Data mining and Security in Big data

Muni raj Choopa, Department of Computer Science, University of Bridgeport

***Abstract*— Data has been growing rapidly due to digitization of the world. Vast amount of data accumulating in every field. Organizations are storing this vast amount of data in their databases as big data which has been maintaining as clusters. But these clusters are not providing security and confidentiality of data being stored. In recent years privacy preserving data mining has been emerged as popular research area for the security of data stored as big data. We have many algorithms for preserving security in data mining. In this paper, we will focus on extraction of knowledge by including security measures which we need to apply in different phases of the data mining. Different phases include, Source of data from which we store in target data base and data miner and final end user. In every phase we will see the risks in handling vast amount of data and corresponding measures to protect data in each phase.**

***Index Terms*— Big data, Privacy-preserving data mining.**

I. INTRODUCTION

Data mining has become more popular from the past decade due to accumulating vast amount of data called big data in the databases of the organization .Data mining is the process of extracting use full information from the vast amount of data present in databases and there by recognizing the patterns in data which tells us the facts there by making decisions to improve business[1] . This data mining has vast applications and useful in many domains like health care to recognize the disease patterns and in business intelligence to understand the customer choices and many domains like scientific discovery and also in libraries which stores terabytes of journals and books. Data accumulation has been increasing year by year and volume of data being increased is predicted to increase by about 650 % over the next few years[2].The companies which have their big data are using data mining techniques to extract the knowledge from vast amount of data with compromising on the security point of view[4].So, it is important that ,we need to keep security considerations in data mining and the owners who are giving output of the data mining should make sure, not to reveal identity of the clients .

The privacy preserving mechanisms has so many considerations [3]. Firstly, we need to consider the masking. of raw data which are sensitive fields like name, id, password and some confidential personal information. Secondly, by using some data mining techniques, we need to derive knowledge from sensitive information without compromising on privacy.

In this paper we propose multi-level security with masking algorithm for providing security of data present as big data.

Manuscript received March, 2015.
Muni raj Choopa, *Computer science, University of Bridgeport., Bridgeport, United states of America, (+1-773-456-2091)*

Also, as the data is huge, we need to derive some knowledge from it .While mining the data from big data, we need to maintain privacy of the customer's records .For example computing the customer satisfaction of the goods by abstracting the total information of all the customers while data mining. The final goal is to implement algorithm which preserves privacy of the customers and derive knowledge from that by providing privacy.

II. DATA MINING ON BIG DATA

Data mining is the process of finding interesting patterns from the large collection of data .The name "big data "itself infers that, it is a collection of large volume of terabytes of data and in addition to volume, it has variety and velocity [5] by which we can read and write the data. It consist of large quantities of structured and unstructured data[6] .we need to derive knowledge from these all varieties ,and simply storing these data is not sufficient, we need to derive knowledge from this data to use in the business decisions.

Big data analytics plays a main role to perform some real time decisions and to take decisions on business choices. Data mining is the process of extracting knowledge or pattern from big data. In mining process, different users will participate in phase wise. Users include data providers, data collector, and data miner and decision maker.

Data provider: the user who owns the data that was ready for the data mining task.

Data collector: the user who collects data from the data provider and gives it to data miner.

Data miner: the user who performs data mining task for the data collected from the data collector.

Decision maker: After mining of the data, data miner sends the output to data miner. From the knowledge which got from data miner, decision maker will take decisions

III. SECURITY CONCERNS ON BIG DATA

New enormous data applications are getting to be a piece of security administration programming in light of the fact that they can help to clean, plan and question information in heterogeneous, deficient and clamor arranges proficiently. Examination of huge data gives organizations the ability to recognize drifts and enhance business. Big data analytics changes their scene to enhance data security and circumstance mindfulness. As the time of enormous data start, data assumes an imperative part, so we need to give security for it. Huge data analytics can examine budgetary exchanges, log records,

and system movement to distinguish anomalies [7]. Fraud detection is a standout amongst the most unmistakable utilization for huge data analytics, so there is a prerequisite for framework structural planning which offer security to enormous data. Security may not have been as vital enormous information groups were gotten to just by some little gatherings of software engineers, yet for more extensive undertakings or associations it is hard to impart all data to all levels of employee. Big data framework and tools are presently commoditizing the arrangement of substantial scale, solid groups and in this way are empowering new chances to process and analyze the data.

As IT rises, the size of databases builds too quickly, so it is hard to handle such an immense size of information. The enormous data incorporates organized, unstructured and semi-organized data. In this paper we examine security prospect of huge data. Presently we are in data rich circumstance where data assumes a vital part and consider as "Gold" [14]. So security is vital to these data from unauthorized used. For any business, they deal with their clients and their personal details which if reveal it will make issues for its survival. Privacy of individuals is to be preserved within the organization.. Data mining or different methods are used on big data.

IV. RELATED WORK

In this section, we report some of the relevant works on privacy preserving scheme.

Agrawal and Srikant's scheme [8] considered a decision tree classifier from training data in which the values of individual records have been perturbed by adding random values from probability distribution. After this the data records look very different from original records and distribution of data values also looks very different from original. Then there is a problem to accurately estimate the original values in individual's data records, for this problem they proposed a novel reconstruction procedure to accurately estimate the distribution of original data values with some loss of information. But the authors say that this is acceptable for practical situation.

Oliveira and Zane [9] considered some geometric data transformation to study the feasibility of achieving PPC..
They revealed that basic transformation is feasible only after normalization of data because data transform through this methods would change similarity between data points. So clustering of data is useless. Distortion methods adopted to successfully balance privacy and security in statically databases are limited when the perturbed attributes are considered as a vector in the n-dimensional space.

Inan, Saygin, et al's scheme [10] ensures accuracy based on the dissimilarity matrix construction using a secure comparison protocol for numerical, alphanumerical and categorical data. Here the communication cost is high because of the involvement of the third party.

Teng and Du [11] gives an approach which takes advantage of the strength of both SMC (Secure Multi-party Computation) and randomization approaches to balance the accuracy and efficiency constraints. They implemented method for ID3 decision tree algorithm and association rule mining problem. This approach achieves a better accuracy compared to the only randomization approach and more efficient than the SMC approach.

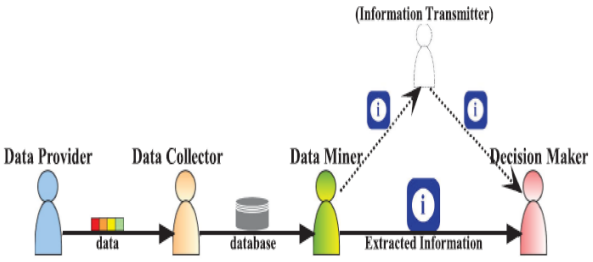
Secure Multi –party Computation [12] based on clustering vertically partition data. In vertically partitioning data, the attributes are split across the partitions. This work ensures the privacy while limiting communication cost.

Kalita, Bhattacharyya et al's scheme [16] used three transformations- translation, rotation and reflection successfully in combination. The authors established a secure and accurate scheme after applying the hybrid perturbation technique. In this technique, reflection based transformation is helpful to improving the intruders' complexity significantly.

There are a several methods for taking care of the issue of protection of data mining. The principle objective is to give security to the data

V. OUR APPROACH ON MULTI-LEVEL SECURITY WITH MASKING ALGORITHM

In the data mining a user represents either a person or an organization. So in each level, user has their corresponding operations to do. In total data mining phase we have users as represented in the below figure 1.1



1.1 Data Mining Phases

By separating the four distinctive user parts, we can investigate the protection issues in data mining in a principled way. All users think about the security of sensitive information, but every user part sees the security issue from its own viewpoint. What we have to do is to recognize the security issues that every user is worried about, and to appropriate arrangements the issues. Here we briefly depict the protection concerns of every user.

DATA PROVIDER:

The real concern of an data supplier is whether he can control the sensitivity of the information he gives to others. On one hand, the supplier ought to have the capacity to make his exceptionally private information, specifically the

information containing data that he doesn't need any other individual to know, blocked off to the information gatherer. Then again, if the supplier needs to give some information to the information authority, he needs to shroud his sensitive data

DATA COLLECTOR:

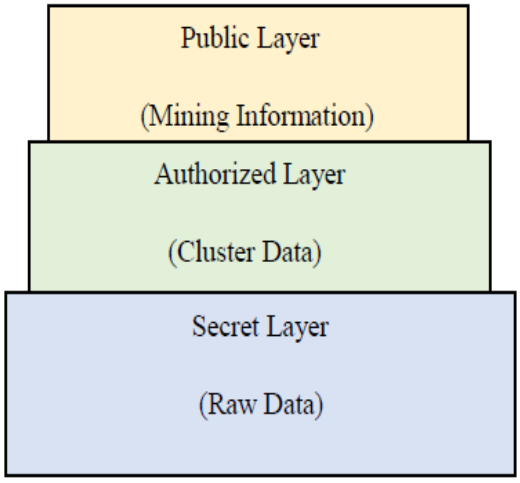
The data gathered from information suppliers may contain people's sensitive data. Straightforwardly discharging the information to the data miner will damage data suppliers' protection, henceforth information modification is needed. Then again, the information ought to still be helpful after modification, generally gathering the information will be insignificant. Therefore, the significant concern of data gatherer is to ensure that the modified information contain no sensitive data yet at the same time preserve high utility.

DATA MINER:

The data miner applies mining algorithms to the information gave by data collector, and he wishes to concentrate valuable data from information in a security safeguarding way. We consider the data authority should assume the significant liability of ensuring sensitive information, while data miner can concentrate on instructions to conceal the sensitive mining results from un trusted parties.

DECISION MAKER

Decision maker can get the data mining results specifically from the data miner, or from some Information Transmitter. It is likely that the data transmitter changes the mining results deliberately or accidentally, which may cause genuine misfortune to the decision maker. Hence, what the decision maker concerns is whether the mining results are credible. After getting the mining results, the decision maker is responsible for making correct decisions and, if we get any wrong data by altering in the middle that will cause entire business to collapse.



1.2 Layers of data

The above figure 1.2, depicts the layering approach we are using to provide security in each layer according to the user. In secret layer, we has raw data which was very sensitive and if some unauthorized user access the raw data ,then this makes big effect in entire business, so we will apply algorithms to encrypt some sensitive information before sending this data to data collector, that is cluster data in above figure. In cluster data, we maintain data in different subjects. From this cluster, we will send the secured data with some encryption techniques to the next public layer.

VI. APPLYING ENCRYPTION LAYER WISE

Data provider has access to all the big data present in the databases and there need to be some sought of masking for all the sensitive information present in it. Now our algorithm will identify the sensitive columns in the data. For example, card number is a sensitive column in a record, then we will maintain two types of data for that column. One is scrambled data and one is clear data.

Step: 1.Read all data← file

Step: 2.Identify the sensitive data

Step: 3.

For (all the sensitive data)
  Add data scrambled for data.
  Add data clear for data.
End

Step: 4.Repeat the step 3 for all the sensitive data.

Step: 5.Store the modified data to database.

By above steps we get encryption for all the sensitive data in the big data and we can send this data to next phase that is data collector without compromising the privacy of data.

Data collector gathers information from data provider keeping in mind the end goal to backing the consequent data mining operations. The first data collected from data provider more often than not contain delicate data about people. On the off chance that the data provider doesn't take sufficient safeguards before discharging the information to open or information mineworkers, those sensitive data may be unveiled, despite the fact that this is not the collector's original intention. Case in point, on October 2, 2006, the U.S. online motion picture rental administration Nettflix[15] released a data set containing movie ratings of 500,000 subscribers to the public for a challenging competition called "the NetflixPrize". The objective of the opposition was to enhance the personalized movie recommendations..The released data set was supposed to be privacy-safe, since each data record only contained a subscriber ID (irrelevant with the subscriber's real identity), the movie info, the rating, and the date on which the subscriber rated the movie. However, soon after the release, two researchers [17] from University of Texas found that with a little bit of auxiliary information about an individual subscriber, e.g. 8 movie ratings (of which 2 may be completely wrong) and dates that may have a 14-day error, an adversary can easily identify the individual's record

From above example we can see that, it is necessary for the data collector to modify the original data before releasing them to others, so that sensitive information about data providers can neither be found in the modified data nor be inferred by anyone with malicious intent. Generally, the modification will cause a loss in data utility. The data collector should also make sure that sufficient utility of the data can be retained after the modification otherwise collecting the data will be a wasted effort.

Data collector gets the data from the data provider in the form scrambled and cleared data. But here, data collector may send the data miner which has cleared information, there by disclosing sensitive information to the data miner. To avoid this we hide sensitive information and send the cleared information to other data base where strict rules are applied on that with only authorized access to that databases. Now only the scrambled data will go in to the data miner with minimal information.

In data miner, the different mining algorithms were performed on deriving knowledge from the big data. But before deriving the knowledge, the data mining will connects to sensitive data base and applies decryption by getting de-tokenized data and provides output to decision maker in reliable design patterns for decision making.

Decision maker is the end user who makes decision on the business improvement .For example, consider wall mart which is selling product. By mining the data from the past one year, we can get the most sold product ,there by decision maker can make decisions on increasing the stock this year by some percentage or invest on improvement on quality for this product. Here ,the decision maker should make sure about the data getting from the data miner. So for reliability in the data, we check the data with the sensitive information database and compare the data before making any decision by the decision maker. Below is the algorithm which used to compare.

Data_mined←get the data from the data miner.

For all the data present in data_mined

Compare with the information present in sensitive database

If (data matches)

Handover the knowledge to the decision maker.

Else

Return back to the data miner and representing false information

Decision maker is the last phase in the data mining in big data. Once the reliable data enters the last phase, then decision maker will take decision on the mined data and improve the business accordingly.

VII. EXPERIMNETAL RESULTS

All these layers which constitute of large amount of data are designed in keeping in mind with TB's (Terabytes).So while transferring the data or processing of data, we run these algorithms with the data present on these layers. You can see the results of the processing time and amount of data processed in the below table which has 300 node cluster .

File Size	DATA PROVIDER Layer	DATA COLLECTOR Layer	DATA MINER Layer	DECISION MAKER Layer	Total Time
1 GB	0:00:20	0:00:08	0:00:30	0:00:05	0:01:02
100 GB	0:05:34	0:04:28	0:04:24	0:01:02	0:15:28
500 GB	0:20:24	0:15:01	0:15:14	0:08:58	0:59:37
800 GB	0:30:40	0:18:16	0:16:13	0:10:40	1:13:49
1 TB	0:35:12	0:20:26	0:25:44	0:19:51	1:41:04

Table 1.1 processing times of each layer

VIII. CONCLUSION

In business, information is the main asset .In this environment having authorized access to information is important. End users will take decisions on these information and these critical decisions are a must for the survival. As the data is growing bigger and bigger, security breaches and data is exposing to unauthorized users .It is not possible for an organization to protect every record from security .In data mining, this is the main concern on providing data security for knowledge derived and in every phase from source data to decision maker. We introduced multi-layer with masking that protects the data from the breaching of security and access of data by un authorized user. Future research on this concept will go further with improved algorithms.

IX. ACKNOWLEDGEMENT

The author would like to express their sincere gratitude to the editor and all those who help to review and give their valuable suggestions that significantly improved the quality of this paper

X. REFERENCES

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. an Mateo, CA, USA: Morgan Kaufmann, 2006.
[2] IDC, Digital data to double every 18 months, worldwide marketplace model and forecast, Framingham, MA. available at www.idc.com May 2009
[3] Data mining using Matrix Algebraic Approach",doi:10.4156/jcit.vol4.issue3.5.
[4] Arie Friedman,"Privacy preserving data mining"pp.4,January 2011
[5] P. Russom, Big Data Analytics, Best Practices Report, Fourth Quarter, The DataWarehouse Institute , Renton, WA, September 18 2011.
[6] Big Data Analytics for Security Intelligence,september 2013

[7] R. Agrawal, R. Srikant, "Privacy-preserving data mining", In: Proceedings of the 2000ACM-SIGMOD on management of data, Dallas, TX, USA, May 15-18, 2000
[8] S.R.M. Oliveira, O.R. Zaiane, "Privacy Preserving Clustering By Data Transformation", In Proc. Of the 18th Brazilian Symposium on Databases, Manaus, Brazil, October 2003, pages 304-318
[9] A. Inan, Y. Saygin, E. Savas, A. Hintoglu, A. Levi.: Privacy-Preserving Clustering on Horizontally Partitioned Data. Data Engineering Workshops, 2006
[10] Z. Teng, W. Du, "A hybrid multi-group approach for privacy preserving decision tree building", In: Proceedings of the 11th Paci_c-Asia conference on knowledge discovery and data mining (PAKDD2007).
[11] B. Pinkas, "Cryptographic techniques for privacy-preserving data mining", SIGKDD Explore, 2002, 4(2): 12-19
[12] "Big security for big data", available at www8.hp.com/ww/en/secure/pdf/4aa4-4051enw.pdf
[13] N.I.Hussain,P.Saikia,B.Choudhury, S.Rakshit,"Study of a Decision tree approach to analyse Big data",pp.2,published on Micro-2014
[14] Brian Lent, Arun Swami, Jennifer Widom,"Clustering Association Rule".
[15] http V ==en:wikipedia:org=wiki=Netflix
[16] M. Kalita, D.K. Bhattacharyya, M. Dutta, "Privacy Preserving Clustering - A Hybrid Approach", In: Proceedings of the ADCOM'08, Chennai, December 2008
[17] A. Narayanan and V. Shmatikov, ``Robust de-anonymization of large sparse datasets," in Proc. IEEE Symp. Secur. Privacy (SP), May 2008,pp. 111_125

Authors

Muni raj Choopa is currently pursuing masters in computer science at University of Bridgeport .He has around 5 years of Information technology work experience in business intelligence domain. Technical skills included, Hadoop (Map reduce,HDFS,Pig,Hive),Teradata, Oracle, Java,C++, Informatica, Cognos and he Teradata 12 Certified Professional. Area of interest is Hadoop framework

