International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 15 (2018) pp. 11824-11829

Implementing Big Data Privacy with MapReduce for Multidimensional Sensitive Data

Vineetha Venugopal[1], S.Maria Celestin Vigila[2]

Research Scholar, Department of Software Engineering, Noorul Islam centre for Higher Education, Kumaracoil 629180, Tamilnadu, India.

Associate Professor, Department of Information Technology, Noorul Islam centre for Higher Education, Kumaracoil 629180, Tamilnadu, India.

Abstract

Data privacy is one of the most vital concerned issues while processing large datasets in Big Data Applications. The term Big data describes combination of large amount of structured data such as numerals, dates etc; semi-structured data such as XML documents and unstructured data such as video and audio files. Big data processes include data retrieval, storage, and analytics. Due to increase in all areas of technological perspective, Big data is more prone to security threats. Current anonymization techniques can provide solutions to these issues to an extent. But these methods have certain drawbacks. They lack scalability and performance efficiency. In this paper, a framework is introduced that incorporates Java programming language in Hadoop system. It provides multi-dimensional anonymization for sensitive data. A sample medical data set is used which is stored in MySQL database. The result of the proposed work is compared with existing anonymization techniques and it showed better performance in accuracy and execution speed.

Keywords: Big data, Anonymization, Data privacy, Hadoop, MapReduce.

INTRODUCTION

Big data can be used to extract potential information. The structured data are in a clearly defined manner while the unstructured data are in an unorganized or undefined way. Definition of big data is 4Vs. i.e. volume, variety, velocity, and variability of data. Data processing has a critical influence on synchronous data performance. Large number of scientists are inspired to deal with this confront. The client permits developers to develop who has no prior practice in parallel systems and distributed systems. It allows them to utilize easily the huge distributed system devices which are beneficial for the system. The way is hugely quantifiable and huge data processing is simply parallelized.

Traditional data uses anonymity techniques that are used to veil private information. It causes side area attacks. Anonymization techniques, such as k-anonymization, have been broadly engaged to tackle such purposes. Anonymization methods falls under two two main sections. The first series includes methods to simplify data based on the classification tree's top one. They are referred as bottom-up generalization (BUG) techniques. The second approach passes from taxonomic wood top to base. They are called as top-down specialization (TDS) techniques [1-2].

Conventional anonymization techniques lack scalability and do not deal with up huge volume data, and their performance shrinks. There is a necessity to scale traditional methods. The k-anonymity techniques categorize a set of fields, known as Quasi-identifiers i.e. Q-ID [3]. Quasi-identifiers are columns which contain personal details, where adversaries use these attributes which uncover secret details by connecting them to related external rudiments.

The currently used anonymous methods are to combine half identifiable columns using steps that use the taxonomic tree that intend to improve data privacy. Quasi-Identifier attribute values replace the taxonomic tree values or intervals. The best value of the interval is defined as the best reduction. In Top-down specialization, measurement is used to calculate the highest results of semi-identification for better cut. Former techniques calculate middle value for sorting the best cut within the arithmetical Quasi-Identifier columns. But, the quantifiability issues are not adequately addressed in these techniques.

Large size data is a barrier for these methods, since they should fit data into the accessible restricted memory. Big data processing cause damage to the hardware and it results in inefficiency of performance in parallel computing environments. To beat such pitfalls, there are methods that divide huge sized data randomly into small chunks of data blocks. Such techniques are used such that those are fit in the memory for making required computations. But the solution is not really scalable, as the masking of anonymized data is increased and it deteriorates the performance.

In this paper, a novel Multi dimensional sensitivity based anonymization method is introduced. It uses Bottom up Generalization in k-anonymity. It addresses the measurability and effectiveness. This method utilizes the anonymization for multi-dimensional sensitive data (MDSBA). An important goal of this method is a beneficial enhancement of improving the anonymization process and data efficiency. However, most of the proposed methods for evaluating K values are expensive solutions, and data owners will need to scan multiple data again and again to allow suitable K-Finding. Also, proposed methods do not investigate the increase in the

number of Q-id characteristics for the day's identification improvement.

In this work, a solution to find the optimal k value in k-anonymity method is being introduced. Most of the anonymization methods used currently may divide huge data into blocks of smaller chunks to reduce usage of heap memory. The uncertainty linked with dividing data into sections may result in reduced effectiveness of the data processed. The proposed technique simply divides the data, and the technique is not the basis of random data division. The continuous division process proposed in this method provides an appropriate data for parallel and distributed sphere. Moreover, there is an alternative heuristic method that provides the appropriate value for k. This task is based on the large number of workloads and the anonymity that enters the business, which creates a framework of this work.

RELATED WORKS

Data analytics and their utilization in big data environments witnessed a rapid growth in the past few years. Several undesirable side-effects have appeared, it relates to data disclosure and privacy violations risks. This trend imposes to find privacy methods with a scale-up ability to cope with the big data growth. Data anonymization is one of the pioneer privacy solutions that can minimize such risks. Therefore, many researches focused on securing big data using anonymization techniques. In this section, some of the highlights of the relevant work in this area are outlined.

Benjamin C.M Fung, et.al [4] introduces a new framework that introduces a SQL-like Hadoop ecosystem with additional data division, including PigLatin. The solution provides delicate masking and hiding, based on user-level access privileges. The article introduces a simple classification technique that accurately measures the level of anonymity of anonymous information. The author Al Zobbi, et.al[2]proposed K-anonymity protection model. K-anonymity is an important protection model because it forms the basis of real systems. The proposed method resolves the performance and anonymization loss concern, and provides a multi-level access control.

The author L.Sweeney[18] introduced a method where users are required to reveal their private details in health records stored digitally in cloud services for data analysis and mining, which brings concern about secrecy of data. Jisha Jose Panackala[12] has proposed to resolve the problem with a compromising method based on a mining association and compatibility inconsistency based on public utility. Initially, this model was tested by sampling samples of basic data from the National Family Health Survey (NFHS-3), and this document proves estimation performance of adaptive utility based anonymization using data sets carries out the anonymization process without compromising the quality of the data.

Author P. Jain, et.al [11] focuses on privacy and security in large data, identifying confidentiality and security requirements in large data. This paper focuses on confidentiality by using existing methods such as K-anonymity, T-proximity, and L-diversity and its application in business. The author Karim Abderrahim [8] in his article refers to the major risks associated with big data and existing security mechanisms. The document focuses on the security of Hadoop and its components and it has the framework of maintaining and processing large data.

The author Puneet Goswami[22] in this document gives a look at the large data overview, the related challenges, the privacy and security concerns and the difference between privacy and security requirements in large data. In addition, different confidential models that can be extended to larger domains focus on the advantages and disadvantages of the privacy models of anonymous data. The author Ki Liang [14], provides text sharing mechanism that runs confidentiality to achieve the above qualities. It combines the proxy mechanism and accuracy in encrypting anonymous technique and the advantages are that texts can be accessed safely and conditionally many times, as well as identification of the main message, such as text-sending sender/ recipient information.

Author D. Shiva, et.al [8] introduces a scalable two-stage top-down specialisation (TDS) to provide large-scale data using the map redesign framework. The cloud is supported. A new group of two levels of this method map actions to achieve a specific calculation in a highly scalable way. This method receives input data and is divided into smaller datasets. Then the anonymization of small data sets will result in intermediate productivity results applied. Then, small data is merged and anonymity is replicated. Then every field that sensitizes the data set prefers this sensitive field. Then this sensitive field anonymization depends on the graph.

The author T.Karle and et.al [23] in their article discussed various anonymous techniques and algorithms. The paper focuses on generalization and suppression techniques and describes the Datafly and Mondrian algorithm and discusses their comparison. Larger security and privacy issues have arisen with Big Data, which is probably not resolved by traditional security solutions. Therefore, the author Nikunj Joshi[21] in this article aims to provide an overall perspective on the image of Big Data Security and Privacy Issues.

In the documentary, Matturdi B. [20] first criticized the benefits of big data and security and privacy in Big Data. There are some possible methods and techniques to ensure the safety and privacy of Big Data. The author Elisa Bertino[9] introduces a large-value data security and privacy research program. The document addresses the challenges and guidance in the field of research on data confidentiality and reliability in the context of large data. The author Boel Nelson[7] in the review aimed at exploring security and privacy research in large data, outlining and providing a structure for what is currently under way. Moreover, which documents link security and privacy with large data and which categories cover these documents are being investigated.

In this paper, the author Lifei Wei, et.al [19] proposed prohibition on fraud regarding privacy and security audit protocol in SecCloud. This is the first report linked to the secure Computing Cloud Audit and a few signatures of depreciation fraud related to secrecy by competent, inspection

accounts and model potential techniques. A detailed analysis is given to get the ideal sample size to reduce the cost.

Author Kang Soo, et.al [20] proposed a privacy protection data mining technique in Hadoop to resolve confidentiality violation without the degradation of utilities. In this work, author Al Zobbi M, et.al [2] compares his proposed method with one of the recently proposed methods known as multidimensional top-down specialization. The comparison shows the limitations and contamination of the big data structure by applying the top-down specialization method, in contrast to the proposed method that adapts the parallel mistrust structure during the anonymization operation.

The publication of data to preserve privacy provides methods and tools to publish useful information while preserving the privacy of the data. Joneston Dhas et.al [13] proposed a framework for health record to securely store in big data. The author Jacques Bughin [10] finds that distribution asymmetry arises from a few telecommunication companies that are able to follow key management practices and big data. To provide privacy to the data is a major issue so that the third party is not able to access the sensitive information. Owing to these existing works on anonymization and its popularity, it is proposed to implement sensitive anonymization of multi-dimensional data. The framework efficiently anonymizes big data, by incorporating existing anonymization methods with MapReduce.

PROPOSED ARCHITECTURE

In the proposed work, Multidimensional sensitivity based anonymization method is implemented by MapReduce operations. This paper, introduces K-Anonymity for a Defensive Mode against Privacy Approval. K-anonymity identifies an equal number of data records. Instead of setting the k-anonymity parameters, it provides guidelines for the data owner. This article proposes a framework for robotic anonymous control. It provides anonymity for large data in the lower way, provides better access control to the framework by splitting Quassi-Identifier properties into vertical groups with two or four features for each cluster.

The basic method of access is introduced through three methods; Probable Value of Q-Identifier Features, Ownership Level K, and Q-Identification Group Method. The probability of Q-ID is calculated by counting the number of unique values in a specific attribute. Ownership level K is a key factor in identifying user access levels. K value is a factor indicating k-anonymity, which determines the number of equal anonymous records. Figure 1 shows the architecture of Multidimensional sensitivity based anonymization.

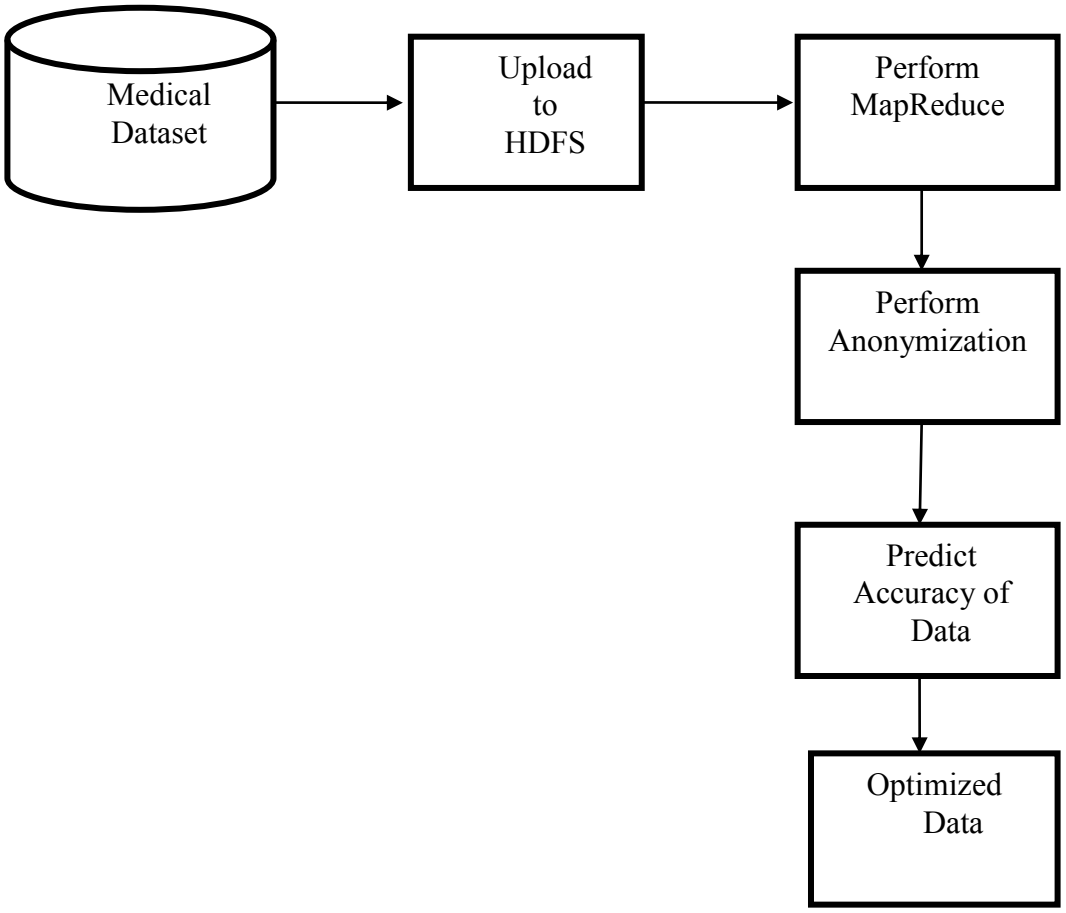

Figure 1. Architecture of Multi Dimensional Sensitivity Based Anonymization

Multi dimensional sensitivity based technique is used to protect sensitive data. It can protect large size data. Here the input data is a sample medical data of a hospital. The data set is loaded into the hadoop distributed file system(HDFS). Then Map reduce operations are performed. In Map Reduce operation the data is split as various blocks as per the capacity of hadoop distributed file system. In Hadoop distributed file system data is split into each block size, each of 64KB. Then shuffling of data of data is performed. i.e data is re-ordered.

After shuffling, data is sorted and the block size of each data block is reduced by the reducer operation. Then k-anonymization is performed. Probability of sensitive Quasi-identifier attributes is calculated. High probability values and low probability values based on Q-ID attributes are calculated. The result is an optimized data with calculated results of probability. The steps involved in multidimensional sensitivity based anonymization are splitting, shuffling, sorting, reducing and anonymization process.

Step 1: Splitting Process

Multidimensional sensitivity based anonymization framework breaks the input data into 3 blocks of data and stores it different nodes. The size of each block of data is 64 Kb. Each block is stored in a separate Hadoop Distributed File System folder. Hadoop distributed file system can store each block with 64kb size of data.

Step 2:Shuffling Process

In the shuffling process the output after the mapping function to transferred to the reducer. There shuffling process takes place. Sorting process takes place in the mapper. Then it sends the sorted data as input to the reducer.

Step 3: Sorting Process

The key-value pairs which are created as output in the mapper are not in a sorted format. They can be in any order. The key-value pairs in the Mapper are mostly sorted by keys and not by values. If it was in sorted format it would have saved the time for reducer.

Step 4: Reducing Process

Reducer takes the key-value pairs and sorts it and produces the sorted key-value pairs as output. The process starts in the mapper. The mapper does not sort the input. It transfers the unsorted data as output to the reducer. The reducer performs the sorting and shuffling process. The output of the reducer is a set of sorted key-value pairs.

Step 5: K-Anonymization Process

The main goal of multidimensional sensitivity based anonymization process is to divide the data technically and decrease the volume of data for the anonymization process. It perpendicularly bundles some Quasi-Identifiers and divides them into small groups. Almost all the datasets hold more than one class attribute and a significant number of Quasi Identifiers. The bigger number of Quasi Identifiers may produce huge calculation cost and data overflow, which suddenly terminates the program. Also, bundles of groups of Quasi-Identifiers assist the execution of authorization and access control. Hence, the method supports the anonymization efficiency and scalability.

PERFORMANCE ANALYSIS

The proposed framework efficiently anonymizes big data, by incorporating existing anonymization methods with MapReduce in NetBeans IDE. It is done in Java programming language by implementing Hadoop distributed file system and MapReduce in the framework. A sample medical dataset is used in MySQL and it is uploaded to the HDFS. HDFS can accelerate the MapReduce tasks, by gathering, shifting and parting of information. The proposed work adopts both bottom-up and top-down approaches to perform anonymization.

The results through implementation of proposed Multidimensional Sensitivity based Anonymization by Map Reduce operations using big data tool Hadoop and Medical Dataset is compared with existing K-Anonymity algorithm with bottom-up generalization (BUG), top-down specialization (TDS) and quasi-identifiers (Q-IDs) technique. The performance analysis of proposed anonymization method and existing anonymization method is carried out using the metrics of accuracy and execution time.

Accuracy

Accuracy is the comparative measurement of the original attribute values of the dataset with the anonymized values.

$$\text{Accuracy} = (\text{Attribute DataValue} \not\text{Qid} * \text{Number of Tuples}) * 100 \quad (1)$$

Table 1 gives the values of accuracy calculated for the proposed Multidimensional Sensitivity Based Anonymization by Map Reduce operations and the existing K-Anonymity algorithm as stated by the formula specified in Equation (1).

Table 1. Comparative results of accuracy

Input	Accuracy (%)	
Data Value	K-anonymity	MDSBA
1	5	10
2	20	15
3	30	35
4	40	55
5	65	82

Fig 2 shows of Accuracy for existing algorithm K-Anonymity with proposed MDSBA Multi Dimensional Sensitivity Based Anonymization method by Map Reduce operations. In generated graph the X axis is taken as the anonymization of medical data and Y axis is taken as accuracy.

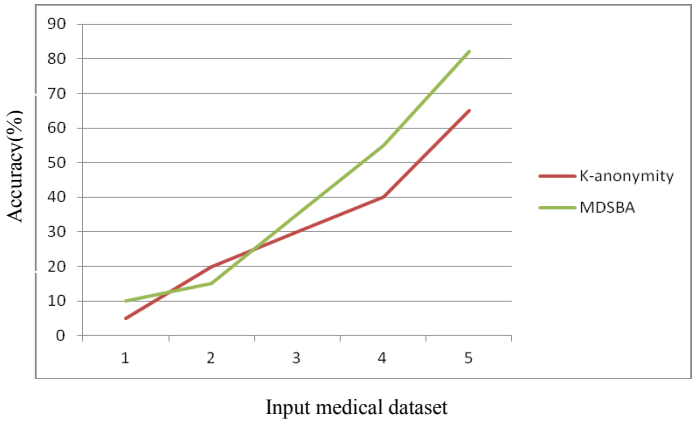

Figure 2. Comparison of Accuracy

Execution Time

Execution time is the time the anonymization method takes to execute. Table 2 gives the values of execution time calculated for the proposed Multidimensional Sensitivity Based Anonymization by Map Reduce operationsand the existing K-Anonymity algorithm as stated by the formula specified in Equation (2).

Execution Time=End time-Start time in milliseconds (2)

Table 2. Execution Time Comparison

Input	Execution Time (ms)	
Medical	K-	MDSBA
1	10224	124
2	1267	335
3	1345	432
4	1430	523
5	1984	1619

Fig 3 shows of Execution time for existing algorithm K-Anonymity with proposed Multi Dimensional Sensitivity Based Anonymization method by Map Reduce operations. In generated graph the X axis is taken as the anonymization of medical data and Y axis is taken as Execution time.

Figure 3. Comparative results of Execution time

The proposed Multi Dimensional Sensitivity Based Anonymization method by Map Reduce operations implements iterated split and filter technique for data records with mapping, shuffling, and reducing The K-anonymization then is applied to the highest probability and lowest probability attribute for each split, which gives the better performance in comparison. The proposed work adopts both bottom-up and top-down approaches to perform anonymization. The experiments, with the existing dataset, have confirmed that accuracy and execution speed is much faster and better than k-anonymity.

CONCLUSION

In this paper, sensitive anonymization is implemented for a multi-dimensional data. The framework efficiently anonymizes big data, by incorporating existing anonymization methods with MapReduce in NetBeans IDE. It is done in Java programming language by implementing Hadoop distributed file system and MapReduce in the framework. A sample medical dataset is used in MySQL and it is uploaded to the HDFS. HDFS can accelerate the MapReduce tasks, by gathering, shifting and parting of information. The proposed work adopts both bottom-up and top-down approaches to perform anonymization. The experiments, with the existing dataset, have confirmed that accuracy and execution speed is much faster and better than k-anonymity. It has demonstrated a low execution time on anonymization process with low estimations of k. Future tasks extend experiments to add more complexity and advanced settings to delicately pass the proposed method. The aim is to establish a full multi-dimensional framework by providing support mechanisms that improve access to service providers and data services

REFERENCES

[1] Aditya Dev Mishra, 2016, “Big Data Analytics for Security and Privacy Challenges”, International Conference on Computing, Communication, and Automation, pp 50-53.

[2] Al Zobbi M., Seyed Shahrestani, Chun Ruan, 2017,"Multi-Dimensional sensitivity-based anonymization of Big Data", Journal of Big Data,Vol 16,pp 415-430.

[3] Al-Zobbi, Seyed Shahrestani, Chun Ruan , 2017,"Improving MapReduce privacy by implementing multi-dimensional sensitivity-based anonymization", Journal of Big Data,Issue 1.

[4] Benjamin C.M, Ke Wang, Rui Chen, Philip S.Yu ,2010, "Privacy preserving Data Publishing : A survey of recent developments", Journal of ACM Computing Surveys, Vol 42,Issue 4.

[5] "Anonymizing Classification Data for Privacy Preservation", 2007,Journal of Security in Computing and Communications, pp 99-109.

[6] Benjamin C.M, Ke Wang, Rui Chen, Philip S.Yu , 2010, "Privacy preserving Data Publishing : A survey of recent developments", Journal of ACM Computing Surveys, Vol 42,Issue 4.

[7] Boel Nelson, 2016 ," Security and Privacy for Big Data: A Systematic Literature Review", IEEE International Jounal on Big Data,Vol 34,Issue 5.

[8] D.Shiva, P.Karthiga, X.J.J Anitha, 2016, "A System to analysis real-time big data using Top down Specialization", International Education and Research Journal.

[9] Elisa Bertino, 2015,"Big Data-Security and Privacy", Jounal name-A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, pp 425-439.

[10] Jacques Bughin, 2016 ,"Reaping the benefits of Big Data in telecom" ,Journal of Big Data, Issue 1.

[11] Jain P, Manasi Gyanchandani, Nilay Khare, 2016, "Big data privacy: A technological perspective and review", Journal of Big Data, Issue 1.

[12] Jisha Jose Panackala, Anitha S Pillai, 2015 ,"Adaptive Utility-based Anonymization Model:Performance Evaluation on Big Data Sets", 2nd International Symposium on Big Data and Cloud Computing.

[13] Joneston Dhas J. L., Maria Celestin Vigila S. and Ezhil Star C., 2017, "A Framework on Security and Privacy-Preserving for Storage of Health Information using Big Data", International Journal of Control Theory and Application, Vol. 10, No. 10, pp. 91-100.

[14] Kai Liang, Willi Susilo, Joseph K. Liu, 2015 ,"Privacy preserving ciphertext Multi-sharing control for Big Data Storage", IEEE transactions on Information Forensics and Security.

[15] Kang Soo , Sehwa Park, Seog Park,2014, "Hiding a needle in a Haystack: privacy preserving Apriori Algorithm in MapReduce Framework", Proceedings of the First International Workshop on Privacy and Secuirty of Big Data, Pages 11-17.

[16] Karim, Abderrahim, Hayat, Mostafa, 2002, "Big Data Emerging Issues: Hadoop Security and Privacy", International Journal of Uncertainity, Fuzziness, and Knowledge-based Systems,Vol.5.

[17] Ke Wang, Philip S Yu, Sourav Chakraborty, 2004, "Bottom-up generalization: A Data Mining Solution to Privacy Protection", IEEE International Conference on Data Mining.

[18] Latanya sweeney, 2002," k-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, Issue 5.

[19] Lifei Wei, Haojin Zhu, Zhenfu Cao, Xiaolei Dong, Yunlu Chen, 2013, "Security and Privacy for Storage and Computation in Cloud Computing", Journal of Information Sciences, Vol. 258, Pages 371-386.

[20] Matturdi Bardi, Zhou Xianwei, Li Shuai, Lin Fuhong,2014, "Big Data Security and Privacy: A Review", Vol 11, Issue 14, Pages 135-145.

[21] Nikunj Joshi, Bintu Kadhiwala,2017, "Big Data Security and Privacy Issues-A Survey", International Conference on Innovations in Power and Advanced Computing Technologies.

[22] Puneet Goswami, 2014,"A Survey on Big Data & Privacy preserving publishing techniques", Journal of Big Data, Vol 10, pp. 395-409.

[23] T. Karle, Deepali Vora, 2017, "Privacy Preservation in Big Data using anonymization techniques", International Conference on Data Management, Analytics and Innovation.