Anonymization Based Fisher–Yates Shuffle Method for Streaming of Twitter Data

AR.Arunachalam, G.Michael, D. Vimala

***Abstract*: In this era of Big Data, many organizations are functioning with personal data, that has to be preserved for privacy reason. There are hazards to identify the individual details by using Quasi Identifier (QI). So to preserve the privacy, anonymization points us to convert the personal data into unidentified personal data. There are many organizations that produce the large data in real time. With the help of Hadoop components like HDFS and MapReduce and with its ecosystems, large volume of data can be processed in real time. There are many basic data anonymization techniques like cryptographic, substitution, character masking, shuffling, nulling out, date variance and number variance.**

Here privacy preservation is achieved for streaming data by using one of the anonymization techniques called 'shuffling' with Big data concept. K-anonymity, t-closeness, l-diversity are usually used technique for privacy concern in a data. But in all these techniques information loss and data utility are not preserved very well. Dynamically Anonymizing Data Shuffling (DADS) technique is used to overcome this information loss and also to improve data utility in streaming data.

***Keywords*: Big data, Privacy preservation, data anonymization, data masking, shuffling, Hadoop, Flume, Twitter.**

I. INTRODUCTION

In the past, the world's data doubled every century. Now it doubles every two years. This flood of data is driven by the Internet of Anything, including more data from the Internet, mobile devices, server logs, geolocation coordinates, social network, machine and sensors [8]. In some cases the streaming of data has to process in real time before it is analyzed. During process, there is a great challenge of maintaining privacy while still retaining the utility of the data. Researchers and business analysts require data which is reliable and intelligible; encrypting these data with cryptographic algorithms will not give them the data with their completeness/truthfulness. A better way of hiding those unique and joined attributes (quasi-identifiers) from identifying individuals is to use anonymization methods.
There are number of anonymization method like cryptographic, substitution, character masking, shuffling, nulling out, date variance and number variance [24]. In spite of all those methods, shuffling method can achieve high data utility and less information loss. Investigation is the thing that makes huge information wake up. Examination, containing various diverse computational advances, is the thing that fills the big-data transformation. Examination is the thing that makes the new incentive in enormous datasets, endlessly more than the total of the estimations of the parts. Each of these analyses will produce some insights which are to be favorable to both analyst as well as customers. In DADS method of data masking process, streaming of data can be analyzed in real time without revealing any sensitive values. Considering about the data size and speed, it can be only processed with the help of big data concepts. There are many Hadoop components and ecosystem to process this much of volume and velocity of data and DADS method makes data to solve re-identification of original sensitive values and maintain the utility of data in balanced state.

Revised Manuscript Received on July 22, 2019.

Dr.AR.Arunachalam, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai , India. Email: ararunachalam78@gmail.com

G.Michael, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai , India. Email: micgeo270479@gmail.com

D. Vimala, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai , India. Email: vimalamuthu3@gmail.com

II. RELATED WORKS

Many anonymization processes are done in traditional Bigdata. But applying in streaming data is more needed in day-to-day's life of many organizations and shuffling method can give more privacy as well as data utility. By combining this two streaming and shuffling will overcome the disadvantages of many existing systems and leads to develop a new method called Dynamically Anonymize Data using Shuffling (DADS).

III. DADS

Privacy for the streaming data can be preserved in this DADS method. There were many anonymization processes like generalization, suppression, etc. Contrast with those techniques, DADS strategy accomplish more protection than different strategies. Data veiling in Big Data is required for the present examinations of the certified structure; behind which are changes of necessities with the desire for complimentary usage of individual data by making them obscure. It is possible to process singular data into an express that has generally abstained from refinement joined into data. In PK anonymization method, it clarifies that the replace of entire attribute with the probability maximum value, which was done in Map Reduce process to handle Big data. Unfortunately, it cannot achieve full anonymous data and its data quality became low. So to find insight from this data cannot give correct output. In DADS method, it is resolved by effectively applying shuffling method in streaming of big data using Hadoop components and its ecosystem.

IV. ARCHITECTURE DIAGRAM

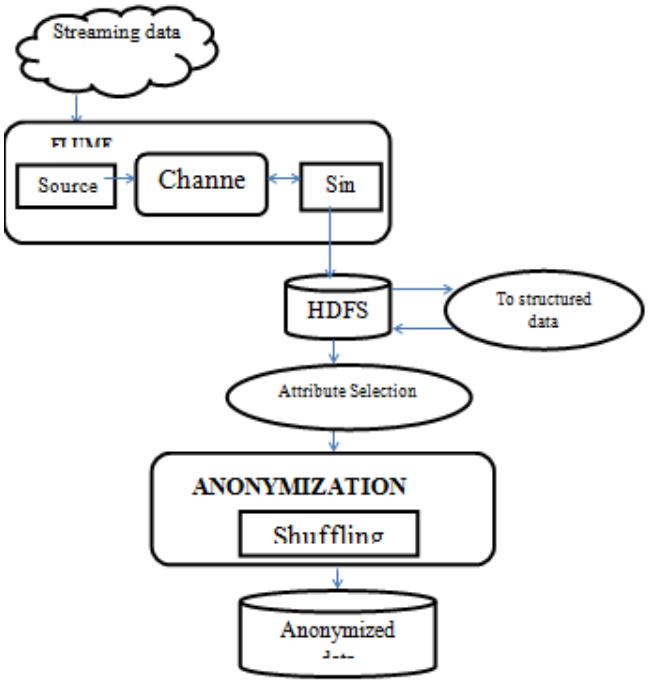

Figure 1: Architecture diagram of DADS

The streaming data source is one of the main parts of this DADS method. The streaming source generates the real time data to the Flume process. The data owner specifies the privacy requirement and then has to submit framework of the data masking technique. Figure 1 specifies the overall process done by DADS method. First the process needs the real time data and it is streamed into the distributed and scalable database. During retrieving the real time data some of the data are stored in unstructured or semi-structured data. These data cannot be processed very well as compared to structured data. So these data have to be converted to structured data and once again stored in the database.

After conversion of data, it under goes a selection process called attribute selection. In data sets, there may be 3 types of data that is sensitive data, insensitive data and quasi identifier. During anonymization process the sensitive data has to be left out for customer safety and insensitive data are not used for anonymization[13],[15],[17]. Among these data quasi identifier attribute is very much considered for data anonymization process. So only quasi attributes are considered for this DADS method for data masking.

Next step is applying shuffling method for selected data. Shuffling can be achieved in many ways. In DADS method, shuffling is done with a method called 'Fisher–Yates shuffle'. This method is quiet efficient, sequential, time complexity is optimal[14],[16],[18]. Using shuffling will decrease the information loss in the data set and increase the data utility. The output from this shuffling method is discussed with data owner and then released to data analysts or any research scholar without revealing the customers details.

V. RESULTS

SD-Streaming data
StD-Structured data
AS-Attribute selection data
AN-Anonymized data

1. Input the credential details to access the stream data(SD).
2. Using Apache Flume Source, grab SD to processing phase.
3. Redirect SD to store in database.
4. If SD is unstructured or semi-structured
4.1. Convert to structured data (StD)
5. Else
5.1. SD=StD.
6. Apply attribute selection
7. Find sensitive attribute, quasi identifier, insensitive attribute
8. If StD is sensitive attribute discard that.
9. Else if StD is insensitive attribute discard that.
10. Else StD=AS
11. Apply shuffle method
11.1. For (i=AS.length-1;i>0;i--)
11.2. A=AS[index]
11.3. AS[index]=AN[i]
11.4. AN[i]=A
12. Return Anonymized data(AN)

VI. DISCUSSIONS

From vast variety of streaming data, Twitter data is one of the free real time data which is available. While getting tweets from Twitter using Flume, it enables to search for a particular keyword. In twitter, it has many topics that are discussed in every day but not in same process. The trends in Twitter are changing continuously based on the events occurred in the world or to a specific location. So to search a key word try to give trends name in the key word to generate many tweets in short time. If other than trend, the data generates slowly compare to trends. Figure 2 describe about difference between trend and other key word.

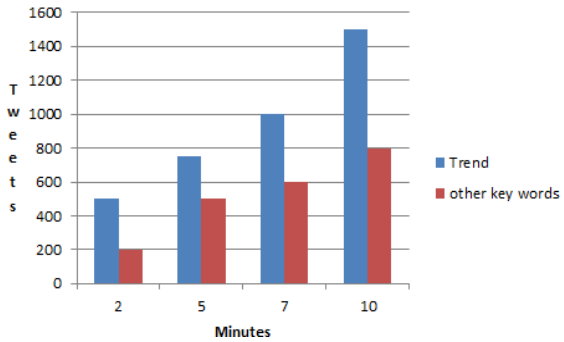

Figure 2: Difference between key words during retrieving data from twitter

In giving details about key words can also increase the tweet size. If key word is given as one or two then the tweet size is less[31],[33],[35]. If key words are given more than two, then the size will increase in every minute compare to other. Figure 3 describe about the difference between no of key word given to search the tweets.

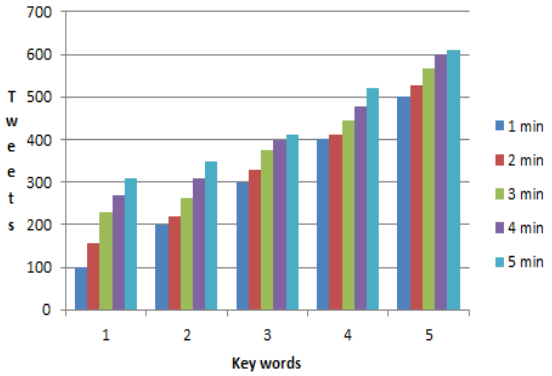

Figure 3: Difference between the no of key word give to search tweet and its timing

A definitive point of the venture is to give analysts the capacity to break down information from over the globe. Consequently, a fruitful information insurance instrument must guarantee that the anonym zed information gives results that are like that utilizing the first information. One significant part of this assessment is that the information was veiled without learning of the consequent investigations that will be done on the information[32],[34],[36]. Subsequently, this assessment gives a progressively broad evaluation of the adequacy of the veiling strategy. The data utility measured using the difference between original data and anonymized data. In this, DADS give cent percent equal information. So there will be no information loss in this DADS method. While compare to other methods there is information loss. The below chart describes about the information loss between the different anonymization methods[37],[39],[41].

- K anonymity
- L diversity
- DADS

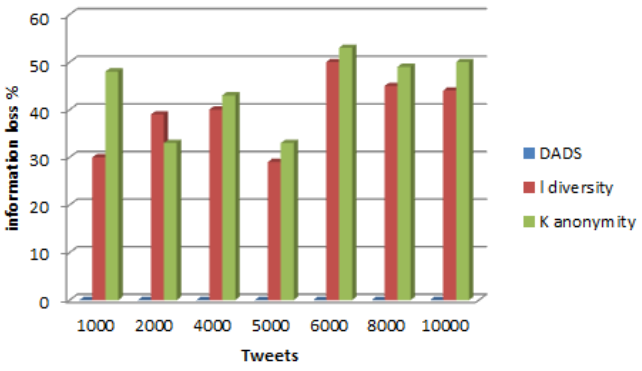

Figure 4: Difference between information loss among anonymization methods

Figure 4 describes about the information loss of the different anonymization technique. Those techniques are DADS, K- anonymity and l- diversity. In chart, the X-axis describes about the no. of tweets and Y-axis describes about the percentage of information loss. The insight of the graph reveals that DADS method has no information loss while compare to other two techniques.

VII. CONCLUSION

The interest for information from reviews, censuses or registers containing reasonable data on individuals or endeavors has expanded altogether throughout the most recent years. In any case, before information can be given to the general population or to specialists, privacy must be regarded for any informational collection potentially containing reasonable data about individual units. Privacy and utility can be accomplished by applying DADS (Dynamically Anonymize Data utilizing Shuffling) strategy to the information so as to diminish the divulgence danger of information without influencing the first information. Information rearranging is generally perceived as a hearty veiling method for private information. In the data world the information are created continuously and furthermore to examine the information progressively. Tragically the past model can't accomplish full information utility and can't process continuously. Be that as it may, DADS strategy can be connected for ongoing spilling information and furthermore for a wide range of information to protect mystery and decrease the data misfortune.

REFERENCES

[1] Kumaravel A., Rangarajan K.,Algorithm for automaton specification for exploring dynamic labyrinths,Indian Journal of Science and Technology,V-6,I-SUPPL5,PP-4554-4559,Y-2013
[2] P. Kavitha, S. Prabakaran "A Novel Hybrid Segmentation Method with Particle Swarm Optimization and Fuzzy C-Mean Based On Partitioning the Image for Detecting Lung Cancer" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5, June 2019
[3] Kumaravel A., Meetei O.N.,An application of non-uniform cellular automata for efficient cryptography,2013 IEEE Conference on Information and Communication Technologies, ICT 2013,V-,I-,PP-1200-1205,Y-2013
[4] Kumarave A., Rangarajan K.,Routing alogrithm over semi-regular tessellations,2013 IEEE Conference on Information and Communication Technologies, ICT 2013,V-,I-,PP-1180-1184,Y-2013
[5] P. Kavitha, S. Prabakaran "Designing a Feature Vector for Statistical Texture Analysis of Brain Tumor" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5, June 2019
[6] Dutta P., Kumaravel A.,A novel approach to trust based identification of leaders in social networks,Indian Journal of Science and Technology,V-9,I-10,PP--,Y-2016
[7] Kumaravel A., Dutta P.,Application of Pca for context selection for collaborative filtering,Middle - East Journal of Scientific Research,V-20,I-1,PP-88-93,Y-2014
[8] Kumaravel A., Rangarajan K.,Constructing an automaton for exploring dynamic labyrinths,2012 International Conference on Radar, Communication and Computing, ICRCC 2012,V-,I-,PP-161-165,Y-2012
[9] P. Kavitha, S. Prabakaran "Adaptive Bilateral Filter for Multi-Resolution in Brain Tumor Recognition" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN:

2278-3075, Volume-8 Issue-8 June, 2019
[10] Kumaravel A.,Comparison of two multi-classification approaches for detecting network attacks,World Applied Sciences Journal,V-27,I-11,PP-1461-1465,Y-2013
[11] Tariq J., Kumaravel A.,Construction of cellular automata over hexagonal and triangular tessellations for path planning of multi-robots,2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016,V-,I-,PP--,Y-2017
[12] Sudha M., Kumaravel A.,Analysis and measurement of wave guides using poisson method,Indonesian Journal of Electrical Engineering and Computer Science,V-8,I-2,PP-546-548,Y-2017
13] Ayyappan G., Nalini C., Kumaravel A.,Various approaches of knowledge transfer in academic social network,International Journal of Engineering and Technology,V-,I-,PP-2791-2794,Y-2017
[14] Kaliyamurthie, K.P., Sivaraman, K., Ramesh, S. Imposing patient data privacy in wireless medical sensor networks through homomorphic cryptosystems 2016, Journal of Chemical and Pharmaceutical Sciences92.
[15] Kaliyamurthie, K.P., Balasubramanian, P.C. An approach to multi secure to historical malformed documents using integer ripple transfiguration 2016 Journal of Chemical and Pharmaceutical Sciences92.
[16] A.Sangeetha,C.Nalini,"Semantic Ranking based on keywords extractions in the web", International Journal of Engineering & Technology, 7 (2.6) (2018) 290-292
[17] S.V.GayathiriDevi,C.Nalini,N.Kumar,"An efficient software verification using multi-layered software verification tool "International Journal of Engineering & Technology, 7(2.21)2018 454-457
[18] C.Nalini,ShwtambariKharabe,"A Comparative Study On Different Techniques Used For Finger – Vein Authentication", International Journal Of Pure And Applied Mathematics, Volume 116 No. 8 2017, 327-333, Issn: 1314-3395
[19] M.S. Vivekanandan and Dr. C. Rajabhushanam, "Enabling Privacy Protection and Content Assurance in Geo-Social Networks", International Journal of Innovative Research in Management, Engineering and Technology, Vol 3, Issue 4, pp. 49-55, April 2018.
[20] Dr. C. Rajabhushanam, V. Karthik, and G. Vivek, "Elasticity in Cloud Computing", International Journal of Innovative Research in Management, Engineering and Technology, Vol 3, Issue 4, pp. 104-111, April 2018.
[21] K. Rangaswamy and Dr. C. Rajabhushanamc, "CCN-Based Congestion Control Mechanism In Dynamic Networks", International Journal of Innovative Research in Management, Engineering and Technology, Vol 3, Issue 4, pp. 117-119, April 2018.
[22] Kavitha, R., Nedunchelian, R., "Domain-specific Search engine optimization using healthcare ontology and a neural network backpropagation approach", 2017, Research Journal of Biotechnology, Special Issue 2:157-166
[23] Kavitha, G., Kavitha, R., "An analysis to improve throughput of high-power hubs in mobile ad hoc network" , 2016, Journal of Chemical and Pharmaceutical Sciences, Vol-9, Issue-2: 361-363
[24] Kavitha, G., Kavitha, R., "Dipping interference to supplement throughput in MANET" , 2016, Journal of Chemical and Pharmaceutical Sciences, Vol-9, Issue-2: 357-360
[25] Michael, G., Chandrasekar, A.,"Leader election based malicious detection and response system in MANET using mechanism design approach", Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016 .
[26] Michael, G., Chandrasekar, A.,"Modeling of detection of camouflaging worm using epidemic dynamic model and power spectral density", Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016 .
[27] Pothumani, S., Sriram, M., Sridhar, J., Arul Selvan, G., Secure mobile agents communication on intranet,Journal of Chemical and Pharmaceutical Sciences, volume 9, Issue 3, Pg No S32-S35, 2016
[28] Pothumani, S., Sriram, M., Sridhar , Various schemes for database encryption-a survey, Journal of Chemical and Pharmaceutical Sciences, volume 9, Issue 3, Pg NoS103-S106, 2016
[29] Pothumani, S., Sriram, M., Sridhar, A novel economic framework for cloud and grid computing, Journal of Chemical and Pharmaceutical Sciences, volume 9, Issue 3, Pg No S29-S31, 2016
[30] Priya, N., Sridhar, J., Sriram, M. "Ecommerce Transaction Security Challenges and Prevention Methods- New Approach" 2016 ,Journal of Chemical and Pharmaceutical Sciences, JCPS Volume 9 Issue 3.page no:S66-S68 .
[31] Priya, N.,Sridhar,J.,Sriram, M."Vehicular cloud computing security issues and solutions" Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016
[32] Priya, N., Sridhar, J., Sriram, M. "Mobile large data storage security in cloud computing environment-a new approach" JCPS Volume 9 Issue 2. April - June 2016
[33] Anuradha.C, Khanna.V, "Improving network performance and security in WSN using decentralized hypothesis testing "Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016 .
[34] Anuradha.C, Khanna.V, "A novel gsm based control for e-devices" Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016 .
[35] Anuradha.C, Khanna.V, "Secured privacy preserving sharing and data integration in mobile web environments " Journal of Chemical and Pharmaceutical Sciences(JCPS) Volume 9 Issue 2, April - June 2016 .
[36] Sundarraj, B., Kaliyamurthie, K.P. Social network analysis for decisive the ultimate classification from the ensemble to boost accuracy rates 2016 International Journal of Pharmacy and Technology
[37] Sundarraj, B., Kaliyamurthie, K.P. A content-based spam filtering approach victimisation artificial neural networks 2016 International Journal of Pharmacy and Technology83.
[38] Sundarraj, B., Kaliyamurthie, K.P. Remote sensing imaging for satellite image segmentation 2016 International Journal of Pharmacy and Technology8 3.
[39] Sivaraman, K., Senthil, M. Intuitive driver proxy control using artificial intelligence 2016 International Journal of Pharmacy and Technology84.
[40] Sivaraman, K., Kaliyamurthie, K.P. Cloud computing in mobile technology 2016 Journal of Chemical and Pharmaceutical Sciences92.
[41] Sivaraman, K., Khanna, V. Implementation of an extension for browser to detect vulnerable elements on web pages and avoid click jacking 2016 Journal of Chemical and Pharmaceutical Sciences92.

AUTHORS PROFILE

Dr.AR.Arunachalam Associate Professor, Department of Computer Science & Engineering, Bharath Institute of Higher Education and Research, Chennai, India

G.Michael Assistant Professor, Department of Computer Science & Engineering, Bharath Institute of Higher Education and Research, Chennai, India

D. Vimala Assistant Professor, Department of Computer Science & Engineering, Bharath Institute of Higher Education and Research, Chennai, India