IJCSC 0973-7391 Volume 7 • Number 1 Sept 2015 - March 2016 pp. 205-209 available online at www.csjournalss.com

Security and Privacy with Perturbation Based Encryption Technique in Big Data

Prajapati Het I., Patel Shivani M., Prof. Ketan J. Sarvakar
IT Department, U. V. Patel college of Engineering Ganapat University, Gujarat

Abstract: In era of information age, due to different electronic, information & communication technology devices and process like sensors, cloud, individual archives, social networks, internet activities and enterprise data are growing exponentially .[1]The most challenging issues are how to effectively manage these large and different type of data .Big data is one of the term named for this large and different type of data .Due to its extraordinary scale, privacy and security is one of the critical challenge of big data . Many techniques have been suggested and implemented for privacy preservation of large data set like Anonymization based, encryption based and others but unfortunately due to different characteristic (large volume, high speed, and unstructured data) of big data all these techniques are not fully suitable. In this paper we have deeply analyzed, discussed and suggested the Perturbation Based Encryption Technique. In that to provide privacy data is perturbed and after that the security algorithm is going to apply for the more security.
Keywords: Big Data, Perturbation Technique, Anonymization, Big Data Privacy, Encryption.

I. INTRODUCTION

Any data that is difficult to Capture, Curate, Store, Search, Share, Transfer, Analyze and to create visualization. "Big Data" is the data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Big data is a buzzword, or catch phrase, used to describe a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

Big Data Analysis programs are done for government, business, healthcare, law, cyber security, research and development, etc. National governments have recently announced significant programs on Big Data applications. An example of big data might be Petabytes (1,024 terabytes) or Exabyte (1,024 petabytes) of data consisting of billions to trillions of records of millions of people all from different sources (e.g. Web, sales, customer contact centre, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible.

II.APPROACHES FOR BIG DATA PRIVACY PRESERVING

Big data privacy can be preserved by two different approaches, one is to impose the rules and legality to individual and organization and another way is developing Privacy by design or maybe both of them together. Better privacy can only be achieved by developing an approach which is capable of handling both technical and legal aspects. This approach may not be suitable in all cases. Another approach is hiding the sensitive information but due to presence of many re-identification techniques, privacy can easily be violated [2] [3].The best approach which guaranties the privacy, better utility and most suitable for big data (statistical) is noise based approach. Here are the approaches.

A. **Anonymization Based:** It mainly studies anonymization approaches for publishing useful data while preserving privacy. The original data is assumed to be a private table consisting of multiple records. Each record consists of the following 4 types of attributes: Identifier (ID), Quasi-identifier (QID), Sensitive Attribute (SA), Non Sensitive Attribute (NSA). Before being published to others, the table is anonymized, that is, identifiers are removed and quasi-identifiers are modified. As a result, individual's identity and sensitive attribute values can be hidden from adversaries.[4]
B. **Encryption Based:** In the encryption based approach from the many encryption algorithms such as symmetric key algorithms or asymmetric key algorithms the encryption is applied to the data set. The main problem with this approach is degrading the efficiency of system if we are using Asymmetric key algorithm for big data. So it is better to use symmetric key algorithm for Big Data.

C. **Noise Based:** It a privacy approach which probability of output of two different data set will nearly be same. When two different data set produces nearly same output then the adversary can't determine the actual targeted data set by any quasi identifier. It is perturbation technique which we are going to use.[1]

III. PROPOSED SOLUTION

Here we are going to apply Perturbation Based Encryption Technique. In the Perturbation Based Encryption we are adding the garbage value in the different way. In this the garbage value is added in two parts. In the data set there are odd and even rows. So, in the odd number of rows we are multiply some garbage value as noise and in the even number of rows we are adding some garbage value. After the injection of garbage value in the dataset we'll apply encryption on it. For the encryption technique we are using AES encryption algorithm.

ALGORITHM
Algorithm: Perturbation Based Encryption()
Input: Original data
Output : Perturb and Encrypted data

1. Begin
2. Declare int i, int tc, float rn, float temp, float j;
3. Set tc=total number of rows in data;
4. Set i=1;
5. While i <= tc do
   if(i%2 != 0){
   temp = j * rn;
   }
6. Print the perturb value with multiplying the garbage value.
7. else {
   temp = j + rn;
   }
8. Print the perturb value with adding the garbage value.
9. rn++;
10. End if;
11. i++;
12. End while;
13. After applying the perturbation technique we are going to apply encryption technique with AES encryption algorithm.
14. END.

IV.IMPLEMENTATION WORK

Here, we are implementing Perturbation Based Encryption Technique. For that we are taking an example of Stock Dataset. Below are the results for that. In the dataset there are columns of Name, Price, Previous_Open, Previous_Close, Low, High etc. In that Name is an identifier. Other are the sensitive attributes so that we are going to hide them with the Perturbation Technique. Below are the three tables. In which first is the main dataset. Next is the perturbed dataset and the last is encrypted dataset.

Name	Price	Previous_Open	Previous_Close	Low	High
NASDAQ Bank	2770.27002	2760.399902	2736.550049	2748.580078	2771.110107
NASDAQ Industrial	3814.1001	3779.360107	3738.889893	3765.820068	3817.300049
NASDAQ Insurance	6924.68018	6911.410156	6869.669922	6890.649902	6930.290039
NASDAQ Computer	2364.38989	2338.72998	2310.350098	2335.219971	2364.77002
NASDAQ Financial	3128.47998	3107.149902	3083.219971	3096.26001	3130.179932
NASDAQ Composite	4620.16016	4574.379883	4517.319824	4559.180176	4620.160156

(a)

Name	Price	Previous_Open	Previous_Close	Low	High
NASDAQ Bank	1123.07723	1119.07585	1109.407035	1114.284051	1123.417805
NASDAQ Industrial	3814.505501	3779.765511	3739.295296	3766.225472	3817.705452
NASDAQ Insurance	4539.727947	4531.0283	4503.663959	4517.418154	4543.405699
NASDAQ Computer	2365.045479	2339.385567	2311.005684	2335.875557	2365.425606
NASDAQ Financial	2833.682578	2814.362438	2792.687429	2804.498703	2835.222342
NASDAQ Composite	4621.065926	4575.285653	4518.225594	4560.085946	4621.065926

(b)

Name	Price	Previous_Open	Previous_Close	Low	High
NASDAQ Bank	byZsOIL15/VrE/...	2IU5a80IdcW5D2oH...	K/tsKcRvRYn+yZ...	vmdSK1x5nJ4Bb0b...	AhboSaXGLW2A...
NASDAQ Industrial	k7tGkXrO9Ya/Ia...	PD6mGLNQj84kO/...	Pdp7/h2B+EcuZ...	KOsrrngHBsCTOlKi...	LZJQsmDT8ORto...
NASDAQ Insurance	WTzhD2Ygk6ZA...	8vkWlcn+t7ogmeC...	xZZUXILTzKJLwo...	kjqQwB3X4xFFdsz...	ZDlrUMf2uzpO...
NASDAQ Computer	WGJNgmU2+O...	jDOe1HNCZqtKwCm...	2rctbe/poM3bS...	fZreb5oDta2+ucfa...	hioHuSAQeKP6...
NASDAQ Financial	78tNQahuvv9p...	DbaAOww19aglJQP...	ICRx+Z3VhiVrjY...	qtsT8gJswEThbW...	c6W9f9Z11DrAo...
NASDAQ Composite	z9hHH7gWfbdU...	wZbaV6o5052fol36...	5Wx1yKakkcPlJ...	dbOUd1Z4N8oJV2...	z9hHH7gWfbdU...

(c)

Table: 1-(a) Original Dataset, (b) Perturbed Dataset, (c) Encrypted Dataset.

Here, in the table it is a result of this technique. In the first table there is an Original dataset. In the second table there is processed dataset. In the original dataset the garbage value is added. And the perturbed data is generated in that table. The perturbed data is for the privacy. In the third table the AES encryption technique is applied. So, that it is encrypted value is there in the table which concerns security of data.

V.EXPERIMENT EVALUATION

After applying the perturbation on the dataset we are comparing the variance of original data and perturbed dataset. As, the difference of variance of original and perturb data is high we get the percentage of accuracy for this technique. Below is the value of variances.

	Original data	Perturb data
Price	2779222.207	1867058.09
Previous_Open	2780389.34	1843702.47
Previous_Close	2754051.17	1809280
Low	2763361.77	1831531.63
High	2784592.57	1869123.35

Table: 2- Variance of original n personal dataset.

Here, in Table-2 there are the value of variances of original dataset and perturbed dataset.

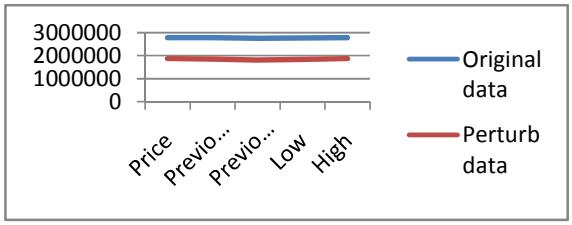

Figure: 1- Comparison of variance.

There is high difference in that so we can say that we are provided privacy in that point. For the privacy preserving we have to give data in hiding form. Here with the perturbation technique we get the higher difference in variances. So we can say that it is not easy to get the data for any attacker. The difference of variances of original data and perturbed data is show in graphical from also in Fig. 1. Hence it is a first evaluation for this technique.

After this comparison we also compare the changes in the previous and current values of stock. These changes are calculated in percentage. And also the changes of percentage are perturbed. We get the difference in the graph after perturbing the dataset. So we can say the privacy is achieved with this technique. Below are the changes in percentage of dataset with its graph.

Symbol	Name	Change in percentage of original dataset	Change in Percentage of perturbed dataset
ANK	NASDAQ Bank	1.230000019	0.498646343
INDS	NASDAQ Industrial	2.00999999	2.415403515
INSR	NASDAQ Insurance	0.800000012	0.524469336
IXCO	NASDAQ Computer	2.339999914	2.995586574
IXF	NASDAQ Financial	1.470000029	1.331481581
IXIC	NASDAQ Composite	2.279999971	3.185769737

Table:3- Change in percentage.

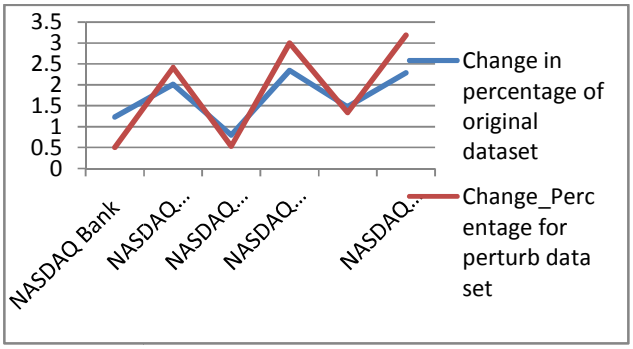

Figure:2- Graph for the changes in percentage.

VI.FUTURE ENHANCEMENT

In the future we are going to apply this technique with different type of dataset. Other than that we are going to apply authentication for the security and to save the data from the adversaries. In the authentication we will do with the Email and password or we will do it with the OTP(One Time Password) technique.

VII.CONCLUSION

In this study, we have seen the background of Big Data. We have seen the technique to provide security and privacy to big data in detail. We are proposing the approach which is Perturbation Based Encryption Technique. The flow of the approach is discussed. For the encryption we are taking AES algorithm.

REFERENCES

[1] K. M. P. Shrivastva, M. a. Rizvi, and S. Singh, "Big Data Privacy Based on Differential Privacy a Hope for Big Data," 2014 Int. Conf. Comput. Intell. Commun. Networks, pp. 776–781, 2014.

[2] Hirsch, Dennis D. "The Glass House Effect: Big Data, the New Oil, and the Power of Analogy." Maine Law Review 66 (2014): 2.

[3] Martin, David J., Daniel Kifer, Ashwin Machanavajjhala, Johannes Gehrke, and Joseph Y. Halpern."Worst-case background knowledge for privacy-preserving datapublishing." In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on, pp. 126-135. IEEE, 2007.

[4] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information Security in Big Data: Privacy and Data Mining," J. rapid open access Publ., vol. 2, pp. 1149–1176, 2014.