Security of Big Data: Threats and Different Approaches Towards Big Data Security

Yashi Chaudhary and Heman Pathak

Abstract In the present era, the use of the Internet has extended abruptly. With this abrupt increase, massive data is being created, resulting in big data. Big data means more diverse, more impetus, and more complex data streams. Data is being produced in abundance in exabytes and zettabytes by electronic devices, power grids, and modern software. This big data brings different challenges such as incompleteness, inconsistency, heterogeneity, and security with itself. The presented paper targets the security challenge as it is a very significant feature overseen by various data analysts; thus, data must be secured from dwindling in the wrong hands. This paper discusses the approaches and mechanisms mainly based on anonymization, access control, and encryption.

Keywords Access control · Anonymization big data · Big data life cycle · Big data security · Encryption · High volume · Storage · Introduction

1 Introduction

The term **big data** is used to define humongous data that could either be of structured, semi-structured, or unstructured type. The extensively large data makes processing difficult by using traditional available databases and software technologies. Heavy parallel software devices running on thousands of servers can be used for processing [1].

Big data initially was categorized by the four Vs—volume, variety, velocity, and veracity. However, with time other categorizations are also made as the data is emerging vastly with each passing day [2, 3].

Y. Chaudhary (✉) · H. Pathak
Gurukul Kangri University, Haridwar, India
e-mail: mohita.chaudhary5@gmail.com

D. P. Agrawal et al. (eds.), *Cyber Security, Privacy and Networking*, Lecture Notes
in Networks and Systems 370, https://doi.org/10.1007/978-981-16-8664-1_31

1.1 Ten Vs of Big Data

Big data is mainly characterized by the Vs that define the different traits of the data one is dealing with the under mentioned figure and describes significant ten Vs that make big data different from the traditional data.

Volume. Generated data's quantity is referred to as volume. Value and potential of the data are examined with the size of the data and whether it can be put into the category of big data or not it is also examined through the size of the available data. The huge scale increment makes the analysis of data a difficult process if one is using traditional available tools.

Variety. The category of the data it belongs to is known as variety. Data can come in any form. It may be structured data, unstructured data, or semi-structured data coming out of various sources like e-mails, videos, audios, transactions, etc.

Velocity. How fast the data is generated and processed to meet the required demand refers to as velocity. The speed with which information is gathered and processed to meet the required demands of the intended users.

Veracity. It can be referred to as the trustworthiness of the data that is being used. Analysis correctness depends heavily on the veracity of the source data. Data captured quality can differ immensely.

Variability. It refers to different things. It focuses on adequately understanding and interpreting the correct meaning of raw data that depends on its context. It can also be defined as the inconsistency of the speed at which data is generated and stored. Validity means how accurate the data is for its specified use. If one wants to use the results in some decision-making, subsequent analysis must be accurate enough.

Vulnerability. Vulnerability means a flaw that can leave a system open to attack. The vulnerability may also be referred to as any type of lapse in a computer system, in a set of procedures, or in anything that can hinder the security of the system.

Volatility. In the world of real-time analysis, it is important for the decision-makers to analyse till when the data provided is relevant. This relevancy of data validity is known as volatility.

Visualization. Data visualization means how the data is presented in a graphical format that is easily understood and interpreted by its users. Various complex representations like heat maps and fever charts are included here that help decision-makers to identify the hidden patterns and correlations.

Value. It is the most important trait of the data. Without this, other characteristics are of no use if we are not able to deduce the business value from the data. Big data helps in decision-making in the organization by measuring the importance of the data.

1.2 Big Data Analytics

Big data analytics means exploring large data sets containing a variety of datatypes in big data—to reveal hidden data patterns, unknown interactions, market trends, customer preferences, and other useful business information. To make companies more knowledgeable by enabling data scientists is the first goal of big data analysis. Big data comprises structured, semi-structured, and unstructured data. Tools that are used for advanced analytics such as predictive analysis, data mining, and text analysis can be used in big data analysis as well. Data visualization tools along with some mainstream BI tools can also be very effective in the analysis process of big data [4].

Big data analytics life cycle—As we are using vast data repositories to gain information that will be useful for analytics purposes, we need to refine the available data. The refinement process includes various steps as defined in Fig. 1.

In all the above phases, there are multiple threats that are required to take care of.

Data Collector. Data comes from various sources and with different formats, i.e. structured, semi-structured, and unstructured. In this phase, information is gathered to address various things that can be used by an organization for various purposes. From the security point of view, securing big data from the first phase is very important. Limited access control and encryption of data fields can be done here to ensure privacy here [5].

Data Storage. Data storage mainly addresses the volume challenge by making use of distributed, shared nothing architectures. Data is stored and prepared here that will be used in the next phase. Here produced data may be sensitive, so it is vital to take care of it. Data anonymization, permutation, data partitioning, etc., are some techniques that can be applied to ensure security [6].

Data Analytics. The primary aim of big data analysis process is to disintegrate the significant data from the bunch of data and to provide decisions and recommendations based on the findings after investigating the whole data. This phase is used to create

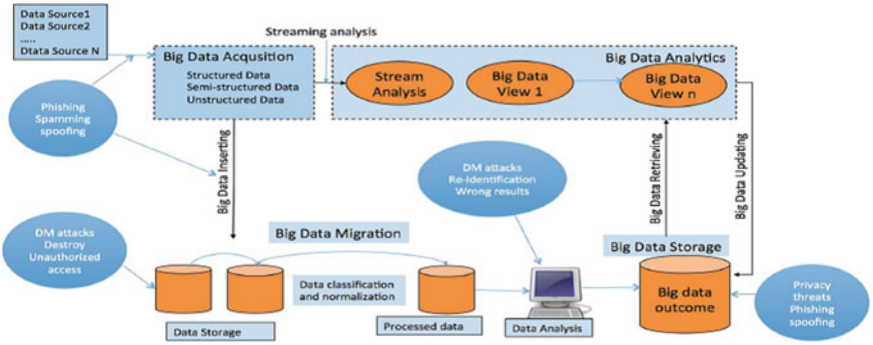


Fig. 1 Phases of big data analytics

Table 1 Threats on various phases of big data analytics

Phase	Threats	Description
Data collector	Spoofing	These attacks are performed to get access to the data collection phase
	Spamming	
	Phishing	
Data storage	Data mining-based attacks	Targets the data sets to extract knowledge
	Attacks on data storage devices	These include stealing the hard discs
	Unauthorized data access	People access the data illegally
Data analytics	Data mining attacks	Uses data mining methods to extract valuable information
	Re-identification attacks	Includes personal threat identification
Knowledge creation	Privacy threats	Releasing the resulted knowledge
	Phishing	Decision-makers are targeted
	Spoofing	Decision-makers are targeted

knowledge. Various data mining methods can be used here. Data miners use powerful algorithms that can extract sensitive data. A security breach may also happen here [7].

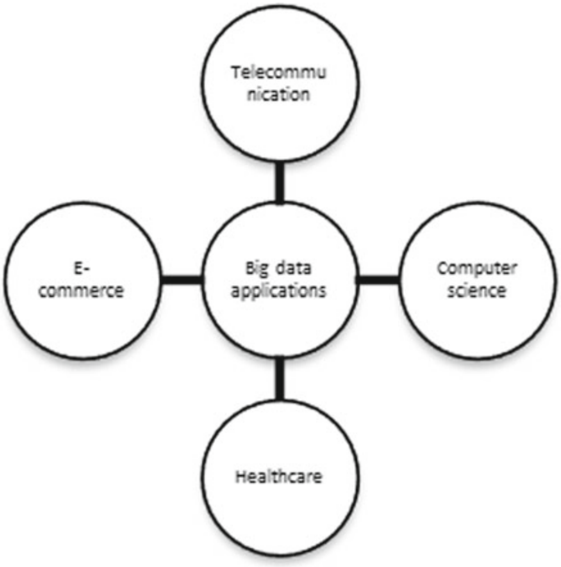
Knowledge Creation. This is the final phase. Conversion of the data into some useful information is done at this step. If data seizing and sensing are done right, then big data repositories can be created in the form of knowledge repositories. It is used by decision-makers. New information and valued information are created here. Knowledge is sometimes considered sensitive here [8].

Threats Associated with Big Data Life Cycle
Various threats associated with different phases of the big data life cycle have been summarized in Table 1 [2].

1.3 Applications of Big Data

In the present era, the use of the Internet has extended abruptly. Due to the vast usage of the Internet anywhere and everywhere, big data applications are also increased due to their decision-making ability. Big data is no more just a buzzing word, but its use is everywhere today. All credit goes to the technology that is nowadays not

Fig. 2 Various applications of big data



just confined to the urban areas, but rural and underdeveloped areas are also taking its advantage. Big data applications range from the water supply, smart cities, crime, health care, education, electricity, etc. (Fig. 2).

1.4 Challenges with Big Data

Generated from different devices at a very fast pace, big data brings the following challenges with itself:

Security. Data is generated at a high pace in huge quantities every day. Big data analytics will not be considered a reliable system if security algorithms will not be taken into account. Security issues can be further categorized: input, analysis of data, and output, system communication.

Inconsistent Data. More inconsistent data and incompatible data will easily appear since data is being gathered from different systems. So it will also be a challenge while doing big data analytics.

Privacy. It is different from the issue of security as it deals with the fact that whether it is possible to restore the personal information of the system with the help of big data analytics, even though the input variables are anonymous. With big data analytics being widely used, it is quite possible that private information may get exposed to other people after the analysis process. So it is also a challenge in big data analytics.

Heterogeneity. Insights of data can be achieved through the richness and nuances of the data. Though, machine algorithms cannot understand nuances as they expect comparable data. So structuring the data carefully is the first step of big data analysis. Even after applying data cleaning and data correction methods, some incompleteness of data may be there. Managing this is a great challenge.

Timeliness. In a dynamic and rapidly growing world, a second or even a microsecond between one reading and the other may led to mismatching against each other. So timeliness is a very fundamental concept while dealing with real-time data.

Communication Between the Systems. Since most of the tasks of the big data analytics system will be designed for parallel computing, big data analytics and other systems communication will impact the performance of the system immensely. So managing the cost of communication and making connections reliable are two open challenges to deal with [2].

2 Big Data Security: A Multifaceted Challenge

Big data security is a cumulative term for all the techniques and tools that are used in securing the data against all malicious activities such as theft of data, attacks, or any activity that affects negatively. The other threats include DDoS attacks, ransomware, and online stored information stealing [9, 10].

The prime reason for security concerns in big data is because big data can be accessed widely nowadays. Data is shared on a large scale by scientists, doctors, business officials, government agencies, and normal people. The current approaches are inadequate when dealing with big data security. The present technology has weak security capability for maintenance. So intruders can easily breach those. Thus, reassessment and updation of current approaches should be performed to prevent data leakage [11]. There are various challenges when one is dealing with big data security. A few of them are mentioned below.

2.1 Issues

Vulnerability to fake data generation—Before dealing with all the operational security challenges of big data, the concerns of counterfeit data generation should be kept in mind. To purposively undermine the quality of the big data analysis, cybercriminals can forge the data. For instance, if a manufacturing company uses sensor data to detect malfunctioning production processes, cybercriminals can penetrate that system and make sensors show fake results, say, wrong temperatures. This way, one can fail to notice alarming trends and miss the opportunity to solve problems that can cause severe damage. Such issues can be addressed by applying the fraud detection approach [12].

Untrusted mapper's presence—After collection, big data firstly undergoes parallel processing. MapReduce paradigm is used here and when data splits a mapper processes that and allocates a position for storage to the data. If anyone from outside knows your mapper's code, he/she can change it. In this way, it is ruining the information processed very effectively. Outsiders can get inside to access sensitive information [13].

Mining of sensitive information—Perimeter-based security ensures data protection at entry and exit levels. But inside the system, the work of IT professionals is a mystery. Such a lack of control over big data solutions can allow corrupted IT professionals to mine the data and sell it for their benefit. As a result, the organization will suffer huge losses. Here, data can be made more secure by adding values to it. Anonymization can also benefit the system's security. The private details with absent names, telephones, etc., practically will not harm if someone acquires this information with malicious intentions [14].

Real-time protection of data—It is hard for organizations to maintain orderly checks as data is generated vastly on a real-time basis. However, security checks in real time or almost in real time will prove beneficial [15].

Access control granularly—Granular access control allows people to access the required sets of data but can view the only part of data they are allowed to see. The whole valuable content will not be visible to them. Vastly, it can be very useful in health care where sensitive information like names and phone numbers will remain hidden while other information may be useful for medical researchers to find new insights [16].

Privacy protection of non-relational database—Various security vulnerabilities are faced by datastores such as NoSQL that lead to privacy threats. At the time of logging and tagging, it is unable to encrypt the data and so is the case with the distribution of data to different groups while it is streamed and collected [16].

3 Security-Based Literature Survey

There are three major security considerations outline that has been taken into account while dealing with big data: anonymization, encryption, and access control [17] (Fig. 3).

Diversified data sources, data streams, data formats, and infrastructures may impose unique security vulnerabilities (Table 2).

3.1 Existing Approaches to Handle the Big Data Security

Listed below are the different approaches to manage security as discussed by various authors.

Table 2 Summary of the literature survey

S. No.	Authors name	Problem identified by authors	Security algorithm discussed	Security aspects
1.	Li et al.	Low computation time over the cloud	Security-aware efficient distributed storage (SA-EDS) model	Encryption
2.	Aljawarneh et al.	High computational cost	The amalgamation of Feistel encryption scheme, an advanced encryption standard (AES), and genetic algorithms	
3.	Yan et al.	Data deduplication security issue	Proxy re-encryption algorithm	
4.	Dong et al.	Sensitive data	Proxy re-encryption algorithm based on heterogeneous ciphertext	
5.	Hu et al.	Access control of data	ABAC algorithm	Access control
6.	Zeng et al.	High overhead of conventional algorithms while dealing with big data	Content-based access control (CBAC)	
7.	Khuntia et al.	User's private information leakage	Hidden policy-ciphertext policy-attribute-based encrption: HP-CP-ABE	
8.	Siffah et al.	High risk of data leakage	MeDShare: blockchain for sharing trustless medical data	
9.	Jasim et al.	Zero trust between models	Transaction's manager model algorithm	
10.	Zhang et al.	Low scalability due to high I/O cost	MONDRIAN WITH MapReduce (MRMONDRIAN)	Anonymization
11.	Al Zobi et al.	Ignorance of generalizations	MDSBA (expanded K-anonymity algorithm)	
12.	Ferrer et al.	Overlapping populations and increase in quasi-identifiers	Advanced K-anonymity algorithm	
13.	Mehta et al.	Loss of information	Improved scalable l-diversity	
14.	Cui et al.	Confidentiality of shared data	Attribute-based storage system	

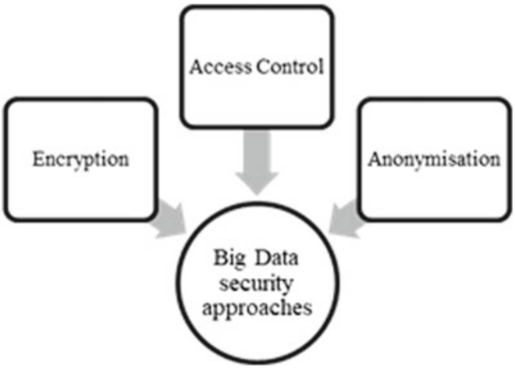


Fig. 3 Approaches towards big data security

Security by encryption—Enabling only the authorized user's access to the information by encoding the information is known as encryption. Li et al. proposed an algorithm to avoid cloud operators reaching the user's sensitive data. It is the amalgamation of AD2, SED2, and efficient data conflation algorithms entitled as security-aware efficient distributed storage model [18]. Aljawarneh et al. proposed a system for multimedia big data against real-time tampered data attacks. The proposed scheme is made by merging the Feistel network, AES, S-Box, and genetic algorithm. The scheme is applied over the data set of JUST university hospital [19]. Yan et al. proposed a scheme based on deduplication of encrypted data and proxy re-encryption. Deduplication is an important practice to achieve successful cloud storage, especially for big data storage. It allows only the authorized users to access the information. It supports flexible data updates offline as well [20]. Dong et al. presented a scheme for heterogeneous ciphertext transformation. It is a proxy algorithm that works on a virtual-based monitor which provides support for the realization of system functions. It is designed to protect and secure user's data effectively. It also provides the data owner the total control over their data for modern information security [21].

Security by access control—One of the most important security components is access control systems. Due to misconfiguration of the access control policies, the security and privacy of the system are often compromised. Hu et al. have proposed a scheme for distributed big data processing clusters. The scheme aims to authorize the protection of big data processing from internal attacks [22]. Wnorong et al. have introduced the mechanism for content access. The proposed mechanism is very suitable for the content-sharing of information in big data. CBAC is used for access control decisions based on semantic similarity between the requester's credentials and the content [15]. Siffah et al. proposed an off-chain-based sovereign blockchain where transactions are made between parties through a virtual container. Then blockchain network is used to store the output [23]. Kumar et al. proposed a scheme based on ciphertext policy with an attribute—encryption along with less computation overhead [24]. Khuntia et al. proposed a scheme for privacy preserving in the cloud to ensure big data access control. To reduce computational overhead, authors have used the concept of multi-sharing here [25]. Jasim et al. proposed a three-tier approach including cloud architecture, transaction manager, and clients. Zero trust is the basis of communication between the models [26].

Security by anonymization of the data—Control over private information gathering and its usage is information privacy. The ability to stop information from becoming public either by a group or an individual is known as information privacy. The assimilation of private information over the Internet during its transmission is one of the issues faced by the users. Privacy protection is one of the most bothering issues in big data and cloud applications, so there is an urgent need for strong customer privacy preservation techniques. Data anonymization is one of the efficient and effective ways towards privacy preservation [27]. Zhang et al. proposed a technique based on MapReduce on the cloud. A combination of highly scalable median finding algorithm and histogram technique is used here to propose for achieving cost effectiveness. Scalability is also measured here using multivariate partitioning [28]. Zhang et al. have pointed out the scalability issue in the cloud over big data. For this, a hybrid approach of top-down specialization and bottom-up generalization is used. K-anonymity parameter with workload sharing is used for selecting the component to achieve a highly scalable environment if compared with the existing approaches [29]. Al Zobi et al. have proposed a novel framework MDSBA. According to the authors, the loss of important information is the result of the avoidable generalized identical details. Through the proposed scheme, authors have expanded the k-anonymity and applied the bottom-up approach to avoid the identical widespread records more methodically and efficiently [30]. Ferrer et al. have focused their work towards dealing with the two important issues while using k-anonymity, i.e. the quasi-identifier attributes and the data controllers attributes by proposing a k-anonymity algorithm that avoids the dimensionality problem and by using mean and median to avoid the risk of disclosure by replacing the generalization method with the alternative aggregation method which is comparatively less sensitive, respectively [31]. Cui et al. proposed a deduplication-based system for a hybrid cloud used for attribute-based storage. The authors also discuss the ways to achieve semantic security along with keeping in mind the context of confidentiality to share the data with other users [32]. Mehta et al. proposed a scheme with the name improved scalable l-diversity approach based on K-anonymization. The run-time of this scheme is very less, and the loss of information is also less in comparison with other schemes [33].

4 Conclusion

Data is increasing with each passing moment over the Internet, making it impossible for traditional approaches to deal with the data. Out of the available bulky and raw data, extracting the relevant information is the important task of big data analytics. However, while dealing with the data, security is the major threat that is being faced by the analysts. The present paper discusses some of the novel approaches that can be used to ensure the security of big data. Moreover, we have noted that all the present traditional schemes cannot be applied over big data, but with certain advancements, in the future, the schemes can be improved and applied.

References

1. Arora M, Bahuguna H (2016) Big data security—the big challenge. Int J Sci Eng Res 7(12)
2. Tarekegn GB, Munaye YY (2016) Big data: security issues, challenges and future scope. Int J Comput Eng Technol 7(4):12–24
3. Siddique M, Mirza MA, Ahmad M, Chaudhry J, Islam R (2018) A survey of big data security solutions in healthcare. In: International conference on security and privacy in communication systems, Aug 2018. Springer, Cham, pp 391–406
4. Bhadani AK, Jothimani D (2016) Big data: challenges, opportunities, and realities. In: Effective big data management and opportunities for implementation. IGI Global, pp 1–24
5. Elgendy N, Elragal A (2014) Big data analytics: a literature review paper. In: Industrial conference on data mining, July 2014. Springer, Cham, pp 214–227
6. Zuech R, Khoshgoftaar TM, Wald R (2015) Intrusion detection and big heterogeneous data: a survey. J Big Data 2(1):1–41
7. Ruiz-Rosero J, Ramirez-Gonzalez G, Williams JM, Liu H, Khanna R, Pisharody G (2017) Internet of things: a scientometric review. Symmetry 9(12):301
8. Sagiroglu S, Sinanc D (2013) Big data: a review. In: 2013 international conference on collaboration technologies and systems (CTS), May 2013. IEEE, pp 42–47
9. Mujawar S, Kulkarni S (2015) Big data: tools and applications. Int J Comput Appl 115(23):7–11
10. Joseph Charles P, Carol I, MahaLakshmi S (2018) Big data security—an overview. IRJET 11(2)
11. Tarekgen GB, Munaye YY (2016) Big data: security issues, challenges and future scope. IJCET 4(7):12–24
12. Sisense. https://www.sisense.com/glossary/big-data-security. Accessed 2019/12/03
13. Joseph A, Cherian M (2018) The quest for privacy and security in various big data applications: a survey. IJCESR 3(5):1–8
14. Lafuente G (2015) The big data security challenge. Netw Secur 2015(1):12–14
15. Begoli E, Horey J (2012) Design principles for effective knowledge discovery from big data. In: 2012 joint working IEEE/IFIP conference on software architecture and European conference on software architecture, Aug 2012. IEEE, pp 215–218
16. Xu L, Jiang C, Wang J, Yuan J, Ren Y (2014) Information security in big data: privacy and data mining. IEEE Access 2:1149–1176
17. Zhang D (2018) Big data security and privacy protection. In: 8th international conference on management and computer science (ICMCS 2018), vol 77, Oct 2018. Atlantis Press, pp 275–278
18. Li Y, Gai K, Qiu L, Qiu M, Zhao H (2017) Intelligent cryptography approach for secure distributed big data storage in cloud computing. Inf Sci 387:103–115
19. Aljawarneh S, Yassein MB (2017) A resource-efficient encryption algorithm for multimedia big data. Multimed Tools Appl 76(21):22703–22724
20. Yan Z, Ding W, Yu X, Zhu H, Deng RH (2016) Deduplication on encrypted big data in cloud. IEEE Trans Big Data 2(2):138–150
21. Dong X, Li R, He H, Zhou W, Xue Z, Wu H (2015) Secure sensitive data sharing on a big data platform. Tsinghua Sci Technol 20(1):72–80
22. Hu VC, Ferraiolo D, Kuhn R, Friedman AR, Lang AJ, Cogdell MM, Scarfone K (2013) Guide to attribute based access control (ABAC) definition and considerations (draft). NIST Spec Publ 800(162):1–54

23. Zeng W, Yang Y, Luo B (2013) Access control for big data using data content. In: 2013 IEEE international conference on big data, Oct 2013. IEEE, pp 45–47
24. Xia QI, Sifah EB, Asamoah KO, Gao J, Du X, Guizani M (2017) MeDShare: Trust-less medical data sharing among cloud service providers via blockchain. IEEE Access 5:14757–14767
25. Khuntia S, Kumar PS (2018) New hidden policy CP-ABE for big data access control with privacy-preserving policy in cloud computing. In: 2018 9th international conference on computing, communication and networking technologies (ICCCNT), July 2018. IEEE, pp 1–7
26. Jain P, Gyanchandani M, Khare N (2016) Big data privacy: a technological perspective and review. J Big Data 3(1):1–25
27. Jasim AC, Tapus N, Hassoon IA (2018) Access control by signature-keys to provide privacy for cloud and big data. In: 2018 5th international conference on control, decision and information technologies (CoDIT), Apr 2018. IEEE, pp 978–983
28. Zhang X, Qi L, Dou W, He Q, Leckie C, Ramamohanarao K, Salcic Z (2017) Mrmondrian: scalable multidimensional anonymisation for big data privacy preservation. IEEE Trans Big Data
29. Zhang X, Liu C, Nepal S, Yang C, Dou W, Chen J (2014) A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud. J Comput Syst Sci 80(5):1008–1010
30. Al-Zobbi M, Shahrestani S, Ruan C (2016) Sensitivity-based anonymization of big data. In: IEEE 41st conference on local computer networks workshops (LCN workshops). IEEE, pp 58–64
31. Domingo-Ferrer J, Soria-Comas J (2016) Anonymization in the time of big data. In: International conference on privacy in statistical databases, Sept 2016. Springer, Cham, pp 57–68
32. Cui H, Deng RH, Li Y, Wu G (2017) Attribute-based storage supporting secure deduplication of encrypted data in cloud. IEEE Trans Big Data 5(3):330–342
33. Mehta BB, Rao UP (2019) Improved l-diversity: scalable anonymization approach for privacy preserving big data publishing. J King Saud Univ Comput Inf Sci