Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Developing a framework for securing blockchain-driven systems with large amounts of data

Mahmoud Ahmad Al-Khasawneh [1, 2], Marwan Mahmoud [3, *]

[1]*Department of Computer Science, College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia*
[2]*Applied Science Research Center, Applied Science Private University, Amman, Jordan*
[3]*The Applied College, King Abdulaziz University, Jeddah, Saudi Arabia*

ARTICLE INFO

Article history:
Received 24 March 2024
Received in revised form
19 January 2025
Accepted 12 February 2025

Keywords:
Big data
Blockchain technology
Design science research
Integrity
Privacy

ABSTRACT

The rapid expansion of big data has boosted advancements in fields such as healthcare, finance, and marketing. However, handling and storing large amounts of sensitive data have raised significant concerns due to security and privacy risks. Research suggests that blockchain technology could help address these challenges to some extent. This study aims to create a framework for securing big-data systems powered by blockchain, using the design science method. The framework includes seven key components: authentication and access control, data encryption and key management, privacy and confidentiality, data integrity and authenticity, data provenance and audit trails, intrusion detection and prevention, and incident response and recovery. This framework allows organizations to harness the potential of big data without risking data integrity or privacy. The findings indicate that this framework offers comprehensive guidelines for safely using big data across different sectors. Combining blockchain and big data can safeguard sensitive information, ushering in a new era of secure, data-driven innovation and trust.

1. Introduction

"Big data" has become a term often used in the industry to describe the generation of large amounts of data that must be analyzed to be effectively used (Rawat et al., 2019). A tremendous amount of data has been generated over the last few years at an alarming rate, with the amount of data increasing every day. This is because the Internet has grown rapidly over the last few years. According to Laney (2001), the three elements associated with big data are volume, velocity, and variety. The volume of data being generated, the velocity at which the data is being generated, and the variety at which the data is being generated, all represent enormous amounts of data being generated. The authors in Miloslavskaya (2017) stated that the easiest way to summarize Big Data is as a type of data at rest. Furthermore, they explained big data, data lakes, and fast data are three different concepts. A data lake is a huge storage area for raw data in its original format, which contains a large amount of raw data. On the other hand, fast data is time-sensitive, which may be structured or unstructured, and which is usually acted upon immediately after being received. Therefore, it is important to secure big data from the perspective of privacy and integrity.

Over the past few years, blockchain has been one of the most interesting topics of discussion across a wide range of industries. It was first introduced in 2000 when the concept of blockchain technology was first developed. In this technology, every new record is created within a block, with a link to the previous record, which, in other words, creates a linking block of data (Rittinghouse and Ransome, 2017). Blockchain is a technology that enables the use of cryptocurrencies, and Bitcoin is the name of one of the most popular cryptocurrencies worldwide. Cryptocurrency can be seen as a digital currency as well as a component of virtual currencies (Jaiswal, 2020). With the aid of cryptography, cryptocurrency acts as a medium of exchange, which helps final transactions be secured. Cryptocurrencies not only have the advantage of decentralized control but also rely on the principle of distributed ledgers, which is used for establishing a public database of financial transactions. A decentralized cryptocurrency, such as Bitcoin, is

* Corresponding Author.
Email Address: mmamahmoud@kau.edu.sa (M. Mahmoud)
https://doi.org/10.21833/ijaas.2025.03.002
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-0787-8225

regarded as the first type of cryptocurrency that can be classified as decentralized. As a result of the technology behind Bitcoin, the digital token is formed by Bitcoin, and the blockchain is used to keep track of all the users who possess the digital token, which is the ledger that keeps track of who has what digital token (Tumasjan, 2024). Hyperledger is a modular framework for building blockchain applications that can be coupled with Bitcoin and other cryptocurrencies, despite the fact that it does not support Bitcoin or other cryptocurrencies (Badr, et al., 2018). Regarding the features, Hyperledger stands out mainly for its ability to allow a smart contract to be developed. Smart contracts are composed of lines of codes stored on a blockchain, and when certain conditions and terms are met, they are executed according to the codes. Compared to cloud servers, which store data in a single location on a computer, blockchains break everything down into small blocks and spread them throughout the entire computer network without requiring any third parties to be involved (Zhang et al., 2021). As a result, the way users approach big data management is also changing as a result of both privacy and data security being ensured, as well as data quality being improved. In big data, some blockchain use cases range from ensuring that a transaction has been vetted to ensuring that it is trustworthy as it has been vetted carefully.

Therefore, the aim of this study is to develop a comprehensive framework for securing big-data blockchain-driven systems using design science research, with a focus on enhancing data integrity and privacy. The developed framework consists of seven main components: Authentication and access control, data encryption and key management, privacy and confidentiality, data integrity and authenticity, data provenance and audit trails, intrusion detection and prevention, and incident response and recovery. By implementing this framework, organizations can confidently leverage the power of big data without compromising data integrity or privacy. The combination of blockchain and big data technologies offers a robust and scalable solution to the problem of sensitive data security, paving the way for a new era of data-driven innovation and trust.

The rest of the article is arranged as follows: Related work is presented in Section 2, then the methodology in Section 3. The results and discussion of the study are given in Section 4. Finally, the conclusion and recommendations for future work are presented in Section 5.

2. Related works

The literature consists of several approaches to securing big data for different sectors with different perspectives. For example, Gottwalt and Karduck (2015) developed a security model for enhancing security information management based on the principles of big data analysis. The developed model is able to detect and mitigate threats and improve incident response collaboration and communication. Nonetheless, to manage and process big data, the model requires a lot of sensitive information.

Dong et al. (2015) suggested a framework for securely sharing sensitive data on big data platforms. It provides effective protection for the sensitive data of users and shares the data in a safe and secure manner. However, farmwork has a limitation where shared information can be accessed by unauthorized people or accidentally disclosed.

Soceanu et al. (2015) focused on the eHealth data sector and provided a framework for securely sharing sensitive data using encryption and attribute-based access control. The framework supported secure hierarchical healthcare data access with cryptographic tools. However, it focused on electronic health records only.

Las-Casas et al. (2016) proposed a big data framework architecture for cybersecurity applications. It focused on phishing characterization using Big Data Processes, ensuring continuous data analysis. It may however be necessary to have specialized expertise in data science and statistics to be able to extract meaningful insights from large volumes of security data.

Zhong et al. (2016) developed an automated cybersecurity data triage system to triage cybersecurity analyst traces of operations automatically. The developed system automatically generates triage automata from operation traces and reduces generation costs of data triage automatons orders of magnitude compared to existing methods. However, it lacks implementation in the real world.

Chen et al. (2016) focused on the big data infrastructure. They developed a cybersecurity system that prevents vulnerabilities from occurring. It is possible to obtain the data from public sources, transform (clean) and load it, cluster, visually represent it, and curate it. However, it has been found that their developed system lacks the ability to be implemented and is limited to ad hoc applications.

Yang and Zhang (2016) proposed a meta-model of a family of machine-learning algorithms for detecting network anomalies. A major focus of the study was on analyzing the network traffic data to detect anomalies in the network. Several features of the proposed metamodel, such as its scalability, flexibility, model selection capabilities, and model integration capabilities, make it an excellent method for gaining valuable insights from large datasets. However, owing to the complexity and scalability of the proposed metamodel, its flexibility and extensibility are limited.

Xiao et al. (2016) offered an encryption control model with multiple levels of intelligence to investigate the confidentiality of multimedia big data. This model can be used to optimize the resource distribution at the system level in multiple streams from multiple sources. Further, it has been demonstrated that the proposed scheme can be adapted to real-time applications as a result of experiments conducted. Nevertheless, it is only

suitable for multimedia big data security, which has some limitations.

Zaki et al. (2017) developed a model demonstrating that big data can be used to generate phishing sites on Enron E-mail. They found that those who employ phishing and hacking techniques through the manipulation of big data can create significant security threats by analyzing email users' behaviors. The developed model, however, does not cover how to prevent the misuse of big data by hackers. Casas et al. (2017) proposed a big data analytics framework to detect the anomaly of network monitoring applications. The proposed framework can process both stream data and batch data from heterogeneous sources of unstructured and structured heterogeneous data. Nevertheless, it has a limited capability of detecting anomalies. Uchibeke et al. (2018) developed an ecosystem for blockchain access control to manage and protect large data sets against breaches. By using blockchain technology, data can be made transparent and traceable within a network, which eliminates centralized and traditional challenges of access control. The developed ecosystem, however, lacks the ability to be implemented in a real-world environment. Guan et al. (2018) proposed a trusted big data collection and trade system to assure fairness and trust across the big data world for all participants. This proposed system allows for the trading of private data and decreases the conflicts between the invasion of privacy and the lack of it, which have arisen in the past. In spite of this, it hasn't been implemented yet. In Keshk et al. (2019), the researchers developed a framework to protect data in smart power networks based on blockchain technology and deep learning. It was found that the framework outperformed the rivals regarding performance quality and that by manipulating the original datasets, it prevented both data poisoning and inference attacks at the same time. However, its scalability and utility have not been evaluated yet. In Makhdoom et al. (2020), a blockchain-based approach was proposed in order to prepare smart cities to share sensitive data and personal information with one another to avoid harming the most sensitive aspects of the data. The blockchain-based solution guarantees the integrity and privacy of user data in a smart city and enables the "right to forget" for users relating to their personal information through the use of a blockchain-based solution. The model has some limitations, for example, it requires the committing peers to maintain a large number of ledgers, which can lead to massive resource demands if there is an extensive network of smart cities. Younis et al. (2021) proposed a new blockchain-based telehealth architecture system that provides secure and privacy-preserving access to telehealth data based on blockchain technology. As a result of the system's overall design, telehealth attacks such as impersonation, message replay, and data forgery, which are common in the healthcare sector, can be mitigated. However, one of the limitations of this work is that it does not allow the storage and retrieval of data. Kumar et al. (2021) proposed a framework for integrating blockchain and deep-learning techniques to enhance privacy and security in C-ITS infrastructure. Using the proposed framework, C-ITS participants can securely communicate data among themselves and prevent original datasets from being poisoned. The framework, however, is not capable of evaluating utility and scalability in a real-world C-ITS environment. In Rahman et al. (2022), a hierarchical blockchain-based system was proposed to protect IoT data by integrating blockchains in smart cities. The proposed model addressed a number of data management challenges, ensuring data integrity and providing interoperability in smart cities to meet the needs of users. However, none of the smart city interoperability solutions discussed in that study allow cross-service communication. Juma et al. (2023) proposed a trusted consortium Blockchain framework, which could build a modular Hyperledger fabric that could provide a secured and verifiable solution for the integrity of big data in the ecosystem. This proposed framework is combined with consensus protocols, which are intended to validate the signing of evidence of big data recorded over the course of real-time HFM operations. Currently, there is no implementation of the idea in the real world. In Liu et al. (2024), an industrial big data analytics framework was suggested, which considers privacy and security protection to provide references for big data analytics that maintain privacy and security in the industrial sector. However, the proposed framework lacks implementation whereas it is the first framework proposed for preserving privacy and securing industrial big data analytics. Table 1 displays the existing studies that have focused on the security of big data using blockchain technology. The present paper analyzed the existing studies on different security measures such as authentication and access control, data encryption and key management, privacy and confidentiality, data integrity and authenticity, intrusion\detection and prevention, and blockchain techniques. The analysis of the collected data (Table 1 and Fig. 1) revealed that to secure big data, the authors have focused on the encryption and key management of data, the authentication and access control of data, and privacy and confidentiality. In addition, the security features related to the integrity and authenticity of data and intrusion detection and prevention have received the same amount of attention. Blockchain has been employed recently by the authors to secure large amounts of collected data.

3. Methodology

The present research used design science methodology (Peffers et al., 2007; Alotaibi et al., 2023) to develop a framework for securing big-data blockchain-driven systems. The design science research methodologies are considered part of the

design science paradigm and are closely related to it (Peffers et al., 2007; Alotaibi et al., 2023). As a result, design science provides a series of specific guidelines for evaluating and iterating research projects in a structured manner, which includes the methodologies of various disciplines of research, such as mathematics, computer science, and information technology. Therefore, design science focuses on an approach to research that incorporates the methods of various disciplines of research (Peffers et al., 2007; Alotaibi et al., 2023). The adopted methodology consisted of 5 phases:

- Phase 1: Recognizing and identifying the popular online databases: This phase involved the identification of six online databases that are widely known for their content: IEEE Xplore, Scopus, Springer Link, Web of Science, Science Direct, and Google Scholar.
- Phase 2: Assigning search protocols: This phase defined the rules/protocols governing the searching behavior in the identified online databases, such as keywords, the time period, and language. Several keywords were used in this study, including "Big Data," "Blockchain," and "Securing Big Data." And the period of search was set to range from 2015 to 2024. The English language was selected as the search language, and articles written in other languages were excluded.
- Phase 3: Collecting data from identified online databases: The required data was collected from the six databases based on the rules set in Phase 2. The results of searching are displayed in Fig. 2. In total, 613 articles were collected, among which 11 articles were retrieved from Scopus, 400 from Google Scholar, 27 from IEEE Xplore, 34 from Science Direct, 139 from Springer Link, and finally two articles from Web of Science.
- Phase 4: Analyzing the collected data: The 613 articles collected from the databases were analyzed based on the following including and excluding criteria: Repeated articles, books, book chapters, irrelevant articles, and articles without results were excluded from the study. This study included journals and conference articles that had focused on the domain, been reviewed in detail, and had clear results, advantages, disadvantages, output, methodology, and conclusions. Table 2 displays the summary of the articles that have concentrated on securing big data using blockchain technology. Table 2 indicates that the security of big data with the help of blockchain technology is diverse and lacks a comprehensive framework that is able to address all the issues of big data from a security perspective in a systematic manner.
- Phase 5: Developing the framework: In this phase, the framework for securing big-data blockchain-driven systems was developed. It consists of seven main components (Fig. 3): Authentication and access control, data encryption and key management, privacy and confidentiality, data integrity and authenticity, data provenance and audit trails, intrusion detection and prevention, and incident response and recovery. In the following, the components are introduced in detail.

1. Authentication and access control: There are two important features in any security system: authentication and access control. These features ensure that any sensitive data can be accessed and modified at the hands of only authorized individuals or entities in the context of blockchain-driven systems. The framework provides a variety of authentication methods for authenticating individuals, such as usernames and passwords, biometrics, and multi-factor authentication, so that a robust and secure service can be provided.
2. Data encryption and key management: The encryption of sensitive data is clearly one of the most important aspects of protecting sensitive data against being accessed by unauthorized individuals. The framework encrypts both data in transit and data at rest using powerful encryption algorithms. There are a number of other critical components that are included in this framework in addition to key management. Taking reasonable steps to ensure that the keys are managed appropriately ensures that those who require access to the private keys in order to decrypt the data will only have access to those keys. Key management systems, such as key escrow or key federation, are used in order to protect keys and prevent them from being disclosed to unauthorized parties without their permission.
3. Privacy and confidentiality: A sensitive data collection process involves handling sensitive data that requires a high level of privacy and confidentiality. This framework provides mechanisms to ensure that user data is kept private and confidential, thereby ensuring the security of the framework. To minimize the risk of individual users being identified, data anonymization techniques, such as data masking and pseudonymization, are applied. The SSL protocol provides a mechanism for maintaining the privacy of communication between clients and servers by encrypting it, while at the same time preventing an eavesdropper from gaining access to the communication. Other cryptographic algorithms, such as TLS or SSL, are used to provide further security.
4. Data integrity and authenticity: It is fundamentally important that both the integrity and authenticity of data are assured to ensure the accuracy and trustworthiness of the data. There are many aspects of the framework that are key to its effectiveness; among them is its ability to rely on mechanisms to verify the integrity of data, as well as to prevent it from being altered or tampered with. The technology behind blockchains is naturally guaranteed to provide data integrity because of its distributed nature, as well as the consensus mechanisms that are inherent to its technology. Using a consensus algorithm, which ensures that all transactions and modifications to the data are recorded and verified at all times, the

framework brings transparency and accountability to users.

5. Data provenance and audit trails: For blockchain-driven systems to maintain their integrity and accountability, provenance and audit trails are vital components. A comprehensive audit trail can be generated using the framework by incorporating mechanisms that trace the origins and histories of data records throughout the entire process. By maintaining an audit trail of every transaction, you can verify that data is not being tampered with or modified by authorized parties so that investigations and audits can be conducted. In addition to facilitating regulatory compliance and forensic investigations, provenance information can be used to identify and track data lineages, helping to identify and track data chains.
6. Intrusion detection and prevention: The intrusion detection process plays a major role in preventing unauthorized access to the network and systems as part of the intrusion detection process. Various intrusion detection and prevention mechanisms are included in this framework to minimize the risk of potential threats being detected and mitigated. A host-based intrusion detection system, a network-based intrusion detection system, a firewall, and a web application firewall are all examples of similar mechanisms you may find in this mechanism. These mechanisms are capable of detecting and blocking malicious activities or suspicious behaviors based on the monitoring and analysis of the network traffic they receive and the activities that are taking place in the network.
7. Incident response and recovery: Security frameworks that can be considered effective must be capable of responding to security incidents and resolving them. The developed framework contains several components, such as procedures, tools, and mechanisms that enable the organization to respond quickly and efficiently to security incidents and breaches of data. There are several measures that can be taken to prevent an incident from taking place, including reporting the incident, investigating the incident, containing it, and recovering from it. Having an incident response plan that has been well thought out is essential if an organization is to minimize the impact and the recovery time of security incidents and the impact of security breaches in the future.

4. Results and discussions

This section discusses and analyzes the results of the study and compares them with the existing ones. This study used six popular online databases to collect the relevant articles based on systematic criteria. A total of 613 articles were collected, among which 21 articles were selected for further analysis based on their coverage of the topic of using blockchain to secure big data, as well as the clarity of their studies and results. The developed framework consists of seven key components that have not been covered by the existing models and frameworks (Table 3). In comparison with existing frameworks, the developed framework is more comprehensive and introduces three new security features.

For example, the authors of Gottwalt and Karduck (2015) covered four components of the developed framework, such as data encryption and key management, privacy and confidentiality, intrusion detection and prevention, and incident response and recovery, while Dong et al. (2015) addressed only two components namely authentication and access control, and intrusion detection and prevention. In their paper Las-Casas et al. (2016), the authors covered only one component, intrusion detection and prevention. As discussed by the authors in Zhong et al. (2016), three main components of data security were addressed: Authentication and access control, data encryption and key management, and data privacy. Authentication and access control, data encryption and key management, as well as privacy and confidentiality, were covered by the authors of Chen et al. (2016). It should also be noted that the authors of Yang and Zhang (2016) only mentioned one component, which is the detection and prevention of intrusions. According to the researchers of Liu et al. (2016), the researchers covered two components: Authentication and access control, data encryption, and key management. There are four components addressed by Naik et al. (2016) the first of which is authentication and access control, the second of which is data encryption and key management, the third of which is privacy and confidentiality, and the fourth of which is intrusion detection and prevention. Based on the findings of this study, it has been determined that most of the existing studies have focused on the privacy and confidence component, the encryption component, and the key management component, respectively, but there are few studies that are dedicated to data provenance and audit trails, as well as incident response recovery as shown in Fig. 4. Therefore, the security components addressed by different authors in their developed models vary. While some models cover a broader scope, others may focus on specific components. The developed framework for securing big-data blockchain-driven systems, on the other hand, provides a comprehensive approach by incorporating authentication and access control, data encryption and key management, privacy/confidentiality, and intrusion detection and prevention. This approach ensures a holistic and well-rounded security solution for big-data blockchain-driven systems, safeguarding them from potential threats and risks.

5. Limitations and open directions

The field of securing big-data blockchain-driven systems has gained significant attention in recent years due to its potential to revolutionize various domains such as finance, healthcare, and supply chain management.

Table 1: Summary of existing studies

Reference	Focus	Purpose	Methodology	The security of big data features					
				Authentication and Access Control	Data encryption and key management	Privacy and confidentiality	Data integrity and authenticity	Intrusion detection and prevention	Blockchain technique
Gottwalt and Karduck (2015)	Security information management environments applied to big data principles	To integrate big data analytics into security information management	Design science research	√	√	√	×	√	×
Dong et al. (2015)	Big data Platform.	To provide a framework for securely sharing sensitive data	Design science research	√	√	×	×	√	×
Soceanu et al. (2015)	eHealth Data	To secure and protect clinical data	Encryption and attribute-based access control	×	√	√	×	×	×
Las-Casas et al. (2016)	Phishing characterization using big data	To mitigate cybersecurity risks	Case study	×	×	×	×	√	×
Zhong et al. (2016)	Automate cybersecurity data triage	To automatically triage cybersecurity analyst traces of operations	Case study	√	√	√	×	×	×
Chen et al. (2016)	Big data infrastructure	To prevent vulnerabilities	Extracting information from public data sources	√	√	√	×	√	×
Yang and Zhang (2016)	Network traffic data analyses	To detect the network anomalies	Metamodeling approach.	×	×	×	×	√	×
Liu et al. (2016)	RC4 security	To examine RC4 security	Transport layer security	√	√	×	×	×	×
Naik et al. (2016)	Windows desktop users	To explore how desktop users can identify malicious attacks in their untouched content and prevent them	Computational intelligence techniques, EmEditor, and R	√	√	√	×	√	×
Xiao et al. (2016)	Multimedia big data security	To investigate the confidentiality of multimedia big data	Sensing system with resource constraints	×	√	√	×	×	×
Zaki et al. (2017)	Enron E-mail	To demonstrate that big data can be used to generate phishing sites on Enron E-mail	A case study on the Enron email dataset	√	×	×	×	×	×
Casas et al. (2017)	Network traffic monitoring and analysis	To detect the anomaly of network monitoring applications	Off-the-shelf big data storage and processing engines	×	×	×	×	√	×
Uchibeke et al. (2018)	Access control management	To manage and protect large data sets against breaches	Blockchain technology	√	√	√	×	×	√
Guan et al. (2018)	Blockchain and TSM	To assure fairness and trust across the big data world for all participants	Physical Unclonable Function (PUF), and Trusted Security Module (TSM)	√	√	√	×	×	√
Keshk et al. (2019)	Smart power networks	To detect potential attacks on smart power network	Input data collection, privacy-preserving modules, and LSTM-based anomaly detection modules	√	√	√	√	×	√
Makhdoom et al. (2020)	Smart cities that preserve privacy and protect data	To protect sensitive and personal information in smart cities	Private-data collection	√	√	√	√	×	√
Younis et al. (2021)	Smart healthcare	To provide secure and privacy-preserving access to data	Design science approach	√	√	√	√	×	√
Kumar et al. (2021)	Intelligent transport system	To provide security and privacy in C-ITS	Blockchain-enabled deep-learning	√	√	√	√	×	√
Rahman et al. (2022)	IoT data integrity in smart cities	To protect IoT data and interoperating blockchains in smart cities.	Hyperledger fabric and Ethermint	√	√	√	√	×	√
Juma et al. (2023)	Industrial IoT	To create a modular Hyperledger fabric that provides a secured and verifiable solution to big data integrity	Design science research	√	√	√	√	×	√
Liu et al. (2024)	Industrial big data analytics	To provide references for big data analytics that maintain privacy and security in the industrial sector	Survey and DSR	×	×	√	√	×	×

Table 2: Summary of analyzed articles regarding the approach of securing big data using blockchain technology

Reference	Advantages	Disadvantages	Findings
Gottwalt and Karduck (2015)	Enhances threat detection and response, improves incident response, and enhances collaboration and communication.	Managing and processing big data analytics requires a lot of sensitive information.	A proposal was presented for enhancements to security information management based on the principles of Big Data Analysis.
Dong et al. (2015)	Provides effective protection for users' sensitive data and shares these data in a safe and secure manner.	Shared information can be accessed by unauthorized people or accidentally disclosed.	Secure delivery, storage, sharing, and destruction of sensitive data are among the features presented in this paper.
Soceanu et al. (2015)	Supports secure hierarchical healthcare data access with cryptographic tools.	Focuses on electronic health records only.	eHealth Data Privacy and Security Model was proposed.
Las-Casas et al. (2016)	Processes large amounts of diverse security data and ensures continuous data analysis.	To extract meaningful insights from large volumes of security data, a specialist expert may be required in data science and statistical analysis.	A big data framework architecture was presented for cybersecurity applications.
Zhong et al. (2016)	Automatically generates triage automata from operation traces and reduces generation costs of data triage automatons orders of magnitude compared to existing methods.	It lacks the implementation in the real world.	Researchers have demonstrated that patterns can be inferred by analyzing and mining data triage operations of analysts.
Chen et al. (2016)	Data can be found and retrieved from public sources, transformed (cleaned) and loaded, clustered and visualized, and curated.	It lacks the implementation and is limited for ad hoc fashion.	A Proactive Cybersecurity System was developed.
Yang and Zhang (2016)	This approach works robustly in extracting valuable insights from large datasets because of its scalability, flexibility, model selection capabilities, and model integration capabilities.	It is limited in flexibility and extensibility due to complexity and scalability.	A meta-model of a family of machine learning algorithms was developed for detecting network anomalies.
Liu et al. (2016)	It obtains the key distribution from a large set of keys.	Any secure protocol using RC4 should be deprecated as soon as possible because of its inability to provide large enough security margins.	A new method was presented for analyzing the security of RC4 based on big data processing.

Naik et al. (2016)	It can be successfully implemented in a modest desktop configuration.	Only simulated data based on a few protocols, firewall rules, and IP addresses is available on the platform.	A Computational Intelligence (CI)-based big data security analysis approach was suggested for Windows desktop users.
Xiao et al. (2016)	Models proposed in this study can optimize resource distribution at the system level in multiple streams. Moreover, experiments have proven that real-time applications can be accommodated by the presented schemes.	It is limited for Multimedia Big Data Security.	An encryption control model was proposed with multiple levels of intelligence.
Zaki et al. (2017)	As a result, big data creates great joy and new opportunities.	It does not cover preventing such big data security threats through e-mail.	The researchers found that phishers and hackers can create big data security threats by understanding the behaviors of email users through big data analysis.
Casas et al. (2017)	The Big-DAMA is capable of both stream and batch processing of unstructured and structured heterogeneous data sources.	It's limited for anomaly detection.	Big-DAMA, a Big Data Analytics Framework
Uchibeke et al. (2018)	Through blockchain technology, data is transparent and traceable within a network, which solves access control challenges, both traditional and centralized.	It lacks the implementation in the real word.	An ecosystem was developed for blockchain access control.
Guan et al. (2018)	This framework allows for trading of private data and reduces the conflicts between invading privacy and not having it.	It lacks implementation.	The Trusted Big Data Collection and Trade (TBDCT) system was developed.
Keshk et al. (2019)	This approach outperforms the previous ones, and by manipulating the original datasets, it prevents data poisoning and inference attacks as well.	It lacks the evaluation of scalability and utility.	A framework was developed to protect data in smart power networks based on blockchain technology and deep learning.
Makhdoom et al. (2020)	Secures user data integrity and privacy from most of the external and internal attacks in a smart city and provides users with the "right to forget" about their personal information through a blockchain-based solution.	This model has some limitations, most notably the fact that committing peers must maintain numerous ledgers, which can cause massive resource demands if there is an extensive network of smart cities.	"PrivySharing" was introduced as a blockchain-based way for smart cities to share data.
Younis et al. (2021)	As a result of the system's overall design, the most common attacks on telehealth systems are mitigated, such as impersonation, message replay, and data forgery.	Data storage and retrieval is the only limitation of this work.	A new blockchain-based telehealth architecture was presented.
Kumar et al. (2021)	As a result of the proposed framework, C-ITS participants are able to securely communicate data among themselves and to protect original datasets from data poisoning attacks.	It lacks the capability of evaluating the framework in a real-world C-ITS environment and applying different real-world datasets for gauging utility and scalability and evaluating framework scalability and utility with different real-world datasets.	A framework was suggested for integrating blockchain and deep-learning techniques to enhance privacy and security in C-ITS infrastructure.
Rahman et al. (2022)	Providing data management challenges, ensuring data integrity, and enabling interoperability in smart cities are addressed by the proposed model.	None of the smart city interoperability solutions discussed in the paper allows cross-service communication.	A hierarchical blockchain-based model was proposed.
Juma et al. (2023)	The proposed framework is combined with consensus protocols to validate the signing of evidence of big data recorded over real-time HFM operations.	There is no real-world implementation of it.	The Trusted consortium blockchain framework was developed.
Liu et al. (2024)	It assists in the construction of an industrial big data analytics platform that preserves privacy and is secure.	It lacks implementation whereas it is the first proposed work for privacy-preserving and secure industrial big data analytics.	An industrial big data analytics framework was proposed, which considers privacy and security protection.

Table 3: Comparison between the proposed framework and the existing ones in previous studies

Proposed framework's components	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Authentication and access control	×	√	×	×	√	√	×	√	√	×	×	√	×	×	√	√	√	×	×	√	×
Data encryption and key management	√	×	√	×	√	√	×	√	√	√	×	×	×	×	√	√	√	×	√	√	×
Privacy and confidentiality	√	×	√	×	√	√	×	×	√	√	×	×	×	×	√	√	√	×	√	√	√
Data integrity and authenticity	×	×	×	×	×	×	×	×	×	×	×	×	×	×	√	√	√	×	√	√	√
Data provenance and audit trails	×	×	×	×	×	×	×	×	×	×	×	×	√	×	×	×	×	×	×	×	×
Intrusion detection and prevention	√	√	×	√	×	√	√	×	√	×	×	√	×	×	×	×	×	×	√	×	×
Incident response and recovery	√	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

1: Gottwalt and Karduck (2015); 2: Dong et al. (2015); 3: Soceanu et al. (2015); 4: Las-Casas et al. (2016); 5: Zhong et al. (2016); 6: Chen et al. (2016); 7: Yang and Zhang (2016); 8: Liu et al. (2016); 9: Naik et al. (2016); 10: Xiao et al. (2016); 11: Zaki et al. (2017); 12: Casas et al. (2017); 13: Uchibeke et al. (2018); 14: Guan et al. (2018); 15: Keshk et al. (2019); 16: Makhdoom et al. (2020); 17: Younis et al. (2021); 18: Kumar et al. (2021); 19: Rahman et al. (2022); 20: Juma et al. (2023); 21: Liu et al. (2024)

Fig. 1: Analyzed existing studies

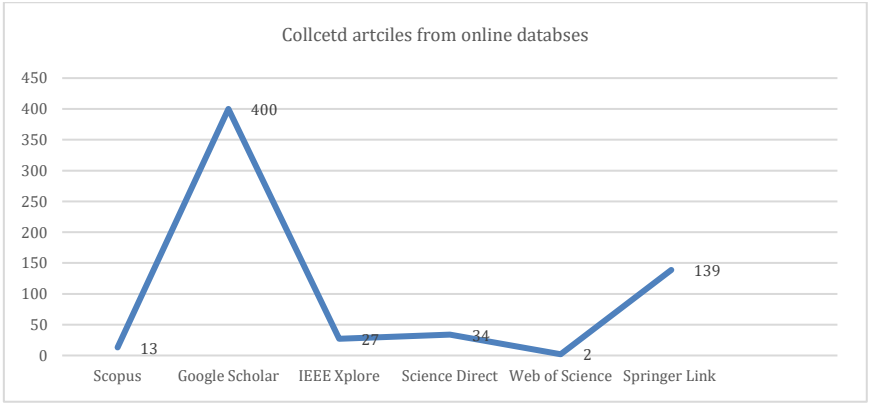

Fig. 2: Collected articles from the six online databases

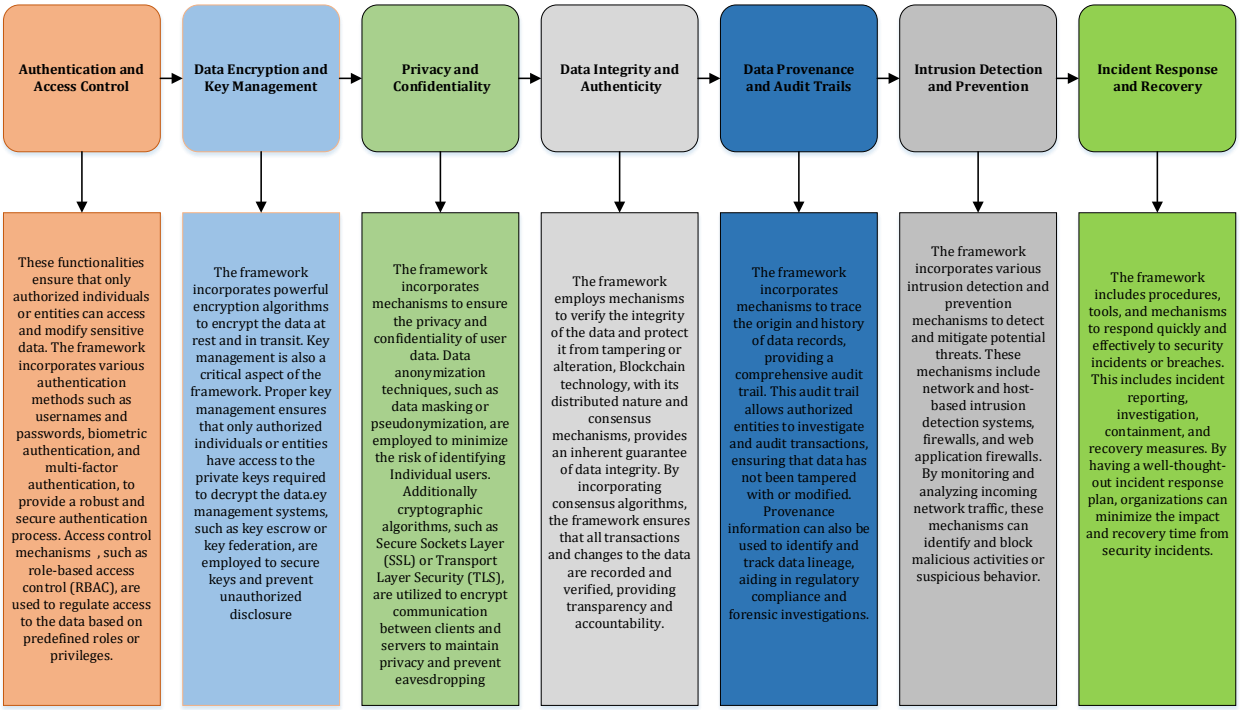

Fig. 3: Developed framework for securing big-data blockchain-driven systems

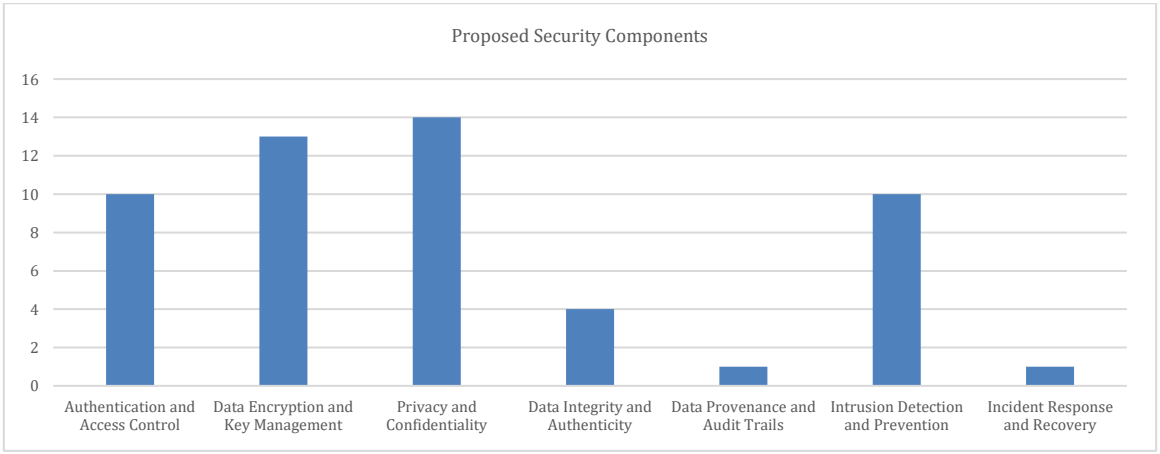

Fig. 4: The proposed security components against the existing security components

However, despite advancements in this area, there are still several limitations and open directions that need to be addressed. The Big-data blockchain-driven systems present a unique set of challenges and complexities, which can make it challenging to develop secure and efficient solutions. These systems often involve the integration of multiple components, including blockchain protocols, big-data analytics, and data storage mechanisms. This complexity can lead to vulnerabilities and hinder the effectiveness of security measures. By focusing on advanced security mechanisms, privacy-enhancing blockchain architectures, secure data sharing and access control, and privacy-by-design approaches,

we can pave the way for more secure and efficient big-data blockchain-driven systems.

6. Conclusions

This study developed a framework for securing big-data blockchain-driven systems using design science methodology to enhance the integrity and privacy of data. The developed framework contains seven key components: authentication and access control, data encryption and key management, privacy and confidentiality, integrity and authenticity of data, data provenance and audit trails, intrusion detection and prevention, and incident response and recovery. Using this framework, companies could confidently leverage the power of big data without compromising the integrity of the data or the privacy of their clients and without having to worry about the security of their data. The framework developed in this paper can help solve a wide range of security issues associated with big data in different sectors using both blockchain technology and big data knowledge. In a future study, the developed framework could be implemented in the real world to verify its effectiveness and capabilities.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Alotaibi FM, Al-Dhaqm A, Yafooz WM, and Al-Otaibi YD (2023). A novel administration model for managing and organising the heterogeneous information security policy field. Applied Sciences, 13(17): 9703. https://doi.org/10.3390/app13179703

Badr B, Horrocks R, and Wu XB (2018). Blockchain by example: A developer's guide to creating decentralized applications using Bitcoin, Ethereum, and Hyperledger. Packt Publishing Ltd., Birmingham, UK.

Casas P, Soro F, Vanerio J, Settanni G, and D'Alconzo A (2017). Network security and anomaly detection with Big-DAMA, a big data analytics framework. In the IEEE 6th International Conference on Cloud Networking, IEEE, Prague, Czech Republic: 1-7. https://doi.org/10.1109/CloudNet.2017.8071525

Chen HM, Kazman R, Monarch I, and Wang P (2016). Predicting and fixing vulnerabilities before they occur: A big data approach. In the 2nd International Workshop on BIG Data Software Engineering, Association for Computing Machinery, Austin, USA: 72-75. https://doi.org/10.1145/2896825.2896829

Dong X, Li R, He H, Zhou W, Xue Z, and Wu H (2015). Secure sensitive data sharing on a big data platform. Tsinghua Science and Technology, 20(1): 72-80. https://doi.org/10.1109/TST.2015.7040516

Gottwalt F and Karduck AP (2015). SIM in light of big data. In the 11th International Conference on Innovations in Information Technology, IEEE, Dubai, UAE: 326-331. https://doi.org/10.1109/INNOVATIONS.2015.7381562

Guan Z, Zhao Y, Li D, and Liu J (2018). TBDCT: A framework of trusted big data collection and trade system based on blockchain and TSM. In the IEEE International Conference on Smart Cloud, IEEE, New York, USA: 77-83. https://doi.org/10.1109/SmartCloud.2018.00021

Jaiswal M (2020). Cryptocurrency an era of digital currency. International Journal of Creative Research Thoughts, 8(1): 60-70.

Juma M, Alattar F, and Touqan B (2023). Securing big data integrity for industrial IoT in smart manufacturing based on the trusted consortium blockchain (TCB). IoT, 4(1): 27-55. https://doi.org/10.3390/iot4010002

Keshk M, Turnbull B, Moustafa N, Vatsalan D, and Choo KKR (2019). A privacy-preserving-framework-based blockchain and deep learning for protecting smart power networks. IEEE Transactions on Industrial Informatics, 16(8): 5110-5118. https://doi.org/10.1109/TII.2019.2957140

Kumar R, Kumar P, Tripathi R, Gupta GP, Kumar N, and Hassan MM (2021). A privacy-preserving-based secure framework using blockchain-enabled deep-learning in cooperative intelligent transport system. IEEE Transactions on Intelligent Transportation Systems, 23(9): 16492-16503. https://doi.org/10.1109/TITS.2021.3098636

Laney D (2001). 3D data management: Controlling data volume, velocity, and variety, Application delivery strategies. META Group Inc., Stamford, USA.

Las-Casas PH, Dias VS, Meira W, and Guedes D (2016). A big data architecture for security data and its application to phishing characterization. In the 2nd International Conference on Big Data Security on Cloud; IEEE International Conference on High Performance and Smart Computing; and IEEE International Conference on Intelligent Data and Security, IEEE, New York, USA: 36-41. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.44

Liu C, Cai Y, and Wang T (2016). Security evaluation of RC4 using big data analytics. In the 7th IEEE International Conference on Software Engineering and Service Science, IEEE, Beijing, China: 316-320. https://doi.org/10.1109/ICSESS.2016.7883075 **PMid:27468472**

Liu L, Li J, Lv J, Wang J, Zhao S, and Lu Q (2024). Privacy-preserving and secure industrial big data analytics: A survey and the research framework. IEEE Internet of Things Journal, 11(11): 18976-18999. https://doi.org/10.1109/JIOT.2024.3353727

Makhdoom I, Zhou I, Abolhasan M, Lipman J, and Ni W (2020). PrivySharing: A blockchain-based framework for privacy-preserving and secure data sharing in smart cities. Computers and Security, 88: 101653. https://doi.org/10.1016/j.cose.2019.101653

Miloslavskaya N (2017). Security intelligence centers for big data processing. In the 5th International Conference on Future Internet of Things and Cloud Workshops, IEEE, Prague, Czech Republic: 7-13. https://doi.org/10.1109/FiCloudW.2017.68

Naik N, Jenkins P, Savage N, and Katos V (2016). Big data security analysis approach using computational intelligence techniques in R for desktop users. In the IEEE Symposium Series on Computational Intelligence, IEEE, Athens, Greece: 1-8. https://doi.org/10.1109/SSCI.2016.7849907 **PMid:27036598**

Peffers K, Tuunanen T, Rothenberger MA, and Chatterjee S (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3): 45-77. https://doi.org/10.2753/MIS0742-1222240302

Rahman MS, Chamikara MAP, Khalil I, and Bouras A (2022). Blockchain-of-blockchains: An interoperable blockchain platform for ensuring IoT data integrity in smart city. Journal

of Industrial Information Integration, 30: 100408. https://doi.org/10.1016/j.jii.2022.100408

Rawat DB, Doku R, and Garuba M (2019). Cybersecurity in big data era: From securing big data to data-driven security. IEEE Transactions on Services Computing, 14(6): 2055-2072. https://doi.org/10.1109/TSC.2019.2907247

Rittinghouse JW and Ransome JF (2017). Cloud computing: Implementation, management, and security. CRC Press, Boca Raton, USA. https://doi.org/10.1201/9781439806814

Soceanu A, Vasylenko M, Egner A, and Muntean T (2015). Managing the privacy and security of ehealth data. In the 20th International Conference on control systems and computer science, IEEE, Bucharest, Romania: 439-446. https://doi.org/10.1109/CSCS.2015.76

Tumasjan A (2024). The promise and prospects of blockchain-based decentralized business models. In: Glückler J and Panitz R (Eds.), Knowledge and digital technology: 203-224. Springer Nature, Cham, Switzerland. https://doi.org/10.1007/978-3-031-39101-9_11

Uchibeke UU, Schneider KA, Kassani SH, and Deters R (2018). Blockchain access control ecosystem for big data security. In the IEEE International Conference on Internet of Things; IEEE Green Computing and Communications; IEEE Cyber, Physical and Social Computing; and IEEE Smart Data, IEEE, Halifax, Canada: 1373-1378. https://doi.org/10.1109/Cybermatics_2018.2018.00236

Xiao C, Wang L, Jie Z, and Chen T (2016). A multi-level intelligent selective encryption control model for multimedia big data security in sensing system with resource constraints. In the 3rd International Conference on Cyber Security and Cloud Computing, IEEE, Beijing, China: 148-153. https://doi.org/10.1109/CSCloud.2016.37

Yang B and Zhang T (2016). A scalable meta-model for big data security analyses. In the 2nd International Conference on Big Data Security on Cloud; IEEE International Conference on High Performance and Smart Computing; and IEEE International Conference on Intelligent Data and Security, IEEE, New York, USA: 55-60. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.71

Younis M, Lalouani W, Lasla N, Emokpae L, and Abdallah M (2021). Blockchain-enabled and data-driven smart healthcare solution for secure and privacy-preserving data access. IEEE Systems Journal, 16(3): 3746-3757. https://doi.org/10.1109/JSYST.2021.3092519

Zaki T, Uddin MS, Hasan MM, and Islam MN (2017). Security threats for big data: A study on Enron e-mail dataset. In the International Conference on Research and Innovation in Information Systems, IEEE, Langkawi, Malaysia: 1-6. https://doi.org/10.1109/ICRIIS.2017.8002481 **PMid:27868424**

Zhang C, Xu Y, Hu Y, Wu J, Ren J, and Zhang Y (2021). A blockchain-based multi-cloud storage data auditing scheme to locate faults. IEEE Transactions on Cloud Computing, 10(4): 2252-2263. https://doi.org/10.1109/TCC.2021.3057771

Zhong C, Yen J, Liu P, and Erbacher RF (2016). Automate cybersecurity data triage by leveraging human analysts' cognitive process. In the 2nd International Conference on big data security on cloud; IEEE International Conference on high performance and smart computing; and IEEE International Conference on intelligent data and security, IEEE, New York, USA: 357-363. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.41 **PMCid:PMC4725632**