and expel the need to import products (Bendre *et al*., 2015; Gupta *et al*., 2020; Crampton *et al*., 2015). With the help of big data, we can improve the production of crops with accurate crop prediction mechanisms as shown in Fig. 9.

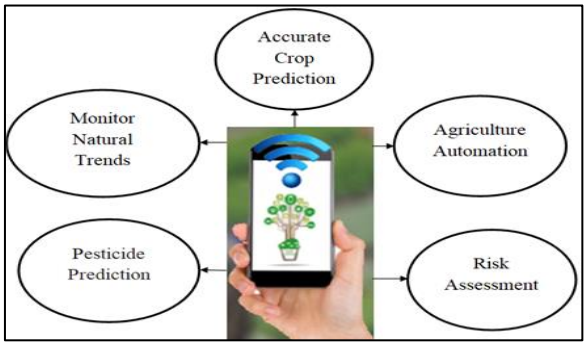

Fig. 9: Big data in agriculture

National Security

Innovation has molded our lives and improved with its colossal conceivable outcomes. Big data is liable for the accomplishment of these items. In many police powers, big data is utilized to improve their work process and activities all around the globe. Developed nations like the USA and UK have actualized big data in their social and security exercises sometime in the past, but some developing nations have likewise begun accepting the advantages of utilizing big data now (Crampton *et al*., 2015; Klein *et al*., 2016; Brewster *et al*., 2015).

Disaster Management

Consistently disasters like typhoons, floods and seismic tremors cause immense harm and claim numerous lives. Researchers are not able to foresee the possibility of catastrophe and play it safe by the legislatures. It is the critical drawback of the huge impact. In spite of the fact that the employment of big data in calamity is not new, the ongoing improvement of AI, information mining and representation are helping meteorologists to gauge climate conditions all the more precisely (Puthal *et al*., 2016; Choi and Bae, 2015).

E-Commerce

E-commerce makes a high-performing advertising model that sets a start-up apart from the current one and becomes fruitful. Web-based business proprietors can distinguish the most seen items and the pages that showed up the maximum number of times. It assesses the client's conduct and proposes comparable items. It builds a high quantity of deals for the customer and produces income. When an item is added to the cart but is not yet conclusively purchased by a client, Big data analysis can naturally send limited-time special offers to that specific client. Big data applications can create and sort reports according to the customer's age, location, gender and so on (Braik *et al*., 2016; Silahtaroglu and Donertasli, 2015).

Health Care

A couple of years back, the job of big data in the medical field was not mentionable. Be that as it may, information science is overwhelming to improve social insurance these days. Big data acquainted with recognizing treatment as well as improved the way toward rendering human services as shown in Fig. 10. Big data greatly affects decreasing misuse of cash and time. Nearby, legislatures are utilizing big data to grow new foundations and emergency services (Sun and Reddy, 2013; Belle *et al*., 2015; Lv and Qiao, 2020; Pramanik *et al*., 2020).

Challenges of Big Data

Figure 11 shows different big data challenges.

Scalable and Interoperable Computing Infrastructure

A vast and diverse array of data is gathered from all corners, encompassing both dynamic streaming and non-streaming data, along with structured and unstructured data. This continuous flow of data, ranging from machine-to-machine interactions to machine-to-human communications, presents a formidable challenge in terms of data storage, sharing and processing. To effectively address these challenges, there is a pressing need for a scalable and interoperable computing infrastructure.

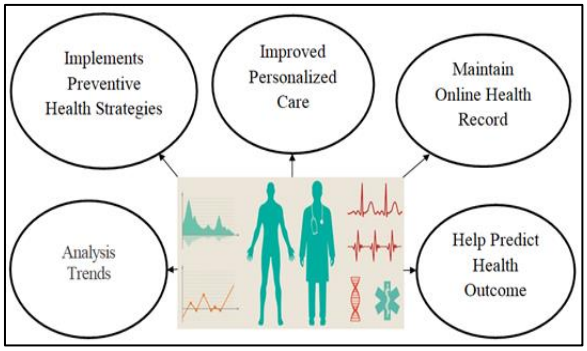

Fig. 10: Big data in the healthcare sector

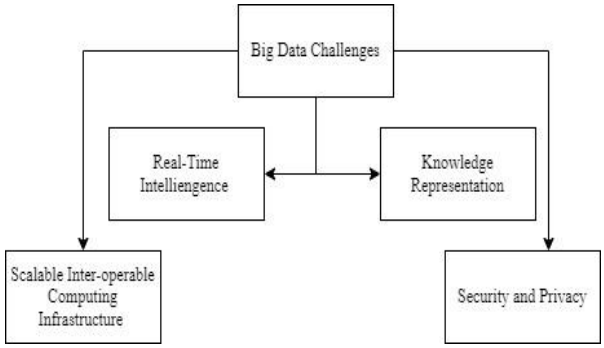

Fig. 11: Big data challenge taxonomy

Real-Time Intelligence

Intelligent decision-making necessitates the analysis of both current and historical data. However, due to the substantial volume and diverse nature of this data, processing it is already a complex task. When coupled with the requirement for real-time processing, the design of new algorithms capable of delivering real-time insights from such extensive datasets becomes an even more daunting challenge.

Knowledge Representation

Artificial intelligence and novel machine learning theory are needed for big data analytics. The lack of intuitive physical interpretation in machine learning and artificial intelligence processes and outputs is well established (Wagstaff, 2012). It is crucial to close this knowledge gap by offering appropriate knowledge interpretation in order to use artificial intelligence and machine learning to make wise decisions.

Security and Privacy

Data often contains sensitive information, whether it pertains to an organization's confidential data or an individual user's private information. Crucially, this data can influence decisions that impact the secure operation of critical infrastructure. As a result, security and privacy are paramount concerns.

Even though they are crucial, traditional security techniques are inadequate for big data systems. Big data security presents new, distinctive problems that involve data and applications. For instance, the main focus of big data security platforms currently in use is fine-grained security achieved by in-depth data analysis. However, such models inadvertently open the door to potential misuse of user data by applications and service providers. This concern has given rise to the concept of differential privacy, which strives to protect sensitive user information while still supporting valuable data analytics.

Big Data Security and Privacy

Big data systems need traditional security measures, but they are insufficient in this case. Big data security presents certain particular difficulties in terms of both applications and data. As an illustration, modern big data security solutions put a lot of effort into offering granular security through in-depth analysis of stored data. However, these models inadvertently enable the misuse of user information by services and application developers. Figure 12 displays the main security issues that need to be resolved.

Infrastructure security: Data is kept on equipment owned by others. User data loss could result from intrusions, assaults, memory corruption, physical damage, etc.

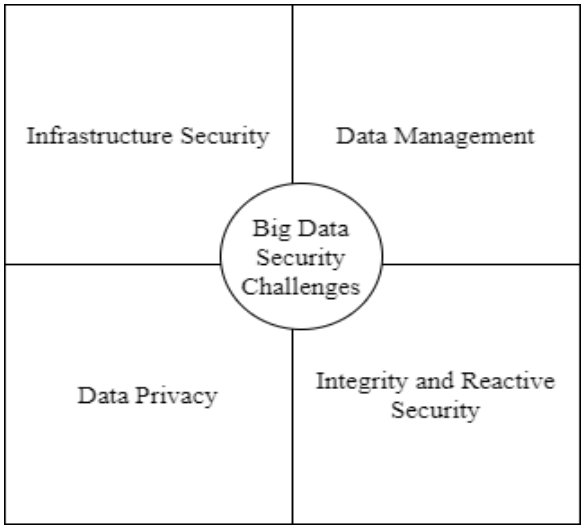

Fig. 12: Security challenges in the big data

Data management: To facilitate quick recovery, high availability and fault tolerance, data should be stored and copied on many machines.

Data privacy: Even though users access data from machines owned by other parties, service providers shouldn't have access to the queries, files, or data that users issue or access.

Integrity and reactive security: Data owners should be able to check that the uploaded data is still present on the server and that no unauthorized parties have accessed or modified it. Real-time monitoring of attacks on programs and devices used to collect data is necessary.

Security and privacy is a big concern for big data. Numerous organizations utilize some sort of big data system, yet many don't get the security essentials right. Just like the case with numerous new innovations, security is regularly disregarded or is, best-case scenario, a reconsideration. The outcomes of not completely considering a big data security setting can anyway be incredibly harmful. We discuss the estimation of information and information has a place with the individuals who sort out this information and work with this information. So if this information gets lost, at that point, you may have lost worth or notoriety only one company, possibly the biggest retailer, target, is aware of the terrible reputational loss brought on by missing information. Target shops all around the country were impacted by a significant hack in late 2013- just before the start of the holiday shopping season. As well as information from about 40 million visa applications, individual data from about 70 million consumers were collected. Target has paid out compensation to the impacted parties totaling just under $300 million as a result of the breach. It resulted in an immediate decline in

sales, an incalculable loss of client confidence and reputational harm.

A Breach at Terracom and its subsidiary Yourtel America accidentally released 310,000 records of potential organization clients. The data was seen through Google for the vast majority of the year 2013. In 2014, a penetration at AT&T uncovered 280,000 client records. Organizations presently face fines particularly if information penetrates emerged from organization carelessness. The Federal Communications Commission or FCC found that Terracom, Yourtel had put away the client data on an unsecured server. The organizations were fined $10 million, which was brought down to 3.5 million. AT&T employed a few considered focuses with a side business of selling client data. For their absence of oversight, they were fined $25 million. Yahoo itself didn't recognize the spillage of a huge number of its records. In 2018, its proprietor, Albata was fined $35 million. In 2018, an IBM report assessed that information breaks cost each organization a normal $148 per private record that they spilled. It tends to be profoundly troubling to discover that data relating to your well-being has been spilled. This kind of data is inherently private and secret. A large portion of us would not need it shared past our medical care experts and maybe a couple of friends and family. There are wide-going situations where private or classified clinical data may advance into the public area or to outsiders without the individual concerned. This can regularly occur because of carelessness or malicious attacks, for example, when an individual or expert relationship breaks down and one individual reveals private data, inspired by outrage or looking for vengeance. A case of this is the situation of Cooper vs Turrell EWHC 3269, which concerned the deliberate posting on the web of data identifying the soundness of the petitioner by a previous representative.

Every year, cybercriminals target medical clinics and healthcare centers, extracting a significant number of clinical records, many of which end up being sold on the darknet for substantial sums of money. HHS' Office of Inspector General investigated almost 400 cases of clinical data breaches. In 2018, Protenus, a cybersecurity firm, identified around 222 instances of medical record data breaches, marking a 25% increase since 2017. Hackers are drawn to clinical records because they contain a patient's complete personal information, including their name, address history, financial data and social security numbers. This trove of information is enough for hackers to apply for credit lines or set up fraudulent accounts in the patients' names, potentially leading to identity theft and financial fraud. Hospitals and healthcare organizations are often perceived as easy targets due to their relatively low level of security, making it straightforward for cybercriminals to access vast amounts of personal data for nefarious purposes. Increasingly, hackers are not only stealing this data but also selling it on the black market. Buyers of such data may use it to create fake IDs for purchasing medical equipment or drugs or to file false insurance claims. The stolen records can carry significant price tags, with a patient's full medical history potentially fetching up to $1,000. In comparison, social security numbers and credit card data typically sell for much lower prices. For instance, one hacker known as "the dark overlord" attempted to sell 655,000 clinical records from three healthcare organizations for nearly $700,000 on the darknet. However, when the case gained notoriety, the hacker attempted to return the unsold records to the healthcare organizations. The repercussions for patients whose identities are stolen can be catastrophic. They often face a tumultuous process of rectifying the damage caused by the misuse of their personal information, including financial losses and medical procedure charges. Even after the perpetrator is caught, victims may continue to grapple with the consequences, such as damage to the integrity of their medical records. These cases primarily occur outside India.

As medical data gets digitized in the Indian setting and with a large number of advanced medical records being created each day, medical services suppliers likewise need to take a gander at network protection truly. In the most recent information leak identified with the user in India, over a million clinical records and 121 million clinical pictures of Indian patients, including X-rays and scans, have been released online to be open by anybody. As indicated by German cybersecurity organization Greenbone networks, the patient records and scans from India additionally incorporate details, for example, the name of the patient, their date of birth, the public ID, name of the clinical foundation, their clinical history, doctor names and other details that are meant to be classified. Among the spilled information are clinical records from Mumbai's top breach candy hospital and Utkarsh Scans, a generally notable clinical imaging supplier. Upon audit, Inc42 found that the connection where the information has transferred likewise permits anybody to download the patient's clinical details. From this, we identified that securing Healthcare records is a hot topic of research. Since we are using third-party storage systems like the cloud, security and privacy are an important concern. From the above case, we can understand the importance of the security mechanism needed to secure data storage. Big data-enabled cloud computing is a solution for storing a large volume of data at a low cost. Cloud computing is a method of utilizing I.T. that has these five similarly significant attributes. To begin with, you get figuring assets on request and self-service. You should simply utilize a basic interface and you get the processing power, storage and networking you need, with no requirement for

human intervention. Second, you can access these assets over the net from any place you need. Third, the supplier of those assets has a big pool of them and allocates them to clients out of that pool. That permits the supplier to get economies of scale by purchasing in mass and giving the savings to the clients. Users don't need to know or care about the specific physical area of those assets. Fourth, the assets are flexible. If we need more resources, we can get more quickly. In the event that you need less, you can downsize. Furthermore, last, the clients pay just for what they use or reserve as they go. In the event that they quit utilizing assets, they quit paying. Let's now focus on security architecture on the cloud.

The big data sector makes considerable use of cloud computing, a relatively new paradigm in digital technology (Griebel *et al*., 2015). It enables the simple exchange or transfer of medical data among numerous parties and offers efficient information storage. Cloud services offer significant advantages in terms of the efficient and effective processing, updating and cost-effective storing of information. The fact that the data is stored on a large network of remote servers that are connected to one another and run as a single ecosystem that can be accessed by several people from various locations puts both privacy and security at risk Furthermore, storing data on servers owned by third parties unintentionally increases the risks involved because the majority of information is very sensitive and personal. Given the fragility of information in the public domain, there is an urgent need to establish a more secure, efficient and effective platform for data access and exchange across stakeholders. Figure 13 shows the security and privacy of big data taxonomy.

Security within the realm of big data has emerged as a subject of significant interest for researchers. This heightened attention is primarily attributed to the fact that big data storage systems present lucrative targets for potential intruders (Santos and Masala, 2019; Jeong and Shin, 2016). Big data, defined by its three characteristics High volume, high velocity and high variety of information, gives rise to complex challenges in storage and introduces vulnerabilities that are challenging to address in real-world scenarios (Stergiou and Psannis, 2017b).

Furthermore, it's essential to recognize that cloud servers and other big data servers cannot be entirely trusted. For instance, in critical applications like healthcare, where secure large-scale storage is imperative due to the need to safeguard extensive genome data, the size of which can reach up to 140 gigabytes (Stergiou and Psannis, 2017b). This inherent need for security often compels data owners to entrust their private and sensitive data to cloud or big data servers, despite the associated security concerns (Wei *et al*., 2019).

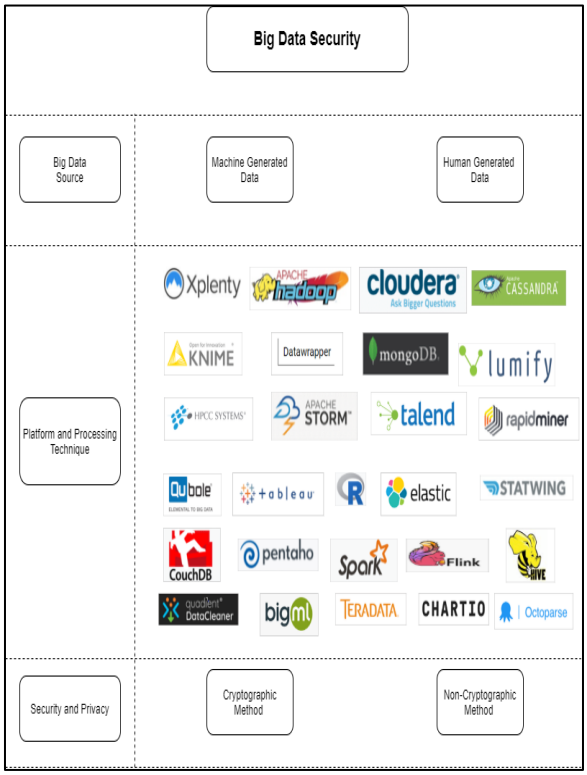

Fig. 13: Taxonomy of security and privacy-preserving big data

Within the domain of big data storage systems, various security threats loom, including password guessing attacks, brute force attacks, stolen verifier attacks and more. Prior security approaches have sought to protect data by transmitting it in encrypted form. However, these methods have often fallen short of providing adequate confidentiality and privacy for both data owners and users (Li *et al*., 2017). Consequently, the realm of big data has introduced critical concerns regarding data security.

From a security standpoint, it is essential because of:

- When ciphertext is modified, access policies are not stated and in this case, user legitimacy is compromised, which implies who wants to access the data is still a major worry in big data
- There is no authorized organization to oversee the outsourcing of data storage and sharing

Real-time security (authentication, data confidentiality and integrity) monitoring is crucial for massive data storage systems.

Literature Survey

Companies and organizations in the digital age struggle to manage their complex data effectively. Outsourcing the data to a cloud is a wise move given the development of cloud storage. The two major issues with

big data are how to securely store the data and how to offer access control over the stored data (Jadon and Mishra 2019). This is because large data frequently contains a significant amount of personally identifiable information. We mostly summarise the current situation of protecting massive data stored in clouds in this section.

This study delves into the application of various techniques in the context of big data within a cloud environment, taking into account both cryptographic and non-cryptographic approaches. Furthermore, it explores methods for upholding data security, privacy and anonymity in the cloud. To effectively query encrypted data stored on third-party cloud servers, specialized Searchable Encryption (SE) approaches are elucidated. Traditional search methods are ineffective due to data encryption. As a solution, Searchable Symmetric Encryption (SSE) is introduced. Unlike other surveys, our research comprehensively covers all aspects and techniques related to cloud privacy and security.

Cryptographic and non-cryptographic procedures are the two main categories that are taken into consideration. Public key encryption, symmetric key encryption and other cryptographic primitives are some of the encryption techniques used in cryptographic schemes. Access control technologies like Role-Based Access Control (RBAC), Attribute-Based Access Control (ABAC) and others are included in non-cryptographic systems. Figure 14 shows how privacy-preserving techniques are categorized. The most recent developments in big data security in a cloud context are covered in this section. This section is divided into two classes: Data security and retrieval in large data clouds using cryptographic methods and user privacy (authentication) in big data clouds using non-cryptographic methods.

Cryptographic Method in Secure Cloud Big Data

The access control method can be basically classified as shown in Fig. 15.

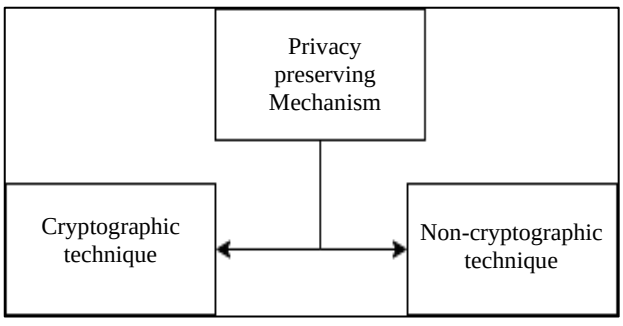

Fig. 14: Classification of cryptography

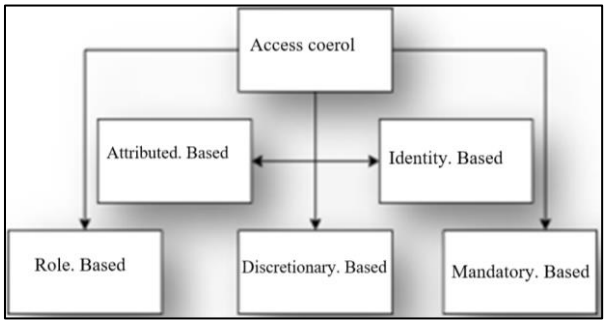

Fig. 15: Access control method

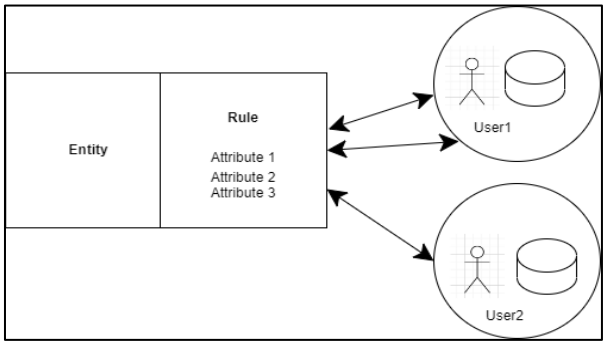

Fig. 16: Attribute-based access control structure

Attribute-Based AC

Attributed based access control method is shown in Fig. 16. Dutta *et al.* (2020). Employing Semantic Web technologies, an attribute-based access control model executes access control decisions by performing dynamic reasoning over these attributes and context-driven rules. The above framework represents the physical context that is derived from context-driven regulations and sensed data (attributes). The system makes decisions about access control by taking into account the device type, the details of the information collected by the cloud service provider and the context of the user. The system's access control decisions are supplemented by another sub-system that uses behavioral and network data to identify breaches into smart home systems. The integrated technique helps identify signs that a smart home system is being attacked and restricts the amount of data leaks that such attacks can cause.

A semantically rich access control system is proposed by Joshi *et al.* (2017). that makes use of an access broker module to assess decisions based on rules created using the organization's confidentiality policy. Before deciding whether to grant access, the suggested system evaluates the multi-valued properties of the user making the request and the requested document, which is kept on a cloud service platform. Additionally, our system uses oblivious storage techniques to ensure an

end-to-end oblivious data transaction between the organization and the cloud service provider. As a result, a company can utilize our technology to encrypt its documents and hide its access patterns from unreliable cloud service providers.

In this research, researchers create a smart farming ontology to protect the ecosystem that Chukkapalli *et al*. (2020) have created for smart farms. The ontology represents a variety of physical entities, including sensors, farm laborers and their interactions. We implement an Attribute-based Based Access Control (ABAC) system to dynamically assess access control requests using the expressive ontology.

A safe authentication technique using a tree-based signature in a hierarchical attribute authorization structure has been presented by Shen *et al*. (2017). It is used in multi-level structures for user authentication. It defends against replay attacks and forgery attacks while still preserving privacy. Greater temporal complexity and storage problems result from hierarchical attribute authorized structures.

The decentralized access control technique was introduced by Ruj *et al*. (2014) and it can offer users anonymous authentication while thwarting replay assaults. And supporting data creation, data modification, reading data stored and user revocation.

Colombo and Ferrari (2017a), have introduced an Attribute-Based Access Control (ABAC) framework for NoSQL data stores, utilizing SQL++. This framework provides robust access control mechanisms to enhance privacy protection, but it may pose challenges for administrators in terms of setting up access control policies. This complexity is particularly evident when dealing with NoSQL systems, which are schema-less and characterized by heterogeneous data structures, requiring fine-grained access control policies.

Gupta *et al*. (2017), have introduced extensions to the existing authorization capabilities within the Hadoop core and its related ecosystem projects, such as Apache Ranger and Apache Sentry. They introduce a fine-grained attribute-based access control model known as HeABAC, specifically tailored to meet the security and privacy requirements of multi-tenant environments operating within the Hadoop ecosystem.

In a separate work, Longstaff and Noble (2016) have developed an efficient implementation of Attribute-Based Access Control (ABAC) for large-scale applications that utilize a variety of data storage technologies, including Hadoop, NoSQL and relational database systems. There are two important phases in the ABAC authorization procedure. First, a set of permissions is generated to specify which data a user can access during a transaction. Then, query modification is used to add code that enforces ABAC controls to the user's transaction.

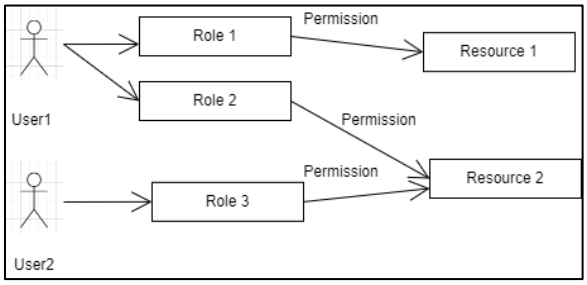

Fig. 17: Role-based access control structure

Role Based AC

Role-based access control method is shown in Fig. 17.

Ulusoy *et al*. (2015), have introduced the GuardMR framework, designed to implement fine-grained Role-Based Access Control (RBAC) within the popular big data platform, Hadoop, running on top of MapReduce. By separating and perhaps altering the key-value pairs that are extracted from a target data asset using a MapReduce task and supplied as input to the Map function, GuardMR secures data.

Colombo and Ferrari (2017b) have integrated the RBAC model into the MongoDB platform, enhancing it with support for specifying and enhancing purpose-based policies. These policies were initially used in relational database systems to regulate access at the document level.

Gupta *et al*. (2017), have introduced object tagged RBAC, an RBAC model that combines RBAC role-based authorization tasks with support for object attributes. They have implemented a prototype of this model by incorporating role support into Apache Ranger.

Nabeel and Bertino (2014), have implemented an access control technique within Cassandra, employing an efficient RBAC authorization design that operates within Cassandra's distributed architecture. This approach serves as an example of a platform-specific solution built on platform-specific features.

Identity Based Access Control

In the identity-based access control domain, (Xiong *et al*., 2022), have presented a mechanism to safeguard cloud data confidentiality by introducing cloud-based fine-grained access control systems. Their scheme allows verifiable outsourcing, enabling computationally expensive operations at the receiver to be offloaded to an untrusted cloud server. The scheme features efficient revocation functionality (R-IBSC) based on a binary tree structure. Belchior *et al*. (2020), have put forth an access control system rooted in the self-sovereign identity paradigm. Verifiable Credentials (VCs) represent qualities, while players' identities are represented by their Decentralized Identifiers (DIDs) in this system. With the help of blockchain technology and conventional access control mechanisms, the Self-Sovereign Identity Access

Control (SSIBAC) paradigm enables cross-organization identity management through decentralized authentication and centralized authorization. Notably, sensitive user data is not stored as part of this access control procedure. Gupta *et al.* (2018), have introduced a framework that incorporates user smart cards to ensure secure access to Cloud-based services and data in a distributed IoT environment. They employ an identity-based access control mechanism to ensure secure access for authenticated users. Unlike many conventional cloud access models that require separate account management for various services from the same provider, this study focuses on resource-constrained IoT devices in a distributed cloud computing IoT environment. They propose a novel framework that enables users to access different cloud services using the same password credentials and a smart card.

Additionally, a novel lightweight identity authentication-based access control scheme for the cloud has been introduced (Shen *et al*., 2016). This scheme utilizes an authorized agency to assist in authentication and key distribution. By employing XOR and hash functions to obscure parameters and verify identities, this scheme boasts a low computational cost. Furthermore, it shifts the primary computational load to the authorized agency, rendering it more efficient than other access control schemes.

With the rapid advancements in IoT, numerous IT-based services and applications are being developed for user convenience, and cloud computing adoption will significantly shape future Internet applications. The paper (Gupta and Quamara, 2018) proposes a framework using smart cards for secure, identity-based access to cloud services and data in a distributed IoT environment, including informal security and functional analysis.

Mandatory Access Control (MAC)

In a paper (Hu *et al*., 2011a), the authors present a comprehensive method for property verification in Mandatory Access Control (MAC) models. By creating a uniform framework for MAC models, this method makes property verification easier and makes it possible to generate test cases automatically. The process entails describing generic access control properties using a property language and representing MAC models in the specification language of a model checker. The integrity, coverage and confinement of these features within the MAC models are then evaluated using the model checker. Ultimately, a combinatorial covering array is used to produce test cases for the system implementations that are based on these models.

Zhang *et al*. (2005). introduce a model-checking algorithm designed to evaluate whether a MAC policy can fulfill a user's access request while preventing unauthorized access to malicious objectives. Unlike a generic model language, this method mandates that MAC system policies and agent goals be described using the Access Control (AC) description and specification language introduced as RW in their previous work. However, this language has limitations in specifying dynamic or historical aspects of MAC models and lacks support for general access constraint descriptions.

Mandatory Access Control (MAC) is a policy model that dictates access rights, with ordinary users unable to modify these access rules, relying on centralized administration (Bell and LaPadula,1976; Xiaolei Qian and Lunt, 1996). It enforces security labels to regulate access to sensitive data, categorizing data as top-secret, secret, or confidential and granting access based on the user's trust in the Certificate Authority (CA). The bell-Lapadula model is used to control information flow and ensure that lower-security data is not disclosed to higher-security entities. However, MAC is a rigid security management model controlled by a central entity and lacks adaptability, making it unsuitable for resource-constrained IoT environments.

Mandatory Access Control (MAC) is a control mechanism where the system specifies subject categories, security levels and their relationships with objects (Bell and LaPadula, 1996). There are two primary implementation models of MAC. The first is the Bell-Lapadula (BLP) model, where subjects with low-security levels have write permissions for objects with high-security levels, while subjects with high-security levels have read permissions for objects with low-security levels. The other model is the Biba model, in which subjects with high-security levels have write permissions for objects with low-security levels and subjects with low-security levels have read permissions for objects with high-security levels.

Discretionary Access Control

The discretionary-based access control method is shown in Fig. 18.

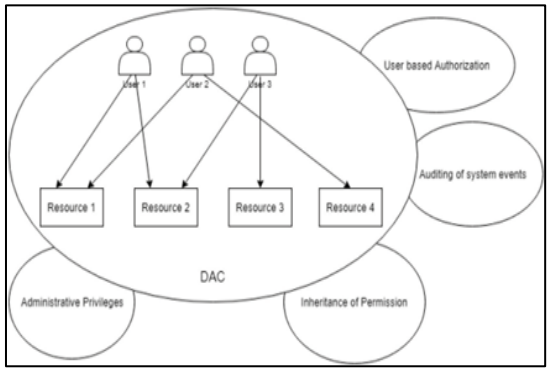

Fig. 18: Discretionary access control structure (Downs *et al*., 1985)

Graham and Dennig created Discretionary Access Control (DAC), an access control that serves as the foundation for security systems. Figure 18 depicts the DAC structure. Centralized access management which is one of the types that is often used for access control is Discretionary Access Control (DAC) (Bell and LaPadula, 1996). Based on their identities or user groups, DAC provides authorized users access to items. Users have the autonomy to delegate their rights or power to any other user. The link between users and objects was the foundation upon which DAC was created. The subject's access privileges have an impact on the access control decisions. The assigned access privileges for each object are shown in the access matrices. DAC is a simple and flexible AC, therefore which is made use in real life in IoT deployments, such as where IoT resources are identified by its Media Access Control (MAC) address.

Users control access to resources in the Discretionary Access Control (DAC) models Graham and Denning (1971) and they can grant permissions to their resources by including them in Access Control Lists (ACL). Users (or groups of subjects) are given authorization to access resources by each entry in the access control list (Jayant *et al*., 2014) Objects typically store the permissions. Via DAC, users decide the access rights to the resources they belong to, as opposed to MAC, where permissions are supplied by the administrator via established policies. UNIX, FreeBSD and Windows-based operating systems currently use DAC. Different approaches are shown in Figs. 19-21.

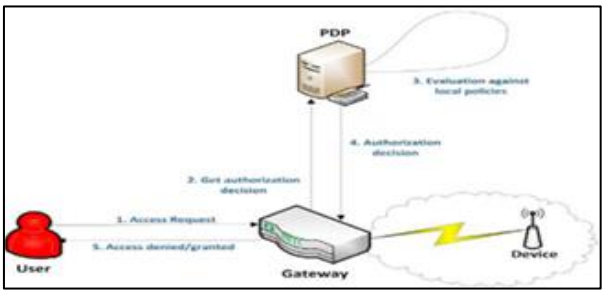

Fig. 19: Central approach

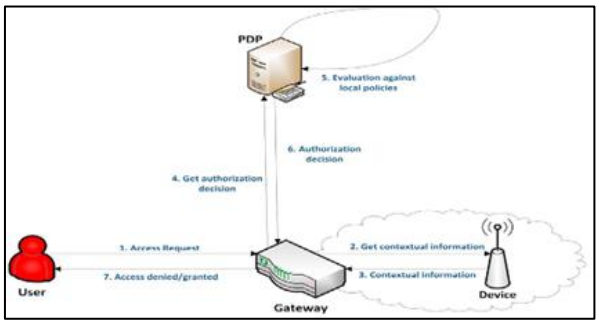

Fig. 20: Hybrid approach

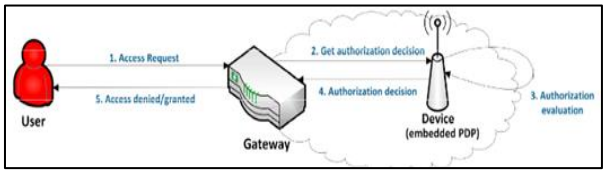

Fig. 21: Distributed approach

Access control is a widely utilized technical method for upholding the confidentiality and integrity of data across various aspects of systems and information security. Access control primarily deals with regulating resource access rights between subjects and objects. However, it's noteworthy that in the domain of virtualization, security challenges often arise due to unauthorized resource access (Lampson, 1974).

Cryptographic Method in Secure Cloud Big Data

Cryptography, which means hidden writing, involves the design of protocols to safeguard secret messages from being accessed by unauthorized third parties. Cryptographic methods encompass both symmetric key cryptography and asymmetric key cryptography (refer to Figs. 22-23). In symmetric key cryptography, the same key is used for both encryption and decryption, whereas in asymmetric key cryptography, distinct keys are employed.

(Prasetyo *et al*., 2014) Proposed the application of a symmetric encryption algorithm for IoT. They implemented the Blowfish encryption algorithm using VHDL on FPGA resources. Their evaluation considered performance metrics such as security, encryption time, avalanche effect and throughput, yielding positive outcomes. It's worth noting that this assessment focused exclusively on text inputs and did not take multimedia inputs into account.

Kazim *et al.* (2018) introduced an innovative framework for ensuring secure and dynamic access to IoT services within a multi-cloud environment. This protocol, developed on a cloud platform, facilitates collaboration in the IoT multi-cloud ecosystem. The framework encompasses several stages, including service matchmaking, authentication and SLA management. The SLA management component ensures that services are executed in an external cloud in accordance with agreed Service Level Agreements (SLAs) and monitors provider compliance.

Mollah *et al.* (2017) proposed a secure data storage scheme for IoT systems integrated with the cloud. This approach combines secret key encryption and public key encryption and offloads security operations to nearby servers, reducing processing overhead. The scheme also incorporates a secure search feature that enables authorized users to retrieve encrypted and shared data from the cloud, ensuring data integrity throughout the process.

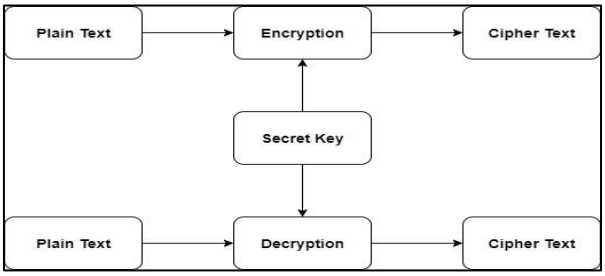

Fig. 22: Symmetric key cryptography process

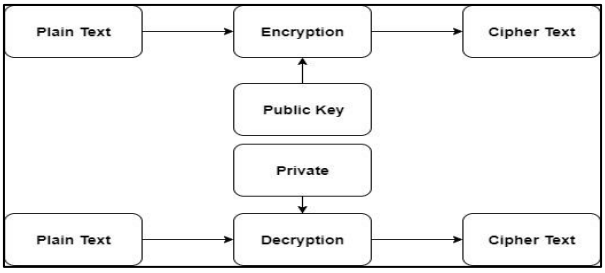

Fig. 23: Asymmetric key cryptography process

In "secure big data storage and sharing scheme for cloud tenants" (Cheng *et al*., 2015), presents an alternative approach that divides big data into sequential parts stored across multiple cloud providers. This scheme primarily focuses on safeguarding the mapping of data rather than the data itself, thereby reducing the high cost associated with encrypting large datasets.

Matturdi *et al*. (2014) reviewed the security and privacy aspects of big data, highlighting their paramount importance. They emphasize the utilization of big data for implementing solutions that enhance the security, reliability and safety of distributed systems.

A novel technology known as the integrated Rule-Oriented Data System (iRODS) is proposed as a solution for ensuring security and privacy in big data (Jensen., 2013). It recognizes the importance of technology in addition to legal regulations in addressing security concerns, owing to the rapid pace of technological advancements and regional variations in regulations.

Vorugunti (2016) introduced the PPMUS framework, specifically designed for privacy-preserving mobile user authentication, making use of big data characteristics like storage capacity and robust management. This framework employs fuzzy hashing and Fully Homomorphic Encryption (FHE) algorithms to maintain user privacy, although it may have vulnerabilities in scenarios involving user password typing.

Zhao *et al*. (2018) presented a password-based secure authentication scheme for users across multiple servers, utilizing Elliptic Curve Cryptography (ECC) for user authentication. ECC offers robust security against Impersonation and offline password-guessing attacks, but it comes with the drawback of small key sizes and increased computational and communication overhead.

Jiang *et al*. (2018a-b) discussed user privacy within the context of sending queries over encrypted multi-dimensional big metering data. They sent encrypted data to several large data storage systems and used the Locality Sensitive Hashing (LSH) algorithm to execute a similarity search. Policies based on Attribute-Based Encryption (CP-ABE) ciphertext were used to restrict access and safeguard search results. Although this method performed well in terms of data privacy and secrecy in a semi-trusted cloud environment, it took a while to search through multi-dimensional big metering data.

Key-Policy Attribute-Based Encryption (KP-ABE) Yu *et al*. (2010) and Ciphertext-Policy Attribute-Based Encryption (CP-ABE) (Bethencourt *et al*., 2007) are two components of the Attribute-Based Encryption (ABE) algorithm. This encryption technique includes ABE decryption rules, which eliminate the need for routine key distribution in ciphertext access control. However, data owners must re-encrypt the data when access control parameters change dynamically. In Reference (to Li *et al*., 2010), an approach based on PRE is suggested, allowing re-encryption of ciphertext by a semi-trusted agent with a proxy key, without access to the associated plaintext or decryption key.

Reference entry. (2009) introduces the Fully Homomorphic Encryption (FHE) technique, which enables certain algebraic operations on ciphertext while keeping the result encrypted. The encrypted data can be retrieved, compared and analyzed without decryption throughout the process. However, FHE involves substantial computation and may pose challenges in implementation with current technology. Ciphertext retrieval solutions in the cloud are discussed in References (Ananthi *et al*., 2011; Hu *et al*., 2011b; Cao *et al*., 2014), focusing on safeguarding data privacy during ciphertext retrieval.

A new cryptographic access control technique, Attribute-Based Access Control for Cloud Storage (ABACCS), is suggested (Hong *et al*., 2010). Data is encrypted with an attribute condition, allowing decryption only if a user's attributes satisfy the condition. Each user's private key is tagged with a set of attributes.

In reference to Lv and Qiao (2020), the challenge of providing fine-grained access control for a large user base in the cloud is addressed and a secure and efficient revocation approach based on a modified CP-ABE algorithm is proposed. Shamir's secret sharing principle is applied to create fine-grained access control with user revocation. Single Sign-On (SSO) enables authorized users to access the cloud storage system through a standard common application interface.

Xu *et al*. (2014) introduced a novel approach to Privacy-Preserving Data Mining (PPDM) that involves collaboration among multiple users. In this method, data

providers protect sensitive information by limiting access, introducing false data and making privacy concessions for mutual benefit. An alternative technology, DNT (Mengke *et al*., 2016) shows promise in addressing privacy concerns. Data collectors emphasize privacy preservation through Privacy-Preserving Data Publishing (PPDP). Data miners employ techniques such as privacy-preserving association rule mining, privacy-preserving classification and privacy-preserving clustering to enhance privacy protection.

The current landscape of information security presents challenges, including the risk of data leakage (Machanavajjhala and Reiter, 2012) susceptibility to cyberattacks and overall safety and security concerns related to the use of big data. Information leaks may encompass personal privacy violations or threats to national security.

Fan *et al*. (2018) have elaborated on secure key management in the context of user privacy and data confidentiality within big data networking environments. They propose a multi-layered approach to key generation, with upper-layer keys used to encrypt lower-layer keys to bolster user security. While this hierarchical key management scheme is secure and efficient, it does entail increased complexity and computation requirements. Large data sizes result in extended encryption and decryption times.

Win *et al*. (2018) conducted security analytics on virtualized infrastructure stored in the Hadoop Distributed File System (HDFS). Their two-step machine learning algorithm combines logistic regression and belief propagation for attack probability computation. Third-party auditors (TPAs) play a pivotal role in cloud-enabled big data environments. (Zhan *et al*., 2018) have introduced a trusted verifier that monitors data owners in a multi-layered outsourced big data environment. This addresses situations where data owners do not directly audit or manage data in storage systems. The work includes the proposal of two policy methods and a chain of trust for MapReduce applications. However, this complexity arises from the use of two policy methods.

In centralized environments, malicious users may tamper with data without the data owner's permission. (Neela and Kavitha., 2018) Have focused on enhancing big data security for data owners, primarily employing the cyclic shift transportation algorithm and a hash-based timestamp to prevent real-time security breaches. While their approach involves partitioning the original file into matrices and implementing shifting operations, it does not account for insider attacks, which are crucial to data recovery.

Wu *et al*. (2018) have introduced a security situational awareness-based big data analysis strategy for smart grid applications. They combine game theory, reinforcement learning and a fuzzy cluster-based analytic model to analyze security. Real security factors are input into a neural network, incorporating a game theory element where legitimate users and insider attackers participate. The complexity arises from the incorporation of deep learning and game theory approaches.

Mall *et al*. (2018) have introduced a new security model for cloud computing environments. They split input data into fixed-size blocks and use a Genetic Algorithm (GA) for block encryption. Each encrypted file is stored in the cloud at various locations. However, the use of a genetic algorithm increases processing time for individual files and larger block sizes result in a lack of security.

Goyal and Kant (2018) have proposed a hybrid cryptography algorithm for data encryption, combining symmetric and asymmetric algorithms such as AES, ECC and SHA-1. Their approach involves four phases, including data owner registration, data storage, user authentication and data auditing. While the hybrid encryption algorithm exhibits better performance, the combination of hash, symmetric and asymmetric algorithms introduces complexity and ECC's slow key generation may pose challenges.

Hababeh *et al*. (2019) have discussed methods for big data classification and security in a cloud environment, categorizing data into public and confidential classes based on risk levels. Security measures are applied to confidential data within the Hadoop Distributed File System (HDFS). However, the designed security algorithm may not be suitable for handling large volumes of data with varying characteristics.

Adnan and Ariffin (2019) have introduced a 3D-AES algorithm for big data security, which incorporates multiple functions and iterations to enhance complexity, security and performance. It offers promising results when compared to AES, particularly in terms of randomness and performance. However, further comparisons with alternative randomness algorithms are necessary to validate its effectiveness.

Cryptography algorithms provide a satisfactory solution to data confidentiality concerns, but they may be complex and computationally intensive for encrypting large data volumes. As a result, researchers have explored alternative solutions to address this challenge.

Hybrid Method (*Combination of More Than One Method*)

In the realm of hybrid methods, (Ulusoy *et al*., 2014) address approval channels through per-client task records coded in Java. GuardMR, on the other hand, assigns channels based on roles and proposes a flexible approach to channel definition. This approach allows for the specification and modification of standards at a high level of reflection using the object constraint language.

Yang and Ren (2015) have devised an attribute-based access control scheme with dynamic policy updates for big data. This scheme minimizes the need to transfer encrypted data between data owners and cloud servers by utilizing previously encrypted data under older access policies. The cloud server only requires a policy updating key from the data owner to update access policies directly.

Authentication schemes like passport and OpenID, while convenient for mobile users, involve fully trusted third parties in each authentication phase, potentially creating security bottlenecks (Jiang *et al*., 2016).

Aditham and Ranganathan (2018) have introduced a two-step attack detection algorithm that functions as a secure communication protocol for monitoring system execution processes. This involves the creation of system controls for each process and matching instructions with replica nodes. In secure data communication, data nodes generate random keys, potentially posing a risk to user privacy and data security.

Reddy (2018) has described access control and anomaly detection for big data processing in the cloud. They utilize Kerberos as a third-party authentication protocol for non-secure data access. The combination of access control and anomaly detection measures is used to safeguard data from malicious users, with spikes employed for monitoring and control. However, the system has limitations in supporting diverse data sources, including text, images, audio and video.

Jiang *et al.* (2016) introduced an anonymous authentication scheme for medical environments based on the cloud, providing data confidentiality and message authenticity. The paper asserted that the proposed scheme's security was demonstrable in the standard model and demonstrated its suitability for cloud-based telecare medical information systems when compared to competitive protocols.

In the study presented in the paper Narayanan *et al*., (2020), a comprehensive examination of encryption algorithms in mobile data applications was conducted. Experimental results comparing the lightweight algorithm with existing state-of-the-art alternatives showcased its excellent performance when handling large-scale data.

The realm of big data requires additional security and privacy measures in data collection, storage, analysis (Narayanan *et al*., 2017a-b; Unnikrishnan *et al*., 2017) and transmission. In a paper (Narayanan *et al*., 2022b), two critical concerns, user privacy and data security within a big data-enabled cloud environment, are addressed. The paper presents three key strategies: Big data outsourcing from data owners, big data sharing with data users and big data management in the cloud.

The paper introduces the Decentralized Blockchain-based Security (DeBlock-Sec) scheme Narayanan *et al*., (2022a), aiming to address security challenges prevalent in resource-constrained IoT environments. It highlights the limitations of centralized authentication and complex encryption schemes, emphasizing the need for innovative solutions. DeBlock-sec operates in three phases: Authentication, data encryption and data retrieval, employing novel protocols and algorithms to ensure high-level security. The authentication phase utilizes a decentralized blockchain-based protocol to verify the legitimacy of IoT devices and users. Data encryption is performed in the spark environment using the lightweight SALSA20 algorithm, with encryption levels determined by the scores method. Encrypted data is then stored in a spark-enabled cloud with an index generated using DenFT indexing. The retrieval phase enables fast searches through the index, with secret key exchange secured using the revised diffie-hellman algorithm. Experimental results demonstrate improved performance metrics, including reduced encryption and decryption times, improved storage efficiency, throughput and search times. The paper concludes by reaffirming the significance of DeBlock-Sec in mitigating security risks in IoT systems, particularly in industrial IoT and highlights its potential for real-time applications. Overall, DeBlock-Sec represents a significant advancement in IoT security, offering a promising solution to enhance security and privacy in modern communication environments.

Considering the impending transition to Post-Quantum Cryptography (PQC), it's crucial for papers addressing privacy concerns to acknowledge this paradigm shift. As PQC gradually supplants traditional cryptographic methods like ECC and RSA, every facet of security applications, ranging from smartphones to blockchain technology, will feel its impact. Therefore, it's advisable to include relevant literature on post-quantum cryptographic techniques to ensure comprehensive coverage of contemporary security challenges. This proactive approach not only enriches the discourse on privacy but also prepares stakeholders for the forthcoming cryptographic landscape.

Two defect diagnosis techniques for lightweight BLAKE constructions are presented in the work (Kermani *et al*., 2019). A complementary strategy is one approach that, despite a 17.4% throughput reduction, accomplishes full fault identification for both transient and persistent defects. In comparison to time-redundancy techniques that do not make use of the sub-pipelining technique described in this research, this reduction is noticeably smaller. Furthermore, the suggested RERO-based method achieves an area overhead of 8.7%, a throughput deterioration of 7.1% and a defect detection capability of over 99.8%. Either of these methods can be used to improve the robustness of BLAKE algorithm hardware constructs, depending on the application's overhead tolerance and error detection needs.

The study addresses the threat posed by quantum computing to classical cryptography by focusing on the Supersingular Isogeny Key Encapsulation (SIKE) (Kermani *et al*., 2019) mechanism, a leading post-quantum cryptographic solution. SIKE is valued for its compact key sizes, which minimize bandwidth and memory requirements. The research aims to enhance the speed performance of SIKE by implementing optimized finite field arithmetic techniques tailored for ARMv7-M architecture. Handcrafted assembly code for modular multiplication and squaring functions achieves notable speed improvements compared to previous implementations. Integration of these optimizations into the SIKE protocol results in significant latency reductions across various SIKE instances, demonstrated through experimentation on the NIST-recommended STM32F407VG discovery board.

Software/hardware codesign approaches tackle the challenge of implementing purely hardware-dedicated Post-Quantum Cryptography (PQC) algorithms by designing various hardware accelerators for lattice-based Key Encapsulation Mechanisms (KEMs) (Canto *et al*., 2023). Nevertheless, strong defenses are required because these hardware accelerators are vulnerable to differential fault analysis attacks. This study uses recomputing with negated, shifted and scaled operands to propose various error detection algorithms for FrodoKEM, saber and NTRU hardware accelerators. Additionally, in order to evaluate overheads and performance deterioration, the suggested fault detection architectures are incorporated into the original designs. The Xilinx FPGA Kintex Ultrascale + xcku5p-sfvb784-1LV-i is used for evaluation and it shows good error coverage. With the NTRU hardware accelerator, the suggested error detection approaches have a maximum area overhead of 39.6% and with the FrodoKEM hardware accelerator, the worst-case latency overhead is less than 33%. Power overhead is minimized, with a worst-case scenario of under 4% observed with the NTRU hardware accelerator. Comparative analysis with other fault detection methods underscores the reasonable overhead of the techniques proposed in this study.

The article (Kaur *et al*., 2024) introduces novel error detection schemes, namely normal signature, interleaved signature and the (7, 4) Hamming code, specifically tailored for squaring matrices, trace functions and the PB multiplier components of the stream cipher WG-29. These schemes are robust in detecting Single Bit Upsets (SBUs) and Multiple Bit Upsets (MBUs), providing protection against both manufacturing flaws and intentionally introduced faults. These schemes were not previously investigated for WG-29. Error coverage simulations are used to assess the performance of the suggested strategies on the Xilinx Kintex-7 and Xilinx spartan-7 FPGA families using Xilinx Vivado 2020.2. Across the two Xilinx FPGA families, the protected WG-29 has FPGA overheads that range from 28.87-36.12% for area, 11.04-12.34% for power and 4.22-7.04% for delay, all while keeping error coverage near 100%. The benchmark results show that the given techniques produce high error coverage with reasonable overheads when compared to other cutting-edge WG and cryptographic ciphers. Additionally, these techniques can be modified for additional cryptographic hardware implementations.

The emergence of quantum computing has underscored the necessity for cryptographic algorithms that are not only low-power and low-energy but also resilient against potential attacks facilitated by quantum capabilities. In response to this post-quantum era, various solutions have been explored, among which code-based cryptography stands out as a viable option (Cintas-Canto *et al*., 2023). The hardware architectures of such cryptographic systems have garnered significant attention within the NIST standardization process, progressing to the final round scheduled for conclusion between 2022 and 2024. However, despite the robust error correction properties exhibited by constructions like McEliece and Niederreiter public key cryptography, prior research has exposed vulnerabilities in their hardware implementations to faults arising from environmental factors and deliberate attacks, such as differential fault analysis. Prior research has shown that the efficiency of error detection techniques can be affected by the selection of codes, either reduced or classical (using quasi-cyclic alternant codes or quasi-dyadic Goppa codes). To overcome these drawbacks, this study presents the first efficient fault detection systems, which include interleaved parity, normal parity and two different Cyclic Redundancy Checks (CRC), called CRC-2 and CRC-8. Although we experiment mostly with the McEliece variation, it's crucial to remember that the suggested techniques work with other code-based cryptosystems as well. To verify the feasibility of these approaches, we carry out evaluations of error detection performance and apply them to a field-programmable gate array Kintex-7 device (xc7k70tfbv676-1). Furthermore, we evaluate the performance degradation and overheads of the proposed schemes to demonstrate their suitability for resource-constrained embedded systems.

Advances in quantum technologies threaten classical cryptography, necessitating the adoption of Post-Quantum (PQ) schemes. The work (Anastasova *et al*., 2022) achieves a new speed record for the Supersingular Isogeny Key Encapsulation (SIKE) protocol by implementing optimized low-level finite field arithmetic on the ARMv7-M architecture, resulting in significant latency reductions.

The study (Kermani *et al*., 2019) addresses the emergence of secure deeply embedded systems, exemplified by implantable and wearable medical devices, which present heightened security risks

compared to conventional embedded systems like smart cards and nano-sensor networks. Security breaches in medical devices can have life-threatening consequences, making traditional solutions impractical due to their stringent constraints. While cryptographic engineering research has begun tackling these challenges, educational programs lag behind, hindered by the multidisciplinary nature of emerging security issues. This study introduces a strategy for integrating research and education to address security concerns in deeply embedded systems, with a focus on medical devices. Implementation of this strategy at the graduate level, emphasizing fault analysis attacks, demonstrates its effectiveness while highlighting challenges compared to traditional embedded system security education. Notably, the proposed integration approach is adaptable to other critical infrastructures.

Research Finding

This article comprehensively addresses a range of critical issues in the realm of big data, including but not limited to data processing, heterogeneity, data life cycle management, scalability, data visualization security and privacy. Even more, it delves deeply into the security and privacy aspects, exploring key facets such as key management, integrity, confidentiality, availability, monitoring and auditing. The main objective of this article is to furnish readers with a comprehensive overview of the security and privacy challenges associated with big data. It underscores the significant efforts that have been invested in addressing these challenges while emphasizing that the unique characteristics of big data necessitate innovative solutions beyond the scope of traditional or current security measures.

Table 1 displays the comparison of security properties aof different papers presented in the literature. Figure 24, presents an analysis of the prevalence of big data across a multitude of papers, shedding light on its extensive utilization. Additionally, Fig. 25 provides insights into our survey methodology, categorizing the selected papers based on their year of publication. This article is intended to contribute valuable insights to the field of big data security and privacy, furthering the understanding and exploration of these critical aspects.

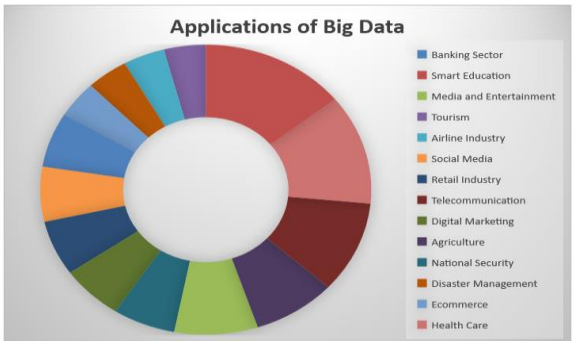

Fig. 24: Application of big data

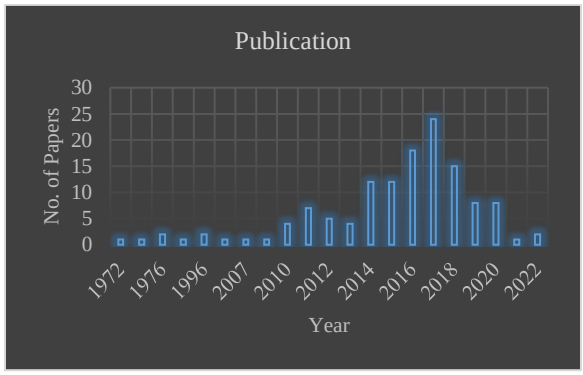

Fig. 25: Paper based on year of publications

Table 1: Comparison of security features with state-of-the-art studies

Sl. No.	Ref	Privacy requirement					Weakness	Strength
		IN	AU	CO	NR	AC		
1	Dutta *et al*. (2020)	*	√	*	*	√	Insider attack	Detects intrusions
2	Bell and LaPadula (1996)	*	√	*	*	√	Computation complexity is higher	Security level
3	Prasetyo *et al*. (2014)	√	*	√	√	-		FPGA resource
4	Mollah *et al*. (2017)	√	√	√	√	√	Computation overhead	Use private key encryption and public key encryption
5	Vorugunti (2016)	√	√	√	√	√		Searchable encryption
6	Zhao *et al*. (2018)	√	√	√	√	√	Smaller key size	Addresses two security attacks such as Impersonation attack and offline password guessing attack
7	Fan *et al*. (2018)	√	√	√	√	√	Encryption time and decryption time is a large for sized data	Hierarchical key management scheme
8	Goyal and Kant (2018)	√	√	√	√	√	Computation complexity is a higher	Hybrid cryptography algorithm for data encryption
9	Reddy (2018)	√	√	√	*	*	Failed to support variety of data sources such as images, audio and videos	Access control was invoked as data access
10	Jiang *et al*. (2016)	*	√	√	√	√	Susceptible to user impersonation attacks in the registration phase	Three-factor authentication combining passwords, for mobile device

Table 1: Continue

								and biometrics perfectly match this requirement by providing high security strength
11	Narayanan *et al*. (2020)	√	√	√	*		Offline password-guessing attack	Lightweight encryption
12	Narayanan *et al*. (2020)	√	√	√	√	√	Centralized method	Data security and user privacy in big data-enabled cloud environment
13	Narayanan *et al*. (2022a-b)	√	√	√	√	√	Nil	Decentralized blockchain-based security

Note: √ denotes the scheme's ability to offer the matching security characteristic, while * denotes its inability to do so. Integrity (IN), Authenticity (AU), Confidentiality (CO), Non-Repudiation (NR) and Accountability (AC)

Conclusion

The advent of big data has left an indelible mark on diverse sectors and industries, fundamentally transforming the landscape. This study seeks to provide a succinct overview of the profound significance of big data while simultaneously addressing the formidable challenges it introduces. Notably, the sheer volume of big data necessitates a heightened focus on security and privacy measures, spanning the entire spectrum of data processes, encompassing collection, storage, analysis and transmission. Within the scope of this study, we embark on a comprehensive comparative analysis of existing studies pertaining to big data security and privacy. Drawing from the wealth of available literature, several salient recommendations emerge as guiding principles. To begin, network traffic demands robust encryption using established standards. The access to devices must be subjected to rigorous authentication protocols, ensuring that only authorized personnel can gain entry to systems. Furthermore, data analysis should be conducted on anonymized data to safeguard individual privacy. The establishment of secure communication channels is paramount in thwarting data leakage and continuous network monitoring assumes a pivotal role in the early detection of potential threats. The paramount concern is that the issues of privacy, safety and security in the realm of big data will continue to occupy a prominent place in future discussions. Consequently, it becomes imperative to develop novel techniques, technologies and solutions that enhance human-computer interactions. Existing technologies should also undergo refinement to yield more precise and efficient results. The ultimate objective is to address the entirety of the problem through integrated solutions, eschewing isolated successes in isolated areas.

In essence, an integrated engineering approach is indispensable to comprehensively manage the security of big data. This study aspires to enrich the understanding of big data and its ecosystem, paving the way for the development of superior systems, tools, structures and solutions, not merely for the present but also to meet the evolving demands of the future. While our examination has shed light on the privacy and efficiency challenges in the broader context of big data analytics, it is equally critical to intensify research efforts aimed at addressing the unique privacy issues that surface within specific big data analytics domains. By leveraging decentralized blockchain technology, innovative protocols and algorithms, DeBlock-sec effectively mitigates the limitations of centralized authentication and complex encryption schemes. Through its three-phase approach of authentication, data encryption and data retrieval, DeBlock-sec ensures high-level security for devices and users. The authentication phase, utilizing a decentralized blockchain-based protocol, verifies the legitimacy of devices and users, enhancing overall system integrity. Data encryption, performed in the Spark environment using the lightweight SALSA20 algorithm and scoresen method, ensures confidentiality and reliability of data transmission. Storage efficiency is further improved through the use of DenFT indexing, facilitating fast and efficient data retrieval. Experimental results demonstrate significant improvements in performance metrics, including reduced encryption and decryption times, improved storage efficiency, throughput and search times. These results validate the effectiveness of DeBlock-sec in enhancing security and efficiency. Moreover, the relevance of DeBlock-sec extends to real-time applications, where security, reliability and availability of information are paramount. By offering a robust security framework that adapts to modern communication environments and application deployment scenarios, DeBlock-sec represents a significant advancement in security.

Acknowledgment

I would like to express my gratitude to my supervisor, Professor Varghese Paul, for his unwavering support, guidance and invaluable feedback throughout this research.

Funding Information

This research received no financial support from public or private sources.

Author's Contributions

Uma Narayanan: Provided essential guidance and oversight throughout the project, designed the research plan and organized the study, coordinated the data

analysis and contributed to the written of the manuscript.

Nithin Puthiya Veettil, Ratheesh Thottungal Krishnankutty and Leya Elizabeth Sunny: Participated in all experiments, coordinated the data analysis and contributed to the written of the manuscript.

Varghese Paul: Supervised study contributed to the written of the manuscript.

Ethics

The present study is an original research effort and the lead author confirms that all co-authors have reviewed and approved the manuscript, with no ethical concerns raised.

Conflict of Interest

The authors declare no conflicts of interest.

References

Aditham, S., & Ranganathan, N. (2018). A System Architecture for the Detection of Insider Attacks in Big Data Systems. *IEEE Transactions on Dependable and Secure Computing*, *15*(6), 974–987. https://doi.org/10.1109/tdsc.2017.2768533

Adnan, N. A. N., & Ariffin, S. (2019). Big Data Security in the Web-Based Cloud Storage System Using 3D-AES Block Cipher Cryptography Algorithm. In *Soft Computing in Data Science: 4th International Conference, SCDS 2018, Bangkok, Thailand, August 15-16, 2018, Proceedings* (1st Ed., Vol. *4*, pp. 309–321). Springer Singapore. https://doi.org/10.1007/978-981-13-3441-2_24

Agrawal, D., Das, S., & El Abbadi, A. (2011). Big data and cloud computing. *Proceedings of the 14th International Conference on Extending Database Technology*, 530–533. https://doi.org/10.1145/1951365.1951432

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1), 1–24. https://doi.org/10.1186/s40537-019-0191-6

Ananthi, S., Sendil, M. S., & Karthik, S. (2011). Privacy Preserving Keyword Search over Encrypted Cloud Data. In *Advances in Computing and Communications: First International Conference, ACC 2011, Kochi, India, July 22-24, 2011* (1st Ed., pp. 480–487). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-22709-7_47

Anastasova, M., Azarderakhsh, R., & Kermani, M. M. (2022). Time-Optimal Design of Finite Field Arithmetic for SIKE on Cortex-M4. In I*nternational Conference on Information Security Applications* (1st Ed., pp. 265–276). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-25659-2_19

Arsenault, A. H. (2017). The datafication of media: Big data and the media industries. *International Journal of Media & Cultural Politics*, *13*(1–2), 7–24. https://doi.org/10.1386/macp.13.1-2.7_1

Belchior, R., Putz, B., Pernul, G., Correia, M., Vasconcelos, A., & Guerreiro, S. (2020). SSIBAC: Self-Sovereign Identity Based Access Control. *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 1935–1943. https://doi.org/10.1109/trustcom50675.2020.00264

Bell, D. E., & La Padula, L. J. (1976). *Secure Computer System: Unified Exposition and Multics Interpretation*. Defense Technical Information Center. https://apps.dtic.mil/sti/citations/ADA023588

Bell, D. E. B., & LaPadula, L. J. (1996). Secure Computer Systems: A Mathematical Model. Volume II. *Journal of Computer Security*, *4*(2/3), 229–263. https://dblp.org/rec/journals/jcs/BellL96.html

Belle, A., Thiagarajan, R., Soroushmehr, S. M. R., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big Data Analytics in Healthcare. *BioMed Research International*, *2015*(1), 370194. https://doi.org/10.1155/2015/370194

Bendre, M. R., Thool, R. C., & Thool, V. R. (2015). Big data in precision agriculture: Weather forecasting for future farming. *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, 744–750. https://doi.org/10.1109/ngct.2015.7375220

Bethencourt, J., Sahai, A., & Waters, B. (2007). Ciphertext-Policy Attribute-Based Encryption. *2007 IEEE Symposium on Security and Privacy (SP '07)*, 321–334. https://doi.org/10.1109/sp.2007.11

Braik, W., Morandat, F., Falleri, J.-R., & Blanc, X. (2016). Real time streaming pattern detection for eCommerce. *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, 916–922. https://doi.org/10.1145/2851613.2851653

Brewster, B., Kemp, B., Galehbakhtiari, S., & Akhgar, B. (2015). Cybercrime: Attack Motivations and Implications for Big Data and National Security. In *Application of Big Data for National Security* (1st Ed., pp. 108–127). Elsevier. https://doi.org/10.1016/b978-0-12-801967-2.00008-2

Brown, R. L., & Harmon, R. R. (2014). Viral geofencing: An exploration of emerging big-data driven direct digital marketing services. *IEEE*, 3300–3308. https://ieeexplore.ieee.org/abstract/document/6921234

Cantabella, M., Martínez-España, R., Ayuso, B., Yáñez, J. A., & Muñoz, A. (2019). Analysis of student behavior in learning management systems through a Big Data framework. *Future Generation Computer Systems*, *90*, 262–272. https://doi.org/10.1016/j.future.2018.08.003

Canto, A. C., Sarker, A., Kaur, J., Kermani, M. M., & Azarderakhsh, R. (2023). Error Detection Schemes Assessed on FPGA for Multipliers in Lattice-Based Key Encapsulation Mechanisms in Post-Quantum Cryptography. *IEEE Transactions on Emerging Topics in Computing*, *11*(3), 791–797. https://doi.org/10.1109/tetc.2022.3217006

Cao, N., Wang, C., Li, M., Ren, K., & Lou, W. (2014). Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data. *IEEE Transactions on Parallel and Distributed Systems*, *25*(1), 222–233. https://doi.org/10.1109/tpds.2013.45

Cerchiello, P., & Giudici, P. (2016). Big data analysis for financial risk management. *Journal of Big Data*, *3*(1), 1–12. https://doi.org/10.1186/s40537-016-0053-4

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188. https://doi.org/10.2307/41703503

Chen, M., Mao, S., Zhang, Y., & Leung, V. C. M. (2014). *Big Data* (1st Ed., Vol. *100*). Springer. https://doi.org/10.1007/978-3-319-06245-7

Cheng, H., Rong, C., Hwang, K., Wang, W., & Li, Y. (2015). Secure big data storage and sharing scheme for cloud tenants. *China Communications*, *12*(6), 106–115. https://doi.org/10.1109/cc.2015.7122469

Choi, S., & Bae, B. (2015). The Real-Time Monitoring System of Social Big Data for Disaster Management. *Computer Science and Its Applications: Ubiquitous Information Technologies*, 809–815. https://doi.org/10.1007/978-3-662-45402-2_115

Chukkapalli, S. S. L., Piplai, A., Mittal, S., Gupta, M., & Joshi, A. (2020). A Smart-Farming Ontology for Attribute Based Access Control. *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, 29–34. https://doi.org/10.1109/bigdatasecurity-hpsc-ids49724.2020.00017

Cintas-Canto, A., Kermani, M. M., & Azarderakhsh, R. (2023). Reliable Architectures for Finite Field Multipliers Using Cyclic Codes on FPGA Utilized in Classic and Post-Quantum Cryptography. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *31*(1), 157–161. https://doi.org/10.1109/tvlsi.2022.3224357

Colombo, P., & Ferrari, E. (2017a). Enhancing MongoDB with Purpose-Based Access Control. *IEEE Transactions on Dependable and Secure Computing*, *14*(6), 591–604. https://doi.org/10.1109/tdsc.2015.2497680

Colombo, P., & Ferrari, E. (2017b). Towards a Unifying Attribute Based Access Control Approach for NoSQL Datastores. *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, 709–720. https://doi.org/10.1109/icde.2017.123

Crampton, J. W. (2015). Collect it all: national security, Big Data and governance. *GeoJournal*, *80*(4), 519–531. https://doi.org/10.1007/s10708-014-9598-y

Downs, D. D., Rub, J. R., Kung, K. C., & Jordan, C. S. (1985). Issues in Discretionary Access Control. *1985 IEEE Symposium on Security and Privacy*, 208–208. https://doi.org/10.1109/sp.1985.10014 https://www.statista.com/statistics/254266/global-big-data-market-forecast/

Das, S. (2020). Innovations in Digital Banking Service Brand Equity and Millennial Consumerism. In *Digital Transformation and Innovative Services for Business and Learning* (1st Ed., pp. 62–79). IGI Global. https://doi.org/10.4018/978-1-7998-5175-2.ch004

Das, S. (2021). *Search engine optimization and marketing: A recipe for success in digital marketing* (1st Ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429298509

Dam, R. V. D. (2013). Big Data a Sure Thing for Telecommunications: Telecom's Future in Big Data. *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 148–154. https://doi.org/10.1109/cyberc.2013.32

Denning, D. E. (1976). A lattice model of secure information flow. *Communications of the ACM*, *19*(5), 236–243. https://doi.org/10.1145/360051.360056

Drigas, A. S., & Leliopoulos, P. (2014). The use of big data in education. *International Journal of Computer Science Issues (IJCSI)*, *11*(5), 58-63. https://www.researchgate.net/publication/274890131_The_Use_of_Big_Data_in_Education

Dutta, S., Chukkapalli, S. S. L., Sulgekar, M., Krithivasan, S., Das, P. K., & Joshi, A. (2020). Context Sensitive Access Control in Smart Home Environments. *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, 35–41. https://doi.org/10.1109/bigdatasecurity-hpsc-ids49724.2020.00018

Elia, G., Solazzo, G., Lorenzo, G., & Passiante, G. (2019). Assessing learners' satisfaction in collaborative online courses through a big data approach. *Computers in Human Behavior*, *92*, 589–599. https://doi.org/10.1016/j.chb.2018.04.033

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*, *69*(2), 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001

Fan, K., Lou, S., Su, R., Li, H., & Yang, Y. (2018). Secure and private key management scheme in big data networking. *Peer-to-Peer Networking and Applications*, *11*(5), 992–999. https://doi.org/10.1007/s12083-017-0579-z

Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society*, *3*(1), 205395171664582. https://doi.org/10.1177/2053951716645828

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations–A case from Sweden. *Journal of Destination Marketing & Management*, *3*(4), 198–209. https://doi.org/10.1016/j.jdmm.2014.08.002

Gupta, B. B., & Quamara, M. (2018). An identity based access control and mutual authentication framework for distributed cloud computing services in IoT environment using smart cards. *Procedia Computer Science*, *132*, 189–197. https://doi.org/10.1016/j.procs.2018.05.185

Goyal, V., & Kant, C. (2018). An Effective Hybrid Encryption Algorithm for Ensuring Cloud Data Security. In *Big Data Analytics: Proceedings of CSI 2015* (1st Ed., pp. 195–210). Springer Singapore. https://doi.org/10.1007/978-981-10-6620-7_20

Graham, G. S., & Denning, P. J. (1971). Protection: principles and practice. *Proceedings of the May 16-18, 1972, Spring Joint Computer Conference*, 417–429. https://doi.org/10.1145/1478873.1478928

Griebel, L., Prokosch, H.-U., Köpcke, F., Toddenroth, D., Christoph, J., Leb, I., Engel, I., & Sedlmayr, M. (2015). A scoping review of cloud computing in healthcare. *BMC Medical Informatics and Decision Making*, *15*(1), 1–16. https://doi.org/10.1186/s12911-015-0145-7

Gupta, M., Abdelsalam, M., Khorsandroo, S., & Mittal, S. (2020). Security and Privacy in Smart Farming: Challenges and Opportunities. *IEEE Access*, *8*, 34564–34584. https://doi.org/10.1109/access.2020.2975142

Gupta, M., Patwa, F., & Sandhu, R. (2017). Object-Tagged RBAC Model for the Hadoop Ecosystem. In *IFIP Annual Conference on Data and Applications Security and Privacy* (1st Ed., pp. 63–81). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-61176-1_4

Gupta, M., Patwa, F., & Sandhu, R. (2018). An Attribute-Based Access Control Model for Secure Big Data Processing in Hadoop Ecosystem. *Proceedings of the Third ACM Workshop on Attribute-Based Access Control*, 13–24. https://doi.org/10.1145/3180457.3180463

Hababeh, I., Gharaibeh, A., Nofal, S., & Khalil, I. (2019). An Integrated Methodology for Big Data Classification and Security for Improving Cloud Systems Data Mobility. *IEEE Access*, *7*, 9153–9163. https://doi.org/10.1109/access.2018.2890099

He, Y., Yu, F. R., Zhao, N., Yin, H., Yao, H., & Qiu, R. C. (2016). Big Data Analytics in Mobile Cellular Networks. *IEEE Access*, *4*, 1985–1996. https://doi.org/10.1109/access.2016.2540520

Hong, C., Zhang, M., & Feng, D. G. (2010). AB-ACCS: A cryptographic access control scheme for cloud storage. *Journal of Computer Research and Development*, *47*(1), 259–265.

Holst, A. (2020). *Big Data market revenue forecast worldwide 2011-2027"*. Technology & Telecommunications–IT Services.

Hu, H., Xu, J., Ren, C., & Choi, B. (2011a). Processing private queries over untrusted data cloud through privacy homomorphism. *2011 IEEE 27th International Conference on Data Engineering*, 601–612. https://doi.org/10.1109/icde.2011.5767862

Hu, V. C., Kuhn, D. R., Xie, T., & Hwang, J. (2011b). Model checking for verification of mandatory access control models and properties. *International Journal of Software Engineering and Knowledge Engineering*, *21*(1), 103–127. https://doi.org/10.1142/S021819401100513X

Jayant, D., B., Swapnaja A, U., Sulabha S, A., & Dattatray G, M. (2014). Analysis of DAC MAC RBAC Access Control based Models for Security. *International Journal of Computer Applications*, *104*(5), 6–13. https://doi.org/10.5120/18196-9115

Jadon, P., & Mishra, D. K. (2019). Security and privacy issues in big data: A review. In V. Rathore, M. Worring, D. Mishra, A. Joshi, & S. Maheshwari (Eds.), *Emerging Trends in Expert Applications and Security* (Vol. 841). Springer. https://doi.org/10.1007/978-981-13-2285-3_77

Jensen, M. (2013). Challenges of Privacy Protection in Big Data Analytics. *2013 IEEE International Congress on Big Data*, 235–238. https://doi.org/10.1109/bigdata.congress.2013.39

Jeong, Y.-S., & Shin, S.-S. (2016). An Efficient Authentication Scheme to Protect User Privacy in Seamless Big Data Services. *Wireless Personal Communications*, *86*(1), 7–19. https://doi.org/10.1007/s11277-015-2990-1

Jiang, Q., Khan, M. K., Lu, X., Ma, J., & He, D. (2016). A privacy preserving three-factor authentication protocol for e-Health clouds. *The Journal of Supercomputing*, *72*(10), 3826–3849. https://doi.org/10.1007/s11227-015-1610-x

Jiang, Q., Ma, J., & Wei, F. (2018a). On the Security of a Privacy-Aware Authentication Scheme for Distributed Mobile Cloud Computing Services. *IEEE Systems Journal*, *12*(2), 2039–2042. https://doi.org/10.1109/jsyst.2016.2574719

Jiang, R., Lu, R., & Choo, K.-K. R. (2018b). Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data. *Future Generation Computer Systems*, *78*, 392–401. https://doi.org/10.1016/j.future.2016.05.005

Joshi, M., Mittal, S., Joshi, K. P., & Finin, T. (2017). Semantically Rich, Oblivious Access Control Using ABAC for Secure Cloud Storage. *2017 IEEE International Conference on Edge Computing (EDGE)*, 142–149. https://doi.org/10.1109/ieee.edge.2017.27

Kasturi, E., Devi, S. P., Kiran, S. V., & Manivannan, S. (2016). Airline Route Profitability Analysis and Optimization Using BIG DATA Analyticson Aviation Data Sets under Heuristic Techniques. *Procedia Computer Science*, *87*, 86–92. https://doi.org/10.1016/j.procs.2016.05.131

Kaur, J., Canto, A. C., Mozaffari Kermani, M., & Azarderakhsh, R. (2024). Hardware Constructions for Error Detection in WG-29 Stream Cipher Benchmarked on FPGA. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *43*(4), 1307–1311. https://doi.org/10.1109/tcad.2023.3338108

Kazim, M., Liu, L., & Zhu, S. Y. (2018). A Framework for Orchestrating Secure and Dynamic Access of IoT Services in Multi-Cloud Environments. *IEEE Access*, *6*, 58619–58633. https://doi.org/10.1109/access.2018.2873812

Kermani, M. M., Bayat-Sarmadi, S., Ackie, A.-B., & Azarderakhsh, R. (2019). High-Performance Fault Diagnosis Schemes for Efficient Hash Algorithm BLAKE. *2019 IEEE 10th Latin American Symposium on Circuits & Systems (LASCAS)*, 201–204. https://doi.org/10.1109/lascas.2019.8667597

Khan, Y., Shafiq, S., Naeem, A., Ahmed, S., Safwan, N., & Hussain, S. (2019). Customers Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry. *International Journal of Advanced Computer Science and Applications*, *10*(9), 1–607. https://doi.org/10.14569/ijacsa.2019.0100918

Klein, J., Buglak, R., Blockow, D., Wuttke, T., & Cooper, B. (2016). A reference architecture for big data systems in the national security domain. *Proceedings of the 2nd International Workshop on BIG Data Software Engineering*, 51–57. https://doi.org/10.1145/2896825.2896834

Lampson, B. W. (1974). Protection. *ACM SIGOPS Operating Systems Review*, *8*(1), 18–24. https://doi.org/10.1145/775265.775268

Leeflang, P. S. H., Verhoef, P. C., Dahlström, P., & Freundt, T. (2014). Challenges and solutions for marketing in a digital era. *European Management Journal*, *32*(1), 1–12. https://doi.org/10.1016/j.emj.2013.12.001

Li, J., Zhao, G., Chen, X., Xie, D., Rong, C., Li, W., Tang, L., & Tang, Y. (2010). Fine-Grained Data Access Control Systems with User Accountability in Cloud Computing. *2010 IEEE 2nd International Conference on Cloud Computing Technology and Science*, 89–96. https://doi.org/10.1109/cloudcom.2010.44

Li, Y., Gai, K., Qiu, L., Qiu, M., & Zhao, H. (2017). Intelligent cryptography approach for secure distributed big data storage in cloud computing. *Information Sciences*, *387*, 103–115. https://doi.org/10.1016/j.ins.2016.09.005

Longstaff, J., & Noble, J. (2016). Attribute Based Access Control for Big Data Applications by Query Modification. *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, 58–65. https://doi.org/10.1109/bigdataservice.2016.35

Lv, Z., & Qiao, L. (2020). Analysis of healthcare big data. *Future Generation Computer Systems*, *109*, 103–110. https://doi.org/10.1016/j.future.2020.03.039

Machanavajjhala, A., & Reiter, J. P. (2012). Big privacy. *XRDS: Crossroads, The ACM Magazine for Students*, *19*(1), 20–23. https://doi.org/10.1145/2331042.2331051

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munoz-Gama, J. (2018). Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. *Computers in Human Behavior*, *80*, 179–196. https://doi.org/10.1016/j.chb.2017.11.011

Mall, S., & Saroj, S. K. (2018). A New Security Framework for Cloud Data. *Procedia Computer Science*, *143*, 765–775. https://doi.org/10.1016/j.procs.2018.10.397

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition and productivity*. McKinsey Global Institute. http://dln.jaipuria.ac.in:8080/jspui/bitstream/123456789/14265/1/mgi_big_data_full_report.pdf

Matturdi, B., Zhou, X., Li, S., & Lin, F. (2014). Big Data security and privacy: A review. *China Communications*, *11*(14), 135–145. https://doi.org/10.1109/cc.2014.7085614

Mengke, Y., Xiaoguang, Z., Jianqiu, Z., & Jianjian, X. (2016). Challenges and solutions of information security issues in the age of big data. *China Communications*, *13*(3), 193–202. https://doi.org/10.1109/cc.2016.7445514

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A Big Data Analytics Method for Tourist Behaviour Analysis. *Information & Management*, *54*(6), 771–785. https://doi.org/10.1016/j.im.2016.11.011

Mollah, M. B., Azad, Md. A. K., & Vasilakos, A. (2017). Secure Data Sharing and Searching at the Edge of Cloud-Assisted Internet of Things. *IEEE Cloud Computing*, *4*(1), 34–42. https://doi.org/10.1109/mcc.2017.9

Nabeel, M., & Bertino, E. (2014). Privacy Preserving Delegated Access Control in Public Clouds. *IEEE Transactions on Knowledge and Data Engineering*, *26*(9), 2268–2280. https://doi.org/10.1109/tkde.2013.68

Narayanan, U., Paul, V., & Joseph, S. (2017a). Different analytical techniques for big data analysis: A review. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 372–382. https://doi.org/10.1109/icecds.2017.8390139

Narayanan, U., Unnikrishnan, A., Paul, V., & Joseph, S. (2017b). A survey on various supervised classification algorithms. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2118–2124. https://doi.org/10.1109/icecds.2017.8389824

Narayanan, U., Paul, V., & Joseph, S. (2020). A light weight encryption over big data in information stockpiling on cloud. *Indonesian Journal of Electrical Engineering and Computer Science*, *17*(1), 389–397. https://doi.org/10.11591/ijeecs.v17.i1.pp389-397

Narayanan, U., Paul, V., & Joseph, S. (2022a). 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*, *34*(6), 3121–3135. https://doi.org/10.1016/j.jksuci.2020.05.005

Narayanan, U., Paul, V., & Joseph, S. (2022b). Decentralized blockchain based authentication for secure data sharing in Cloud-IoT. *Journal of Ambient Intelligence and Humanized Computing*, *13*(2), 769–787. https://doi.org/10.1007/s12652-021-02929-z

Neela, K. L., & Kavitha, V. (2018). Enhancement of data confidentiality and secure data transaction in cloud storage environment. *Cluster Computing*, *21*(1), 115–124. https://doi.org/10.1007/s10586-017-0959-4

Oh, C. G. (2017). Application of Big Data Systems to Aviation and Aerospace Fields; Pertinent Human Factors Considerations. *19th International Symposium on Aviation Psychology*, 214. https://corescholar.libraries.wright.edu/isap_2017/32/

Olanrewaju, R. F., Ul Islam Khan, B., Naaz Mir, R., Mehraj Baba, A., & Anwar, F. (2016). DFAM: A distributed feedback analysis mechanism for knowledge based educational big data. *Jurnal JurnalTeknologi (Sciences & Engineering)*, *78*(12–3), 31–38. https://doi.org/10.11113/jt.v78.10020

Pramanik, Md. I., Lau, R. Y. K., Azad, Md. A. K., Hossain, Md. S., Chowdhury, Md. K. H., & Karmaker, B. K. (2020). Healthcare informatics and analytics in big data. *Expert Systems with Applications*, *152*, 113388. https://doi.org/10.1016/j.eswa.2020.113388

Prasetyo, K. N., Purwanto, Y., & Darlis, D. (2014). An implementation of data encryption for Internet of Things using blowfish algorithm on FPGA. *2014 2nd International Conference on Information and Communication Technology (ICoICT)*, 75–79. https://doi.org/10.1109/icoict.2014.6914043

Puthal, D., Nepal, S., Ranjan, R., & Chen, J. (2016). A Secure Big Data Stream Analytics Framework for Disaster Management on the Cloud. *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 1218–1225. https://doi.org/10.1109/hpcc-smartcity-dss.2016.0170

Reddy, Y. (2018). Big Data Processing and Access Controls in Cloud Environment. *2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS)*, 25–33. https://doi.org/10.1109/bds/hpsc/ids18.2018.00019

Ruj, S., Stojmenovic, M., & Nayak, A. (2014). Decentralized Access Control with Anonymous Authentication of Data Stored in Clouds. *IEEE Transactions on Parallel and Distributed Systems*, *25*(2), 384–394. https://doi.org/10.1109/tpds.2013.38

Santos, N., & Masala, G. L. (2019). Big Data Security on Cloud Servers Using Data Fragmentation Technique and NoSQL Database. In G. De Pietro, L. Gallo, R. J. Howlett, L. C. Jain, & L. Vlacic (Eds.), *Intelligent Interactive Multimedia Systems and Services: Proceedings of 2018 Conference* (1st Ed., Vol. *11*, pp. 5–13). Springer International Publishing. https://doi.org/10.1007/978-3-319-92231-7_1

Schmidt, B., & Hildebrandt, A. (2017). Next-generation sequencing: Big data meets high performance computing. *Drug Discovery Today*, *22*(4), 712–717. https://doi.org/10.1016/j.drudis.2017.01.014

Sedkaoui, S., & Khelfaoui, M. (2019). Understand, develop and enhance the learning process with big data. *Information Discovery and Delivery*, *47*(1), 2–16. https://doi.org/10.1108/idd-09-2018-0043

Shen, J., Liu, D., Liu, Q., Sun, X., & Zhang, Y. (2017). Secure Authentication in Cloud Big Data with Hierarchical Attribute Authorization Structure. *IEEE Transactions on Big Data*, *7*(4), 668–1. https://doi.org/10.1109/tbdata.2017.2705048

Shen, J., Liu, D., Liu, Q., Wang, B., & Fu, Z. (2016). An authorized identity authentication-based data access control scheme in cloud. *2016 18th International Conference on Advanced Communication Technology (ICACT)*, 56–60. https://doi.org/10.1109/icact.2016.7423271

Silahtaroglu, G., & Donertasli, H. (2015). Analysis and prediction of E-customers' behavior by mining clickstream data. *2015 IEEE International Conference on Big Data (Big Data)*, 1466–1472. https://doi.org/10.1109/bigdata.2015.7363908

Srivastava, A., Singh, S. K., Tanwar, S., & Tyagi, S. (2017). Suitability of big data analytics in Indian banking sector to increase revenue and profitability. *2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA) (Fall)*, 1–6. https://doi.org/10.1109/icaccaf.2017.8344732

Stergiou, C., & Psannis, K. E. (2017a). Efficient and secure BIG data delivery in Cloud Computing. *Multimedia Tools and Applications*, *76*(21), 22803–22822. https://doi.org/10.1007/s11042-017-4590-4

Stergiou, C., & Psannis, K. E. (2017b). Recent advances delivered by Mobile Cloud Computing and Internet of Things for Big Data applications: A survey. *International Journal of Network Management*, *27*(3), e1930. https://doi.org/10.1002/nem.1930

Sun, J., & Reddy, C. K. (2013). Big data analytics for healthcare. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1525. https://doi.org/10.1145/2487575.2506178

Suri, M., & Singh, S. N. (2018). The Role of Big Data in the Media andEntertainment Industry. *2018 4th International Conference on Computational Intelligence & Communication Technology (CICT)*, 1–5. https://doi.org/10.1109/ciact.2018.8480281

Tsou, M.-H. (2015). Research challenges and opportunities in mapping social media and Big. *Cartography and Geographic Information Science*, *42*(1), 70–74. https://doi.org/10.1080/15230406.2015.1059251

Ulusoy, H., Colombo, P., Ferrari, E., Kantarcioglu, M., & Pattuk, E. (2015). GuardMR: Fine-grained Security Policy Enforcement for MapReduce Systems. *ASIA CCS '15: Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, 285–296. https://doi.org/10.1145/2714576.2714624

Ulusoy, H., Kantarcioglu, M., Pattuk, E., & Hamlen, K. (2014). Vigiles: Fine-Grained Access Control for MapReduce Systems. *2014 IEEE International Congress on Big Data*, 40–47. https://doi.org/10.1109/bigdata.congress.2014.16

Unnikrishnan, A., Narayanan, U., & Joseph, S. (2017). Performance analysis of various supervised algorithms on big data. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2293–2298. https://doi.org/10.1109/icecds.2017.8389861

Vorugunti, C. S. (2016). PPMUAS: A privacy preserving mobile user authentication system for cloud environment utilizing big data features. *2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, 1–6. https://doi.org/10.1109/ants.2016.7947803

Wagstaff, K. (2012). Machine learning that matters. *ArXiv Preprint*, 529–536. https://doi.org/10.48550/arXiv.1206.4656

Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, *34*(2), 77–84. https://doi.org/10.1111/jbl.12010

Warren, J., & Marz, N. (2015). *Big Data: Principles and best practices of scalable realtime data systems.Simon and Schuster* (1st Ed.). Simon and Schuster.

Wei, J., Hu, X., Liu, W., & Zhang, Q. (2019). Forward and backward secure fuzzy encryption for data sharing in cloud computing. *Soft Computing*, *23*(2), 497–506. https://doi.org/10.1007/s00500-017-2834-x

W. E. F. (2016). *Digital Transformation of Industries: Media, Entertainment and Information*. Word Economic Forum. https://www.weforum.org/publications/digital-transformation-of-industries/

West, D. M. (2012). Big data for education: Data mining, data analytics and web dashboards. *Governance Studies at Brookings*, *4*(1), 1–10. https://www.insidepolitics.org/brookingsreports/education%20big%20data.pdf

Win, T. Y., Tianfield, H., & Mair, Q. (2018). Big Data Based Security Analytics for Protecting Virtualized Infrastructures in Cloud Computing. *IEEE Transactions on Big Data*, *4*(1), 11–25. https://doi.org/10.1109/tbdata.2017.2715335

Wu, J., Ota, K., Dong, M., Li, J., & Wang, H. (2018). Big Data Analysis-Based Security Situational Awareness for Smart Grid. *IEEE Transactions on Big Data*, *4*(3), 408–417. https://doi.org/10.1109/tbdata.2016.2616146

Xiaolei Qian, & Lunt, T. F. (1996). A MAC policy framework for multilevel relational databases. *IEEE Transactions on Knowledge and Data Engineering*, *8*(1), 3–15. https://doi.org/10.1109/69.485625

Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information Security in Big Data: Privacy and Data Mining. *IEEE Access*, *2*, 1149–1176. https://doi.org/10.1109/access.2014.2362522

Xiong, H., Choo, K.-K. R., & Vasilakos, A. V. (2022). Revocable Identity-Based Access Control for Big Data with Verifiable Outsourced Computing. *IEEE Transactions on Big Data*, *8*(1), 1–13. https://doi.org/10.1109/tbdata.2017.2697448

Yadranjiaghdam, B., Yasrobi, S., & Tabrizi, N. (2017). Developing a Real-Time Data Analytics Framework for Twitter Streaming Data. *2017 IEEE International Congress on Big Data (BigData Congress)*, 329–336. https://doi.org/10.1109/bigdatacongress.2017.49

Yang, K., Jia, X., & Ren, K. (2015). Secure and Verifiable Policy Update Outsourcing for Big Data Access Control in the Cloud. *IEEE Transactions on Parallel and Distributed Systems*, *26*(12), 3461–3470. https://doi.org/10.1109/tpds.2014.2380373

Yu, S., Wang, C., Ren, K., & Lou, W. (2010). Attribute based data sharing with attribute revocation. *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security*, 261–270. https://doi.org/10.1145/1755688.1755720

Zhan, J., Fan, X., Cai, L., Gao, Y., & Zhuang, J. (2018). TPTVer: A trusted third party based trusted verifier for multi-layered outsourced big data system in cloud environment. *China Communications*, *15*(2), 122–137. https://doi.org/10.1109/cc.2018.8300277

Zhang, N., Ryan, M., & Guelev, D. P. (2005). Evaluating Access Control Policies Through Model Checking. *Information Security: 8th International Conference, ISC 2005, Singapore, September 20-23, 2005. Proceedings 8*, 446–460. https://doi.org/10.1007/11556992_32

Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data* (1st Ed.). McGraw-Hill Osborne Media. ISBN-10: 0071790535. https://doi.org/10.1142/s021819401100513x

Zhao, Y., Li, S., & Jiang, L. (2018). Secure and efficient user authentication scheme based on password and smart card for multiserver environment. *Security and Communication Networks*, *2018*(1), 9178941. https://doi.org/10.1155/2018/9178941