International Journal of Pure and Applied Mathematics
Volume 116 No. 10 2017, 281-286
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

A STUDY ON PRIVACY PRESERVING TECHNIQUES IN BIG DATA ANALYTICS

Dr.C.Nalini[1], Dr.A.R.Arunachalam[2]
[1,2]Professor, Department of Computer Science and Engineering, BIST, BIHER,
Bharath University
[1]nalini.cse@bharathunive.ac.in, [2]arunachalam.cse@bharathunive.ac.in

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. Henceforth this Veracity is turning into another measurement in the bigdata time. In this paper, we display an overview on different methods that force security and protection 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.

1. Introduction

Enormous information is a term connected to the information sets whose size is past the capacity of regularly utilized programming frameworks keeping in mind the end goal to store, oversee and prepare. As of late in today's information driven world the enormous information handling and examination have turned out to be basic to the vast majority of the applications like government and endeavors. In the previous couple of years, the aggregate sum of information produced by human has detonated increment 300 circumstances shape exabytes to octabytes. These data are created form various fields like scientific research, government, finance and business, social networks, photography, video, audio mobile phones etc. [1-7]

Big data has many typical challenges mostly they are assisted as dimensions of big data. The essential difficulties are alluded to as "3 Vs" of huge information in particular – volume, speed, assortment. The volume which implies the span of the information sets, speed implies the speed with which the information has been produced and assortment expresses that the sort of the information that is created put away and handled. The information produced can be of organized, semi-organized and unstructured information. Alternate difficulties which are demonstrating light on huge information are Value alludes huge information have awesome social esteem. Henceforth there emerges the 4Vs model which is broadly perceived as it finds esteem from different information sets that have been produced. [8-12]

The essential difficulties are alluded to as "3 Vs" of huge information in particular – volume, speed, assortment. The volume which implies the span of the information sets, speed implies the speed with which the information has been produced and assortment expresses that the sort of the information that is created put away and handled. The information produced can be of organized, semi-organized and unstructured information. Alternate difficulties which are demonstrating light on huge information are Value alludes huge information have awesome social esteem. Henceforth there emerges the 4Vs model which is broadly perceived as it finds esteem from different information sets that have been produced. [13-15]

Big data security techniques:

Privacy-Preserving Cosine Similarity computing Protocol:

The protection safeguarding cosine closeness registering (PCSC)[10]can productively ascertain the cosine comparability of two vectors without unveiling the vectors to each other. The convention depicts that we can specifically ascertain the cosine similitude in a productive way. When we consider bury huge information handling the immediate cosine closeness calculation (DCSC)would uncover each other's security. Consequently we can apply homomorphic encryption (HE, for example, Pallier encryption (PE) to give security yet this requires tedious exponentiation operations. In this way, PCSC convention for enormous information preparing is utilized in view of

the lightweight multi-party arbitrary covering and polynomial collection methods which does not require tedious operations. [16-19]

The issue with this convention is that it has computational overheads with the expansion long of the vector. Contrasted with DCSC convention. This convention does not address the one of a kind protection which turns into another issue. Be that as it may, it is proficient towards time.

Cryptographic Approaches for Big-Data Analytics:

There are numerous across the board procedures in cryptography which are utilized to give the security to the information. Here we consider a portion of the cryptographic ways to deal with securing huge information examination in the cloud.Homomorphic encryption (HE), Verifiable Computation (VC), Multi-Party Computation (MPC) are the three cryptographic systems which can be conveyed on trusted, semi-trusted and untrusted mists [19].Homomorphic Encryption (HE) permits capacities to be processed on encoded information without unscrambling it first. Given just the encryption of a message, one can get an encryption of a component of that message by registering specifically on the encryption. The cloud hubs are not trusted to secure classification. Input holders encode information before it enters the cloud, and information collectors decode the information after it leaves the cloud. In Verifiable Computation cloud hubs are not trusted to ensure respectability. The figure hubs give evidences of right calculation, and the information beneficiary checks the verification. The dashed line signifies physical confinement from outside systems. The secured Multi-Party Computation (MPC) is conveyed in semi-trusted cloud. The info holders mystery share the information among the register hubs who perform multi-party calculation on the shares. The information collector reproduces the yield. [20]

Of the three MPC gives classification, honesty and even validness. In any case, MPC is highly suited for semi-trusted cloud within the sight of genuine gatherings.

Big Data Security and Privacy: a Review

The notable 3V's model of enormous information which involves Volume, Velocity and Variety. These are likewise called as difficulties happening in huge information yet there is one more test in particular Value shaping the 4V's model for the huge information investigation [6]. The Value alludes huge information have an awesome social esteem. The 4V's model is generally perceived in light of the fact that it shows the most basic issue which is the means by which to find esteem from a gigantic, different sorts, and quickly created datasets in huge information. Associations utilized different strategies for de-recognizable proof to authorize security and protection when sharing and totaling information crosswise over dynamic, circulated information frameworks. More progressed mechanical arrangement is cryptography which have encryption plans like AES and RSA. Virtual obstructions, for example, firewalls, secure attachment layer and transport layer security are intended to confine access to information. A novel mechanical named the coordinated Rule-Oriented Data (iRODS) is proposed to be the answer for guarantee security and protection in enormous information which is utilized for giving every adopter group the capacity to create and convey answers for information administration and sharing that are particular to hierarchical needs. For organized information in huge information k-namelessness securing plan is utilized which concentrate on static and one-time information discharged circumstance. The over all the innovation utilizes relationship of gigantic information is the key which really shapes the premise of utilization of huge information and additionally the reason of huge information security. Be that as it may, these does not address the circumstance like "no relationship" of information which might be acknowledgment of security and protection in huge information.

Secure Big Data Storage and Sharing Scheme for Cloud Tenants:

The cloud is progressively being utilized to store and process enormous information. Thus to provide the security for the information at such cases we utilize the accompanying method. At first the huge information is separated into sequenced parts and after that stores them among different Cloud stockpiling specialist organizations [3]. Rather than ensuring the enormous information itself, this plan secures the mapping of different information components to every supplier utilizing a trap entryway work. In distributed computing, enormous information stockpiling administrations speak to an essential capacity for their occupants. At the point when occupants get to their information, the information parts in various server farms will be gathered together and after that be reestablished into unique shape in view of the sequenced number of every information part. There are no additional security prerequisites for open information, every occupant can get to these information uninhibitedly, on other hand private information ought to dependably be kept mystery and out of reach to immaterial people or associations. The security is given by utilizing the Identity based encryption calculation.

Because of the gigantic size, proprietors of enormous information need to consider the cost of encryption. However, the above plan maintains a strategic

distance from this by part the information among a few cloud suppliers and securing the virtual mapping utilizing a trap entryway work.

Big-Data Processing with Privacy Preserving Map-Reduce Cloud:

A substantial number of cloud administrations obliges clients to share private information for information investigation or mining. Under such cases the protection can be given utilizing a booking component Optimized Balanced Scheduling(OBS) [13] to apply the Anonymization on the delicate field just relying on the planning.

This can deal with high measure of information adequately where it depends on time and size of information sets. Presents an adaptable two-stage beat down specialization which has two periods of the approach depend on the two levels of parallelization provisioned by MapReduce on cloud. In the primary stage the resultant anonymization levels are not indistinguishable. To acquire at long last steady unknown information sets, the second stage is important to incorporate the moderate outcomes and further anonymize whole information sets.
In stage one the information parcel and Anonymization is finished. What's more, in Phase two the combining and Anonymization is done utilizing OBS which delivers the predictable k-mysterious information sets.

Data Restoration and Privacy Preserving of Data Using C4.5 calculation :

Information mining removes learning to bolster an assortment of ranges even in huge information. There is a test to extricate certain sorts of information without abusing the information proprietors' security. This offers ascend to another branch of information mining strategy called security saving information mining calculation (PPDM) [12]. This calculation shields effectively influenced data in information from the extensive measure of information set. The security protection depends on the information set supplements calculation which stores the data of genuine dataset. Subsequently the private information can be sheltered from the unapproved parties. In the event that some bit of the information is been lost, then we can reproduce the first information set from the hidden dataset and annoyed dataset. This specific approach can be utilized for both discrete and constant information. In any case, the constant information ought to be changed over to discrete information utilizing testing. The dataset complementation approach considers the information table, preparing set which is developed by embeddings test information sets into information table, all inclusive arrangement of information table, annoyed information set and hidden preparing set

The security protection through information set complementation fizzles if all preparation information sets are spilled on the grounds that the information set remaking calculation in non specific.

A Theoretical Basis for Perturbation Methods:

The bother is a sort of process which includes covering of information. The most extreme utility is accomplished when the factual attributes of the irritated information are same as that of the first information [7]. This can include factual circulations over the information. Henceforth the first information can be concealed which can be utilized with the end goal of protection safeguarding furthermore to deploy security in a dataset where the dataset are looked after secretly. At the point when the annoyed estimations of the classified factors are created as autonomous acknowledgment from the conveyance of the secret factors adapted on non-private factors, they fulfill the information utility and exposure hazard prerequisites. This considers both all out factors and numerical factors. It displays the exchange off between information utility and exposure dangers when consider diverse discharge arrangements.
There are many covering systems, for example, Random irritation, Matrix veiling, Multiple ascription, Post Randomization Method (PRAM), Model-based approach. These are regularly related with the hypothetical premise of bother strategy in order to give the most elevated information utility.

Generating sufficiency-based non-synthetic perturbed data :

The manufactured annoyed information brings about data misfortune, since they create the irritated qualities without considering the estimations of the secret factors. Thus it is no longer considered as a decent answer for giving privacy. The mean vector and covariance grid are adequate statics when the conveyance is multivariate typical. Another strategy called non-engineered bothered information which keeps up the mean vector and covariance grid of the veiled information to be precisely the same as the first information [8]. This offers a selectable level of likeness amongst unique and bothered information. Here the irritated qualities are produced as an element of the non-secret values and gauges of mean vector, covariance matrix.The level of likeness between the first and the bothered information can be shifted in light of the affectability of the information. In the event that the information is more touchy, then larger amount of annoyance can be picked.

In these cases, in the event that we keep up the mean vector and covariance lattice of the covered information to be as same as the first information, the aftereffects of examination utilizing conceal information will be the same as that utilizing unique information. Consequently the first information is unaffected.

2. 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.

3. Conclusion

Protection and security are among the most imperative necessities in Big Data. Here we saw the difficulties in huge information furthermore the issues that are confronted for giving security because of its tremendous size. We have seen the conceivable techniques and answers for actualizing the security and protection in the enormous information examination. While these strategies gives a decent beginning stage to securing the huge information, additionally research is expected to transform them into down to earth arrangements that can accomplish protection and security in this present reality.

References

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

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

[3]Cheng Hongbing; RongChunming; 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 2015doi: 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] Udayakumar R., Kaliyamurthie K.P., Khanaa, Thooyamani K.P., Data mining a boon: Predictive system for university topper women in academia, World Applied Sciences Journal, v-29, i-14, pp-86-90, 2014.

[6] Kaliyamurthie K.P., Parameswari D., Udayakumar R., QOS aware privacy preserving location monitoring in wireless sensor network, Indian Journal of Science and Technology, v-6, i-SUPPL5, pp-4648-4652, 2013.

[7] BrinthaRajakumari S., Nalini C., An efficient cost model for data storage with horizontal layout in the cloud, Indian Journal of Science and Technology, v-7, i-, pp-45-46, 2014.

[8] BrinthaRajakumari S., Nalini C., An efficient data mining dataset preparation using aggregation in relational database, Indian Journal of Science and Technology, v-7, i-, pp-44-46, 2014.

[9] Khanna V., Mohanta K., Saravanan T., Recovery of link quality degradation in wireless mesh networks, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4837-4843, 2013.

[10] Khanaa V., Thooyamani K.P., Udayakumar R., A secure and efficient authentication system for distributed wireless sensor network, World Applied Sciences Journal, v-29, i-14, pp-304-308, 2014.

[11] Udayakumar R., Khanaa V., Saravanan T., Saritha G., Retinal image analysis using curvelet transform and multistructure elements morphology by reconstruction, Middle - East Journal of Scientific Research, v-16, i-12, pp-1781-1785, 2013.

[12] Khanaa V., Mohanta K., Saravanan. T., Performance analysis of FTTH using GEPON in direct and external modulation, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4848-4852, 2013.

[13] Kaliyamurthie K.P., Udayakumar R., Parameswari D., Mugunthan S.N., Highly secured online voting system over network, Indian Journal of Science and Technology, v-6, i-SUPPL.6, pp-4831-4836, 2013.

[14] Thooyamani K.P., Khanaa V., Udayakumar R., Efficiently measuring denial of service attacks using appropriate metrics, Middle - East Journal of Scientific Research, v-20, i-12, pp-2464-2470, 2014.

[15] R.Kalaiprasath, R.Elankavi, Dr.R.Udayakumar, Cloud Information Accountability (Cia) Framework Ensuring Accountability Of Data In Cloud And Security In End To End Process In Cloud Terminology, International Journal Of Civil Engineering And Technology (Ijciet) Volume 8, Issue 4, Pp. 376–385, April 2017.

[16] R.Elankavi, R.Kalaiprasath, Dr.R.Udayakumar, A fast clustering algorithm for high-dimensional data, International Journal Of Civil Engineering And Technology (Ijciet), Volume 8, Issue 5, Pp. 1220–1227, May 2017.

[17] R. Kalaiprasath, R. Elankavi and Dr. R. Udayakumar. Cloud. Security and Compliance - A Semantic Approach in

End to End Security, International Journal Of Mechanical Engineering And Technology (Ijmet), Volume 8, Issue 5, pp-987-994, May 2017.
[18] Thooyamani K.P., Khanaa V., Udayakumar R., Virtual instrumentation based process of agriculture by automation, Middle - East Journal of Scientific Research, v-20, i-12, pp-2604-2612, 2014.
[19] Udayakumar R., Thooyamani K.P., Khanaa, Random projection based data perturbation using geometric transformation, World Applied Sciences Journal, v-29, i-14, pp-19-24, 2014.
[20] Udayakumar R., Thooyamani K.P., Khanaa, Deploying site-to-site VPN connectivity: MPLS VsIPSec, World Applied Sciences Journal, v-29, i-14, pp-6-10, 2014.
[21] T. Padmapriya and V. Saminadan, “Inter-cell Load Balancing technique for multi-class traffic in MIMO-LTE-A Networks”, International Journal of Electrical, Electronics and Data Communication (IJEEDC), ISSN: 2320- 2084, vol.3, no.8, pp. 22-26, Aug 2015.
[22] S.V.Manikanthan and K.Baskaran “Low Cost VLSI Design Implementation of Sorting Network for ACSFD in Wireless Sensor Network”, CiiT International Journal of Programmable Device Circuits and Systems,Print: ISSN 0974 – 973X & Online: ISSN 0974 – 9624, Issue : November 2011, PDCS112011008.
[23] Rajesh, M., and J. M. Gnanasekar. "Congestion control in heterogeneous wireless ad hocnetwork using FRCC." Australian Journal of Basic and Applied Sciences 9.7 (2015): 698-702.