PRIVACY PRESERVING SECURITY MEASURES IN BIG DATA STORAGE

Megala G[1], Maya Mohan[2], and Sruthy Manmadhan[3]
[1]M.Tech First Year Student, [2,3]Assistant Professor
Department of Computer Science and Engineering,
N.S.S College of Engineering, Palakkad
Email: [1]megalaguruvayurkutty1@gmail.com, [2]mayajeevan@gmail.com, [3]sruthym.88@gmail.com

***Abstract*—Due to the advancement in technology, the amount of data generated by internet, various social networking sites, sensor networks, healthcare applications are rapidly increasing day by day. All these enormous quantities of data produced from various sources in multiple formats with very high speed is referred as Big data. Due to enormous quantity of big data, service providers depends on professionals or speacialized tools to analyze such data. Hence Privacy is one of the major challenges in big data when revealing data for third party to analyze.**

The core applications of big data analysis require sharing of information which leads to challenge privacy. Thus proper preventive measures are necessary inorder to preserve sensitive informations of the individuals before outsourcing or revealing the data. This paper considers privacy as the major issue in big data and provides several measures to ensure the privacy of individuals private information.

***Index Terms*—Differential Privacy, Hiding a needle in a haystack, Data anonymization, Proxy re-encryption**

I. INTRODUCTION

Big data is reffered as the very large data sets that have varied and complex in nature, such that conventional data processing applications are not sufficient because of it variety, volume and velocity. Due to the advancement in technology[5], the amount of data generated by internet, various social networking sites, sensor networks, healthcare applications are rapidly increasing day by day. All these enormous quantities of data produced from various sources in multiple formats with very high speed is referred as Big data. The term big data is defined as new generation of technologies and advancements, designed to seperate values from huge data sets by allowing high-velocity capture, discovery and analysis. These characteristics usually leads to additional difficulties in storing, analyzing and extracting results.

Moreover the growth of IOT also influences the data growth, with increase of these technological changes, which leads to various challenges in big data. One such challenge is the privacy in big data[7]. In short privacy is defined as the access and sharing of information.The core applications of big data analysis require sharing of information which leads to challenge privacy. Privacy is the privilege to have some control on our personal information so that we can decide how it is gatherd and used. It is the one of the important issue, which has legal and technological implications. Most of the Businesses as well as government agencies are generating and enormously collecting large amounts of data[15]. Thus the current focus on enormous quantities of data will certainly create opportunities to understand the processing of huge data sets over numerous varying domains[8]. But, these capabilities of big data come with a price; the users privacy is at danger at any cost. To ensures these privacy terms and rules are incorporated in current big data analytics and mining process.

Though there are various techniques to provide privacy to a certain amount, gradually their demerits led to the advent of newer methods in this field.

II. DIFFERENTIAL PRIVACY

Differential Privacy[1] is a technique that provides professionals and database analysts to obtain the necessary information from the databases that contain sensitive information of people without revealing or exposing the personal identities of the individuals. This is done by producing a little amount of distortion in the information provided by the database system. The distortion which is introduced should be large enough so that they ensures the privacy and also should be small enough so that the information provided to analyst is still useful without data utility. When the Commonwealth of Massachusetts Group Insurance Commission (GIC) revealed their anonymous health record of clients for research inorder to benefit the society. GIC preserved some information like name, street address etc. Inorder to protect their sensitive information. But by making use of the publicly available voter database and database released by GIC, one can successfully identify the health record of users by just comparing and analyzing them. Thus hiding some information cannot only assures the protection of individual personal information.

Thus Differential Privacy(DP) agrees to provide the solution for such problem. To hide an individuals identity,it does the addition of mathematical noise to a sample of the individuals usage pattern. It is mainly done to make their services better, not to collect individual users usage patterns. These mathematical noise are added to the sample using laplace distribution. Apple is starting to use DP technology to discover the usage patterns, starting with ios 10 without compromising individual privacy. As large people share the same pattern, general patterns begins to evolve, which can leads to enhance the user experience in future use. In the DP data base analyst are not allowed to have direct access to the database that contains personal information. An intermediate software is introduced between the database and

the database analyst to protect the privacy. This is termed as the privacy guard.

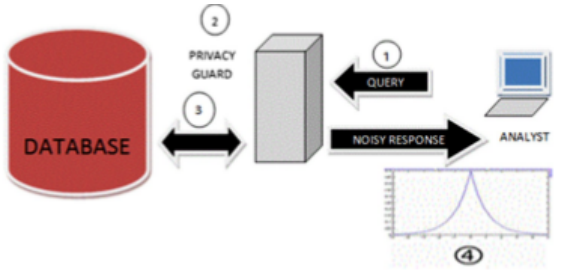

Figure 1. Differential Privacy[1].

Fig 1 shows the various steps to be done in sequence which are defined below.

1) The database analyst will make a query to the database through the privacy guard.
2) The privacy guard gets the query from the analyst and evaluates this query and other previous queries for privacy risk, after evaluating the privacy risk.
3) The privacy guard then gets the data from the database.
4) Add some distortion or noise to it according to the evaluated privacy risk and finally give it back to the analyst.

The amount of noise added in the data is directly proportional to the evaluated privacy risk. If the privacy risk is low, noise added is small enough so that it do not affect the data utility, but should be also large enough that they preserves the individual privacy contained in the database. But if the privacy risk is high then more amount of noise is added which will affect the data utility.

III. HIDING A NEEDLE IN A HAYSTACK

All the previously existing privacy-preserving association rule algorithms change the original transaction data by injecting distortion to the data. These techniques cause the limitation in noise addition because of the need to consider privacy-data utility trade-off. However, this work maintained the original transaction in the noised transaction with the goal as to prevent data utility deterioration while preventing the privacy violations. Hiding a needle in a haystack[1] is based on the idea that finding a small class of data, that is the needles, is very hard to find in a haystack, such as a large size of data. There are several assumptions we adopt at first, we assume that service providers want to use the external cloud service which is basically based on Hadoop framework to achieve association rule mining for big data. Second, is that the data providers want to prevent privacy violations when the external cloud services process their data. Also that external cloud services as the third party which cannot be trusted[11].

Next assumption is that service providers want to consider the data utility for privacy preserving. However, all previous methods to preserve the privacy violation cannot overcome this trade-off. If user want much more stronger privacy protection degree, data utility is degraded, and if user requires the more accurate analysis, it causes the chances of privacy violations.

Figure 2. Process of association rule mining in Hiding a needle in a haystack[1].

Fig 2 shows the process of association rule mining in which service provider does the addition of dummy item as noise to the original data gathered by the data provider. At the same time, a unique code is assigned to the dummy and the original items. The service provider maintains this information to filter the dummy item after the extracting the frequent item by the external cloud platform. There are two types of code information bit string code and prime key code. In bit string code scheme, code consists of bit string such as 00001, 00011 assign to item. Prime number-based code usually assigns unique prime number to the item and then modify the code of the frequent item set by multiplying each item prime number code. Both of bit string and prime number-based code are used as a key for noise filtering.

Apriori algorithm is done at the cloud platform using data which is given by the service provider. The external cloud platform returns back the frequent item set can also support value to the service provider. The service provider filter out the frequent item set that is affected by the dummy item using code information to extract the correct association rule. The process of extraction association rule is not a complex task for the service provider, while consider the amount of calculation required is not much.

IV. A NOVEL GROUP KEY TRANSFER PROTOCOL FOR BIG DATA SECURITY

Group communication is a particular type of many to many communications which goes beyond other communications. Thus confidentiality is a prime concern which is done by encryption. All members in a group share a session key, However since the group membership changes unexpected, group key need to be modified dynamically to confirm both forward and backward secrecy of group sessions. Forward secrecy is ensured if a member who has left from the group and shall not access the content of current and future group sessions. Backward secrecy is ensured for a new member should not access the content of communications of the past sessions. To achieve these objectives, always needs a one time group key such that the key is known to the present group members.

Most of the already existing group key transfer protocols depends on a online mutually trusted key generation centre(KGC), to transfer the group key secretly to all group members. This method needs a trusted server to be set up, and it incurs

communication overhead. A novel group key transfer protocol[2] without an online KGC, which is based on the combination of Diffie Hellman(DH) key agreement and a perfect linear secret sharing scheme(LSS)[14]. In a secret sharing scheme, a secret s is partitioned into n shareholders and shared among them by a mutually trusted dealer such that authorized shareholders can only reconstruct the secret and unauthorized parties cannot get the secret. If unauthorized subset of shareholders does not obtain any information about the secret, then is called perfect. It is said to be linear, if the reconstruction operations are linear. A hybrid of public key approach and a secret sharing scheme is used in this proposed protocol. There are mainly two phases.

A. Secret establishment phase

1) The initiator broadcasts a request containing a random number $r_{np}$and his/her public key puk_n and a list of members to announce the group communication.
2) The initiator broadcasts a request containing a random number $r_{np}$and his/her public key puk_n and a list of members to announce the group communication.
3) After receiving the message from each, the initiator computes $S_i = puk_i^{prk_n r_i r_n} mod p$ and if result is valid, then initiator believes secret is shared with corresponding group member. otherwise the initiator claims that i is fraudlent and then restarts the protocol.

B. Session key Transfer phase

1) The initiator randomly generates a session key K_{Gp}. Then, initiator computes $n-1$ additional values $U_i = (K_G - K_i) mod p$ for each group member where $K_i = (S_i.r_i$ and broadcasts U_i to all group members.
2) For each group member except the initiator, knowing the public value U_i is able to compute the$K_i$and recover the group key $K_G = (U_i + K_i) mod p$. Thus session key is established among all group members.

Adding or removing any user does not require the modification of any existing shared secret. However, for publishing a secret group key, initiator requires to broadcast a message at the time itself each group member need to recover the group key as a function of their own random number and secret shared between them and initiator.Thus the Proposed protocol does not rely on a KGC and it can reduce the overhead of system implementation.

V. PRIVACY AND UTILITY PRESERVING DATA CLUSTERING FOR DATA ANONYMIZATION AND DISTRIBUTION ON HADOOP

Data anonymization is one of the advancing technique in the field of privacy-preserving data mining used to protect users privacy. Linking attack or identity disclosure is the possibility that an attacker reveals the personal attribute of a person with the known information. K-anonymization is the process of anonymizing in a way such that the K records are indistinguishable from each other. This mechanism protects the record from identity disclosure or linking attack but are comprimised to probabilistic and various other attack which are listed below.

1) Homogeneity Attack: When all the records in an anonymized group contain identical sensitive attribute, the process of k-anonymization would become waste of time. This form of attack is called homogeneity attack.
2) Similarity attack: The values in an anonymized group are not identical to each other,but the values may be similar to each other.
3) Background Knowledge Attack: An intruder already knowing some background information the data uses it to inorder to narrow down the possible sensitive values in an equivalence class group.
4) Probabilistic Inference Attack: When any one of the sensitive attribute value is occurring more rapidly than the others in an group, there is a possibility for probabilistic inference attack.

These attacks are overcome by the proposed method to distribute the records with sensitive value equally to all the equivalence classes based on K-Nearest neighbour method. The(G,S) clustering algorithm determines the single best neighbour of each cluster and assigns instances one at a time to the existing cluster. We have changed the algorithm to overcome the skewness in the sensitive value distribution of the resultant clusters by using K-Nearest Neighbour technique. The KNN-(G,S) clustering algorithm[3] finds the KN nearest neighbours from each sensitive value group and then adds the KN records to the clusters at a time. This overcomes the case of probabilistic inference attack. The anonymized data set is then distributed on a Hadoop[10].

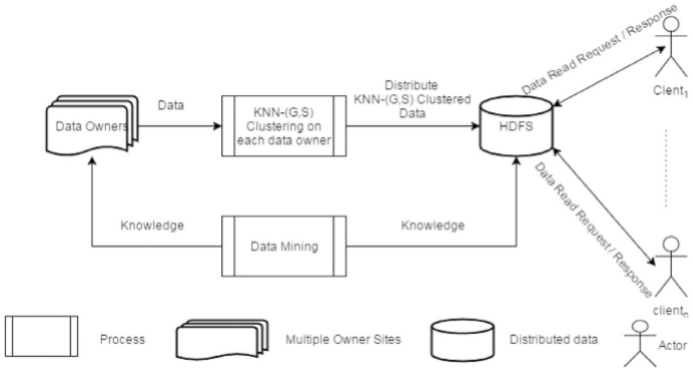

Figure 3. Privacy-preserved data distribution and datamining on Hadoop[3].

Fig 3 describes the privacy preserved data distribution firstly, the data owners apply the KNN-(G,S) clustering algorithm on the input data seperately and distribute the data on Hadoop. Then any kind of data mining algorithm can be applied on these distributed data using Mapreduce task in parallel patterns obtained are shared to the data owners. Any attacker who can view the privacy-preserved data on Hadoop cannot map any data to a particular individual. Thus proposed clustering algorithm can a mechanism to store the data in a privacy-preserving manner on Hadoop cluster without causing complexity and communication overhead. The client requesting for a data, request a message in O(1) and the response would be in O(n) where n is the dataset size.

VI. Proxy Re-Encryption: Analysis of Constructions and its Application to Secure Access Delegation

The growth of the cloud computing has brought great expectations regarding performance, simplifying the business processes, and reduction in cost. At the same time, these advancements come with new security and privacy risks[9]. Threat scenarios occur when moving from resources fully controlled by the data owner to resources processed by third party entities like public clouds. Cloud providers are assumed to be semi-trusted, because their functions are presumed correct with respect to protocol service, but they may have some incentive to read users data without their consent. This kind of approach in cloud is usually called honest-but-curious.

1) Secure Access Delegation Scenario: The need arise when traditional security assumptions get weakened which governs the current security architectures of cloud systems leads to the encryption of data prior to outsourcing as an essential requirement. At that time, it is also important to delegate access for sharing purposes, which is one of the most necessary functionalities. We refer this scenario as the secure access delegation scenario. There are three main separate roles: data producers, data owner, and data consumers. The most generic usage relation scenario is that multiple data producers can generate data which is owned by a data owner, who can share it with multiple data consumers.

A best solution for the secure access delegation problem to use conventional encryption technique(AES, RSA) and to share the decryption key. Symmetric encryption cannot be used alone,because same key is shared between producers, owner and consumers or, producers and owner agree on a key, which is extremely inefficient. While with public-key cryptography, the problem is that the producers do not necessarily need to known in advance who are the intended consumers. Thus it require, the only possibility is that they encrypt the data under some common public key which is controlled by the data owner(public key) this require to data owner has to decrypt the data and subsequently encrypt it with a key known by the intended consumers; this decrypt-and-encrypt solution requires the data owner to be online to re-encrypt the data when needed, which is extremely difficult. The basic concept of a proxy re-encryption[4] scheme is consists the ability of a proxy to transform ciphertexts which are encrypted by the public key of Alice into ciphertexts which can be decryptable by Bob; to do so, the proxy must be in stage to re-encrypt using re-encryption key[13].In addition, the proxy should not learn any information about the encrypted messages. Fig 4 shows the actors who participate in proxy re-encryption.

1) Delegator: This actor is responsible to delegates his decryption rights using proxy re-encryption. In order to do this he creates a re-encryption key, which is sends to the proxy. We refer Alice as the delegator.
2) Delegatee: The delegatee is data consumer who is granted a delegated right to decrypt ciphertexts that, although not provided for him in the first place, but re-encrypted for him with permission from the original recipient(i.e., the delegator). This actor usually refer as Bob.
3) Proxy: It handles the re-encryption process that transforms ciphertexts which are encrypted by the public key of Alice into ciphertexts which can be decryptable by Bob without learning anything.

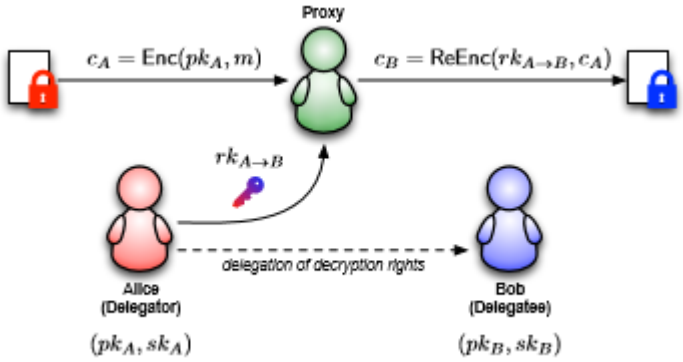

Figure 4. Main actors and interactions in an proxy re-encryption[4].

TABLE I
Comparison Table

TECHNIQUE	ADVANTAGES	DISADVANTAGE
Differential Privacy	Privacy to identity	Affect data utility
Hiding Needle in Haystack	High level of Privacy	High computation cost
Group key Transfer Protocol	Key freshness	Theoretically secure
Data anonymization	Withstand Various attacks	Running time is high
Proxy re-encryption	Against ciphertext attack	Searching is complex

In a PRE solution[6], private data is first encrypted by a data producer(which can be any entity that has the proper public key) and outsourced to a semi-trusted proxy(i.e., storage provider in the cloud). By generating the corresponding re-encryption keys and send it to the proxy, the data owner is inturn authorizing data consumers to access his data. The proxy ensures the secure access delegations through the re-encryption process using the re-encryption keys, while the information that is protected with respect to unauthorized parties and the proxy itself[12].

VII. Discussions and Findings

Table I shows the comparison of various techniques which has been discussed in this paper. Each method has it's own situation where it can be used. Differential Privacy obscure an individual's identity, by adding mathematical noise to a sample of the individual's usage pattern. This is done by adding distortion to the information. If the privacy risk is low, noise added is small so that it do not affect data utility, but should be large enough that they protect the individual privacy of database. But if the privacy risk is high then more noise is added. So that it affect the Quality of answer. Thus it bypass the data utility of answer when privacy risk is high. Hence affect the analysis done by the data base analyst.

Hiding a needle in a Haystack consider both the privacy and data utility trade-off. The service provider maintains the code information to filter out the dummy item after the extraction of frequent item set by an external cloud platform. Thus the possiblity that untrusted cloud service provider infer the real frequent is considered and thus provide privacy without data utility deteoriation. But incurs additional computation cost in the

addition of noise. These both techniques are usually preffered by the service providers. Group key transfer protocol for group oriented applications of big data provide a group key which should be used by current group members without an online KGC. Thus additional server implementation cost overhead can be reduced. This also avoids additional communication cost required in an online KGC. It provides key freshness, key confientiality and key authentication. Thus backward and forward secrecy is ensured in group oriented applications. But the group key transfer phase is only theoretically secure due to network vulnerabilities. Privacy and utility preserving data anonymization and distribution in Hadoop is an anonymization technique which uses KNN algorithm to determine KN nearest neighbour and distributes the sensitive values among all the cluster. This algorithm overcomes similarity, background, Homogenity and Probabilistic inference attack.Also algorithm is precise in terms of cluster size. But the otherside is that the running time of the algorithm in determing K nearest neighbour is $O(n^2)$.

Proxy re-encryption is the better solution to implement privacy in big data analytics. proxy re-encryption scheme is embodied by the ability of a proxy to transform ciphertexts under the public key of Alice into ciphertexts decryptable by Bob; to do so, the proxy must be in possession of a re-encryption key that enables this process. In addition, the proxy cannot learn any information about the encrypted messages, under any of the keys. It considers semi trusted cloud platform and provide a re encryption process. Secure access delegation is taken into consider so that even data owner can be on offline. Thus it achieves better privacy measures in cloud when compare to others as well as taken into consider the complexity. But searching or computing on an encrypted data is slight complex.Thus the big data privacy that is access and sharing of information is considered in a broader aspect.

VIII. Conclusion

Thus various privacy preserving measures in big data storage is discussed. Differential privacy and Hiding a needle in a Haystack are some techniques to implement privacy. Proxy re-encryption is a special type of public key encryption in untrusted cloud providers. A novel group key transfer protocol implements privacy in a group oriented applications of big data. The data anonymization and distribution in Hadoop provides anonymized data sets and distributes on a Hadoop distributed File system and thus third party infereing sensitive values are prevented. Thus the big data privacy that is access and sharing of information is considered in a broader aspect including group communication, cloud environment and Hadoop.

References

[1] Priyank jain, Manasi gyanchandani, Nilaykhare, 2016, "Big data privacy:- a technological perspective and review", *journal of big data*, volume 3, pp 1-25.

[2] Chingfang hsu, Bing zeng , Maoyuan zhang, 2014, "A novel group key transfer for big data security", *Applied mathemathics and computation*, volume 249, pp 436-443.

[3] J.Jesu Vedha Nayahi, V. kavitha, 2016, "Privacy and utility preserving data clustering for data anonymization and distribution on hadoop", *Future generation computer systems*, volume 74, pp 393-408.

[4] David Nunez, Isaac Agudo, Javier Lopez, 2017, "Re-encryption: Analysis of constructions and its applications to secure access Delegation", *Journal of network computer applications*, volume 87, pp 193-299.

[5] Samiya khan, Xiufeng liu, Kashish A shakil, Mansaf Alam, 2017, "A survey on scholary data: from big data perspective", *Information processing and management*, volume 53, pp 923-944.

[6] Kun Wang member ieee, Jiahw yu member ieee, song guo ,2016, "A pre authentication approach to proxy re-encryption in big data context", *IEEE transactions on big data*.

[7] Almeida Fernando, Calistru, Catalin, 2013, "The main challenges and issues of big data management", *International Journal of Research Studies in Computing*, volume 2, pp 11-20.

[8] Min chen, Shiwen Mao, Yunhao Liu, 2014, "Big data: A Survey", *Mobile network applications*, volume 19, pp-171-209.

[9] S. Ananthi , Anjali Periwal , Prince Mary. S, 2016, "Data Security Based On Big Data Storage", *Global Journal of Pure and Applied Mathematics*, Volume 12, pp 1491-1500.

[10] Amrit Pal, Kunal Jain, Pinki Agrawal, Sanjay Agrawal, 2014, "A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data Using Hadoop", *Fourth International Conference on Communication Systems and Network Technologies*.

[11] Giannis Tziakouris, Marios Zinonos, Tom Chothia, Rami Bahsoon, 2016, "Asset-Centric Security-Aware Service Selection", *IEEE International Congress on Big Data*.

[12] Zheng Yan, Senior Member, IEEE, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, Fellow, 2016, "Deduplication on Encrypted Big Data in Cloud", *IEEE transactions on big data*.

[13] Hanshu Hong, Zhixin Sun, 2017, "Towards Secure Data Sharing in Cloud Computing Using Attribute Based Proxy Re-Encryption with Keyword Search", *second IEEE International Conference on Cloud Computing and Big Data Analysis*.

[14] Changxiao Zhao, Jianhua, 2015, "Novel Group Key Transfer Protocol for Big Data Security", *IEEE Conference on big data*.

[15] Feng Xia, Wei Wang, Teshome Megersa Bekele, Huan Liu, 2017, "Big Scholary Data: A Survey", *IEEE Transactions on big data*, volume 3, pp 18-35.