Journal of Advances and Scholarly Researches in Allied Education
Vol. 15, Issue No. 7, September-2018, ISSN 2230-7540

Review on Privacy Preservation on Big Data Using Map Reduce

Amit Kumar[1]* Dr. Manoj Kumar[2]

[1] Research Scholar, Shri Venketashwara University, Gajraula, Uttar Pradesh

[2] Department of Computer Science, Shri Venketashwara University, Gajraula, Uttar Pradesh

Abstratct – Systematic literature review is one of the main research methodologies in research work. It outline the existing information about security dangers and to get the genuine reflection of the security techniques utilized in the current in Cloud Computing, and to give an edge work to recommend counter measures for the future challenges to be looked in Cloud Computing. Systematic reviews are based on a defined hunt procedure that intends to distinguish however much of the pertinent literature as could reasonably be expected. Main point of Privacy preservation in data mining is to find out such solution which will minimize the risk of abuse of the data. There are number of powerful techniques for privacy preserving data mining which have been proposed after broad study of data mining network as of late. In request to play out the privacy preservation, the vast majority of the technique utilizes some type of transformation on the original data. Here for this situation it is imperative to maintain the advantages of privacy preservation even after the changed dataset is made accessible for mining. In this paper we study about the previous studies done in the field of privacy preservation on big data using map reduce.

----------X----------

INTRODUCTION

This is one of the prevalent techniques in privacy preserving data mining thinks about. By adding commotion to the original data, the estimations of the records are made. In this technique the individual estimations of the records can never again be recuperated as commotion added to original data is expansive enough to maintain the privacy. Randomization techniques accomplish both, privacy preservation and knowledge disclosure with the assistance of irregular commotion based perturbation and Randomized Response scheme. In spite of the fact that this technique causes high information misfortune, it is a more effective strategy. Using generalization and suppression anonymization technique, makes indistinguishable records among gathering of records. K-anonymity is known as delegate anonymization technique. To recognize records exceptionally it considers semi identifiers which can be utilized in conjunction with public records. There are such huge numbers of techniques which has been proposed, for example, (a, k)-anonymity, L-assorted variety, t-closeness, M-invariance, Personalized anonymity, P-touchy K-anonymity et cetera. Here anonymization technique results in loss of information to some degree yet guarantees the originality of data.

Sugumar et al (2012), Rengarajan et al (2012), Vijayanand et al (2012) endeavors to take care of the problem of shrouded linkage that can uncover fundamental information even after k-anonymity is connected which encourages the attackers to make decision and in justifying decisions. In request to safeguard the privacy of the customer in data mining process, an assortment of techniques based on K-Anonymity of data records have been proposed as of late. The main motivation of the proposed Association Rule Hiding (ARH) algorithm is that, it has reduced information misfortune by means of hiding those transactions that supports the particular delicate guideline.

Tamas Gal et al (2008), Zhiyuan Chen et al (2008) and Aryya Gangopadhyay et al (2008) endeavor to illuminate the human services privacy issues using k-anonymity and L-Diversity model for various delicate traits. Past research demonstrates that k-anonymity model can be stretched out to various touchy qualities however L-Diversity not on account of L-Diversity for every individual, delicate property doesn't ensure L-Diversity over every delicate trait. So other methodology is required which utilizes the two techniques and ought to be material to numerous delicate traits too. The motivation behind this proposed work is that it creates less distortion then existing methodologies and fulfills L-Diversity for different delicate properties moreover.

PRIVACY PRESERVATION ON BIG DATA USING MAP REDUCE

Most recent Trends and Challenges in Anonymization Technology by **Satoh & Takahashi et al (2014)** "Balancing utility with anonymity of data". It is conceivable to process personal data into an express that has on the whole eliminated individuality included in data leaving only the important minimum measure of information according to the scientific reason. Factual data is a decent model. For this situation, in any case, since the dimension of abstraction of data is too high, the data will often end up unsuited for analysis. This is a problem between the anonymity and utility of personal data. This can be accomplished by a strategy called PK-Anonymity. This makes data incomprehensible regarding which they belong to through randomization, which is processing to change individual data probabilistically. In randomization, records are prepared to be related to a likelihood of 1/k or less. This nature of Anonymization called PK-anonymity (probabilistic k-Anonymity). From that point forward, execute processing to gauge the original condition of the data by using a machine learning technique called Bayesian inference. Thusly, down to earth anonymous data for analysis will be constructed and may state this is pseudo-personal data based on genuine personal data. PK-anonymization is viable for anonymization of big data while retaining a comparable nature to k-anonymization. There is no such thing as adaptable anonymization strategy. It is on a case-by-case premise according to the sorts, highlights and utilization purposes.

SURVEY ON K-ANONYMITY GENERALIZATION ALGORITHM

Kavitha et al (2014), Sivaraman et al (2014) and Raja Vadhana et al (2014) depicts about the Top Down Specialization. For handling the straight out and continuous traits, the top-down specialization approach is the plausible way. By minimizing the privacy specification and maximizing the data utilization, the Top Down methodology utilizes iterative technique to convert the general information into uncommon information. Different anonymity issues can be dealt with this methodology.

Ke Wang et al (2004), Sourav Chakraborty et al (2004) explained about BUG, another idealistic utilization of the data mining technology regardless of whether we mask private information. The bottom-up generalization converts the particular data to less particular however semantically consistent data for privacy preservation and furthermore they concentrated on two main problems, versatility and quality. A bottom-up methodology is additionally called voracious technique. In each iteration, it blends groups indicated the came about weighted certainty, punishment is privately diminished. For higher anonymization in quality than the Multidimensional methodology, the dynamic methodologies by using both bottom-up methodology and furthermore the top-down methodology. Contrasted with bottom-up strategy regarding utility and sharpness, the top-down technique is better.

Tiancheng Li et al (2007), and Ninghui Li et al (2007) depicted a procedure called bottom-up look for locating best anonymization. Once the value of k is little this system works altogether well. They demonstrated the practicability through analyses on genuine evaluation data for this methodology. To find the ideal solution for little k esteems quickly, the bottom up Approach works productively and when k increases, the running time of a generalization scheme increases.

PRIVACY PRESERVATION ON INCREMENTAL CLOUD DATA

Xuyun Zhang et al (2012) have investigated the test about how to productively update tremendous volume incremental data sets to guarantee privacy prerequisites of data proprietors and all the while accomplish high data utility to data clients. They have displayed a, proficient semi identifier index based methodology for privacy preservation over incremental data sets on cloud. In their methodology, QI-groups (QI: semi identifier) were indexed by the domain esteems in the present generalization level, which made it conceivable to get to only a piece of records in a data set within the sight of data updates rather than access all data records as required by existing methodologies. To further enhance the execution of semi identifier indexing, area delicate hashing strategy was incorporated to put comparative QI-groups on similar data stockpiling hubs. Consequently, the quantity of data hubs that a QI-group link crosswise over was reduced considerably with high likelihood. Based on the built up indexes of an anonymized data set, they have planned an effective semi identifier index based privacy preservation algorithm (QUIPP) for their methodology. This technique is contrasted and the existing Xuyun Zhang et al (2012) technique based on the updating time by adding distinctive number of records. The time taken to update is looked at for changed k esteems.

Dilip Singh et al (2010) have introduced a proficient and handy cryptographic based scheme that saved privacy and mine the cloud data which was conveyed in nature. In request to address the classification task, their methodology utilized k-NN classifier. They stretched out the Jaccard measure to find the comparability between two scrambled and appropriated records by conducting a correspondence test. In addition, our methodology quickens mining by finding closest neighbors at nearby and then at worldwide dimension. The displayed methodology abstains from transmitting

the original data and sharing of the key that was required in traditional crypto based privacy preserving data mining solutions.

Jian et al (2009), Yong Cheng et al (2009), Shuo et al (2009) and JiaJin et al (2009) proposed Anonimity Based strategy . Before the smaller scale data is distributed anonymity algorithm will process the data and send these anonymous data to specialist organizations in the cloud .Then the specialist co-op can integrate the helper information to examine the anonymous data in request to mine the knowledge they need. Semi identifiers are anonymized .It is not quite the same as traditional cryptographic strategy which needs key to get to the data. Be that as it may, using anonymity technique no compelling reason to know the key so it is adaptable and safe to secure individual's privacy in cloud computing administrations.

Feng Li et al (2008) and Shuigeng Zhou et al(2008) proposed another generalization principle M-Distinct to defeat the problem in m-invariance adequately anonymize datasets with internal and also outer updates. M-Distinct utilizations m-one of a kind which is utilized to maintain the touchy qualities which are not diverse in independent publication. In request to maintain this indistinguishability of touchy qualities, records must be partitioned painstakingly while releasing new publication. These techniques can be utilized as a superior way to deal with secure a data in a cloud. K-anonymity can be utilized along with M-Distinct to give dynamic anonymization. Then, utilize a key distribution approach.

T. Zurek, and K. Kreplin, (2001) studied traditionally, it has been well accepted that data warehouse databases are updated periodically – typically in a daily, weekly or even monthly basis – implying that its data is never up-to-date, for OLTP records saved between those updates are not included the data area. This implies that the most recent operational records are not included into the data area, thus getting excluded from the results supplied by OLAP tools. Until recently, using periodically updated data was not a crucial issue. However, with enterprises such as e-business, stock brokering, online telecommunications, and health systems, for instance, relevant information needs to be delivered as fast as possible to knowledge workers or decision systems that rely on it to react in a near real-time manner, according to the new and most recent data captured by an organization's information system. This makes supporting near real-time data warehousing (RTDW) a critical issue for such applications.

C. White, (2002) stated that the demand for fresh data in data warehouses has always been a strong desideratum. Data warehouse refreshment (integration of new data) is traditionally performed in an off-line fashion. This means that while processes for updating the data area are executed, OLAP users and applications cannot access any data. This set of activities usually takes place in a preset loading time window, to avoid overloading the operational OLTP source systems with the extra workload of this workflow. Still, users are pushing for higher levels of freshness, since more and more enterprises operate in a business time schedule of 24x7. Active Data Warehousing refers to a new trend where DWs are updated as frequently as possible, due to the high demands of users for fresh data. The term is also designated as Real-Time Data Warehousing for that reason.

T. B. Pedersen (2004) presented in a report from a knowledge exchange network formed by several major technological partners in Denmark refer that all partners agree real-time enterprise and continuous data availability is considered a short term priority for many business and general data-based enterprises. Nowadays, IT managers are facing crucial challenges deciding whether to build a real-time data warehouse instead of a conventional one and whether their existing data warehouse is going out of style and needs to be converted into a real-time data warehouse to remain competitive. In some specific cases, data update delays larger than a few seconds or minutes may jeopardise the usefulness of the whole system.

Labio et. al. (2000) researched that Operational OLTP systems are designed to meet well specified (short) response time requirements aiming for maximum system availability, which means that a RTDW scenario would have to cope with this in the overhead implied in those OLTP systems; (2) The tables existing in a data warehouse's database directly related with transactional records (commonly named as fact tables) are usually huge in size, and therefore, the addition of new data and consequent procedures such as index updating or referential integrity checks would certainly have impact in OLAP systems' performance and data availability. Our work is focused on the DW perspective, for that reason we present an efficient methodology for continuous data integration, performing the ETL loading process.

B. Devlin (2007) stated that the need for data warehousing originated in the mid-to-late 1980s with the fundamental recognition that information systems must be distinguished into operational and informational systems. Operational systems support the day-today conduct of the business, and are optimized for fast response time of predefined transactions, with a focus on update transactions. Operational data is a current and real-time representation of the business state. In contrast, informational systems are used to manage and control the business. They support the analysis of data for decision making about how the enterprise will operate now and in the future. They are designed mainly for ad hoc, complex and mostly

read-only queries over data obtained from a variety of sources. Informational data is historical, i.e., it represents a stable view of the business over a period of time.

W.H. Inmon (2006) investigated that in an organization; it requires a database system for their daily decision making, with better adaptability, top flexibility, and best support. Considering the past decade, the educational (academia) side and the industry side, both have progressively plated different layouts to solve the problems and to present solution to craft an aforementioned system. Adopting the data warehouse technology is one of the solutions to that. DW was defined by Inmon as, ‘‘pooling data from multiple separate sources to construct a main DW”. Proper data analyzing tools can be used by different users to analyze and store required data.

Deng Z H and Lv S L (2014) stated that Data Mining includes lots of tasks and techniques which are essential for critical decision making out of large amount of information. To get the final result of the process of data mining one should go through a collection of these tasks and techniques. One of these techniques is Frequent Pattern Mining. As its name says this technique finds the most frequent itemsets that are present in a database. The databases on which Frequent Pattern Mining technique can be used are called transactional databases. These databases contain transactions in millions and each of these transactions contains a different combination of items. These items can be anything depends upon the context of the transaction. The main theme of Frequent Pattern Mining is to discover the hidden patterns in the given transactional databases. The results produced are not the ultimate result of the data mining process. These results can be the input of another task to get the desired output. Having a very huge set of applications Frequent Pattern Mining stays at the first in the list of techniques used to find the frequent items.

QiuY, LanY J and XieQ S (2004) proposed an algorithm that uses a tree data structure called POC tree and mine frequent itemsets using a different concept called Nodesets. This algorithm makes a huge margin in efficiency in terms of memory and execution time when compared to its previously proposed algorithms. This algorithm is named as FIN algorithm. FIN algorithm is an extended and improved version of the algorithm which uses the Node-lists and N-lists data structures. On the other hand FIN algorithm uses a novel data structure called Nodeset for mining the frequent itemsets. FIN algorithm directly discovers frequent itemsets in a search tree called set- enumeration tree. In this thesis FIN algorithm is used to discover frequent itemsets but in a different way to make the algorithm even better and efficient.

Lin K W and Lo Y C (2013) from the day when the frequent pattern mining came to the picture there are so many methods that are proposed related to the algorithms to mine frequent patterns, parallel and distributed mining process and the privacy preservation of the data which undergo the process of frequent pattern mining. But there is no specific architecture and framework to achieve the efficient way of mining patterns. The main idea of the thesis is to provide an entire framework which includes an efficient Frequent Pattern Mining algorithm. This framework allows user to execute the algorithm on a dataset in parallel processing paradigm. In this case we are using cloud computing environment which contains number of nodes each of which is responsible for the mining process. This proposed system provides an application as a service. By providing this service in a cloud computing environment the memory and execution time required to complete the entire process is divided among the number of hosts that are present and available in the cloud and the resultant memory consumption and execution time is very much reduced.

R. Agrawal and R. Srikant (2015) studied that in the field of data mining, pattern mining has become an important task for a wide range of real-world applications. Pattern mining consists of discovering interesting, useful, and unexpected patterns in databases. This field of research has emerged in the 1990s with the Apriori algorithm which was proposed by Agrawal and Srikant. It is designed for finding frequent itemsets and then extracting the association rules. Note that frequent itemsets are the groups of items (symbols) frequently appearing together in a database of customer transactions. For example, the pattern/products {bread, wine, cheese} can be used to find the shopping behavior of customers for market basket analysis.

Han et. al. (2004) examined that some pattern mining techniques, such as frequent itemset mining (FIM) and association rule mining (ARM), are aimed at analyzing data, where the sequential ordering of events is not taken into account. However, the sequence-based database which contains the embedded time-stamp information of event is commonly seen in many real-world applications. A sequence in a sequence database is an ordered list of items, and sequence is everywhere in our daily life. Typical examples include consumers’ shopping behavior, Web access logs, DNA sequences in bioinformatics, and so on.

Viger et. al. (2017) in the past two decades, pattern mining (i.e., FIM, ARM and SPM) has been extensively studied and successfully applied in many fields. Meanwhile, to meet the demand of large-scale and high performance computing, as mentioned before, parallel data mining has received considerable attention over the past

decades, including parallel frequent itemset mining (PFIM), parallel association rule mining (PARM), parallel sequential pattern mining (PSPM), parallel clustering and so on. Among them, the sequence-based task - PSPM is crucial in a wide range of real-word applications. For example, in Bioinformatics for DNA sequence analysis, it requires a truly parallel computing on massive large-scale DNA. On one hand, the serial sequential pattern mining is computationally intensive. Although a significant amount of developments have been reported, there is still much room for improvement in its parallel implementation. On the other hand, many applications are time-critical and involve huge volumes of sequential data.

CONCLUSION

The Privacy Preservation Enriched Map Reduce for Hadoop Based Big Data Applications, and the corresponding strategies included the privacy portrayal model, anonymizer for datasets, dataset refresh and privacy preserved information administration. The innovative strategy engaged the information clients with the aptitudes to regain the datasets in its unrevealed versions which encourages the client undertaking dispensing with the requirement for publishing vital detail particulars regarding the original information. Moreover, Rongxing Lu et al. (2009) have explained the privacy preserving in enormous information. They initially figure the general engineering of huge information examination, recognize the corresponding privacy necessities, and introduce a proficient and privacy-preserving cosine similitude computing protocol for instance in response to information mining's effectiveness and privacy prerequisites in the enormous information period. Moreover, Mehdi Sookhak et al. (2010) have explained the securing huge information storage in cloud.

REFERENCES

[1] Sugumar et. al. (2012), Rengarajan et. al. (2012), Vijayanand et. al. (2012). Federated data warehousing application framework and platform-as-a-services to model virtual data marts in the clouds. International Journal of Intelligent Information and Database Systems, 8(3), pp. 280-294.

[2] Tamas S. Gal, Baltimore, Zhiyuan Chen, Aryya Gangopadhyay (2008). “A Privacy Protection Model for Patient Data with Multiple Sensitive Attributes”, International Journal of Information Security and Privacy, 2(3), pp. 28-44, July-September 2008

[3] Takahashi et. al. (2014). Fast mining frequent itemsets using Nodesets Expert Systems with Applications 41, pp. 4505–4512

[4] Kavitha, Sivaraman, Raja Vadhana (2014). “A Survey on k-Anonymity Generalization Algorithms“,International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 11, November 2014.

[5] Ke Wang et. al. (2004). Condition Monitoring and Assessment of Power Transformers Using Computational Intelligence; Springer-Verlag: Berlin, Germany, 2011.

[6] Ninghui Li Tiancheng Li, Suresh Venkatasubramanian (2007). t-Closeness: Privacy Beyond k-Anonymity and l-diversity, ICDE 2007, pp. 106–115

[7] Xuyun Zhang, Chang Liu, Surya Nepal, Jinjun Chen (2012). "An efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud, "Journal of Computer and System Sciences.

[8] Meena Dilip Singh, P. Radha Krishna, Ashutosh Saxena (2010). "A Cryptography Based Privacy Preserving Solution to Mine Cloud Data," Proceedings of the Third Annual ACM Bangalore Conference.

[9] Jian, W., L. Yongcheng, J. Shuo and L. Jiajin (2009). "A survey on anonymity-based privacy preserving," International Conference on E-Business and Information System Security, 23-24 May, Coll. of Inf. Sci. Technol., Donghua Univ., Shanghai, pp. 1-4.

[10] Feng Li and Shuigeng Zhou (2008). “Challenging more update. Towards anonymous republication of fully dynamic datasets”, Cornell University,arXIV.Org<cs>arxiv:0806.4703,July 2008.

[11] T. Zurek, and K. Kreplin (2001). “SAP Business Information Warehouse – From Data Warehousing to an E-Business Platform”, 17th International Conference on Data Engineering (ICDE).

[12] C. White (2002). “Intelligent Business Strategies: Real-Time Data Warehousing Heats Up”, DM Preview, www.dmreview.com/article_sub_cfm?articleId=5570.

[13] T. B. Pedersen (2004). “How is BI Used in Industry?”, Int. Conf. on Data Warehousing and Knowledge Discovery (DAWAK).

[14] W. Labio, J. Yang, Y. Cui, H. Garcia-Molina, and J. Widom (2000). "Performance Issues in Incremental Warehouse Maintenance", International Conference on Very Large Data Bases (VLDB).

[15] B. Devlin (2007). Data Warehouse from Architecture to Implementation. Addison-Wesley.

[16] W.H. Inmon (2006). "DW 2.0 Architecture for the Next Generation of Data Warehousing", DM Review, Apr 2006, Vol. 16 Issue 4, pp. 8-25

[17] Deng Z. H. and Lv S. L. (2014). Fast mining frequent itemsets using Nodesets Expert Systems with Applications 41, pp. 4505–4512

[18] Qiu Y., Lan Y. J. and Xie Q. S. (2004). An improved algorithm of mining from FP-tree Proceedings of the Third International Conference on Machine Learning and Cybernetics 26-29

[19] Lin K. W. and Lo Y. C. (2013). Efficient algorithms for frequent pattern mining in many-task computing environment Knowledge-based Systems 49, pp. 10-21

[20] R. Agrawal and R. Srikant (1995). "Mining sequential patterns," in Data Engineering, 1995. Proceedings of the Eleventh International Conference on. IEEE, 1995, pp. 3–14.

[21] J. Han, J. Pei, Y. Yin, and R. Mao (2004). "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," Data mining and knowledge discovery, vol. 8, no. 1, pp. 53–87.

Corresponding Author

Amit Kumar*

Research Scholar, Shri Venketashwara University, Gajraula, Uttar Pradesh