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.2019.6048

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*

# Privacy Preservation on Big Data using Efficient Privacy Preserving Algorithm

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:*** ***While analysing these intermediate data sets, the sensitive information can be accessed by misfeasors. Maintaining the confidentiality of this generated data set is very challengeable. Most of the existing systems uses cryptography methods for privacy preservation. The system consumes more time and cost because of the most often process of encryption and decryption of intermediate data sets which results in inefficiency and are expensive. In this article, we propose anonymization method to protect privacy of data during big data processing. In this article, we analyze a method of hiding sensitive information on big data by reconstruct a dataset according to the anonymization technique applied to clustered data. Unlike the other heuristic modification approaches, firstly, our method clusters a given dataset. Then we replace known values with unknown values in those transactions to hide a given sensitive information. Finally the sanitized dataset is generated. Our experiments show that the sensitive information can be hidden completely on the reconstructed datasets. Information leakage is problem in big data environment. Encryption of data is common method to reduce information leakage. Many cryptographic techniques are developed in past, but these techniques are much complex and time consuming. To improve search efficiency and to provide privacy preservation for big data environment Efficient Privacy Preserving (EPP) Algorithm is used in this article. EPP Algorithm is compared with encryption algorithm called Data Encryption Strategy and proved EPP algorithm performs better based on various criteria.***

***Keywords: Big Data, Anonymization, Privacy Preservation, k-Anonymity, T-Closeness, L-Diversity, Security, Information Loss, Data Encryption Strategy***

## I. INTRODUCTION

The primary goal in privacy preserving is to protect the sensitive data before it is released for analysis [5]. However the data may reside at same place or at different places. In such a scenario appropriate algorithms or techniques should be used which preserves any sensitive information in the knowledge discovery process. To address this issue there are many approaches adopted for privacy preserving data mining [5]. It can be classified as data distribution, data modification, data hiding and privacy preservation.

Anonymization in big data is a challenging job which changes the personal data with non-personal data. The tremendous amount of electronic data floating around us such as operational data, customer data, web data, social data, marketing data, computer data, supply chain data, transaction data, behavioural data etc. The two troublesome patterns at present forcing noteworthy effects on IT industry and research groups are Cloud computing and Big Data [9]. Protection of this information sets being a challenging task. Big Data concerns large-volume, complex, growing data sets with multiple and autonomous sources. Input of Big data is collected from online transactions like bank transactions and online shopping, queries requested in search engines, logs of telephone and mobile calls used in particular area, electronic mails and messages, videos, sensor logs, social media etc. It is stored by distributing along various servers. Big data are now rapidly emerging in all science and engineering domains [9]. Data anonymization approach is based on agglomerative hierarchical clustering algorithm in this research. Agglomerative hierarchical clustering is a bottom-up clustering method. It starts with every single tuple in a single cluster. Then, in each successive iteration, it merges the closest pair of clusters by satisfying some similarity criteria until some stopping rule is satisfied. The critical issue for an agglomerative hierarchical clustering algorithm is to choose the optimal pair of clusters for merging among a large amount of clusters in each iteration. In this method for data anonymization, it make the decision by both of the two factors including information loss and impurity gain [9]. The pair of clusters is chosen for merging, if the merging causes minimum information loss and maximum impurity gain. So, merging index, denoted by MI, is defined in our method to measure the quality of a pair of clusters on both features of information loss and impurity gain. This study mainly focuses on improving privacy on big data analysis. Section-2 briefly summarizes the previous work done by various researchers; In Section-3 various privacy preservation techniques are analysed. Section-4 focused on proposed Anonymization technique for big data analysis. In Section-5 the EPP algorithm is presented with illustration and example. As the detailed analysis of the experimental results on Big Data is analysed.

©IJRASET: All Rights are Reserved

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177*
*Volume 7 Issue VI, June 2019- Available at www.ijraset.com*

## II. RELATED REVIEW

This section presents various research works that are related to the proposed work. Ms. Devangi L. Kotak, Mrs. Shweta Shukla [5], proposed a method of hiding sensitive classification rules from data mining algorithms for categorical datasets. Our approach is to reconstruct a dataset according to the classification rules that have been checked and agreed by the data owner for releasing to data sharing. Unlike the other heuristic modification approaches, firstly, our method classifies a given dataset. Subsequently, a set of classification rules is shown to the data owner to identify the sensitive rules that should be hidden. Then we replace known values with unknown values in those transactions to hide a given sensitive classification rule. Finally the sanitized dataset is generated from which sensitive classification rules are no longer mined. Our experiments show that the sensitive rules can be hidden completely on the reconstructed datasets.

While non-sensitive rules are still able to discovered without any side effect. Namit Gupta [6], survayed various privacy preserving data mining algorithms and comparing different kinds of privacy preserving classification rule mining algorithms. S.Arun Kumar, Dr. M. S. Anbarasi [9], proposed various methods to protect privacy of data during big data processing. Cloud computing provides remote access for data storage and applications to users and organizations. Many Enterprises are utilizing the services of cloud computing to access the data easily without maintaining the data by its own. This makes the enterprise to assure that the system is highly scalable and is available with reduced cost of setup and maintenance. Cross-cloud service is best and suitable approach for large-scale big data processing system as big data processing system required huge data storage and computation power. Complex web based application of big data processing generates huge amount of data sets which are stored in remote location in cloud. While analysing these intermediate data sets, the sensitive information can be accessed by misfeasors. Maintaining the confidentiality of this generated data set is very challengeable. Most of the existing systems uses cryptography methods and stores the encrypted data sets in cloud. The system consumes more time and cost because of the most often process of encryption and decryption of intermediate data sets which results in inefficiency and are expensive.

## III. PRIVACY PRESERVATION TECHNIQUES

In privacy preserving data mining and data publishing, protection of privacy is achieved using Anonymization and Cryptography, among which k-anonymity and k-anonymity based algorithms like Datafly, Incognito and Mondrian are the most commonly used techniques. Our proposed work proved Anonymization method is a best method compared with encryption. The Various Privacy Preservation Techniques involve K- Anonymity, L-Diversity, and T-Closeness. Data anonymization is an one of the method which is also helpful in hiding personal information. It is the process of changing data that will be used or published in a way that prevents the identification of key information. EPPA algorithm is used for anonymization. This algorithm consists of the three main stages [8]. The first stage is for analyzing the frequency of unique attribute values for each quasi-identifier and then finding the crucial values according to frequency analysis. The second stage utilizes a chaotic function to designate new values for the chosen crucial values. In the final stage, data perturbation is performed. The drawback of this algorithm is 100% privacy is not guaranteed, chances of re-identification exists in large data and more information loss.

## IV. PROPOSED WORK

In our proposed work at first phase, big data is clustered using following DBSCAN algorithm. Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996. It is a density-based clustering non-parametric algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away).

DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature. In 2014, the algorithm was awarded the test of time award (an award given to algorithms which have received substantial attention in theory and practice) at the leading data mining conference, KDD.

After clustering process over, the second phase is finding sensitive attribute and adding noise to the sensitive attribute, so that the sensitive information is hided and privacy preservation is achieved. Efficient privacy preserving algorithm works on three stages. The first stage is for analyzing the frequency of unique attribute values for each quasi-identifier and then finding the crucial values according to frequency analysis. The second stage utilizes a chaotic function to designate new values for the chosen crucial values. In the final stage, data perturbation is performed. The following figure 2 depicts our proposed work;

©IJRASET: All Rights are Reserved

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177*
*Volume 7 Issue VI, June 2019- Available at www.ijraset.com*

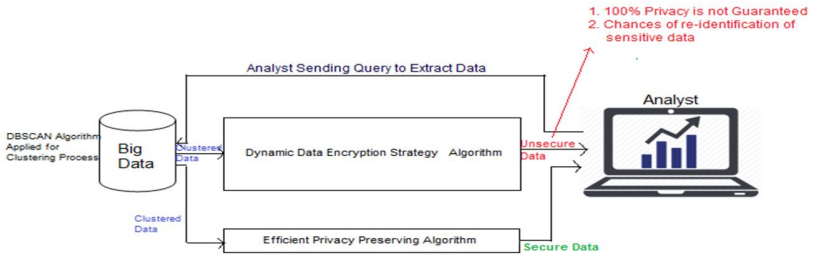



Figure 1: Anonymization based Privacy Preservation

Input: Original input data set D, quasi-identifier attributes QI ($QI_1$, $QI_2$, . . . , $QI_q$), and sensitive attribute SA
Output: Privacy preserved data set $D_p$
Initial assignments: c = 0, ʎ= 3.99, iteration = 400
1: d = |D|
2: q = |QI|
3: for i = 1 to q do
4: $nu_i$ = number of unique values for each $QI_i$
5: for j = 1 to $nu_i$ do
6: $u_{ij}$ = unique values for each $QI_i$
7: $v_{ij}$ = number of records containing the unique value $u_{ij}$
8: end for
9: end for
10: Sort $u_{ij}$ in ascending order based on $v_{ij}$ for each $QI_i$
11: $record_place_i$ = Ø (the size d X $nu_i$ for each $QI_i$)
12: for i = 1 to q do
13: for j = 1 to $nu_i$ do
14: for k = 1 to d do
15: if k-th record value in $QI_i$ == j-th value in sorted $u_{ij}$ then
16: c++
17: $record_place_i$ (c, j) = j
18: else
19: continue
20: end if
21: end for
22:c = 0
23: end for
24: end for
25: for i = 1 to q do
26: $r_i$ = round ($\log_2 nu_i$)
27: end for
28: for i = 1 to q do
29: $x_{i1}$ = 0.1
30: for j = 1 to iteration do
31: $x_{ij+1}$ = ʎ x $X_{ij}$ x(1 - $X_{ij}$)
32: end for
33: end for
34: Determine the new attribute values for the first $r_i$ value in sorted unique values $u_{ij}$ based on the record places $X_{ij}$ for each $QI_i$
35: Replace the chosen record values in D with the determined new values
36: Return $D_p$

Figure 2: Encryption Privacy Preservation Algorithm

©IJRASET: All Rights are Reserved

## V. EXPERIMENTAL RESULT

In this section, the performance metrics used for evaluation of the efficient privacy preserving algorithm compared with dynamic data encryption strategy algorithm and proved efficient privacy preserving algorithm performs better compared with dynamic data encryption strategy algorithm. The proposed algorithm is implemented in MATLAB.

Anonymization technique proved, it is a best method compared with cryptographic technique based on various criteria such as efficiency, scalability, data quality, accuracy, completeness, consistency, hiding failure. In terms of efficiency it performs the assessment of the resources used by a privacy preserving data mining algorithm is given by its efficiency, which represents the ability of the algorithm to execute with good performance in terms of all used resources. Performance is assessed, as usually, in terms of time and space, and, in case of distributed algorithms, in terms of the communication costs incurred during information exchange. Time requirements can be evaluated in terms of CPU time, or computational cost, or even the average of the number of operations required by the PPDM technique. Clearly, it would be desirable that the algorithms have a polynomial complexity rather than an exponential one. Anyway, it can be useful to compare the performance of the privacy preserving method with the performance of the data mining algorithm for which the privacy preserving method has been developed. Our expectation is that the execution times of the hiding strategies be proportional to the execution times of the mining algorithms that extract the sensitive information.

Space requirements are assessed according to the amount of memory that must be allocated in order to implement the given algorithm. Finally, communication requirements are evaluated for those data mining algorithms, which require information exchanges during the secure mining process, as the cryptography based techniques. It is measured in terms of the number of communications among all the sites involved in the distributed data mining task. On other hand Scalability is another aspect that it is important to assess when a PPDM algorithm is analyzed: it describes the efficiency trends for increasing values in data sizes. Therefore, such parameter concerns the increase of both performance and storage requirements together with the costs of the communications required by a distributed technique when data sizes increase. Because of the continuous advances in hardware technology, it is today easy to store large amounts of data. Thus, databases along with data warehouses today store and manage amounts of data which are increasingly large. For this reason, a PPDM algorithm has to be designed and implemented for being scalable with larger and larger datasets. The less rapid is the decrease in the efficiency of a PPDM algorithm for increasing data dimensions, the better is its scalability. Therefore, we have to evaluate the scalability of a PPDM technique as an equally important requirement of such a kind of algorithms.

Data quality is, thus, an important parameter to take into account in the evaluation of a PPDM technique. Accuracy measures the proximity of a sanitized value a' to the original value a. Completeness evaluates the degree of missed data in the sanitized database. Consistency is related to the internal constraints, that is, the relationships that must hold among different fields of a data item or among data items in a database. In hiding failure, the percentage of sensitive information that is still discovered, after the data has been sanitized, gives an estimate of the hiding failure parameter. Most of the developed privacy preserving algorithms are designed with the goal of obtaining zero hiding failure. Thus, they hide all the patterns considered sensitive. However, it is well known that the more sensitive information we hide, the more non-sensitive information we miss. Thus, some privacy preserving data mining algorithms have been recently developed which allow one to choose the amount of sensitive data that should be hidden in order to -find a balance between privacy and knowledge discovery.

The following figure 3 proves that EPPA provides more security compared with DDES algorithm.

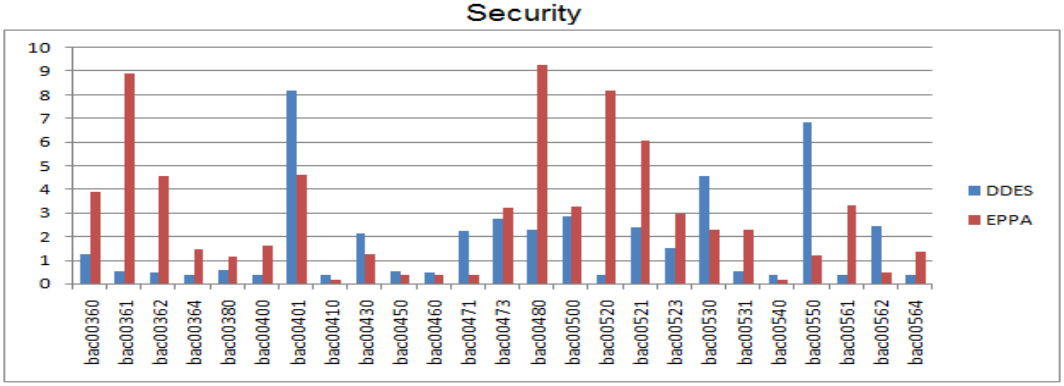



Figure 3: Security (DDES vs EPPA Algorithm)

©IJRASET: All Rights are Reserved

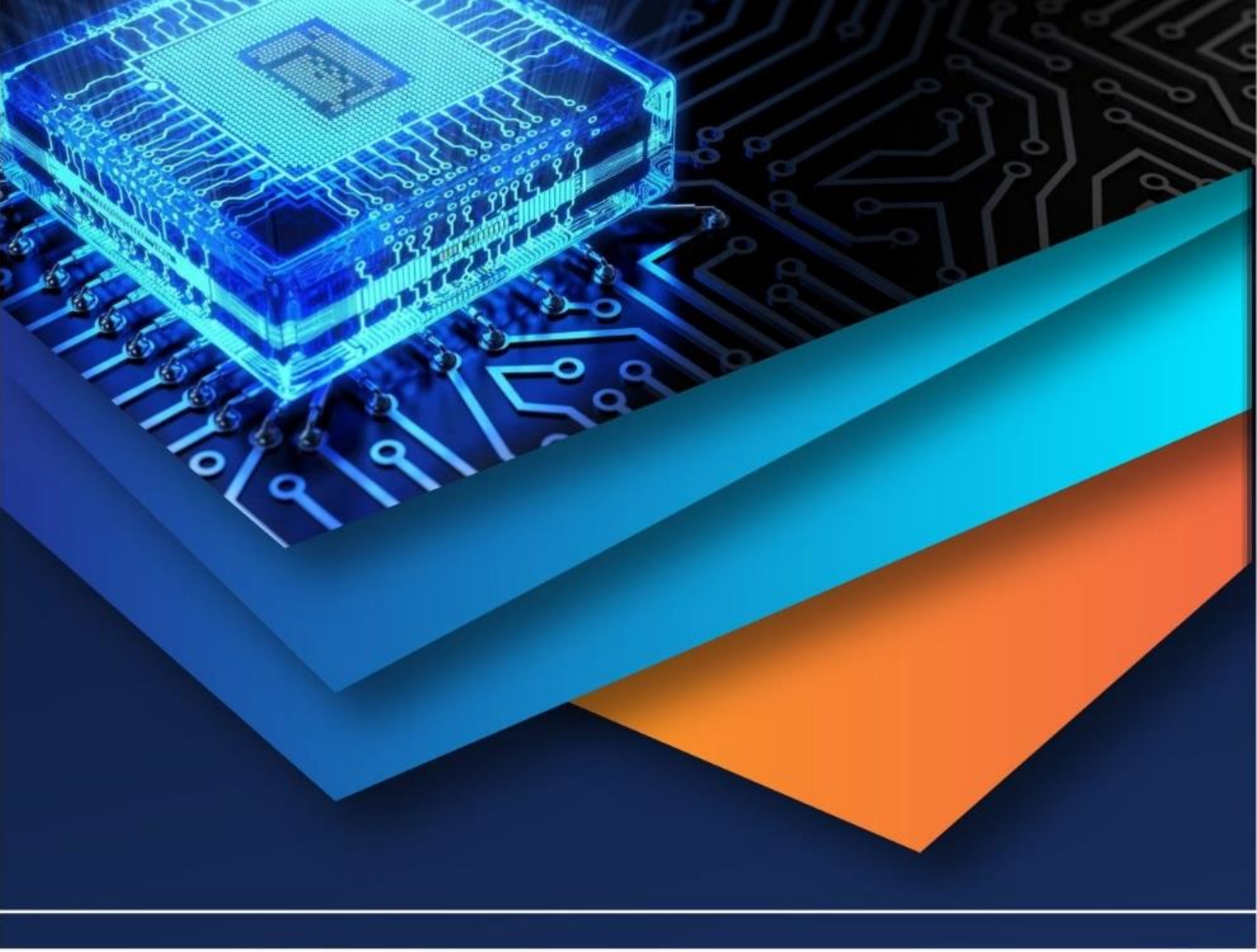
## VI. CONCLUSION

This article solely focused on the privacy issues of big data and considered the practical implementations in Big Data. The EPP Algorithm an anonymization approach was designed to maximize the efficiency of privacy protections. Our approach works on 3 phases. In first phase DBSCAN algorithm is used to cluster the IBM dataset. Then EPP algorithm is implemented on clustered output and finds sensitive attribute. Finally, the sensitive attribute is hidden by adding the noise to it. This work proved the sensitive level of privacy using anonymization which is compared with cryptographic method. The experimental evaluations showed the EPP algorithm had an adaptive and superior performance based on security and information loss.

## REFERENCES

[1] 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.
[2] Dr. Marinos Papadopoulos, "Main Anonymization Techniques for Personal Health Data", data Protection Working Party, opinion 05/2014 on anonymization techniques, adopted on April 10, 2014, 0829.
[3] 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.
[4] E.K.Girisan, Reena Cherian, "A Comprehensive Study on Meta Data Indexing Methods for Big Data and Multi Dimensional Database", Journal of Network Communications and Emerging Technologies (JNCET), Volume 7, Issue 10, October (2017).
[5] Ms. Devangi L. Kotak, Mrs. Shweta Shukla, "Protecting Sensitive Rules Based on Classification in Privacy Preserving Data Mining", International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 11, November - 2013.
[6] Namit Gupta, "Comparative study on classification privacy preserving data mining algorithms", 4th International Conference on System Modeling & Advancement in Research Trends (SMART) College of Computing Sciences and Information Technology (CCSIT) ,Teerthanker Mahaveer University , Moradabad, 2015.
[7] Nidhi Maheshwarkar, Kshitij Pathak et al., "Performance Issues of Various K-anonymity Strategies", International Journal of Computer Technology and Electronics Engineering (IJCTEE) , Volume 1 , Issue 2, ISSN 2249-6343.
[8] Ran Vijay Singh and Agilandeeswari L, "Secure open cloud in data transmission using reference pattern and identity with enhanced remote privacy checking", IOP Conf. Series: Materials Science and Engineering 263 (2017) 04,2006.
[9] S.Arun Kumar, Dr. M. S. Anbarasi, "A Privacy Preservation Framework In Cross-Cloud Services For Big Data Applications", International Journal Of Current Engineering And Scientific Research (IJCESR), ISSN (print): 2393-8374, (online): 2394-0697, volume-5, issue-2, 2018.
[10] 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.
[11] Tamas S. Gal, Zhiyuan Chen et al., "A Privacy Protection Model for Patient Data with Multiple Sensitive Attributes", International Journal of Information Security and Privacy, 2(3), 28-44, July-September 2008.

©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)