Big Data Security and Privacy: A Taxonomy with Some HPC and Blockchain Perspectives

Khalil Alsulbi[1], **Maher Khemakhem**[2], **Abdullah Basuhail**[3] , **Fathy Eassa**[4], **Kamal Mansur Jambi**[5] **and Khalid Almarhabi**[6]

Kalsulbi0004@stu.kau.edu.sa

King Abdul-Aziz University CS Department, KSA [12345], Umm Al-Qura University CS Department, KSA[6]

Abstract

The amount of Big Data generated from multiple sources is continuously increasing. Traditional storage methods lack the capacity for such massive amounts of data. Consequently, most organizations have shifted to the use of cloud storage as an alternative option to store Big Data. Despite the significant developments in cloud storage, it still faces many challenges, such as privacy and security concerns. This paper discusses Big Data, its challenges, and different classifications of security and privacy challenges. Furthermore, it proposes a new classification of Big Data security and privacy challenges and offers some perspectives to provide solutions to these challenges.

Key words:

Big Data, Big Data security, Big Data privacy, cloud storage, high-performance computing (HPC), Blockchain.

1. Introduction

Big Data refers to the massive amounts of data produced by social networks, various websites, multimedia archives, the Internet of Things (IoT), and personalized services. These data can be structured, semi-structured, or unstructured. Moreover, they can be collected, stored, analyzed, and utilized in various platforms [1]. Another definition of Big Data refers to a massive and complex set of data that is difficult to manipulate using traditional database management systems [2]. The storage of these data has become an important topic, and historical data analysis has been used to discover patterns that support decision-making processes.

In recent years, there has been significant growth in the amount of Big Data, which has increased from the petabyte level to the zettabyte level in the past two decades. The International Data Corporation (IDC) and Electro-Motive Company (EMC) predicted that the amount of data would increase to 40 zettabytes by 2020 as shown in Fig.1 [3][4]. New forecasts provided by IDC indicate that the amount of data will reach 175 zettabytes by 2025 [5]. Big Data provides many advantages, such as decision-making strategies, the definition of needs, and the identification of new trends. However, it is also associated with many security and privacy problems, such as confidentiality, availability, integrity, data privacy, monitoring and auditing, and key management [6].

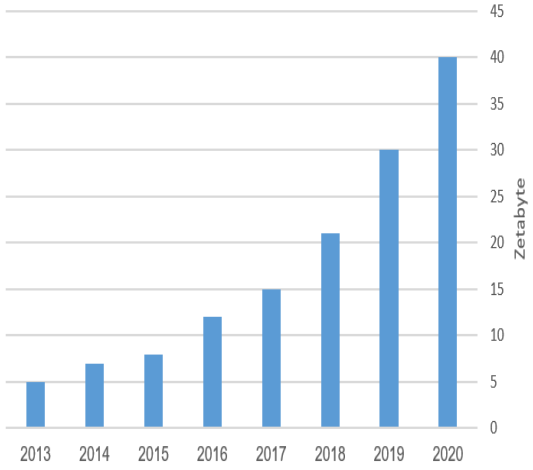


Fig. 1. Big Data in Zettabyte.

Due to the massive amounts of data, their complexity, and the security and privacy issues mentioned above, traditional data management tools are unable to store and process Big Data efficiently [7]. Therefore, many organizations have turned to the use of cloud storage as an alternative solution to the storage and transfer of such data. Cloud storage has many advantages, such as cost saving, quick data accessibility, reliability, the ability to replicate the data to protect it in the event of a disaster, and workload balance [8]. However, cloud storage is accompanied by many challenges because it is an external hosting solution; thus, it may not be trusted by the data owner. Security and privacy are the most critical challenges associated with the storage of Big Data in the cloud [9].

Unfortunately, there are many challenges facing Big Data, with the most important being security and privacy issues. Our information is under the control of

Manuscript received July 5, 2021
Manuscript revised July 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.7.5

banks, social networks, shopping websites, etc., which is an uneasy feeling, as our actions and information are in the hands of others. Furthermore, our data can easily be leaked, especially sensitive data, throughout the Big Data lifecycle.

Many technologies can be used to protect users' data, such as encryption techniques, but with the big companies that are considered Big Data producers and managers (such as Facebook), data privacy cannot be guaranteed using these technologies. Despite the fact that Big Data has become a hot research topic, the existing systems cannot ensure the security and privacy of Big Data. Therefore, it is necessary to summarize the challenges with regard to the security and privacy of Big Data to find practical solutions that address these challenges. This paper addresses these challenges, along with some current solutions. The main contributions of this paper are as follows:

1) We first make an overview of Big Data, including definition, lifecycle, characteristics, and challenges.
2) We list some existing Big Data security and privacy challenge surveys and propose a new classification of Big Data security and privacy challenges based on some requirements.
3) We provide some perspectives with regard to improving security and privacy in Big Data.

The remainder of this paper is divided as follows: Section 2 discusses the life cycle of Big Data. Section 3 outlines the characteristics of Big Data. Section 4 summarizes the Big Data challenges according to Big Data characteristics. Section 5 presents different classification systems for security and privacy challenges. Section 6 proposes a new classification system and offers a survey of some studies that address these challenges, as well as their strengths and weaknesses. Section 7 presents some HPC and Blockchain perspectives in Big Data. Section 8 discusses the previous studies' limitations and the tools needed to address these limitations. Section 9 concludes the paper.

2. Big Data lifecycle

The Big Data life cycle involves several stages of dealing with the data. These stages are as follows: data collection, data integration, data storage and management, and data processing and analysis as shown in Fig.2.

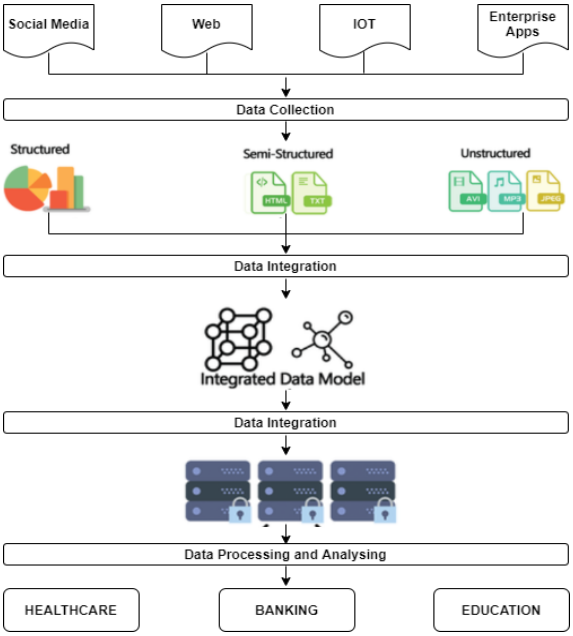


Fig. 2. Big Data life cycle.

Big Data are generated by multiple sources, such as social media, the web, IoT, and enterprise applications. It can be classified into three types: specifically, structured, semi-structured, or unstructured. These types are integrated after cleansing in order to generate an integrated data model. The next stage involves storing the data so that they are ready for processing and analysis. The final stage involves processing and analyzing the data prior to its use for business decisions, such as education, banking, healthcare, government, etc. as shown in Fig.3.

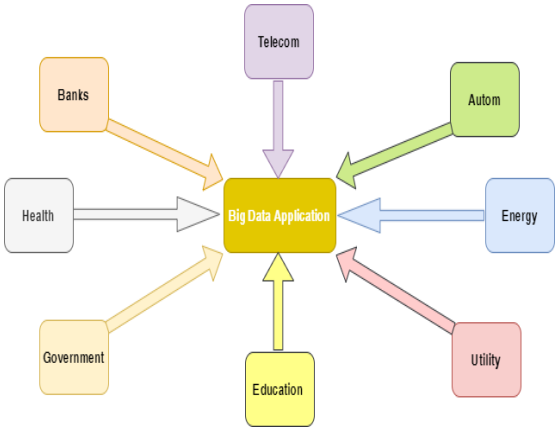


Fig. 3. Big Data applications.

3. Big Data characteristics

The three Vs. of Big Data that relate to their primary characteristics are volume, variety, and velocity [6]. However, with the recent significant technological advances, Big Data have developed different characteristics. Among the most important of these additional dimensions are veracity, validity, volatility, value, vulnerability, valence, variability, and visualization [10][11][12]. In this section, we highlight these characteristics. Table 1 provides a summary of these characteristics and how they affect Big Data.

Table 1: Big Data characteristics

Stage	*Characteristics*	*Definition*
Big Data collection	Variety	Refers to various types of data (structured, semi-structured, and unstructured).
	Velocity	Refers to the speed of data generation.
	Variability	Data are generated by various data sources and stored in a storage facility in variable speeds, formats, or types.
Big Data storage and management	Volume	Refers to exponential growth in the volume of data.
	Volatility	Indicates when the data should be stored and their validity period.
	Valence	Measure that determines the density of data, as it determines the ratio between the data elements connected and the number of connections created within the data collection.
Big Data processing and analysis	Vulnerability	Relates to the security, privacy, and risks arising from collecting personal data through services and products using social networks, IoT devices, and internet applications.
	Value	Refers to ensuring that the organization acquires value for these data. It relates to the measurement of the usefulness of data for decision-making.
	Veracity	Refers to the necessity of correct and accurate data that must be processed to obtain the best outcomes.
	Visualization	Refers to improving data insights and making them easy to understand.
	Validity	This means that the data must be correct and accurate for the intended use. The source of Big Data must be accurate, especially when the results are used for decision-making.

4. Big Data challenges

With the emergence of Big Data, in addition to their various characteristics, many benefits have emerged, as well as many uses in various sectors, such as health, education, and business. However, a number of challenges have also emerged, the most important of which are presented in Fig.4.

A. Heterogeneity

Heterogeneity or variety is one of the characteristics of Big Data. Data are collected from various sources in a variety of forms; they can be structured, semi-structured, or unstructured. Unstructured data account for more than 80% of the data collected by different organizations as shown in Fig.5.

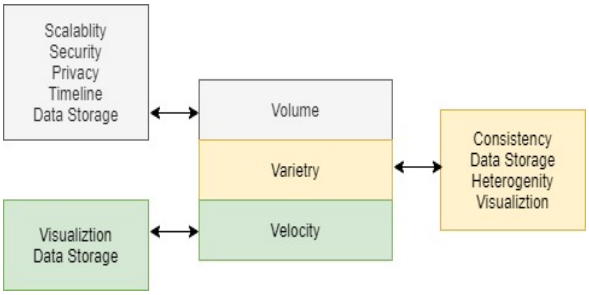


Fig. 4. Big Data challenges.

Variety is a major challenge, as unstructured data are not conducive to the analysis and they are unmanageable. The conversion of unstructured data can be costly. Moreover, it is impossible to convert all data, as most Big Data are unstructured [13]. Heterogeneous data need special data preprocessing methods, such as data cleaning, data integration, and data normalization for processing and analysis.

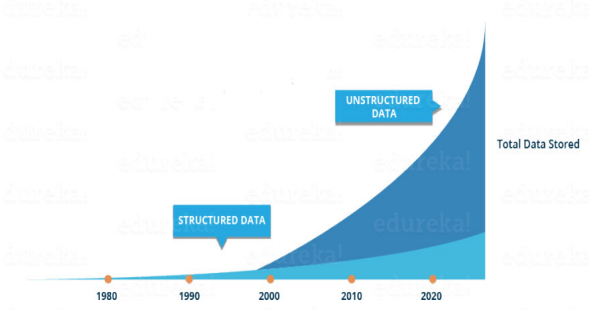


Figure 5. Structured data vs. unstructured data.

B. Data Storage

Big Data are collected from many sources in huge quantities and at high speeds. Not all the data collected are useful, but they should be stored in preparation for

preprocessing. The traditional storage systems are not sufficient for the storage of Big Data. Therefore, the use of cloud storage represents a solution to this problem. On the other hand, the storage of Big Data in the cloud is a significant challenge considering the massive amounts of data, which require long periods of time to be uploaded to the cloud. Moreover, it is not possible to use the cloud for real-time data manipulation due to the velocity characteristics of Big Data [13]. Therefore, the storage problem is one of the most critical issues associated with Big Data, encouraging many researchers to seek practical solutions to this problem.

C. Consistency

Data consistency means that data collected from many sources must be consistent in terms of how we are able to deduce non-contradictory conclusions from them. Data consistency is a challenge in Big Data because they are collected from various sources with varying levels of trustability [14]. The data consistency must be checked during the data cleaning phase, in which some data are modified and some are deleted to improve the overall data quality and consistency.

D. Timeliness

Big Data are continuously growing, and they need to be filtered and stored. Many real-time applications involve Big Data, such as biomedical systems, social networks, intelligent transport systems, fraud detection systems, and traffic control systems. In such applications, decisions must be made in a timely manner. Due to the volume of Big Data, ensuring the timeliness of responses is complicated. Thus, timeliness is one of the most critical Big Data challenges [13][15]. The data value may diminish over time, so they must be collected, processed, and presented instantly in order to be used within an appropriate timeframe.

E. Security

Security is another significant challenge associated with Big Data. Data storage locations such as data warehouses and the cloud must be protected. Criminal groups may attack Big Data repositories to gain confidential and essential data [6][16]. Indeed, security must be ensured throughout all stages of the Big Data process: collection, storage, transfer, analysis, sharing, and knowledge extraction.

F. Privacy

Privacy is one of the most critical concerns in Big Data. It is a technical and social problem. When users share private data, such as locations, pictures, and personal information, this information often reveals their identity and can be used to commit fraud and other criminal activities [16]. When users' data are collected, those who analyze and process these data can benefit from them; consequently, there must be a high level of transparency so that customers are aware of what happens to their data, ensuring their confidence in the Big Data systems that deal with them.

G. Scalability

The scalability challenge is related to volume—namely, the ability to store Big Data that is increasing rapidly and to handle this increase. The scalability challenge arises in the storage and analysis phases. Researchers are currently developing algorithms to improve scalability in Big Data [17].

H. Visualization

Big Data visualization means presenting the data in graphical form so that they are easy to understand and interpret. Big Data are very complex due to their variety and velocity. Data visualization methods still need improvement [13][18]. Good visualization enables users to understand the results and improves decision-making. Big Data visualization offers further benefits, such as recognizing relationships, identifying patterns in order to make predictions, and determining errors and risks.

5. Big Data security and privacy challenges in literature

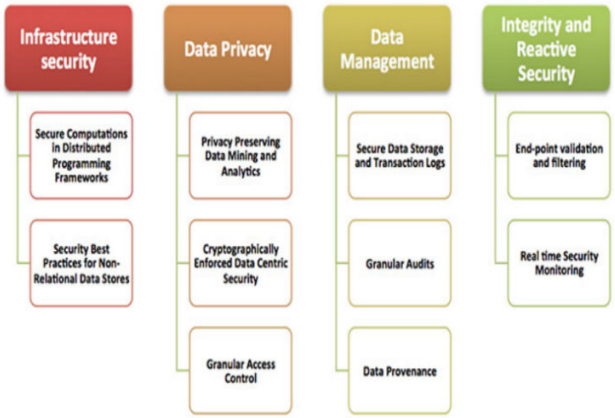
Big Data are an essential source of information in all their forms. The collection and storage of Big Data has attracted interest from industry, business, and government. At the same time, however, it has attracted attackers. Therefore, privacy and security are important challenges associated with Big Data [19]. Robust security mechanisms are needed in order to protect Big Data stores.

The Cloud Security Alliance Big Data security Working Group (CSA) [20] outlined the most important privacy and security challenges as follows:

- Secure computations in distributed programming frameworks;
- Security best practices for non-relational data stores;
- Secure data storage and transaction logs;
- End-point input validation/filtering;
- Real-time security monitoring;

- Scalable privacy-preserving data mining and analytics;
- Cryptographically enforced data-centric security;
- Granular access control;
- Granular audits;
- Data provenance.

The ten challenges can be grouped into four categories, as shown in Fig.6.



Fig, 6. Big Data security and privacy challenges [19].

Another study [21] classified the aspects of privacy and security in Big Data into six main categories: integrity, confidentiality, availability, monitoring and auditing, data privacy, and key management. R.Bao et al. [3] classified security and privacy challenges associated with Big Data into three main categories: storage and management challenges, transmission and sharing challenges, and analytical challenges. An alternative classification scheme is provided in [22]. This paper classifies the security and privacy challenges associated with Big Data into five categories: cloud security, Hadoop security, key management, monitoring and auditing, and anonymization. Moreover, a new study [43] classified the most significant current challenges into seven categories: data breach, misconfiguration and inadequate change control, lack of cloud security architecture and strategy, insufficient identity, access, credentials, and key management, account hijacking, insider threats, and insecure user interfaces (UIs) and APIs. Another study [10] classified the security and privacy challenges in Big Data into three main categories: Big Data acquisition, Big Data storage, and Big Data analytics. Table 2 shows a comparison of these studies.

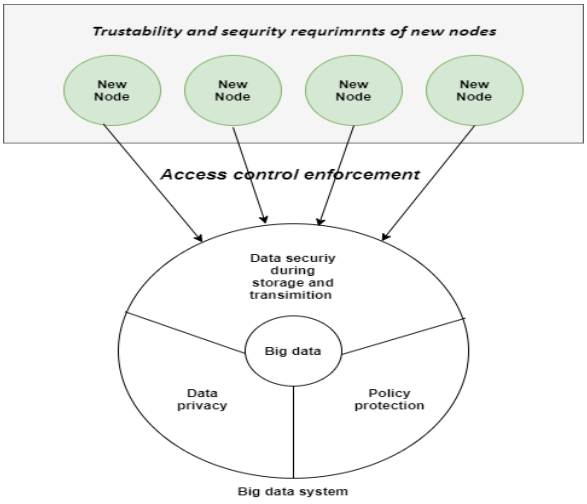
Table 2: Comparison of surveys on security and privacy in relation of Big Data

Ref / year	***Survey Description***	***Big Data Challenges Aspects***	***Distinction with this survey***
[20] / 2013	The authors highlighted the top ten Big Data security and privacy challenges.	Infrastructure Security Data Privacy Data Management Integrity and Reactive Security	The study was limited to listing the ten challenges and mentioning some solutions to these challenges
[21] / 2017	The authors discussed Big Data challenges and reviewed security and privacy issues	Data Privacy Data Management Integrity and Reactive Security	The study was limited to presenting the challenges of Big Data and the challenges of security and privacy without introducing proposals to address these challenges
[3] / 2018	The authors reviewed the challenges of Big Data and surveyed cryptography methods and security schemes.	Data Privacy Data Management	The study was limited to presenting the general challenges of Big Data, with less content focusing on security and privacy challenges.
[22] / 2015	The authors discussed concerns with regard to security and privacy in Big Data and presented a comparison of security approaches in the literature.	Infrastructure Security Data Privacy Data Management Integrity and Reactive Security	The study was limited to presenting the concerns with regard to security and privacy in Big Data and presented a comparison of security approaches in the literature without introducing new proposals to address these challenges.
[43] / 2020	The authors presented an overview of Big Data, current challenges, security and privacy threats. They also discussed previous frameworks that have been implemented to combat those threats.	Infrastructure Security Data Privacy Integrity and Reactive Security	The study was limited to presenting the current challenges, security and privacy threats, previous security and privacy frameworks without introducing new proposals to address these challenges.
[10] / 2019	The authors identified eleven Vs as the Big Data Dimensions, which were	Data Management Integrity and Reactive Security	The study focused on security challenges only.

	then mapped to the phases of the Big Data life cycle. Finally, the authors provided some strategies to address security and privacy challenges.		

6. Proposed classification of Big data security and privacy challenges

Based on the different classification systems presented in the previous section, we propose a new means of classifying the privacy and security challenges associated with Big Data. The new classification system consists of five categories as follows: trustability of new nodes connected to the Big Data environment, the security and integrity of data during storage and transportation, access control enforcement, policy protection, and data privacy as shown in Fig.7. The new classification system is concerned with protecting the environment of Big Data, starting with data collection, where the data must be collected from trusted sources. Another advantage of this classification system is that it mentions the challenges facing Big Data due to the lack of full protection of policies, which allows attackers to access and tamper with the data. We review the related works and consider how these challenges can be addressed, along with the strengths and weaknesses of such solutions.



Fig, 6. Proposed classification of Big Data security and privacy challenges.

A. Trustability of New Nodes Connected to the Big Data Environment

The first step in maintaining data security and privacy in the proposed Big Data system begins with verifying the trustability of the new nodes connected to the system and ensuring that the security requirements are met. Untrusted nodes may threaten the system and allow attackers to exploit them in order to access the data stored in the system. Many nodes and devices can be used to access Big Data systems. Before these entities are allowed to access the Big Data system, they must be checked for their trustability in order to protect the system from attacks. It is important to implement security requirements that must be fulfilled by any new nodes in order for them to become trusted nodes. These requirements may include an up-to-date operating system, active antivirus capabilities, etc.

K. Almarhabi et al. [23] proposed a technique to determine whether new devices meet security requirements. They used a mobile agent to migrate to all connected user devices. It checks devices to determine whether they meet security policy requirements. These requirements are: installing an agent manager, updated antivirus software, a VPN connection, etc. In this way, only trusted devices can be connected to the network. This technique increases the security level and makes the network system more trustable.

Y. B. Reddy [24] proposed a security mechanism based on the Kerberos authentication protocol and the Transport Layer Security (TLS)/Secure Socket Layer (SSL) protocols. Kerberos authentication helps devices to meet the security requirements of trusted users. The proposed mechanism allows secure access control to sensitive data in Big Data systems. Compared to TLS/SSL, Kerberos has better performance due to its simpler user management system and symmetric key operations.

Z. Xiao et al. [25] proposed a new approach based on performing an accountability test (A-test) to detect malicious nodes in real-time, checking all working machines. The A-test is conducted on every node in the system by an auditor group (AG). If malicious behavior occurs in any node, the AG detects it and provides verifiable evidence. The authors proposed the use of P-A accountability in the A-test to improve its performance. The proposed approach was used in the Hadoop platform with MapReduce. The results showed that the system is practical and efficient.

B. Access Control Enforcement

Access control (AC) is a fundamental aspect of the Big Data networking security mechanisms involved in security and privacy. Enforcement of an access control policy is one of the biggest challenges in a Big Data environment. After ensuring the trustability of the new nodes connected to the system, Big Data systems must provide a high-level access control policy. This plays the role of imposing connection control policies as each node or user must fulfill the requirements of strict access policies. Malformed AC policy configurations can led to a loss of privacy and security in a Big Data system [26]. Organizations protect their data through AC policies to address security and privacy permissions. Despite the range of research that addresses access control problems associated with the protection of Big Data and users, there are still many potential issues that must be addressed.

Ashwin et al. [27] proposed a new AC framework based on content sensitivity (CSBAC) to solve the issues of security and privacy in the context of Big Data. The CSBAC framework focuses on granting AC protection to sensitive data only, without using the base set. In addition, it is considered when more data are added if there is a change in the sensitivity of the data. In the CSBAC system, data sensitivity is calculated using the information value equation, as used in the Sensitive Data Detection (SDD) framework. The CSBAC framework is a hybrid access control mechanism that requires the user's participation in access control enforcement and data sensitivity. The CSBAC framework provides complete protection throughout the Big Data lifecycle by guaranteeing the security of sensitive data.

Gupta et al. [28] proposed a new access control model based on the fine-grained attribute (HeABAC), which works with a multi-tenant Hadoop environment to satisfy security and privacy requirements. The authors propose the use of a context enricher to enrich the services, user information, and objects in the access request. The text files for services, users, and objects will be added, with their relevant attributes, into the security administration system. The context enricher will then use these files to add the attributes of the services, users, and objects to the access request. In the proposed model, the central policy server administrates the policies, and Apache Ranger security plugins carry out the enforcement and make the decisions.

Marijuana et al. [29] proposed an access control model for distinct users based on the semantic integration nature of XML data, a standard form of data structure in Big Data environments that use the incorporation approach from lowest to highest and utilize semantic reliance among data. Based on the analysis of the proposed framework's results, it achieves privacy and manages access very efficiently.

C. Security and Integrity of Data During Storage and Transportation

The next step in achieving security and privacy in the Big Data system is to attain data security in the storage and transmission phases to ensure data integrity (DI). To ensure DI, data must be protected from unauthorized changes made by unauthorized users, and it refers to the consistency and accuracy of the stored data. The most important reasons behind DI issues are errors related to hardware, users, software, and intruders [20]. Big Data are sometimes stored in unknown locations; therefore, DI is an important issue and a source of concern. Data must be kept secure during both storage and transmission in order to maintain integrity. Several solutions have been proposed to audit and support DI.

N'TAYA Matisse et al. [30] proposed a new model for Big Data storage and sharing based on Blockchain. The data to be stored are encrypted and indexed to ensure their uniqueness. Authorized users can access and decrypt data. This model does not allow any modification of the data records. However, it creates a new version of the modified record, and this may lead to wasting of resources in the event that data that are requested and altered repeatedly.

Kai Fan et al. [31] proposed to improve Big Data security based on key hierarchical management. There are three levels of keys in the proposed framework: upper key, middle key, and lower key. This approach guarantees the security of all keys. The upper key encrypts the middle key, and the middle key encrypts the lower key. Data are transferred to the cloud after they are encrypted at the client-side.

Zikratov et al. [32] proposed a technique that uses Blockchain to ensure data integrity. Blockchain can ensure the integrity of data through authentication, well-formed transactions, and auditing. By using Blockchain, the threats to data integrity could be decreased. In addition, Blockchain can ensure availability and confidentiality.

D. Policy Protection

Policies are sets of rules designed to control, access, and transfer data from one place to another. Attackers

can alter policies. Data owners will not know if policies have been legally or illegally modified during transit, storage, or processing. Security measures must be implemented to prevent these policies from being tampered with as most of the current research focuses only on data protection and neglects policy protection.
M. Sowmiya et al. [33] proposed a secure cloud storage model based on hidden access control policies by implementing a query-driven approach. The encryption keys are created according to the data owner's access policies and are used to encrypt the data files. Then, the access policies determine which users can access the data. The results show the improvements made to the cloud storage system.

D. D. F. Maesa et al. [34] proposed storing policies based on Blockchain technology. First, the policy is stored in an eXtensible Access Control Markup Language (XACML) file. Sometimes, policies can be large in size. Therefore, storage of XACML files directly on a Blockchain leads to severe problems related to Blockchain size. The proposed solution will store the policy in an external source and keep only a link to this external source in the Blockchain. Through this method, Blockchain features can be exploited, such as transferring access rights from one user to another through Blockchain transactions.
Xue et al. [35] proposed an access control management model based on a private Blockchain. The owner identifies policies of access control. The policies will be stored as a policy file that allows the data owners to check access privileges in the Blockchain header. The proposed solution enhances security and protects the access control policies.

E. *Data Privacy*

Big Data may contain sensitive or personal data of users. Data privacy means not sharing personal data without the explicit consent of their owner. Even after obtaining consent, the use of personal information is limited to a necessary and specific reason. Hence, to protect privacy, data must be de-identified by removing attributes that would identify an individual [36].

Zhang et al. [37] proposed a protocol based on a privacy-preserving multi-keyword search (PRMSM) in a multi-owner cloud model. In this protocol, data owners encrypt files and keywords using different keys. Users can query authenticated data without knowing the secret keys of data owners. In addition, the cloud server returns the search results to users without revealing sensitive data.

Wan et al. [38] proposed a privacy-preserving multi-keyword search schema called VPSeearch. The proposed schema integrates the privacy-preserving multi-keyword search scheme with the Message Authentication Code (MAC) technique. The proposed schema relies on a one-way function; therefore, it is very efficient.

Jiang et al. [39] proposed a privacy-preserving scheme of encrypted data in the cloud. The proposed schema is based on an Inverted Matrix (IM) made up of several index vectors to create a search index. During the index construction, each keyword is linked with an index vector. The authors also use pseudo-random bits to obtain an Encrypted Enlarged Inverted Matrix (EEIM) to blind the index vectors in order to ensure users' privacy.

Table 3 presents a summary of the literature review focusing on the security and privacy issues in the Big Data environment by presenting some existing solutions, as well as their strengths and weaknesses.

Table 3: Summary of the literature review of the security and privacy issues in Big Data

Security and Privacy Issues	***Solution***	***Strengths***	***Weaknesses***
Trustability of new nodes connected to the Big Data environment	Check security requirements by the mobile agent [23]	-Detect trusted devices before connecting to the network. -Increase system performance	The mobile agent itself needs some techniques to improve its security
	Kerberos authentication over TLS/SSL [24]	-An efficient access control mechanism for highly sensitive data -Better performance.	Kerberos protocol has many weaknesses, such as the use of an authenticator to prevent replay attacks.
	Performing an accountability test (A-test) to detect malicious nodes in real time [25].	Detect malicious nodes with high efficiency	The solution is suitable for MapReduce in the Hadoop platform
Access control Enforcement	Access control framework based on content sensitivity (CSBAC)[27]	Dynamic change in data sensitivity when adding more data.	-The solution is suitable for HDFS - Overhead due to extra computations.
	Access control framework that is based	Context enricher in the proposed framework	The solution is suitable for Hadoop.

	on fine-grained attribute (HeABAC), which works with a multitenant Hadoop [28]	is used to enrich the user information and objects and services in the access request.	
	Distinct users access control model based on XML data's semantic integration nature [29].	The proposed model keeps flexible and controllable semantic associations and avoids redundant structure association.	The model needs improvements to prevent similar inference attacks
Security and integrity of data during storage and transportation	Store data indexes in Blockchain [30]	-Highly secure decentralized storage. - Provide traceability of data from the data owners to the warehouse.	The proposed schema requires more computational resources
	Key hierarchical management [31]	Ensure Big Data security by using key hierarchical management to encrypt data.	-Long encryption time. -The system needs to improve the key exchange. -Third-party problem
	Blockchain to ensure data integrity [32]	Ensure data security and integrity through Blockchain technology, which provides authentication, audit, and well-formed transactions.	The possibility of destroying Blockchain systems through the quantum computer.
POLICY PROTECT	Hidden access control policies using query-driven approach [33].	-Hide the users' attributes from the cloud. -Hide the access policies which are associated with individual files.	-It is only allowed to upload a file with a maximum size of 100 MB each time. -Third-party problem.
		-Use homomorphic encryption technology to enhance storage security	
	Store policies based on Blockchain Technology [34]	Use Blockchain to publish access policies and transfer the access rights from one user to another quickly.	A very lengthy formalism creates a space occupation problem using XACML language when storing policies directly in the Blockchain.
	Store policy file on Blockchain [35].	Enhance the security of accessing devices in the smart home.	-Access control policies are not self-enforced. -No mention of how to secure and trace outsourced data.
DATA PRIVAC	Using different keys to encrypt keywords and files [37].	A new efficient protocol for data user authentication and dynamic secret key generation.	-The problem of searching for fuzzy keywords. -Single point of failure. -Third-party problem.
	Integrate a privacy-preserving multi-keyword search scheme with MAC technique [38].	-Not needed to keep a copy of the outsourced data. -Deal with encrypted data.	Not efficient for large-scale database.
	Inverted Matrix (IM) [39].	A novel privacy-preserving model for encrypted cloud data.	Third-party problem.

To evaluate our proposed classification, we collected Big Data security and privacy survey studies and compared our proposal with their classifications. We evaluated the security and privacy challenges covered in each study, as shown in Table 4.

Table 4: Comparison of our proposal with related survey studies

Challenges	[20]	[21]	[3]	[22]	[10]	[43]	Proposal
Trustability of New Nodes Connected to the Big Data Environment					√		√
Access Control Enforcement	√	√			√	√	√
Security and Integrity	√	√	√	√	√	√	√
Policy Protection							√
Data Privacy	√	√	√	√		√	√

7. Some HPC and Blockchain perspectives in Big Data

In addition to Big Data security and privacy monitoring, sensitive data can be very large depending on their sources. The corresponding real-time encryption and decryption during their transfer and storage also represent challenges since they require sufficient computing power. HPC can provide an effective solution to these problems, such as for any real-time Big Data analytics and/or visualizations [40]. Furthermore, Blockchain can be used in different manners to substantially enforce the security and privacy of Big Data [41]. Indeed, the combination of HPC and Blockchain is expected to solve many problems related to Big Data security and privacy monitoring on one hand and Big Data analytics on the other hand [42]. More effective algorithms are required in such cases in order to achieve the expected objectives.

8. Discussion

We presented the different classifications of Big Data security and privacy challenges in Section 5. In Section 6, we introduced our proposed classification, and we conducted a comparative evaluation of the survey studies. Previous studies indicate limitations, particularly in research studies that focus on Big Data security and privacy, such as a lack of comprehensive security and privacy architecture for Big Data, a single point of failure in the centralized approaches, performance, and third-party problems. Most of the Big Data security and privacy challenges are ongoing as they are difficult to fully address with traditional technologies; thus, future work must be conducted in order to resolve them. Further work is needed to find efficient solutions that address these issues. Based on the proposed classification of security and privacy challenges in Big Data, we can clarify some requirements for future solutions to meet these challenges. Any future solution must include the following requirements:

- Build a technique for testing user device trustability and authentication;
- Build a technique for access control policy enforcement in the Big Data system;
- Build a technique for ensuring data security and privacy during transfer and storage phases, which may combine HPC;
- Build a technique for protecting policies;
- Build a distributed model that does not depend on a third party.

The first technique allows only trusted nodes and devices to join the Big Data environment. This technique ensures the collection of data from trusted and secure sources and provides greater security for the Big Data system. In contrast, attackers can exploit untrusted nodes to launch attacks on the Big Data system. The second technique enforces the Big Data system's access control policy. Big Data contains sensitive data that must be protected from illegal access by unauthorized users or even analysts in the analysis phase. Consequently, it is necessary to enforce appropriate policies of access control and implement techniques to prevent out-of-purpose analytics. The enforcement of access policies deters attackers from accessing data and helps to protect data using practices that contribute to data security during storage and transmission. The policy protection technique protects the policies that control the authorization process from changing as a result of an attack. The maintaining and safeguarding of policies from change contribute to increasing data security and privacy. The last requirement of future solutions is eliminating any dependence on a third party in storing data, which contributes to achieving data privacy. We aim to achieve these requirements through building a secure big data management system that achieves those requirements.

9. Conclusion

With the massive growth of data, Big Data are associated with many challenges, the most important being security and privacy. These challenges need to be discussed further in order to develop efficient solutions and techniques that address them and improve Big Data management systems.

This paper presented the most critical challenges associated with Big Data, focusing particularly on the security and privacy challenges and the most critical security and privacy requirements. Traditional approaches were insufficient for Big Data and they did not provide a complete solution to address security and privacy issues. Researchers and Big Data stakeholders should conduct more research and studies that contribute to resolving these challenges and taking advantage of modern technologies such as Blockchain and HPC. Any future solution must develop techniques for testing user device trustability and authentication, ensuring data security and privacy during the transfer and storage phases, enforcing access control policies in the Big Data system, protecting policies, and eliminating the dependence on third parties. Our future work will seek to fulfill all of these requirements in a new framework. We will build our framework, test it, and evaluate the results in a subsequent research.

References

[1] Baig, M. I.; Shuib, L.; Yadegaridehkordi, E. "Big data adoption: State of the art and research challenges," Inf. Process. Manag. 2019, vol. 56, no. 6, p. 1046495.

[2] Padgavankar, M. H.; Gupta, S. R. "Big data storage and challenges," Int. J. Comput. Sci. Inf. Technol. 2014, vol. 5, no. 2, pp. 2218–2223.

[3] Bao, R. Chen, Z.; Obaidat, M. S. "Challenges and techniques in Big data security and privacy: A review," Secure. Priv. 2018, vol. 1, no. 4, p. e13.

[4] Yang, P.; Xiong, N.; Ren, J. "Data Security and Privacy Protection for Cloud Storage: A Survey," IEEE Access, 2020, vol. 8, pp. 131723–131740.

[5] Woodie, Alex. "Global DataSphere to Hit 175 Zettabytes by 2025, IDC Says". Datanami. 2018.

[6] Subbalakshmi, S.; Madhavi, K. "Security challenges of Big Data storage in Cloud environment: A Survey," Int. J. Appl. Eng. Res. 2018, vol. 13, no. 17, pp. 13237–13244.

[7] Venkatraman, S.; Venkatraman, R. "Big data security challenges and strategies," AIMS Math. 2019, vol. 4, no. 3, pp. 860–879.

[8] Akingbade, L. O. "Cloud Storage problems, benefits and solutions provided by Data De-duplication", International Journal of Engineering and Innovative Technology.2016,(5),6, 70-77.

[9] Sun, P. J. "Privacy protection and data security in cloud computing: a survey, challenges, and solutions," IEEE Access. 2019, vol. 7, pp. 147420–147452.

[10] Venkatraman, S.; Venkatraman, R. "Big data security challenges and strategies" AIMS Math. 2019, vol. 4, no. 3, pp. 860–879.

[11] Alaoui, I. E.; Gahi, Y.; Messoussi, R. "Full consideration of Big Data characteristics in sentiment analysis context," in 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2019, pp. 126–130.

[12] Panimalar, A.; Shree, V.; Kathrine, V. "The 17 V's Of Big Data," Int. Res. J. Eng. Technol. 2017, vol. 4, no. 09, pp. 329–333.

[13] Desai, P. V. "A survey on big data applications and challenges," in 2018 second international conference on inventive communication and computational technologies (ICICCT), 2018, pp. 737–740.

[14] Shi, P.; Cui, Y.; Xu, K.; Zhang, M.; Ding, L. "Data Consistency Theory and Case Study for Scientific Big Data," Information.2019, vol. 10, no. 4, p. 137.

[15] Ramasamy A.; Chowdhury, S. "Big Data Quality Dimensions: A Systematic Literature Review," J. Inf. Syst. Technol. Manag. 2020, vol. 17.

[16] Yang, P.; Xiong, N.; Ren, J. "Data Security and Privacy Protection for Cloud Storage: A Survey," IEEE Access, 2020, vol. 8, pp. 131723–131740.

[17] Espinosa, J. A.; Kaiser, S.; Armour, F.; Money, W. "Big Data Redux: New Issues and Challenges Moving Forward", In Proceedings of the 52nd Hawaii International Conference on System Sciences.2019.

[18] Khare, S.; Totaro, M. "Big Data in IoT," in 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1–7.

[19] Bertino, E.; Ferrari, E. "Big data security and privacy," in A Comprehensive Guide through the Italian Database Research over the Last 25 Years, Springer. 2018, pp. 425–439.

[20] BigDataWorkingGroup, "Expanded Top Ten Big Data Security and Privacy Challenges," 2013. [Online]. Available:https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf.[Accessed: 27-Feb-2016].

[21] Joshi, N.; Kadhiwala, B. "Big data security and privacy issues—A survey," in 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), 2017, pp. 1–5.

[22] Terzi, D. S.; Terzi, R.; Sagiroglu, S. "A survey on security and privacy issues in big data" in 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST), 2015, pp. 202–207.

[23] Almarhabi, K.; Jambi, K.; Eassa, F.; Batarfi, O. "Survey on access control and management issues in cloud and BYOD environment,".2017, Vol.6, pg. 44-54.

[24] Reddy, Y. B. "Access control mechanisms in Big Data processing," Softw. Eng. Appl. Adv. Power Energy Syst. 2015, pp. 2015–2829.

[25] Xiao, Z.; Xiao, Y. "Achieving accountable MapReduce in cloud computing," Futur. Gener. Comput. Syst. 2014, vol. 30, pp. 1–13.

[26] Centonze, P.; Kim, D. Y.; Kim, S. "Security and privacy frameworks for access control big data systems,"

Comput. Mater. Contin. 2019, vol. 59, no. 2, pp. 361–374.
[27] Kumar, T. K. A.; Liu, H.; Thomas, J. P.; Hou, X. "Content sensitivity based access control framework for Hadoop," Digit. Commun. Networks. 2017, vol. 3, no. 4, pp. 213–225.
[28] Gupta, M.; Patwa, F.; Sandhu, R. "An attribute-based access control model for secure big data processing in Hadoop ecosystem," in Proceedings of the Third ACM Workshop on Attribute-Based Access Control, 2018, pp. 13–24.
[29] Wang, M.; Wang, J.; Guo, L.; Harn, L. "Inverted XML access control model based on ontology semantic dependency," Comput. Mater. Contin. 2018, vol. 55, no. 3, pp. 465–482.
[30] Li, L. J. "Secured Cloud Storage Scheme Based On Blockchain," in 2019 IEEE 2nd International Conference on Electronic Information and Communication Technology (ICEICT), 2019, pp. 137–142.
[31] Fan, K.; Lou, S.; Su, R.; Li, H.; Yang, Y. "Secure and private key management scheme in big data networking," Peer-to-Peer Netw. 2018, Appl., vol. 11, no. 5, pp. 992–999.
[32] Zikratov, I.; Kuzmin, A.; Akimenko, V.; Niculichev, V.; Yalansky, L. "Ensuring data integrity using Blockchain technology," in 2017 20th Conference of Open Innovations Association (FRUCT), 2017, pp. 534–539.
[33] Sowmiya, M.; Adimoolam, M. "Secure cloud storage model with hidden policy attribute-based access control," in 2014 International Conference on Recent Trends in Information Technology, 2014, pp. 1–6.
[34] Maesa, D. D. F.; Mori, P.; Ricci, L. "Blockchain-based access control," in IFIP international conference on distributed applications and interoperable systems, 2017, pp. 206–220.
[35] Xue, J.; Xu, C.; Zhang, Y. Private Blockchain-Based Secure Access Control for Smart Home Systems. KSII Trans. Internet Inf. Syst. 2018, 12, 6057–6078.
[36] Nelson B.; Olovsson, T. "Security and privacy for big data: A systematic literature review," in 2016 IEEE international conference on big data (big data), 2016, pp. 3693–3702.
[37] Zhang, W.; Lin, Y.; Xiao, S.; Wu, J.; Zhou, S. "Privacy-Preserving Ranked Multi-Keyword Search for Multiple Data Owners in Cloud Computing," in IEEE Transactions on Computers. 2016, vol. 65, no. 5, pp. 1566- 1577.
[38] Wan Z.; Deng, R. H. "VPSearch: Achieving Verifiability for Privacy-Preserving Multi-Keyword Search over Encrypted Cloud Data," in IEEE Transactions on Dependable and Secure Computing.2016, vol. PP, no. 99, pp. 1-1.
[39] Jiang, X.; Yu, J.; Kong, F.; Cheng, X.; Hao, R. "A Novel Privacy Preserving Keyword Search Scheme over Encrypted Cloud Data," 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), Krakow. 2015, pp. 836-839.
[40] Javier Álvarez et al, "Efficient development of high performance data analytics in Python", Future Generation Computer Systems. October 2020,Volume 111 , Pages 570-581.
[41] N. Deepa et al, "sA Survey on Blockchain for Big Data: Approaches, Opportunities, and Future Directions", availableon:https://www.researchgate.net/publication/344066677_A_Survey_on_Blockchain_for_Big_Data_Approaches_Opportunities_and_Future_Directions.
[42] Geoffrey Charles Fox et al, «Contributions to High-Performance Big Data Computing",Technical report,available on: https://www.researchgate.net/publication/328090399_Contributions_to_High-Performance_Big_Data_Computing.
[43] S. Riaz, A. H. Khan, M. Haroon, S. Latif, and S. Bhatti, "Big data security and privacy: Current challenges and future research perspective in cloud environment" in 2020 International Conference on Information Management and Technology (ICIMTech), 2020, pp. 977– 982.

Khalil Alsulbi is currently a PhD student in the computer sciences department in King Abdu-Aziz university, Jeddah. He obtained his M.S. degrees from king from computer sciences department in King Abdu-Aziz university, Saudi Arabia in 2016 and his B.S degree from king from computer sciences department in King Abdu-Aziz university, Saudi Arabia in 2003. His research interested are in the field of OCR, Blockchain and Big Data.

Maher Khemakhem is full Professor in Computer Science at the University of Sfax, Tunisia. He received his M.S of Science, Ph.D, and Habilitation accreditation degrees, respectively, from the University of Paris 11 (Orsay), France in 1984, 1987 and the University of Sfax, Tunisia in 2008. He is currently Full Professor of Computer Science at the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. His research interests include Arabic OCR, distributed systems, performance analysis, and Networks security.

Abdullah Ahmad Basuhail, received the Ph.D. degree in computer engineering from Florida Institute of Technology, Melbourne, FL, USA in 1419H/1998G. His research interests include: digital image processing, computer vision, the use of computer technologies, applications, information technology in e-teaching, e-learning, e-training and e-management supportive systems. Dr. Basuhail was an ex-member of the Saudi Computer Society, the IEEE, and the IEEE Computer Society.

'athy E. Eassa received his B.Sc degree in electronics and electrical communication engineering from Cairo University, Egypt in 1978, his M. Sc. degree in computers and Systems engineering from Al Azhar University, Cairo, Egypt in 1984, and his Ph.D degree in computers and systems engineering from Al-Azhar University, Cairo, Egypt with joint supervision with the University of Colorado, U.S.A, in 1989. He is a full professor with the computer Science dept, Faculty of Computing and Information technology, King Abdulaziz University, Saudi Arabia. His research interests include agent based software engineering, cloud computing, software engineering, big data, distributed systems, and exascale system testing.

Kamal M. Jampi was born in Makkah, Saudi Arabia in 1960. He received his B.S. in computer science from the University of Petroleum and Minerals, Dhahran, Saudi Arabia in 1982, his M.S. degree in computer science from Michigan State, East Lansing, Michigan, USA in 1986 and his Ph.D degree in computer science from Illinois Institute of Technology, Chicago, IL in 1991. Prof. Jambi has been a professor in the Computer Science Department at King Abdul-Aziz University, Jeddah, Saudi Arabia since 2009. His areas of interest includes OCRs, Image processing, NLPs, and big data. He has also been the chairman of the CS department at FCIT and Vice Dean of Graduate Studies and Scientific Research. He was the PI for several projects funded by KACST.

ıalid Almarhabi is a Assistant rofessor in computer science at mm Alqura University in Makkah. e received his M.S degree in ıformation Technology from ueensland University of Technology ı Brisbane, Australia. Also he ceived his Ph.D. degree in computer iences from King Abdul-Aziz ıdi Arabia.. His research interests are secure BYODs, access control policies, information system management, cloud computing, and e-learning.