RESEARCH ON EMERGING DEVELOPMENTS IN DATA PRIVACY AND SECURITY IN THE TECHNOLOGICALLY ADVANCED CORPORATE SECTOR

[1]Muthu Dayalan
Senior software Developer & Researcher
Chennai,Tamilnadu

Abstract—This research focused on the emerging developments in data privacy and data security in the technologically advanced corporate world. With the improvement in networking technology, there is an increase in the amount of online transactions, and a huge amount of data being exchanged through the internet. There is an increased threat to the data security and privacy, which has led to development of more advanced data security and privacy measures. Four notable emergent developments have been addressed in this paper. The first one is hiding a needle in a haystack, a privacy-preserving Apriori algorithm in the Mapreduce Framework that modifies the original data through adding noise, and second is differential privacy that enables the researchers and database analysts to get vital information from the concerned database, which contains personal information. They also include the user behavior analytics, which tracks the user behavior to identify any kind of threat to the database systems; and fourth, the HybrEx, which allows only the safe public cloud operations, is executed, during the organizational private cloud execution.

Index Terms- Data privacy, data security, emerging developments, technologies,

INTRODUCTION

With the rising social networking technologies and firms, which are increasing the focus on the personalized experience of the customers in the online sphere, data privacy and security has become a critical issue for numerous industries. Traditionally, passwords were used as a means of data privacy and security. However, these means has become obsolete. The risk of data breaches is very high, necessitating more advanced data security and privacy techniques, such as biometric analysis. The emerging technologies and techniques are playing a critical role in revolutionizing the manner in which individuals, governments, and companies address the data security concerns. There are also new legal enactments, which have been adopted as a means of enhancing the data privacy and security. This paper conducts extensive research on the emerging developments in the data privacy and security in the technologically advanced corporate sector.

EMERGING DEVELOPMENTS IN DATA PRIVACY AND SECURITY

Emerging technological developments and discoveries are considered to contribute significantly to the improved data privacy and security protection [15]. This incorporates the reduction of risk and improved fraud detection. Before evaluating the various emerging developments in data privacy and security, it is vital to understand concepts of data privacy and security [4]. The two terms go hand in hand, though they are different concepts. Data security implies the technical and physical requirements, which help in the protection of unauthorized entry to the data system, and the help in maintenance of the data integrity. Data privacy, on the other hand, revolves around the confidentiality of the data and the individual rights of the concerned individuals. It is also concerned with how the data is used, the users and concerned legality [30]. To achieve an improved data management, there have been several emerged technological developments. Some of them are discussed below.

HybrEx

Hybrid execution model (HybrEx) is a data confidentiality, security and privacy applied in the cloud computing [23]. Using the model, only the safe public cloud operations are executed, during the organizational private cloud execution. This implies that public clouds are utilized only for the non-sensitive data and organizational clouds computation [22]. The model, on the other hand, utilizes the organization's private clouds for conducting the sensitive, private, data and computations. Therefore, data safety is critically considered before job's execution, in addition to the provision of integration with safety. Four categories of HybrEx MapReduce utilize new kind of applications that use public and private clouds [19]. These are;

The Map hybrid – it contains the map and reduces phases, where the map phase is executed both in public and private clouds, while the reduce phase is executed in the cloud phase [31]. This is shown in the figure below.

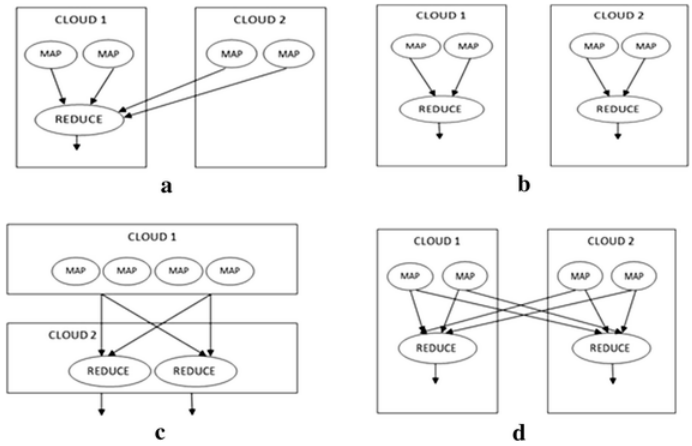

Figure 1: HybrEx methods: a map hybrid b vertical partitioning c horizontal partitioning d hybrid. Showing four categories of HybrEx MapReduce enabling new kinds of application to utilize public and private clouds

The Vertical positioning – as shown in the figure above, vertical positioning involves executing map and tasks in the public cloud by the use of public data as the input [32]. Immediate data is shuffled and result stored in the public cloud [5]. Similarly, the same is done for private data with the private cloud. The two activities are conducted in isolation.

The Horizontal Positioning - As shown in figure 1, the execution of the map phase is done at the public clouds, while the reduce phase is executed at the private cloud [9]

The Hybrid – In this case, the map and reduce phase are executed together at the private and public clouds. Data could also be transmitted among the clouds.

By the use of HybridEx, the full integrity and quick integrity check models are applied. However, it is important to note that HybridEx does not involve the use of key obtained from the public and private clouds during the map phase [8]. It deals with the map phase as an adversary [16].

User Behavior Analytics

User behavior is an emergent development in data privacy and security, which tracks the user behavior to identify any threat to the database systems. In case the username or password of someone is compromised, the attacker can engage in all kinds of malicious behavior. These behaviors are prone to trigger a red flag to the system defenders if the User Behavior Analytics (UBA) is employed. This technology employs the big data analytics, for identification of the anomalous behavior by the user. The user activity is a central concern to the security professionals [18]. This technology is a vital development in the data privacy and security, as it addressed the blind spot in the enterprise security [11]. In the expansion, what follows after a hacker gains access to an enterprise database. Among the first thing they do is to compromise the credentials. Therefore, the question, which arises, is whether it is possible to differentiate between a legitimate user activity, and an attacker, who has gained access to the database and compromised the credentials of the legitimate user [26].

The visibility of the activities, which are different from the norm of a legitimate user, can raise alarms in the middle of the hackers attack chain [17]. If the attack chain is considered to include the initial penetration, the lateral movement, the compromise, theft, and then the exfiltration of the sensitive data, the linking between these attack chains could be explicitly visible to the enterprise security pros. This has precipitated the development of the user behavior analytics [14].

It is vital to note that the comparison of the present and past user behavior is not the only way of UBA identifying the malicious actor. It also incorporates a technique referred to as the 'peer analyses [10]. It involves the comparison of the behavior of an individual, in comparison to the people under the same manager or department [29]. This portrays a clear indication that the concerned person is doing something they are not supposed to be doing, or someone else has taken up their account [1]. Further, it is a valuable tool for training employees in an organization better security measures. Among the critical issues in an organization is the employees' failure to follow the firm's policy. Therefore, it is important to identify those people and mitigate the risk through training them on crucial security issues.

Differential Privacy

Differential privacy is a new technological development enabling the researchers and database analysts to get vital information from the concerned database, which contains personal information. This is done without revealing the personal identities of the concerned individuals [12]. To do this, a minimum distraction is introduced in the provided information within the database system. The distraction introduced is quite small in a way that the information provided by the analyst remains useful and large enough in a manner that the privacy is effectively protected.

In the 90s, the Commonwealth of Massachusetts Group Insurance Commission (GIC) availed anonymous health records of their clients, to be used in research for the benefit of the society. The commission hid some personal information such as name street address, for protecting the privacy of its clients [18]. However, Latanya Sweeney (a PhD. student by then at MIT) managed to identify the health records through comparing the GIC database and voter database [27]. This made it clear that hiding some information is not a guarantee of protecting the individuals' identity [2].

The differential privacy provides a solution to the above problem. As shown in figure 2 below, analysts do not get direct access to the database having persona data. There is an intermediary software, which is put on the analyst and database for privacy protection. This software is referred to as the privacy guard.

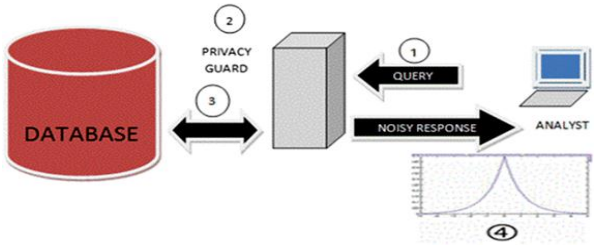

Figure 2: Differential Privacy (DP) acting as a solution to privacy protection

The whole process of inquiry involves some steps. First, the analyst through the intermediary privacy guard makes a query to the database [18]. The privacy guard for privacy risk evaluates the query. The privacy guard retrieves the answer of the query from the database, adds some distortion to ensure privacy and gives it to the analyst [3]. The amount of distortion added to the data is according to the evaluated risk. If the risk is low, little distortion is added in a manner that the quality of the answer is not affected [21]. However, they are large enough to effectively protect the privacy of individuals in the database.

Hiding a Needle in a Haystack

This is a privacy-preserving Apriori algorithm in the Mapreduce Framework. It modifies the original data through adding noise. The original work is maintained I the noise transaction, with the objective of preventing the data utility deterioration, while at the same time, preserving the privacy violation [7]. Though there is the risk of association rule leakage, there is adequate privacy protection through since the algorithm is based on the principle of 'hiding a needle in the haystack' the concept relies on the idea that it is difficult to find the data, as it is to find a needle in the haystack, as shown in figure below [13].

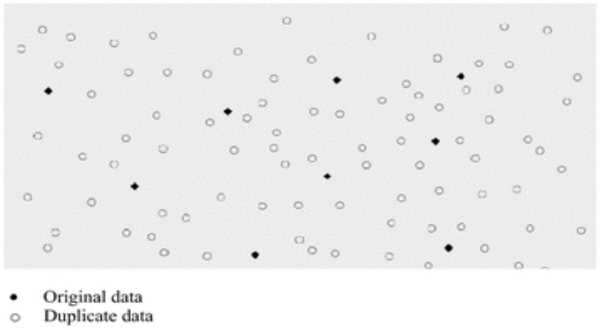

Figure 3: 'hiding a needle in the haystack' concept

However, the noise is not added haphazardly. There is the consideration of the privacy-data utility trade-off, where an additional coat is incurred in adding the noise, which makes "haystack" to hide the "needle." In figure 3 above, the black dots represent the original association rule, while the empty circles represent the noised association rule. The many noises association rule makes it hard to reveal the original rules.
The noise is added at the initiation of the transactions [19]. The service provider adds the dummy variable to the original data collected. There is unique code, which is assigned to the dummy and the original data. The service provider would then maintain the code information so that he can be in a position to filter out the dummy item from the original characters. The Apriori algorithm is executed by the external platform applying the data, which has been sent by the service provider.

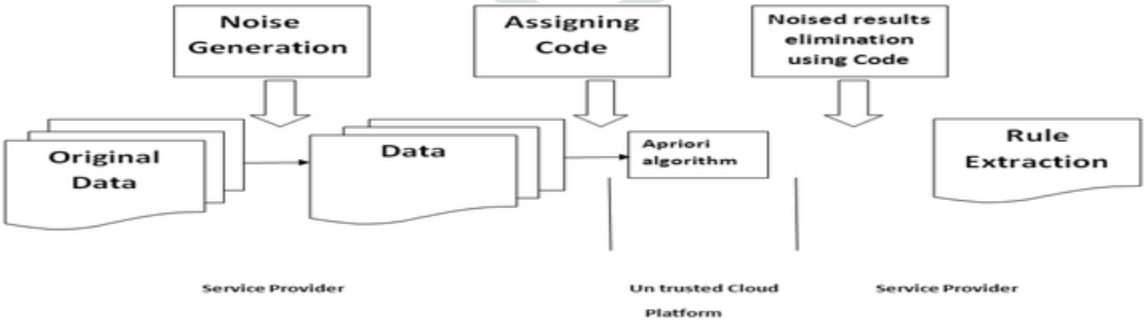

Figure 4: the process of adding a dummy item as noise to the original transaction data by the service provider

The frequent items which are affected by the dummy item are filtered by the code so that the correct association rule is extracted with the help of the frequent item established without the dummy item [6]. The association rule extraction process is not complicated because the amount of calculation needed for association rule extraction is limited [20].

CONCLUSION

It is conclusive from the research that with the advancement in technology, data, both big and small is prone to all sorts of attacks, from viruses to hackers. Similarly, there has been continuous improvement in research and discoveries of new technological developments in the data

privacy and security. Though there is wide range of techniques applied in different sectors and varying firms, this study has addressed four critical emergent development. These include the: hiding a needle in a haystack, privacy preserving Apriori algorithm in the Mapreduce Framework that modifies the original data through adding noise, and differential privacy that enables the researchers and database analysts to get vital information from the concerned database, which contains personal information. They also include the user behavior analytics, which tracks the user behavior to identify any threat to the database systems; and HybrEx, which allows only the safe public cloud operations, are executed, during the organizational private cloud execution.

REFERENCES

[1] Abadi DJ, Carney D, Cetintemel U, Cherniack M, Convey C, Lee S, Stone-braker M, Tatbul N, Zdonik SB. Aurora: a new model and architecture for data stream manag ement. VLDB J. 2003;12(2):120–39.

[2] Bredereck R, Nichterlein A, Niedermeier R, Philip G. The effect of homogeneity on the complexity of k-anonymity. In: FCT; 2011. p. 53–64.

[3] Cheng H, Rong C, Hwang K, Wang W, Li Y. Secure big data storage and sharing scheme for cloud tenants. China Commun. 2015;12(6):106–15.

[4] Fong S, Wong R, Vasilakos AV. Accelerated PSO swarm search feature selection for data stream mining big data. In: IEEE transactions on services computing, vol. 9, no. 1. 2016.

[5] Gantz J, Reinsel D. Extracting value from chaos. In: Proc on IDC IView. 2011. p. 1–12.

[6] Gudipati M, Rao S, Mohan ND, Gajja NK. Big data: testing approach to overcome quality challenges. Data Eng. 2012:23–31.

[7] Hu J, Vasilakos AV. Energy Big data analytics and security: challenges and opportunities. IEEE Trans Smart Grid. 2016;7(5):2423–36.

[8] Jing Q, et al. Security of the internet of things: perspectives and challenges. Wirel Netw. 2014;20(8):2481–501.

[9] Han J, Ishii M, Makino H. A hadoop performance model for multi-rack clusters. In: IEEE 5th international conference on computer science and information technology (CSIT). 2013. p. 265–74.

[10] Jain P, Pathak N, Tapashetti P, Umesh AS. Privacy preserving processing of data decision tree based on sample selection and singular value decomposition. In: 39th international conference on information assurance and security (IAS). 2013.

[11] Kolomvatsos K, Anagnostopoulos C, Hadjiefthymiades S. An efficient time optimized scheme for progressive analytics in big data. Big Data Res. 2015;2(4):155–65.

[12] Liu C, Ranjan R, Zhang X, Yang C, Georgakopoulos D, Chen J. Public auditing for big data storage in cloud computing—a survey. In: Proc. of IEEE Int. Conf. on computational science and engineering. 2013. p. 1128–35.

[13] Liu C, Chen J, Yang LT, Zhang X, Yang C, Ranjan R, Rao K. Authorized public auditing of dynamic big data storage on cloud with efficient verifiable fine-grained updates. In: IEEE trans. on parallel and distributed systems, vol 25, no. 9. 2014. p. 2234–44**.**

[14] Li N, et al. t-Closeness: privacy beyond k-anonymity and L-diversity. In: Data engineering (ICDE) IEEE 23rd international conference; 2007.

[15] Ko SY, Jeon K, Morales R. The HybrEx model for confidentiality and privacy in cloud computing. In: 3rd USENIX workshop on hot topics in cloud computing, HotCloud'11, Portland; 2011.

[16] Lu R, Zhu H, Liu X, Liu JK, Shao J. Toward efficient and privacy-preserving computing in big data era. IEEE Netw. 2014;28:46–50.

[17] Meyerson A, Williams R. On the complexity of optimal k-anonymity. In: Proc. of the ACM Symp. on principles of database systems. 2004.

[18] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. In: Proc. 22nd international conference data engineering (ICDE); 2006. p. 24.

[19] Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. In: IEEE translations and content mining are permitted for academic research. 2016.

[20] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A. Big data: the next frontier for innovation, competition, and productivity. New York: Mickensy Global Institute; 2011. p. 1–137.

[21] Mell P, Grance T. The NIST definition of cloud computing. Natl Inst Stand Technol. 2009;53(6):50.

[22] Middleton P, Kjeldsen P, Tully J. Forecast: the internet of things, worldwide. Stamford: Gartner; 2013.

[23] Porambage P, et al. The quest for privacy in the internet of things. IEEE Cloud Comp. 2016;3(2):36–45.

[24] Qin Y, et al. When things matter: a survey on data-centric internet of things. J Netw Comp Appl. 2016;64:137–53.

[25] Samarati P. Protecting respondent's privacy in microdata release. IEEE Trans Knowl Data Eng. 2001;13(6):1010–27.

[26] Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory; 1998.

[27] Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analytics: a survey. J Big Data Springer Open J. 2015.

[28] Wei L, Zhu H, Cao Z, Dong X, Jia W, Chen Y, Vasilakos AV. Security and privacy for storage and computation in cloud computing. Inf Sci. 2014;258:371–86.

[29] Wang C, Wang Q, Ren K, Lou W. Privacy-preserving public auditing for data storage security in cloud computing. In: Proc. of IEEE Int. Conf. on INFOCOM. 2010. p. 1–9.

[30] Xiao Z, Xiao Y. Security and privacy in cloud computing. In: IEEE Trans on communications surveys and tutorials, vol 15, no. 2, 2013. p. 843–59.

[31] Xu L, Jiang C, Wang J, Yuan J, Ren Y. Information security in big data: privacy and data mining. IEEE Access. 2014;2:1149–76.

[32] Xu K, et al. Privacy-preserving machine learning algorithms for big data systems. In: Distributed computing systems (ICDCS) IEEE 35th international conference; 2015.

[33] Zhang Y, Cao T, Li S, Tian X, Yuan L, Jia H, Vasilakos AV. Parallel processing systems for big data: a survey. In: Proceedings of the IEEE. 2016.