# SECURITY AND PRIVACY IN BIG DATA ANALYTICS

**P.Shobha Rani[1], Vigneswari D[2]**

[1&2] Dept. of CSE., R.M.D Engineering College, Chennai

**Abstract**

The growing need for computing on bigdata is getting higher, the three basic dimensions of big data are (referred as "3V" challenges) high volume, variety and velocity. The other upcoming challenge in the area of bigdata is Veracity, which means the trustworthiness of the data that is how secure the data is received, stored, processed and transmitted. Hence this Veracity is becoming a new dimension in the bigdata era. In this paper, we present a survey on various techniques that impose security and privacy over the bigdata. We elaborate on the techniques like cryptographic algorithms, Privacy-Preserving Cosine Similarity Computing protocol (PCSC), Optimized Balanced Scheduling (OBS), a trapdoor function and mention the differences in terms of performance based on cost and time.

**Keywords-**big data, privacy, security, PCSC, OBS.

## I. INTRODUCTION

Big data is a term applied to the data sets whose size is beyond the ability of commonly used software systems in order to store, manage and process. Recently in today's data-centric world the big data processing and analytics have become critical to most of the applications like government and enterprises. In the past few years, the total amount of data generated by human has exploded increase 300 times form exabytes to octabytes. These data are created from various fields like scientific research, government, finance and business, social networks, photography, video, audio mobile phones etc.

Big data has many typical challenges mostly they are assisted as dimensions of big data. The basic challenges are referred to as "3 Vs" of big data namely – volume, velocity, variety. The volume which means the size of the data sets, velocity means the speed with which the data has been generated and variety states that the type of the data that is generated stored and processed. The data generated can be of structured, semi-structured and unstructured data. The other challenges which are showing light on big data are Value refers big data have great social value. Hence there arises the 4Vs model which is widely recognized as it discovers value from various data sets that have been generated.

One of the most important dimension in big data era is the Veracity – whether the data is protected securely and managed with privacy. Generally providing security to big data is a tedious process because of its large volume. Consequently the privacy and security requirements in big data are very high. There are various approaches that have been proposed and implemented for assuring security in big data. In this paper we present a survey on the different models in section II which addresses the security threats of big data and evaluate them based on the performance overhead and section III reveals the conclusion of this paper.

## II. BIG DATA SECURITY TECHNIQUES

### *A. Privacy-Preserving Cosine Similarity computing Protocol*

The privacy-preserving cosine similarity computing (PCSC)[10]can efficiently calculate the cosine similarity of two vectors without disclosing the vectors to each other. The protocol describes that we can directly calculate the cosine similarity in an efficient way. When we consider inter big data processing the direct cosine similarity computation (DCSC) would disclose each other's privacy. Hence we can apply homomorphic encryption (HE), such as Pallier encryption (PE) to provide privacy but this requires time-consuming exponentiation operations. So, PCSC protocol for big data processing is used based on the lightweight multi-party random masking and polynomial aggregation techniques which does not require time consuming operations.

The issue with this protocol is that it has computational overheads with the increase in length of

the vector. Compared to DCSC protocol. This protocol does not address the unique privacy which becomes another issue. But it is efficient towards time.

### B. *Cryptographic Approaches for Big-Data Analytics*

There are many widespread techniques in cryptography which are used to provide the security for the data. Here we consider some of the cryptographic approaches to securing big data analytics in the cloud. Homomorphic encryption (HE), Verifiable Computation (VC), Multi-Party Computation (MPC) are the three cryptographic techniques which can be deployed on trusted, semi-trusted and untrusted clouds [19]. Homomorphic Encryption (HE) allows functions to be computed on encrypted data without decrypting it first. Given only the encryption of a message, one can obtain an encryption of a function of that message by computing directly on the encryption. The cloud nodes are not trusted to protect confidentiality. Input holders encrypt data before it enters the cloud, and data receivers decrypt the data after it leaves the cloud. In Verifiable Computation cloud nodes are not trusted to protect integrity. The compute nodes provide proofs of correct computation, and the data receiver verifies the proof. The dashed line denotes physical isolation from outside networks. The secured Multi-Party Computation (MPC) is deployed in semi-trusted cloud. The input holders secret-share the data among the compute nodes who perform multi-party computation on the shares. The data receiver reconstructs the output of the three MPC provides confidentiality, integrity and even authenticity. But MPC is much suited for semi-trusted cloud in the presence of honest parties.

### C. *Big Data Security and Privacy: a Review*

The well-known 3V's model of big data which comprises of Volume, Velocity and Variety. These are also called as challenges occurring in big data but there is one more challenge namely Value forming the 4V's model for the big data analytics [6]. The Value refers big data have a great social value. The 4V's model is widely recognized because it indicates the most critical problem which is how to discover value from an enormous, various types, and rapidly generated datasets in big data. Organizations used various methods of de-identification to enforce security and privacy when sharing and aggregating data across dynamic, distributed data systems. More advanced technological solution is cryptography which have encryption schemes like AES and RSA. Virtual barriers such as firewalls, secure socket layer and transport layer security are designed to restrict access to data. A novel technological named the integrated Rule-Oriented Data (iRODS) is proposed to be the solution to ensure security and privacy in big data which is used for providing each adopter community the ability to develop and deploy solutions for data management and sharing that are specific to organizational needs. For structured data in big data k-anonymity protecting scheme is used which focus on static and one-time data released situation.

The above all the technology uses correlation of massive data is the key which actually forms the basis of use of big data as well as the reason of big data security. But these do not address the situation like "no correlation" of data which may be realization of security and privacy in big data.

### D. *Secure Big Data Storage and Sharing Scheme for Cloud Tenants*

The cloud is increasingly being used to store and process big data. Hence for providing the security for the data at such cases we use the following technique. Initially the big data is divided into sequenced parts and then stores them among multiple Cloud storage service providers [3]. Instead of protecting the big data itself, this scheme protects the mapping of various data elements to each provider using a trap door function. In cloud computing, big data storage services represent a basic function for their tenants. When tenants access their data, the data parts in different data centers will be collected together and then be restored into original form based on the sequenced number of each data part. There are no extra security requirements for public data, each tenant can access these data freely, on other hand confidential data should always be kept secret and inaccessible to irrelevant persons or organizations. The security is provided by using the Identity based encryption algorithm.

Due to the enormous size, owners of big data need to consider the cost of encryption. But the above scheme avoids this by splitting the data among several cloud providers and protecting the virtual mapping using a trap door function.

### E. Big-Data Processing with Privacy Preserving Map-Reduce Cloud

A large number of cloud services requires users to share private data for data analysis or mining. Under such cases the privacy can be provided using a scheduling mechanism Optimized Balanced Scheduling(OBS) [13] to apply the Anonymization on the sensitive field only depending upon the scheduling.

This can handle high amount of data effectively where it's based on time and size of data sets. Introduces a scalable two-phase top-down specialization which has two phases of the approach are based on the two levels of parallelization provisioned by Map Reduce on cloud. In the first phase the resultant anonymization levels are not identical. To obtain finally consistent anonymous data sets, the second phase is necessary to integrate the intermediate results and further anonymize entire data sets.

In phase one the data partition and Anonymization is done. And in Phase two the merging and Anonymization is done using OBS which produces the consistent k-anonymous data sets.

### F. Data Restoration and Privacy Preserving of Data Using C4.5 algorithm

Data mining extracts knowledge to support a variety of areas even in big data. There is a challenge to extract certain kinds of data without violating the data owners' privacy. This gives rise to a new branch of data mining method called privacy preserving data mining algorithm (PPDM) [12]. This algorithm protects easily affected information in data from the large amount of data set. The privacy preservation is based on the data set complements algorithm which stores the information of real dataset. Hence the private data can be safe from the unauthorized parties. If some portion of the data is been lost, then we can recreate the original data set from the unrealized dataset and perturbed dataset. This particular approach can be used for both discrete and continuous data. But the continuous data should be converted to discrete data using sampling. The dataset complementation approach considers the data table, training set which is constructed by inserting sample data sets into data table, universal set of data table, perturbed data set and unrealized training set

The privacy preservation through data set complementation fails if all training data sets are leaked because the data set reconstruction algorithm in generic.

### G. Theoretical Basis for Perturbation Methods

The perturbation is a kind of process which involves masking of data. The maximum utility is achieved when the statistical characteristics of the perturbed data are same as that of the original data [7]. This can involve statistical distributions over the data. Hence the original data can be hidden which can be used for the purpose of privacy preservation and also for deploying security in a dataset where the dataset are maintained confidentially. When the perturbed values of the confidential variables are generated as independent realization from the distribution of the confidential variables conditioned on non-confidential variables, they satisfy the data utility and disclosure risk requirements. This considers both categorical variables and numerical variables. It exhibits the trade-off between data utility and disclosure risks when consider different release policies.

There are many masking techniques such as, Random perturbation, Matrix masking, Multiple imputation, Post Randomization Method (PRAM), Model-based approach. These are often related with the theoretical basis of perturbation method so as to provide the highest data utility.

### H. Generating Sufficiency-based Non-Synthetic Perturbed Data

The synthetic perturbed data results in information loss, because they generate the perturbed values without considering the values of the confidential variables. Hence it is no longer considered as a good solution for providing confidentiality. The mean vector and covariance matrix are sufficient statics when the distribution is multivariate normal. A new methodology called non-synthetic perturbed data which maintains the mean vector and covariance matrix of the masked data to be exactly the same as the original data [8]. This offers a selectable degree of similarity between original and perturbed data. Here the perturbed values are generated as a function of the non-confidential values and estimates of mean vector, covariance matrix. The degree of similarity between the original and the perturbed data can be varied based on

the sensitivity of the data. If the data is more sensitive, then higher level of perturbation can be chosen.

In these cases, if we maintain the mean vector and covariance matrix of the masked data to be as same as the original data, the results of analysis using masked data will be the same as that using original data. Therefore the original data is unaffected.

## III. RESULTS

There are many ways to achieve privacy and security in Big Data but how efficient the technique is what matters and the time complexity should also be considered. From the methods mention in this article it is found that the perturbation has greater effect on the privacy and security on data. This is because the technique is simple so it can be implemented at very less time compared to others. And also it is more reliable among all the other methods.

So, perturbation is suggested to be one of the good approaches for the security in Big Data as it uses statistical analysis of the data on which the perturbation method is applied.

## IV. CONCLUSION

Privacy and security are among the most important requirements in Big Data. Here we noticed the challenges in big data and also the issues that are faced for providing security due to its enormous size. We have seen the possible methods and solutions for implementing the security and privacy in the big data analytics. While these techniques provides a good starting point for securing the big data, further research is needed to turn them into practical solutions that can achieve privacy and security in the real world.

## REFERENCES


[1]. R. Agrawal and R. Srikant, ‘‘Privacy-preserving data mining,’’ ACM SIGMOD Rec., vol. 29, no. 2, pp. 439–450, 2000.

[2]. Cavoukian and J. Jonas, “Privacy by Design in the Age of Big Data,” Office of the Information and Privacy Commissioner, 2012.

[3]. Cheng Hongbing; Rong Chunming; Hwang Kai; Wang Weihong; Li Yanyan, "Secure big data storage and sharing scheme for cloud tenants," in *Communications, China* , vol.12, no.6, pp.106-115, June 2015 doi: 10.1109/CC.2015.7122469.

[4]. M. Li et al., “Toward Privacy-Assured and Searchable Cloud Data Storage Services,” IEEE Network, vol. 27, no. 4, 2013, pp. 1–10.

[5]. S. Liu, “Exploring the Future of Computing,” IT Professional, vol. 15, no. 1, 2013, pp. 2–3.

[6]. Matturdi, B.; Zhou Xianwei; Li Shuai; Lin Fuhong, "Big Data security and privacy: A review," in *Communications, China* , vol.11, no.14, pp.135-145, Supplement 2014 doi: 10.1109/CC.2014.7085614

[7]. Muralidhar, K. and R. Sarathy, “A theoretical basis for perturbat ion methods,” Statistics and Computing, October 2003, Volume 13, Issue 4, pp 329-335

[8]. Muralidhar, K. and R. Sarathy, "Generating Sufficiency-Based Non-Synthetic Perturbed Data," *Trans. Data Privacy,* vol. 1, no. 1, pp. 17-33, 2008.

[9]. Muralidhar, K. and R. Sarathy (2005). An enhanced data pertu rbation approach for small data sets. Decision Sciences. 36, 5 13-529.

[10]. Rongxing Lu; Hui Zhu; Ximeng Liu; Liu, J.K.; Jun Shao, "Toward efficient and privacy-preserving computing in big data era," in *Network, IEEE* , vol.28, no.4, pp.46-50, July-August 2014 doi: 10.1109/MNET.2014.6863131.

[11]. Shamir A., “How to share a secret,” Commun. ACM, vol. 22, pp. 612–613, November 1979. [Online]. Available: http://doi.acm. org/10.1145/359168.359176.

[12]. Sharmila A. Harale, “Data Restoration and Privacy Preserving of Data Using C4.5 Algorithm ,”International Journal of Engineering Research & Technology,Vol.3,Issue 1 (January - 2014), ISSN:2278-0181.

[13]. R. Sreedhar, D. Umamaheshwari,” Big-Data Processing with Privacy Preserving Map-Reduce Cloud,” in International Conference on Engineering Technology and Science, Vol 3, Issue 1 (February 2014), ISSN: 2319-8753.

[14]. Soundararajan O M, Jenifer Y, Dhivya S, et al. Data Security and Privacy in Cloud Using RC6 and SHA Algorithms [J]. Networking and Communication Engineering, 2014, 6(5): 202-205.

[15]. S.Subashini, V.Kavitha, A survey on security issues in service delivery models of cloud computing, Journal of Network and Computer Applications ,vol 34, Issue 1, January 2011, pp. 1–11.

[16]. Thomas M. Lenard and Paul H. Rubin, “The Big Data Revolution: Privacy Considerations”, December 2013.

[17]. Y. Wang, L. Xie, B. Zheng, and K. C. K. Lee, ‘‘High utility k-anonymization for social network publishing,’’ Knowl. Inf. Syst., vol. 36, no. 1, pp. 1–29, 2013.

[18]. X. Wu et al. , “Data Mining with Big Data,” IEEE Trans. Knowledge Data Eng. , vol. 26, no. 1, 2014, pp. 97–107.

[19]. Yakoubov, S.; Gadepally, V.; Schear, N.; Shen, E.; Yerukhimovich, A., "A survey of cryptographic approaches to securing big-data analytics in the cloud," in *High Performance Extreme Computing Conference (HPEC), 2014 IEEE*, vol., no., pp.1-6, 9-11 Sept. 2014.