An Efficient Security for Unstructured Big Data using a Reconfigurable Security Suite

Parashiva Murthy B.M.
Assistant Professor
Department of CSE, SJCE, JSSSTU
Mysuru, India, Karnataka - 570006

Sumithra Devi K.A., PhD
Dean Academics & Head
Department of ISE,DSATM
Bengaluru, India, , Karnataka - 560062

ABSTRACT

The unstructured data security is enhanced using a reconfigurable security suite (RSS). The data node security is improved by seeing categories of data & their levels of sensitivity. The efficiency of the system performance is improved by using classification of data on par with the sensitivity levels. Methods: Adequate security is provided to the unstructured data by bearing in mind the various data nodes & their sensitivity. The proposed reconfigurable security suite effectively classifies the data nodes further into adequate security nodes and also enhances the security system overhead. Finding: performance analysis has been carried out on different data types by considering any one of the parameters in common like service code and sensitive code in different algorithms. The proposed reconfigurable security suite is developed by analysis performance of oracle Exadata and Apache mahout on sensitive, confidential and public data. Novelty: the reconfigurable security suite provides the different types of security services, which include each class of data standards and algorithms. The proposed security suite is developed by considering the mean value of sensitive, confidential and public data nodes etc to identify the security suite overhead.

Keywords

Big Data, Oracle Exadata, Apache Mahout, Reconfigurable security suite.

1. INTRODUCTION

Big Data's technologies have been present in enormousvolume, velocity & variety. Various types of R & D tasks are done on Big Data. Currently the investigation on security of Big Data is in the starting phase. Nevertheless, in our inference as such, no precise method to deliver security to Big Data is there. Hence, a method has been established to deliver Big Data security by bearing in mind the diverse kinds of them. In the planned structure the prevailing standards for diverse security services could be combined to provide security to it. We deliberated 2 dimensions of it namely volume &variety. From the time when we could deliver security by seeing these 2dimensions, the velocity can be controlled using parallel programming. In the area of Big Data considering its bulk we see data variations like unstructured, semi-structured, structured etc. Structured data's security could be deliberated by prevailing standards of security by means of SQL queries. Hence, the attention is to deliver security to data which is unstructured that comprise text, video, audio, XML, image and so on. This method designatesexamination of the data which is unstructured by means of data analytics skills, construct a data node of databases that comprises dissimilar kinds of information like text, image, video, XML, electronic mail, audio etc., subsequent phase constructs a security group to deliver security. The analytics of Datacould be finished utilizing dissimilar kinds of skills that is deliberated in the subsequent segment. Post examining the diverse kinds of information, additional study is completed to categorize the information to acquire levels of sensitivity consulting to the security criterions are nominated. To sum up a scheduling algorithm interacts through the security set & deliver suitable security to correct kind of data by seeing the data's sensitivity level.

2. THE PROPOSED RECONFIGURABLE SERCUTIY SUITE MODEL AND ALGORITHM

We observe that Big Data comprises both structured & unstructured data. By examination one could distinct structured data by means of SQL queries and offer security. Numerous kinds of data could be established inside data which is unstructured & hence giving security is a tough job. Subsequently the data which is unstructured comprises categories of data such asvideo,image, text, XML, e-mail etc., later examining the facts we could construct databases that holds dissimilar types of information. A procedure interacts the data node having a security system that has numerous security functions to offer safekeeping to data. We premeditated numerous methods of data analytics procedures & planned a method to deliver security to data which is unstructured by means of the security suite. Hence our method comprises two procedures, first is to perform data analytics and second constructs a security suite. The stage of Data analytics comprises data filtering, clustering & categorization that provides capability to generate an information of databases. We have shown the opinions of the planned method in Figure no.1.

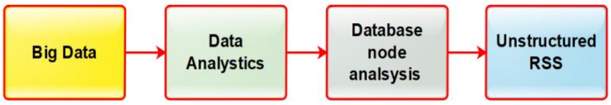

Fig.1. Functional block diagram of proposed Unstructured RSS

3. ANALYTICS OF BIG DATA

In order to do analytics of Big Data, numerous methods could be utilized. The styles of analytics systems for unstructured facts is defined. This could be investigated with numerous methods. Oracle has capability to examine unstructured data by means of Oracle Big Data Appliance. The procedure begins with filtering of data by means of Hadoop and supplies facts using the assistance of Big Data Appliance that contains software tool is Oracle loader for Hadoop. This method is portrayed in figure.2.

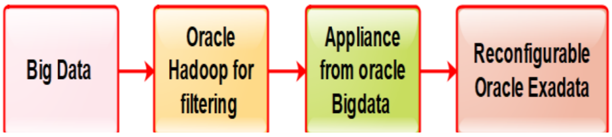
Big Data
Oracle Hadoop for filtering
Appliance from oracle Bigdata
Reconfigurable Oracle Exadata

Fig.2. Fundamental block diagram of data analysis using reconfigurable oracle model.

We see that Oracle loader is similarly utilized to add data to Oracle Exadata. This has infrastructural tools which offer improved interoperability through Hadoop&quicker movement of file. We also observe that Exadata is considered to be a data analytics hub and here diverse kinds of data like image, semantic graphs, text, XML, spatial etc. is categorized and stored. InfoSphere BigInsights of IBM is a platform of analytics utilizing Apache Hadoop which is of open source. It comprises analytics which is built-in that comprises of machine data accelerator of analytics, Big R, accelerator of social data analytics etc. A query language namely Jaql is for JSON. It is presently utilized in InfoSphere Biginsigts. By means of Jaql one could contact &add data through various bases resembling HDFS, HBase, web, local file system, twitter, etc. Apache Mahout is a considered to be one of the production- level platform which is also open-source for conducting Big Data analytics. It is a ML library that comprises cooperative filtering, clustering and classification algorithms. Filtering is done agreeing to operator along with item dependent recommenders and matrix factorization dependent recommenders. We also see that Clustering algorithms namely fuzzy K-means, Kmeans & classifiers like logistic regression. The procedure of data analysis by means of the Apache Mahout is revealed in Figure no.3.

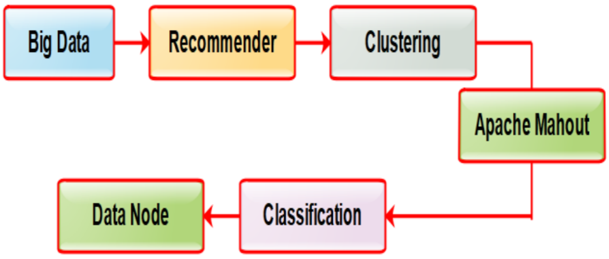

Fig.3. Fundamental block diagram of data analysis by means ofApache Mahout

4. CATALOGUING AS PER SENSITIVITY LEVEL

Post getting dissimilar kinds of data we categorize every kind of data as public, delicate &confidential. Sensitive data comprise data that are secured using privacy guidelines &data secured by sensitivity pacts. Information is represented as public once illegal disclosure output few or zero risk to the atmosphere where utilization of the data is done. Subtle information is valued and requires maximum of security. To offer security towards such course of data, sturdiest security procedures need to be utilized. The data which is Confidential takes intermediate phase of sensitivity and requires security procedure having decent speed of processing. Public data would be exposed for everyone or provide controller for registered users by means of id and password. The data is classified on the basis of sensitivity levels as shown in Table .1 where High security (HS) is most sensitivity data,more security (MS), Security (S), Confidencial(C) and classified (CI).

Table.1 various sensitivity level classes

Class of data	CI	C	S	MS	TS
Required Security Level	0.3	0.45	0.69	0.85	0.96

Table 2. conventional Algorith and its security serivies

Secuity services	**Variours conventional algorithm with respect to sensitivilty level**				
	HS	**MS**	**S**	**C**	**CI**
Privacy	3DES	Snefu-256	CCM	S/MIME	XML Enc
Integrity	DES	Tiger	HMAC-SHA-1	OpenPGP	XML DSig
Authenticity and integrity	UD	UD	UD	UD	SAML

Nevertheless, if we can categorize them as per the sensitivity levels, we could give suitable security amenities to necessary course of data. There are various facilities for data security like digital signature for authentication, cryptographic system for confidentiality, hash function to give data integrity, system with MAC that gives user authentication along with integrity of data & access control system for giving security as per access rights of the consumers to data. UD in Table.2 & Table.3 shows undefined & connected to public class of data. Figure .4 shows the Data analysis with respect to sensivitty Level.

Table 3. variours sensitivity level with respect to percentage data in the clould

Class of data		**CI**	**C**	**S**	**MS**	**TS**
Data in the clould (%)	**D1**	21	11	26	31	16
	D2	11	21	26	26	21
	D3	6	11	41	31	16
	D4	32	31	21	11	11
	D5	11	16	21	31	26
	D6	6	6	16	31	41
	D7	6	6	11	31	51

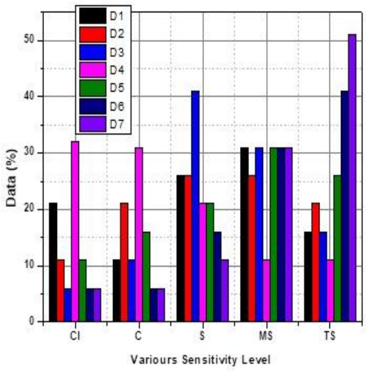

Fig. 4.Data analysis with respect to sensivitty Level

5. PROPSOED RECONFIGURABLE SECURITY SUITE

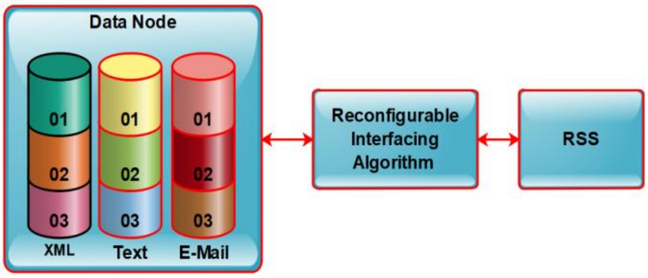

Fig.5. Proposed block diagram of node data analysis using proposed reconfigurable Security Suite.

Figure.5 shows the data node analysis using proposed RSS.Here the best approach has been selected on basis of response on security with respect to the sensitivity level.

Here a security system is constructed to deliver essential & sufficient security to facts. The suite comprises4 large units concerning security features; Initial one is user identification & authentication that comprises digital signature, second relates to confidentiality that has encryption & decryption procedures, the ternary is for integrity contains functions of hash and next is the fourth one for authentication & integrity, that comprises MAC structures & the final one is for access regulator systems. Every unit is subsequently divided into 3 segments that are for 3 data classes as per the sensitivity level. Here is a setting up process that proceeds choice to stimulate suitable security services from the designated unit and give satisfactory security as per the level of sensitivity. The elaborate security view is depicted in Figure.6

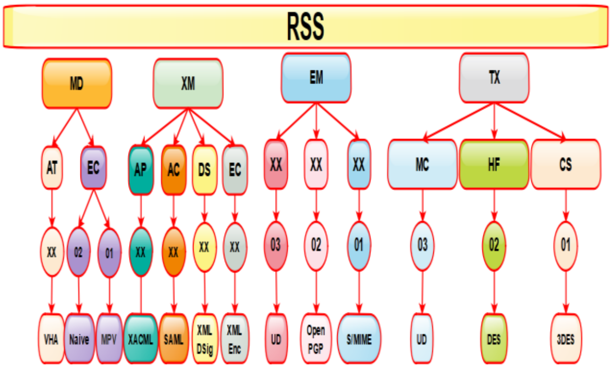

Fig.6. Performance security analysis with respect to sensitivity level using proposed RSS

To construct the security group there is utilized mask for every facility like CS, HF & MC aimed at the privacy, integrity and authenticity with integrity correspondingly. To offer security to every information the system links code related through it & chose procedure after security system. Electronic-mail is considered to be the best extensively cast-off Internet service. PGP is maximum generally utilized criterions established by Phil Zimmerman. PGP comprises authentication & confidentiality of the communication beside the key management. S/MIME is a customary for security improvement to email of MIME. This stood established by RSA data security Inc. On the other hand, digital certificates are utilized in PGP & S/MIME for key management. The security criterions to offer safety to e-mails is depicted in Table 4.

Table 4. Algorithm sensitivity response for Text data.

Code type	Algorithm	Serive code	Sensitivity Code
TX	3DES	CS	01
	DES		02
	UD		03
	Snefru-256	HF	01
	Tiger		02
	UD		03
	CCM	MC	01
	HMAC-SHA-1		02
	UD		03

Digital Signature of XML is utilized to deliver integrity of communication, non-repudiation& authentication. This might be cast-off to sign XML assets along with library resources like JPEG file. SAML is utilized to offer authentication, attribute and authorization evidence. XKMS is a practice planned as a customary preserved by the W3C. It delineates a path to catalogue the public keys & circulation of keys utilized by the XML_SIG requirement. Security could be provided to XML forms having the criterions depicted in Table 5.

Table 5. common security service code for various algorithm in TEXT Data

Data Type	Algorithm	Serive code	Sensitivity Code
EM	S/MIME	XX	01
	OpenPGP	XX	02
	UD	XX	03

The susceptibility of copyright multimedia comes because of copying & modification of content. Hence, the shield & authentication are important. Commonly, digital water marking is broadly utilized method to resolve copyright protection issue of multimedia data in network background. Numerous applications are available for accessing watermarking. VHA is used for authentication. Encryption is utilized to preserve privacy of the video. MP-secure encryption system is essentially AES procedure in a secure mode. Hence, we deliberated VHA for authentication.

6. EFFICIENCY AND SECURITYCOMPARITION ANALYSIS OF VARIOUS FIELDS

Table No. 6 depicts the analysis of the performance of various field with respect to percentage of data in each class. Figure no.7 depicts the analysis of performance of various field response with respect to data in percentage.

Table 6. Analysis of performance of various fields in data percentage

Fields	Public (%)	Confidential (%)	Sensitive (%)
Financial	28	18	62
Research	42	32	32
Software developer	62	28	16
Health care	32	32	41
Educational	67	16	22
Real estate	62	28	16
Mean value	**48.833**	**25.66**	**31.5**

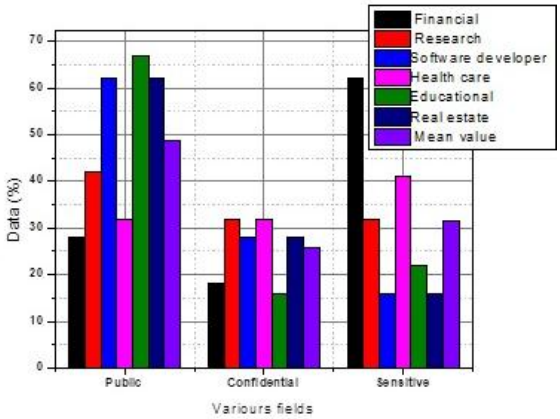

Fig. 7. Performance analysis of variurs field response with respect to data in percentage

Let D_i is represented as data probability, i is described has data class where i = 1, 2, 3,4,5…… . Hence,D_1 is the likelihood of information in sensitive class. As per the Table 6 the health care association information might have 41% data which is sensitive, 32% data which is confidential and32% public class data. Six fields have been deliberated and established 31% sensitive, 25% confidential and 48% public data on an average.

Let R depict the reconfigurable security suite for diverse kinds of security functions that comprise standards and algorithms connected to every data class. Assume O be a task for overhead of the security. In case R(O) =1, it implies the system gains complete overhead required for it. Assume Y_1 be the assessment required for the data security in ith class. Then Y_1 =1, because for the information in subtle class nearly all the facilities need to be utilized. Consider Y_2= 0.7, because to offer data security in confidential class all the facilities might not be needed. Public data assume V3 = 0.2. Now we can calculate R(O) as below:

$$\mathrm{R(O)} = Y_1D_1 + Y_2D_2 + Y_3D_3$$

$$= 1*0.31+0.7*0.25+0.2*0.48$$

$$= 0.31 + 0.175 + 0.096$$

$$= 0.31 + 0.175 + 0.096$$

$$\mathrm{R(O)} = 0.581 \approx 0.6$$

As per the above deliberation we got $\mathrm{R(O)} = 0.6$. Which implies, we require 0.6 of the security system overhead. Though we calculate the assessment for the monetary association, that take the maximum quantity of information in class 01 which is sensitive, we get R(O) for this field equivalent to 0.6. Hence the financial field requires 62% of the security system overhead. It could protect 41% dispensation time if the data are categorized as per the sensitivity level.

7. CONCLUSION

In this paper, Efficient security has been provided to the unstructured big data using proposed RSS compared to the existing security standards and algorithms with respect to the sensitively level. According to the performance analysis, the proposed RSS is capable of providing adequate security and improving the security of unstructured big data processing overheads of the security system. The proposed RSS has enhanced the average of processing overheads to 17% considering several fields on par to the sensitivity levels of the proposed work. The proposed work has been carried out by considering the standards of Apache Mahout and Oracle Exadatagiving security to the unstructured big data & analysis of the performance of a system with all standard data parameter.

Future Scope

In recent days, digitalization is adopted in all fields. due to this process, high security has to be provided to both structured and unstructured data to reduce the overheads in security systems. Hence, the proposed system has to be enhanced further so that it can be capable of handling the huge data in very short duration.

8. REFERENCES

[1] T. -l. Chasupa and W. Paireekreng, "The Framework of Extracting Unstructured Usage for Big Data Platform," 2021 2nd International Conference on Big Data Analytics and Practices (IBDAP), 2021, pp. 90-94, doi: 10.1109/IBDAP52511.2021.9552131.

[2] O. Baker and C. N. Thien, "A New Approach to Use Big Data Tools to Substitute Unstructured Data Warehouse," 2020 IEEE Conference on Big Data and Analytics (ICBDA), 2020, pp. 26-31, doi: 10.1109/ICBDA50157.2020.9289757.

[3] I. Taleb, M. A. Serhani and R. Dssouli, "Big Data Quality Assessment Model for Unstructured Data," 2018 International Conference on Innovations in Information Technology (IIT), 2018, pp. 69-74, doi: 10.1109/INNOVATIONS.2018.8605945.

[4] F. Hamami, I. A. Dahlan, S. W. Prakosa and K. F. Somantri, "Big Data Analytics for Processing Real-time Unstructured Data from CCTV in Traffic Management," 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020, pp. 1-5, doi: 10.1109/ICoDSA50139.2020.9212858.

[5] M. Elsayed, A. Abdelwahab and H. Ahdelkader, "A Proposed Framework for Improving Analysis of Big Unstructured Data in Social Media," 2019 14th International Conference on Computer Engineering and Systems (ICCES), 2019, pp. 61-65, doi: 10.1109/ICCES48960.2019.9068154.

[6] J. McHugh, P. E. Cuddihy, J. W. Williams, K. S. Aggour, V. S. Kumar and V. Mulwad, "Integrated access to big data polystores through a knowledge-driven framework," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 1494-1503, doi: 10.1109/BigData.2017.8258083.

[7] K. Ghane, "Big Data Pipeline with ML-Based and Crowd Sourced Dynamically Created and Maintained Columnar Data Warehouse for Structured and Unstructured Big Data," 2020 3rd International Conference on Information and Computer Technologies (ICICT), 2020, pp. 60-67, doi: 10.1109/ICICT50521.2020.00018.

[8] D. Cansell, J. P. Gibson, and D. Méry, "Refinement: A Constructive Approach to Formal Software Design for a Secure e-voting D. Wu, "A big data analytics framework for forecasting rare customer complaints: A use case of predicting MA members' complaints to CMS," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 3965-3967, doi: 10.1109/BigData.2017.8258406. Interface," *Electron. Notes Theor. Comput. Sci.*, vol. 183, pp. 39–55, Jul. 2007, doi: 10.1016/j.entcs.2007.01.060.

[9] L. Xianglan, "Digital construction of coal mine big data for different platforms based on life cycle," *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, 2017, pp. 456-459, doi: 10.1109/ICBDA.2017.8078862..

[10] Y. Cui, S. Kara and K. C. Chan, "Monitoring and Control of Unstructured Manufacturing Big Data," 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2020, pp. 928-932, doi: 10.1109/IEEM45057.2020.9309975.

[11] K. Adnan, R. Akbar and K. S. Wang, "Towards Improved Data Analytics Through Usability Enhancement of Unstructured Big Data," 2021 International Conference on Computer & Information Sciences (ICCOINS), 2021, pp. 1-6, doi: 10.1109/ICCOINS49721.2021.9497187.

[12] S. Yadav, G. Kumar and S. Kumar, "A graph construction study for graph-based semi-supervised learning: Case study on unstructured text data," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 6254-6256, doi: 10.1109/BigData47090.2019.9006465.

[13] Shivaji, R., Nataraj, K.R., Mallikarjunaswamy, S., Rekha, K.R. (2022). Implementation of an Effective Hybrid Partial Transmit Sequence Model for Peak to Average Power Ratio in MIMO OFDM System. ICDSMLA 2020. Lecture Notes in Electrical Engineering, vol 783. Springer, Singapore. https://doi.org/10.1007/978-981-16-3690-5_129.

[14] Mallikarjunaswamy, S., Sharmila, N., Siddesh, G.K., Nataraj, K.R., Komala, M. (2022). A Novel Architecture for Cluster Based False Data Injection Attack Detection and Location Identification in Smart Grid. In: Mahanta, P., Kalita, P., Paul, A., Banerjee, A. (eds) Advances in Thermofluids and Renewable Energy . Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-16-3497-0_48

[15] Mallikarjunaswamy, S., Sharmila, N., Siddesh, G.K., Nataraj, K.R., Komala, M. (2022). A Novel Architecture for Cluster Based False Data Injection Attack Detection and Location Identification in Smart Grid.Advances in Thermofluids and Renewable Energy . Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-16-3497-0_48..

[16] X. Ge, X. Zhang and P. K. Chrysanthis, "ExNav: An Interactive Big Data Exploration Framework for Big Unstructured Data," 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 503-512, doi: 10.1109/BigData50022.2020.9377741.

[17] L. Yao et al., "Index Method of Unstructured Data in Power System Based on Improved B+ Tree," 2021 International Conference on Wireless Communications and Smart Grid (ICWCSG), 2021, pp. 574-577, doi: 10.1109/ICWCSG53609.2021.00122.

[18] X. Deng, "Big data technology and ethics considerations in customer behavior and customer feedback mining," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 3924-3927, doi: 10.1109/BigData.2017.8258399..

[19] Manjunath T. N., Mallikarjunaswamy S, " An efficient hybrid reconfigurable wind gas turbine power management system using MPPT algorithm,"2021,pp 2501-2510, doi:10.11591/ijpeds.v12.i4.pp 2501-2510.

[20] Mallikarjunaswamy, S., Nataraj, K.R., Rekha, K.R. (2014). Design of High-Speed Reconfigurable Coprocessor for Next-Generation Communication Platform. Emerging Research in Electronics, Computer Science and Technology. Lecture Notes in Electrical Engineering, vol 248. Springer, New Delhi. https://doi.org/10.1007/978-81-322-1157-0_7.

[21] M. Lokanan, "Coding and Analytical Problems with Big Data When Conducting Research on Financial Crimes," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 5386-5388, doi: 10.1109/BigData.2018.8621976.

[22] S. Awaghad, "SCEM: Smart & effective crowd management with a novel scheme of big data analytics," 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 2000-2003, doi: 10.1109/BigData.2016.7840822.

[23] S. K. Sahu, M. M. Jacintha and A. P. Singh, "Comparative study of tools for big data analytics: An analytical study," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 37-41, doi: 10.1109/CCAA.2017.8229827.

[24] Q. Tan, "Research on E-Commerce Security and Data Analysis Platform in the Era of Big Data," 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2020, pp. 410-414, doi: 10.1109/MLBDBI51377.2020.00087.

[25] M. Kantarcioglu and F. Shaon, "Securing Big Data in the Age of AI," 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), 2019, pp. 218-220, doi: 10.1109/TPS-ISA48467.2019. 00035.