iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** VI **Month of publication:** June 2019

DOI: http://doi.org/10.22214/ijraset.2019.6049

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

Differential Privacy Technique for Privacy Preservation on Big Data

Johny Antony P[1], Dr Antony Selvadoss Thanamani[2]
[1]*Research Scholar, Department of Computer Science, NGM College, Pollachi 642001*
[2]*Associate Professor & Head, Department of Computer Science, NGM College, Pollachi*

Abstract: ***Privacy of big data is most important factor for the enterprises so we should have more efficient methods to protect the data. In this article, we proposed a novel mechanism, called Adaptive Firefly Laplace Mechanism (AFLM), to preserve differential privacy on Big Data to protect sensitive information among analyst. We have many existing methods for privacy preservation but each method has its own limitation and drawbacks. To overcome the drawbacks of existing privacy preservation methods such as both cryptographic techniques and data anonymization are analyzed with differential privacy method in this article.Privacy has become crucial in knowledge based applications. Proper integration of individual privacy is essential for data mining operations. This privacy based data mining is important for sectors like Healthcare, Pharmaceuticals, Research, and Security Service Providers etc. There are many algorithms available for clustering the big data. This article focused on the density based clustering algorithms for first phase. Then our proposed algorithm called Adaptive Firefly Laplace Mechanism algorithm is used for both finding sensitive data on clustered output, and to add noisy data instead of sensitive data to hide sensitive information from analyst.***

Keywords: Big data, encryption, security, Anonymization, Privacy Preservation, cryptography, Information hiding, k-Anonymity, T-Closeness, L-Diversity, Security, Information Loss, Adaptive Firefly Laplace Mechanism algorithm.

I. INTRODUCTION

Big data is among one of the emerging technologies that are bringing revolution in the world of data analytics. It has the power to provide insights into the unseen aspects of data analysis. The term big data is used to indicate large volumes of high variety data being generated at high velocity. Security issues are a major challenge in big data analytics. Although data analytics is useful in decision making, it will lead to serious privacy concerns. Hence privacy preserving data analytics became very important. Differential privacy is another big data privacy preservation method that is being widely used. It is a method enabling analysts to extract useful answers from databases containing personal information while offering strong individual privacy protections [1]. It aims to minimize the chances of individual identification while querying the data. As opposed to anonymization, data is not modified in differential privacy. Users don't have direct access to the database. There is an interface that calculates the results and adds desired inaccuracies. It acts as a firewall. These inaccuracies are large enough that they protect privacy, but small enough that the answers provided to analysts and researchers are still useful. The motivation of this research work is to enhance differential privacy method by proposing new algorithm called Adaptive Firefly Laplace Mechanism Algorithm. DBSCAN algorithm is used for clustering the big data. Then, our proposed algorithm applied on these clustered output to extract secured knowledge. Our proposed algorithm compared with Dynamic Data Encryption Strategy Algorithm, Efficient Privacy Preserving Algorithm and Adaptive Laplace Mechanism and proves it is an efficient algorithm which provides low complexity, no-chance of re-identifying the sensitive data and 100% privacy. The remainder of this article is organized as follows. Section 2 reviews different PPDM methods focusing on the evaluation criteria. Section 3 describes the evaluation framework we have developed for assessing various PPDM algorithms. Section 4 shows a proposed work. Finally, Section 5 presents conclusion of privacy preserving data mining.

II. RELATED LITERATURE REVIEW

Privacy preserving data mining is the branch of which includes the studies of privacy concern when mining is applied. Several methods are analyzed to protect the privacy of individuals. Various methods like data hiding, masking, suppression, aggregation, perturbation, anonymization, SMC are studied in literature with regards to PPDM. D. Aruna Kumari and L. Anusha [4] presents some of the technologies to provide security to big data. Big data has the large data and it is an ongoing technology which is used by every organization. It has flexibility to store the structured, semi-structured and structured data. It is faster to store the data than we use before. The data can be generated by connecting devices from pc's and smart phones to sensors. Big data can collect heterogeneous data; it can be in any format like text, document, image, video etc. Since it has a large amount of data, security is

©IJRASET: All Rights are Reserved

fundamental right it should contain. The security should provide to every individual, organization and society data. Without privacy safety, diversity, innovation… etc will be in risk. New technologies may known to other countries which are implemented by own. Like these many more problems effect to country. So to overcome to these problems we are using Big data security. Vignesh Kumar .G Arun Kumar .S [10] proposed various methods to protect privacy of data during big data processing. Cloud computing gives massive computation power and storage capacity which helps users to perform computation and data-intensive applications without infrastructure. Cross-cloud service composition provides a stable approach capable for large-scale big data processing. Along the processing of such complex web based applications, a large volume of intermediate data sets will be generated. However, preserving the privacy of intermediate data sets becomes a challenging problem because misfeasors will access sensitive information by analysing multiple intermediate data sets. Encrypting all data sets in cloud is widely adopted in existing approach to address this challenge. But encrypting all intermediate data sets are neither efficient nor cost-effective as it is very time consuming and costly for big data applications to en/decrypt data sets frequently while performing operations on it.

III. PRIVACY PRESERVATION TECHNIQUES

Data mining technology has been developed with the goal of providing tools for automatically and intelligently transforming large amount of data in knowledge relevant to users [6]. The extracted knowledge, often expressed in form of association rules, decision trees or clusters, allows one to find interesting patterns and regularities deeply buried in the data, that are meant to facilitate decision making processes. Such a knowledge discovery process, however, can also return sensitive information about individuals, compromising the individual's right to privacy.

Moreover, data mining techniques can reveal critical information about business transactions, compromising the free competition in a business setting. Thus, there is a strong need to prevent disclosure not only of confidential personal information, but also of knowledge which is considered sensitive in a given context. For this reason, recently much research effort has been devoted to addressing the problem of privacy preserving in data mining.

As a result, several data mining techniques, incorporating privacy protection mechanisms, have been developed based on different approaches. Cryptography technique, Anonymization and Differential Privacy Preservation are some of the existing methods which is used for privacy preservation on big data [6]. Cryptography is a method where sensitive data is encrypted. Data anonymization is the process of altering the original information to make sure the sensitive information cannot be identified. In this method, sensitive attributes are identified and hided to maintain the privacy of sensitive information. The data can be shown to public after this anonymization process.

Differential privacy method provides strong protection to sensitive data or personal data in big data processing. During data processing analysts can get the required information from statistical database provided with less possibility of identifying sensitive data. This method is different from anonymization, as the data is modified in anonymization method, but in differential privacy the sensitive data is not modified. In differential privacy method users are allowed to access the statistical database directly. There is an interface between the analysts and database like firewall. This query can be passed only through this interface and that will manipulate the output by including inaccuracies. The final result will be passed to data analyst. The sensitive data involved in this process are not visible, at the same time the result given by the interface is useful for the analysts.

IV. PROPOSED WORK

Privacy is considered as an important aspect of preserving information without information loss. The following figure depicts the clear picture of our proposed work. When analyst send query to Big Data to extract knowledge we need to consider preserving sensitive information also. As soon as data has been clustered a special algorithm need to be apply to hide sensitive information. Existing technologies like cryptography and anonymization not provides 100% guarantee on hiding sensitive information. Whereas differential privacy is efficient method which is suitable for privacy preservation. In this research article we proposed differential privacy preservation based new algorithm called Adaptive Firefly Laplace Mechanism Algorithm. This new algorithm is compared with existing algorithms such as cryptographic based Dynamic Data Encryption Algorithm, Anonymization based Efficient Privacy Preserving Algorithm and Differential privacy based Adaptive Laplace Mechanism Algorithm. In our new algorithm we rectified the drawbacks of Adaptive Laplace Mechanism Algorithm by combing Firefly Algorithm with this and enhanced the performance. So that we named it as Adaptive Firefly Laplace Mechanism algorithm.

©IJRASET: All Rights are Reserved

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VI, June 2019- Available at www.ijraset.com

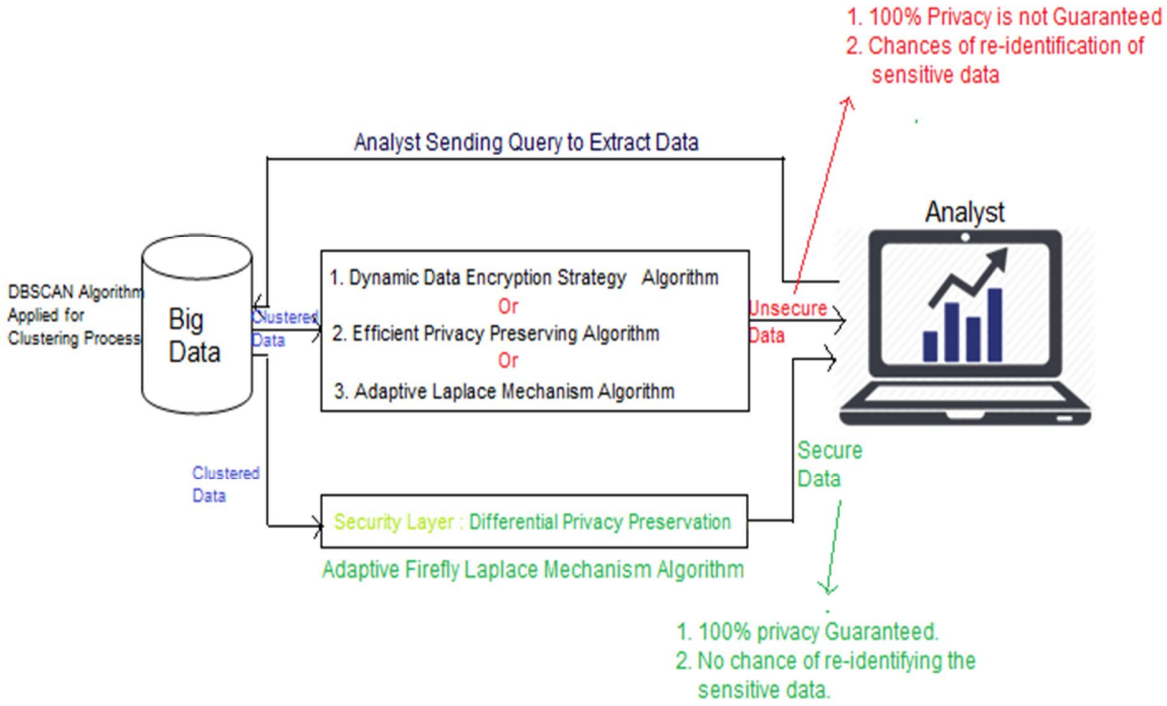


Figure 1: Differential Privacy Preservation

Differential privacy method provides strong protection to sensitive data or personal data in big data processing. During data processing analysts can get the required information from statistical database provided with less possibility of identifying sensitive data. In differential privacy method users are allowed to access the statistical database directly. There is an interface between the analysts and database like firewall. This query can be passed only through this interface and that will manipulate the output by including inaccuracies. The final result will be passed to data analyst. The sensitive data involved in this process are not visible, at the same time the result given by the interface is useful for the analysts. A computation on a set of inputs is said to be differentially private if, for any possible input items, the probability that the computation produces a given output which does not depend much on whether this item is included in the input dataset or not.
The advantages of differential privacy compared with anonymization and encryption methods are as follows:

A. No modification is required over the actual raw data. Generalization and suppression techniques are also not required.
B. By mathematical formulae calculations based on the nature of data, type of questions etc., distortions can be added to the results of those calculation.
C. These distortions are advantageous to analysts because of the hidden sensitive values.

Even though differential privacy is an efficient method to provide secured data, still some drawbacks are there while using Adaptive Laplace Mechanism for adding noise to the sensitive data. Finding sensitive data is an important step because after added noise to the sensitive data, the knowledge should not lose. So to protect sensitive data from analyst without compromising knowledge, we need to enhance Adaptive Laplace Mechanism Algorithm. So to overcome this issue we combined Firefly algorithm with Adaptive Laplace Mechanism and proposed new algorithm called Adaptive Firefly Laplace Mechanism. Firefly algorithm takes place of finding sensitive data and Adaptive Laplace Mechanism takes place of adding more noise to the selected sensitive data.
Firefly algorithm is an evolutionary model derived from the nature and based on collective intelligence algorithms that is on basis of flashing light of fireflies. The algorithm was for the first time presented by Yang in Cambridge University on 2008. Fireflies produce lights that optical pattern of each light is different from another one. They use this light to attract mates and for hunting. The amount of this light is in direct relation with attractiveness of firefly. Through considering amount of light of each firefly as target function, the behaviour of fireflies could be modelled as an optimization algorithm. To ease simulation of life of fireflies, 3 main assumptions are considered in modelling process:

©IJRASET: All Rights are Reserved

1) Fireflies are all from one gender and hence, gender plays no role in attracting them towards each other.
2) The amount of attraction between two fireflies is in reverse correlation with their brightness and with the space between them. Hence, the brighter firefly can attract adjacent fireflies and if no one of them is brighter than others, their movement would be randomly.
3) Brightness of fireflies is determined based on target function related to them.

Adaptive Laplace Mechanism (Database D, hidden layers H, loss function Ƒ(Θ), and privacy budgets $\epsilon 1$, $\epsilon 2$, and $\epsilon 3$, the number of batches T, the batch size $|L|$)
1: compute the average relevance by applying the LRP Alg.
2: $\forall$ j ϵ [1, d]: $R_j(D) = \frac{1}{|D|}, \sum_{x_i \in D} R_{x_{ij}}(x_i)$ #Eq.10#
3: inject Laplace noise into the average relevance of each j-th input feature
4: $\Delta_R = 2d/|D|$ #Lemma 1#
5: for j ϵ [1,d] do
6: $\bar{R}_J \leftarrow \frac{1}{|D|}, \sum_{X_i \in D} R_{x_{ij}(X_i)} + \text{Lap}(\frac{\Delta_R}{\epsilon 1})$
7: $\bar{R}(D) = \{\bar{R}_j\}_{j \epsilon [1,d]}$
8: inject Laplace noise into coefficients of the differentially private layer h_0
9: $\Delta_{h_0} = 2 \sum_{h \in h_0} d$ h #Lemma 3#
10: for j ϵ [1,d] do
11: $\epsilon_j \leftarrow \beta_j$ x ϵ_2 #Eq.16#
12: for $x_i \epsilon$ D, j ϵ [1,d] do
13: $\bar{x}ij \leftarrow x_{ij} + \frac{1}{L} \text{Lap}(\frac{\Delta_{h_0}}{\epsilon_j})$ #perturb input feature x_{ij}#
14: $\bar{b} \leftarrow b + + \frac{1}{L} \text{Lap}(\frac{\Delta_{h_0}}{\epsilon 2})$ #perturb bias b#
15: construct hidden layers {h_1, …… h_k} and normalization layers { $\overline{h_1}$.....$\overline{h_{(k)}}$ }
16: inject Laplace noise into coefficients of the approximated loss function $\hat{\mathcal{F}}$
17: $\Delta_{\mathcal{F}} = M(|\bar{h}_{(k)}| + |\bar{h}_{(k)}|^2)$ #Lemma 5#
18: for $x_i \epsilon$ D, R ϵ |0,2|, l ϵ |1, M| do
19: $\bar{\emptyset}^{(R)}_{x_{l_i}} \leftarrow \emptyset^{(R)}_{x_{l_i}} + \frac{1}{|L|} \text{Lap}(\frac{\Delta_{\mathcal{F}}}{\epsilon_3})$ #perturb coefficients of $\hat{\mathcal{F}}$
20: Initialize 0_o randomly
21: for t ϵ [T] do
22: Take a random training batch L
23: Construct differentially private affine transformation layer
24: $\bar{h}_{ol}(W_0) \leftarrow \{\bar{h}_L(W)\}_{h \in h_0}$
25: s.t. $\bar{h}_L(W) = \sum_{X_i \in L} (\bar{X}_i W^T + \bar{b})$
26: construct differentially private loss function
27: $\bar{\mathcal{F}}_L(0_t) = \sum_{l=1}^{M} \sum_{x_i \in L} \sum_{R=0}^{2} (\bar{\emptyset}^{(R)}_{l_{x_i}} W^T_{l(k)})^2$
28: Compute gradient descents
29: $\theta_{t+1} \leftarrow \theta_t$ - $\eta_t \frac{1}{|L|} \nabla_{\emptyset_t} \bar{\mathcal{F}}_L(\theta_t)$ # η_t is learning rate#
30: Return θ_T # ($\epsilon_1 + \epsilon_2 + \epsilon_3$) -differentially private#

Adaptive Firefly Laplace Mechanism Algorithm

Therefore, whenever analyst pass query to big data, the data is clustered using DBSCAN algorithm. Then our proposed algorithm works on these clustered data in two steps. In first step finding sensitive data takes place. Second step adding more noise to sensitive data takes place. So our new algorithm combines the advantages of both algorithms and performs well compare to existing algorithms. So finally analyst receives secured extracted knowledge from big data. Using this knowledge no one can able to re-identify the sensitive information which is hidden by our proposed algorithm.

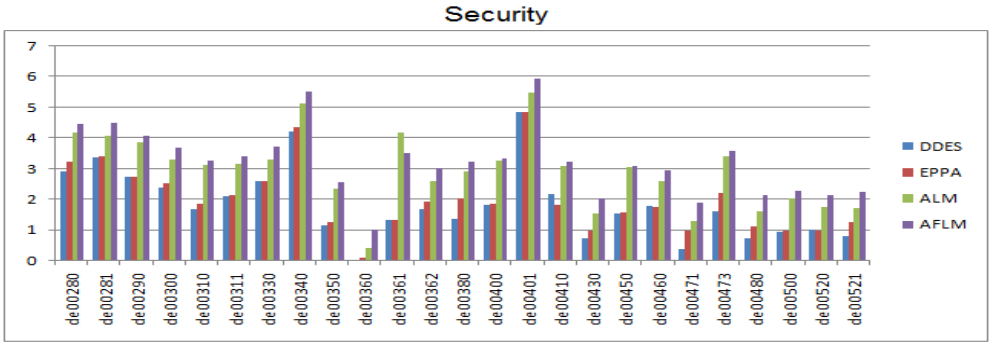
©IJRASET: All Rights are Reserved

V. EXPERIMENTAL RESULT

Current PPDM algorithms do not satisfy all these goals at the same time; for instance, only few of them satisfy the point (2). The above list of goals helps us to understand how to evaluate these algorithms in a general way. The framework we have identified is based on the following evaluation dimensions:

1) *Efficiency:* That is, the ability of a privacy preserving algorithm to execute with good performance in terms of all the resources implied by the algorithm;
2) *Scalability*: Which evaluates the efficiency trend of a PPDM algorithm for increasing sizes of the data from which relevant information is mined while ensuring privacy;
3) *Data Quality*: After the application of a privacy preserving technique, considered both as the quality of data themselves and the quality of the data mining results after the hiding strategy is applied;
4) *Hiding Failure*: That is, the portion of sensitive information that is not hidden by the application of a privacy preservation technique;
5) *Privacy Level*: Offered by a privacy preserving technique, which estimates the degree of uncertainty, according to which sensitive information, that has been hidden, can still be predicted.

The Adaptive Firefly Laplace Mechanism Algorithm was experimented with the IBM Log Dataset. Our proposed algorithm works with three phases. Initially, using DBSCAN clustering algorithm correlation between the data files is extracted; and then sensitive data identification preformed from clustered data using our proposed algorithm. Finally, before sent it to analyst, hiding sensitive data is done by adding more noise. So that analyst can get required knowledge except individuals privacy information. Our work is experienced with IBM Log dataset and it is implemented in MATLAB and the experiments reveal that AFLM algorithm depicts far better performance on various characteristics such as privacy weight, execution speed, efficiency, accuracy, scalability, data quality, dissimilarity, number of rules, information loss, hiding failure, security, reliability and performance compared to DDES, EPPA and ALM algorithms. The limitations in DDES, EPPA and ALM algorithms were also overcome in AFLM algorithm. The following figure clearly depicts our proposed algorithm provides the enhanced security compared with other algorithms such as DDES, EPPA and ALM.



Security

VI. CONCLUSION

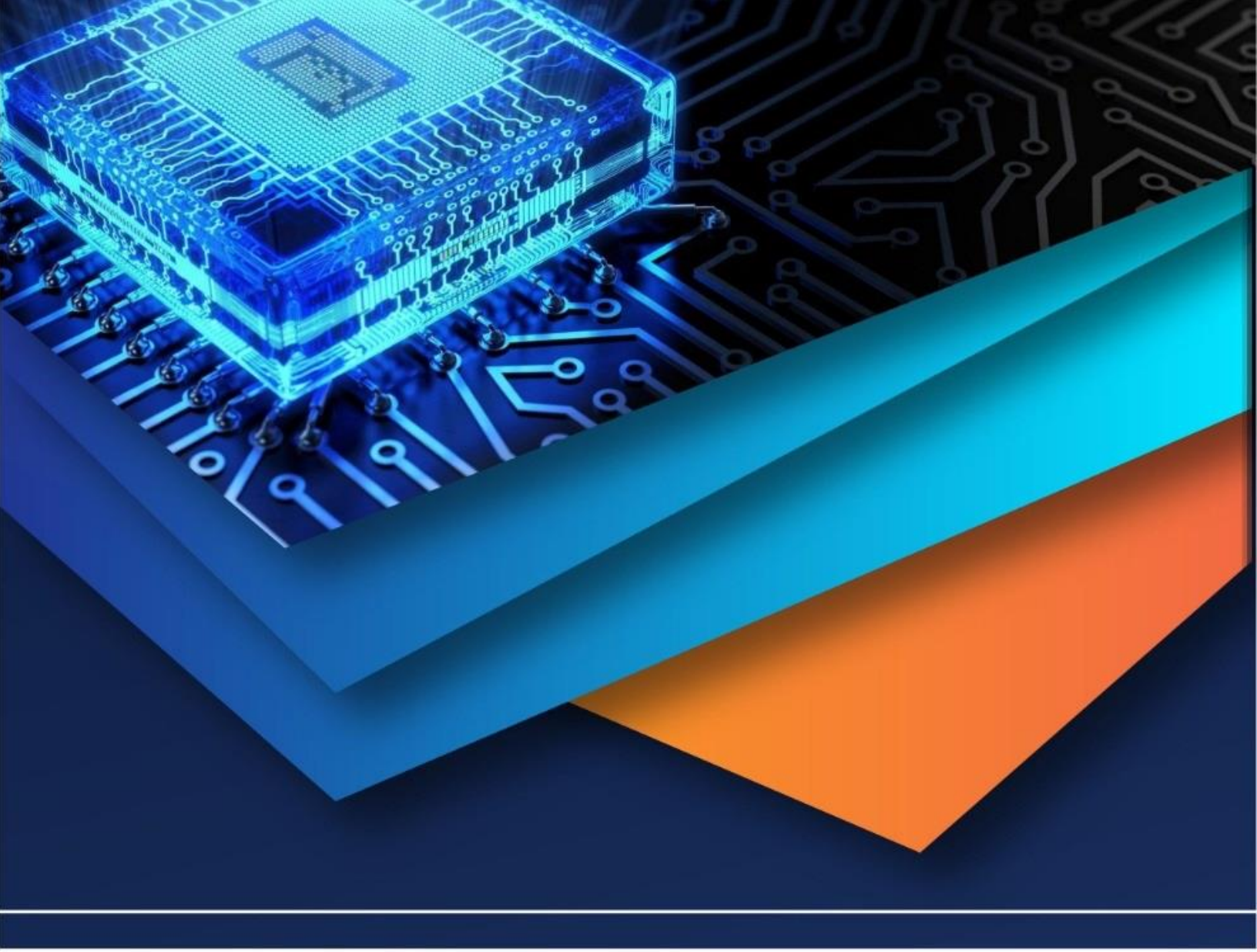
Big data is large amount of data which is unorganized and unstructured. Big data privacy is very important issue in while organizing big data. It is now essential for an organization to promise privacy in big data analytics. Techniques like cryptography, anonymization have limited potential when applied to big data. Differential privacy may be seen as a viable solution for big data privacy. One problem with this method is that analyst should know the query before using the differential privacy model. In this article we proposed Adaptive Firefly Laplace Mechanism Algorithm to preserve differential privacy in Big Data. Initially clustering analysis is conducted on the Big Data by using DBSCAN algorithm. Then our proposed algorithm is implemented with this clustered data. It works in two step process, in first step our algorithm find outs the sensitive data, then in second step it will add more noise to the sensitive data to hide sensitive information. In real time system analyst send query to Big Data to cluster data. Our proposed Adaptive Firefly Laplace Mechanism Algorithm provides required information to analyst without compromising on knowledge and by compared with other algorithms called DDES, EPPA and ALM, our algorithm provides better performance on execution speed, efficiency, accuracy, scalability, data quality, dissimilarity, less information loss, less hiding failure, security and reliability.

©IJRASET: All Rights are Reserved

REFERENCES

[1] Amit Kumar Gupta, Neeraj Shukla, "Privacy Preservation in Big Data using K-Anonymity Algorithm with Privacy Key", International Journal of Computer Applications (0975 – 8887), Volume 153 – No.5, November 2016.
[2] Anjana Gosain, Nikita Chugh, et al "Privacy Preservation in Big Data", International Journal of Computer Applications (0975 – 8887), Volume 100 – No.17, August 2014.
[3] Can Eyupoglu, Muhammed Ali Aydin et al., "An Efficient Big Data Anonymization Algorithm Based on Chaos and Perturbation Techniques", Entropy 2018, 20, 373; doi:10.3390/e20050373.
[4] D. Aruna Kumari and L. Anusha, "Analysis of Bigdata Security", American-Eurasian Journal of Scientific Research 11 (2): 93-97, 2016.
[5] Ester, Martin; Kriegel, et al, "A density-based algorithm for discovering clusters in large spatial databases with noise", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. CiteSeerX 10.1.1.121.9220, ISBN 1-57735-004-9.
[6] Jayashree Patil, , Y.C.Kulkarni, "Privacy Preserving Association Rule in Data Mining", International Journal of Engineering Research and Development, ISSN: 2278-067X, Volume 1, Issue 8 (June 2012), PP.18-21.
[7] NhatHai Phan, Xintao Wuy, et al., "Adaptive Laplace Mechanism: Differential Privacy Preservation in Deep Learning", Machine Learning (cs.LG); Machine Learning (stat.ML), Apr 2018.
[8] P. Ram Mohan RaoEmail, S. Murali Krishna, et al., "Privacy preservation techniques in big data analytics: a survey", Journal of Big Data, 22 September 2018.
[9] Sumit Vikram Tripathi, Ritukar et al., "Privacy-Preserving Data Encryption Strategy for Big Data in Mobile Cloud Computing Environment", International Journal of Innovative Research in Science, Engineering and Technology, 2nd National Conference on Recent Trends In Computer Science & Information Technology, ISSN : 2319 - 8753, Volume 7, Special Issue 6, May 2018.
[10] Vignesh Kumar G Arun Kumar S, "A Privacy Preservation Framework In Cross-Cloud Services For Big Data Applications", International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), ISSN: 0976-1353 Volume 22 Issue 3 – MAY 2016.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)