A PRIVACY PRESERVATION FRAMEWORK IN CROSS-CLOUD SERVICES FOR BIG DATA APPLICATIONS

S.Arun Kumar[1], Dr. M. S. Anbarasi[2]
[1]Research scholar, [2]Assistant Professor
PEC, Pondicherry University, INDIA.

Abstract
Cloud computing provides remote access for data storage and applications to users and organizations. Many Enterprises are utilizing the services of cloud computing to access the data easily without maintaining the data by its own. This makes the enterprise to assure that the system is highly scalable and is available with reduced cost of setup and maintenance. Cross-cloud service is best and suitable approach for large-scale big data processing system as big data processing system required huge data storage and computation power. Complex web based application of big data processing generates huge amount of data sets which are stored in remote location in cloud. While analysing these intermediate data sets, the sensitive information can be accessed by misfeasors. Maintaining the confidentiality of this generated data set is very challengeable. Most of the existing systems uses cryptography methods and stores the encrypted data sets in cloud. The system consumes more time and cost because of the most often process of encryption and decryption of intermediate data sets which results in inefficiency and are expensive. In this paper, we propose various methods to protect privacy of data during big data processing.
Keywords: Cloud, Big Data, Map Reduce, privacy

I. INTRODUCTION

Both Cloud computing and Big Data are intriguing areas of research and IT industries as growth of the data is very big at present compared to past. Cloud computing services aid enterprises to lighten the overhead, mitigate access and trim down the maintenance cost to store and access the data.

Big Data concerns large-volume, complex, growing data sets with multiple and autonomous sources. Input of Big data is collected from online transactions like bank transactions and online shopping, queries requested in search engines, logs of telephone and mobile calls used in particular area, electronic mails and messages, videos, sensor logs, social media etc. It is stored by distributing along various servers. Big data are now rapidly emerging in all science and engineering domains.

The MapReduce framework has been widely used by a large number of companies and organizations to process huge-volume data. Unlike the traditional one, MapReduce incorporated with cloud computing becomes more distributed, parallel processing, reasonably priced, faster, scalable, Support both structural and non-structural languages. A typical example is the Amazon Elastic MapReduce (Amazon EMR) service. This tool processes and analyses the data from different sources that are widely distributed in cloud. Since it is extensively based on the web model, security issues are the key challenge in analytics of big data. Fine analytical output can be obtained by analysing the entire big data, however this leads to the security concern as intruders and attackers steals hidden value from big data i.e, not ensuring to secure big data analytics which may cause great loses to both people and organizations.

So, there is a trade-off between big data availability and big data security and we need to ensure proper balance between these two parameters. Hence, the need for protecting the knowledge in big data during the whole analytics

process. Privacy concerns in Map Reduce platforms are exasperated because the privacy-sensitive information is dispersed among various data sets that can be recovered.

Traditional security mechanisms are unable to handle big data due its large volume, variety and velocity. Among various security aspects of processing big data, privacy is one of the most important concerned issues [1]. This is because big data analysis generally contains analysis on person specific information. Each individual's activities on social media, search engines etc. are recorded and analysed. These data are then pushed back to web and shared with concerned parties on the cloud. Hence, unsecure big data analytics will lead to exposure of Personal Identifiable Information (PII) and results in losing customers as the loyalty of the organizations goes down.

Existing methods like cryptography can be used to protect privacy but they are not sufficient for big data processing because of complex nature of data [2].

Data anonymization or de-identification will be used in the process of hiding personal information. The actual data is altered to secure the key data involved in processing [3]. There are three data anonymization methods that can be used in preserving big data privacy that are: K-Anonymity [3, 4], L-Diversity [3], and T-Closeness [5].

Another method used for enabling consumer privacy is through Notice and Consent [6]. By this method consumer information is shared only after obtaining permission from the user using a notice. This method is usually used when user installs a new application or a new web service installation.

Differential privacy is a latest method of privacy preservation with big data which is extensively adopted. Data analysts obtain the data from statistical database which also includes personal information. This method offers strong individual privacy protections, while analysts get the necessary information from databases.[7].

In this paper we will explain different methods and techniques that are used to resolve big data privacy protection issues. This paper is documented as follows: Section 2 gives a review of existing methods like cryptography and encryption which provides data privacy. Section 3 discusses some privacy preservation methods for big data. Section 4 finally concludes our work.

II. RELATED WORKS

There are lots of traditional data privacy preserving methods are used which are cryptography and attribute based encryption.

Cryptography refers to set of methods and algorithms for protecting data privacy through encryption and decryption The actual data called plain text is converted into cipher text by encryption algorithm. Decryption algorithm does the process in reverse way which converts cipher text into plain text. Key is also another input for both encryption and decryption process. Public key cryptography, digital signatures are examples of cryptographic methods.

Fig. 1. Simple Encryption Scheme

Cryptography can't enforce the privacy which are required by common cloud computing services and big data services [2]. This is because of the big data difference from the traditional large data sets on the basis of 3 V's (velocity, variety, volume) [8, 9].
The features of big data that make processing big data architecture different from traditional architectures. These changes in processing architecture and complex nature of data makes Cryptography method is inappropriate to provide better privacy of big data as it is not scalable for privacy.

The challenge with cryptography is that all or nothing retrieval schemes over encrypted data [10]. The less sensitive data attribute that can be useful in big data analytics must also be

encrypted and the user is not allowed to access it. The data can be either sensitive or non-sensitive, the decryption key is must be used to get original data. Without the decryption key the data cannot be accessed. Another problem is that the sensitive data can be leaked out if cryptographic key is misused by someone. Attribute based encryption can be used to protect only the sensitive information [11, 12]. This method considers inter relationships among the attributes present in data, big data type and organization policies. Attributes are selected from the large number of attributes which requires protection. Only the selected attributes are protected and remaining attributes can be accessed even without cryptography process.

In outer layer, preserving privacy of big data can't be done only using these two techniques cryptography and attribute based encryption. These techniques can be useful for the anonymization of data however can't be directly used for protection of the privacy of big data.

III. PRIVACY PRESERVATION METHODS

Privacy of big data is most important factor for the enterprises so we should have more efficient methods to protect the data. We have many existing methods for privacy preservation but each method has its own limitation and drawbacks. To understand the existing methods, three privacy preservation methods called data anonymization, notice & consent and differential privacy are discussed in this section. We will look into these methods in detail and also drawback of each.

Table 1: Base Dataset

Age	Sex	City	Income
24	M	Delhi	1,00,000
24	M	Gurgaon	18,000
24	M	Gurgaon	25,500
24	M	Delhi	12,000
26	F	Delhi	20,000
26	F	Delhi	50,000
26	M	Delhi	29,000
26	F	Delhi	48,000
32	M	Delhi	26,000
32	F	Gurgaon	45,000
32	F	Gurgaon	34,000
32	M	Delhi	34,000

Data Anonymization:

Data anonymization is the process of altering the original information to make sure the sensitive information cannot be identified [3]. This method is also called as de-identification. The key attributes are identified and hided to maintain the privacy of sensitive information [13]. The data can be shown to public after this anonymization process. Attributes like SSN, passport number, voter id are used to identify the individuals and these attributes will be hided before the data is publically released. The main limitation with anonymization of data is that the identifying of sensitive information can be easily done by linking/combining of multiple data sets or other external data [4]. In [4], it is presented that the anonymized medical records are re-identified by using external voter list data. The quasi attributes such as date of birth, gender, zip code that can be linked with external data for re-identifying of the users.

For example, Table 1 shows the set of data which is to be analysed in tracking down the trends in income of few individuals without revealing the identity information of an individual. Table 2 shows set of data which becomes unknown by eliminating identifier attribute Voter ID. This table becomes anonymous after the elimination but still it can be re-identify these individuals by linking with external data.

Table 2: Anonymous Dataset

Voter ID	Age	Sex	City	Income
	24	M	Delhi	1,00,000
	24	M	Gurgaon	18,000
	24	M	Gurgaon	25,500
	24	M	Delhi	12,000
	26	F	Delhi	20,000
	26	F	Delhi	50,000
	26	M	Delhi	29,000
	26	F	Delhi	48,000
	32	M	Delhi	26,000
	32	F	Gurgaon	45,000
	32	F	Gurgaon	34,000
	32	M	Delhi	34,000

In this paper, mainly 3 privacy-preserving methods based data anonymization are discussed: K-Anonymity [3, 4], L-Diversity [3], and T-Closeness [5].

K-Anonymity:

A raw dataset is known as k-anonymized dataset when any tuple with given attributes in the dataset, there must be at least k-1 other records that should match those given attributes [3, 4]. The K-anonymity model was designed to deal with the possibility of indirect identification of personnel information form public statistical databases, k-anonymity means that each released data record has not less than (k-1) other records in that released record whose values are indistinct.

There are some transformation techniques that is used in K-Anonymity which are generalization, global recording and suppression. [14]. In Generalization method, the quasi identifiers are appended by more general values from few stages up in the hierarchy. In Global Recording method, the quasi identifiers value must be mapped into same generalized values in all available records. It is used in privacy preserving transformation in microdata which are also called as recording. In Suppression method, the quasi identifiers are replaced by some arbitrary constants such as 0, * etc.

One of the major advantage of the data anonymization based information sharing approach is that, once anonymized, data can be freely transferred across multiple parties without having restrictive access controls. This problem leads to another research area called privacy preserving data mining
where multiple parties, each holds some sensitive data and tries to achieve a common aim.

Suppose, in Table 1, Attributes in the input dataset.

- Voter id, Name → Personnel Identifiers attributes.
- Age, DOB, City → Quasi Identifiers attributes.
- Income → Sensitive attribute.

Table 3: 2-Anonymized Dataset (Using Suppression)

Age	Sex	City	Income
2*	M	Delhi	1,00,000
2*	M	Gurgaon	18,000
2*	M	Gurgaon	25,500
2*	M	Delhi	12,000
2*	F	Delhi	20,000
2*	F	Delhi	50,000
2*	M	Delhi	29,000
2*	F	Delhi	48,000
3*	M	Delhi	26,000
3*	F	Gurgaon	45,000
3*	F	Gurgaon	34,000
3*	M	Delhi	34,000

Table 3 shows 2-anonymized version of table 1 using suppression. Here, age attribute has been suppressed and k is equal to 2.

In making the data K-Anonymous the data becomes inefficient as it can still be targeted by the attacks, such as, complementary release attack and temporal attack, unsorted matching attack [4]. Then, L-diversity technique came into advancement for anonymization of the data

L-Diversity

The method of L-diversity is anonymization of data in which diversity is formed inside the sensitive attribute of raw data. Quasi identifiers attribute should have at least L different values of sensitive attribute inside each equivalence class [3].

An equivalence class will be L-diverse and anonymized when there is/are at least 1 “well-represented” values in the tactful attribute. A table is called L-diversity when every equivalence class is L-diversed.

Here in Table 1, the tactful attribute is income field. For data to be L-diversed, income attribute should have L different values that must be linked with each equivalence class. And Table 4 represents 3-diverse version of table 1, as there are 3 alternate vales in each equivalence class for the income that is sensitive attribute.

This method comes with a limitation that it dependent on some range of sensitive attribute. For data to be L-diverse, there must be L different values however there are less than L

different values in tactful attributes. Hence, there is a need to insert false data. The security is provided with this false data but may cause issues in analysing the data.

L-diversity is inadequate to stop disclosure of the attribute. L-diversity method is susceptible to two types of attacks, similarity attack and skewness attack. Hence, it can't prevent the disclosure of sensitive attribute [13, 5].

Table 4: 2-Anonymized Dataset (Using Generalization), 3-Diverse Dataset

Age	Sex	City	Income
24	Person	ncr	1,00,000
24	Person	ncr	18,000
24	Person	ncr	25,500
24	Person	ncr	12,000
26	Person	ncr	20,000
26	Person	ncr	50,000
26	Person	ncr	29,000
26	Person	ncr	48,000
32	Person	ncr	26,000
32	Person	ncr	45,000
32	Person	ncr	34,000
32	Person	ncr	34,000

T – Closeness:

Distribution of sensitive personnel identifying attributes within each quasi identifier group must be adjacent to their distribution in their actual entire statistical database.

An equivalence class is known as t-closeness when the distance between the distribution of a sensitive personnel identifying attribute in this class and the distribution of the sensitive attribute in the entire table is not more than threshold t. A table is said to have t-closeness when all equivalence classes present in the table have t-closeness [5]. The main advantage of t-closeness is that it prevents the disclosure of sensitive attributes.

Anonymization of data can be applied to big data but the issue relies in the fact that as size and variety of the data increases, then the probability of re-identification also increases. Thus, anonymization has a limited capability in the field of big data privacy.

Notice and Consent:

Notice and consent is very popular privacy preservation method and widely used in web services [6]. Notice is an alert displayed to user when the user installs or accesses a new application. This notice is a notification to user to get the permission and based on that the application is allowed to access the private details from user device. The user can either accept (consent) or deny the notice. If the user gave the consent, then the application is allowed to access the data. Application is not allowed to access the data if the user denied the notice. The user may not get some service if the consent is not given by user, as the application may require the data compulsory to provide the service. This approach lets the user to have a privacy rights for their own data [6].

This method can also be implemented in big data processing by providing full control to user for privacy. But the big data processing is not like web services, as the big data processing is done more rapidly in huge size. Whenever the data processed the huge number of notification need to send and get consent from users. It leads to delay in processing data. And also burdens the user by sending more number of notices if a particular user details involved in different data processing.

Differential Privacy:

Differential privacy method provides strong protection to sensitive data or personal data in big data processing [7,15]. During data processing analysts can get the required information from statistical database provided with less possibility of identifying sensitive data. The method of differential privacy is illustrated in fig. 2.

This method is different from anonymization, as the data is modified in anonymization method, but in differential privacy the sensitive data is not modified. In differential privacy method users are allowed to access the statistical database directly. There is an interface between the analysts and database like firewall. This query can be passed only through this interface and that will manipulate the output by including inaccuracies. The final result will be passed to data analyst. The sensitive data involved in this process are not

visible, at the same time the result given by the interface is useful for the analysts.

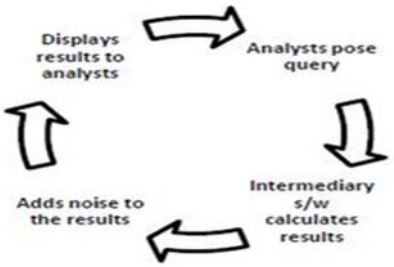

Fig. 2. Differential privacy process

A computation on a set of inputs is said to be differentially private if, for any possible input items, the probability that the computation produces a given output which does not depend much on whether this item is included in the input dataset or not.

The advantages of differential privacy compared with anonymization are as follows:

- No modification is required over the actual raw data. Generalization and suppression techniques are also not required.
- By mathematical formulae calculations based on the nature of data, type of questions etc., distortions can be added to the results of those calculation.
- These distortions are advantageous to analysts because of the hidden sensitive values.

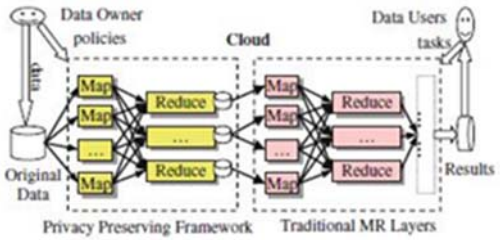

Fig. 3: - Privacy Preservation framework based on MapReduce

The responsibility of the framework is to anonymise original raw data sets according to the given privacy requirements and accordingly anonymise the original data sets. To avoid re-computation on some data, some anonymous data sets will be retained. The management of the intermediate datasets and updating the data when new data join is also done by this framework. Data users can then specify their business application logic in MapReduce jobs and these jobs are executed on the data sets which were anonymized.

We have considerations on numerous operational system requirements that should be fulfilled while designing the framework of privacy preserving. It is as follows.

•***Flexible:*** A User-Interface is provided by this framework in which many privacy preservation techniques are specified by the data owners.

•***Scalable:*** ***C***urrent privacy-preserving approaches are scalable, as the size of data sets are very big to be treated by centralized algorithms. So for managing the anonymization of data the privacy-preserving framework must be scalable. Now then, the data-intensive or computation-intensive operations must be implemented in parallel processes and efficiently in the processing framework.

•***Dynamical:*** In cloud computing, new data sets are collected by most of the applications, e.g., cloud health services usually receive a big size of information from users in some time. With this big size of information, the data sets are created which are huge and bigger, forms of Big Data. Therefore, the framework of privacy preservation should manage dynamical growth of data sets. The privacy of these data sets are invulnerable yet but not sure after each time it is updated.

•***Cost-effective:*** In cloud computing, there is a feature of paying for each usage, saving the cost of IT is one of the task which attract users in this framework. Therefore, it is important for the framework of privacy preservation in saving the costs of privacy preservation to a feasible extent.

Fig. 4; System structure for framework of the privacy-preserving.

A system structure is designed for the framework of the privacy-preserving with four system requirements mentioned above. Similarly, this framework includes four modules which are,

- Privacy Specification Interface (PSI),
- Data Anonymization (DA),
- Data Update (DU).
- Anonymous Data Sets Management (ADM).

According to these modules, the framework of privacy preservation can obtain the four requirements of the system.

Fig.4 represents that DA, DU and ADM are the three important operational modules. Major operations on data sets are done by these modules based on the privacy models that are stated in the PSI module. The DA and DU modules take benefit from the MapReduce framework for anonymizing the data sets or adaptation anonymized data sets when it is updated. The ADM module is used for handling the anonymized data sets so that the cost can be saved from avoiding re-computation. The four modules mentioned above are explained below: -

Privacy Specification Interface

The privacy requirements needed by a data owner are defined as a Privacy Specification (PS). A privacy specification is normally represented by a vector of parameters, i.e.

PS = <PMN, Thr, AT, Alg, Gra, Uti>.

PMN is the privacy model name. From our recent planned privacy models, we use three commonly implemented privacy models in the framework of privacy preservation, such as k-anonymity, l-diversity and t-closeness.

Parameter Thr is the threshold value in the stated privacy model that is, k, l and t which are the threshold in the overhead three privacy principles.

Parameter AT defines application type. Owners of the data can state generalized goals for anonymization of data sets. For e.g., clustering, general use or classification.
These algorithms which implements anonymization are specified by the parameter Algorithm. Many types of algorithms have been implemented for different principles of privacy and types of application.

Parameter Gra defines the granularity of the PS (Privacy Specification). It represents the actual range of the privacy preservation.

The data utility parameter Uti is an optional parameter. Owners of the data can state the extent of data utility they need to allow for exposing to the data users.

Data Anonymization

Anonymization of data can be performed using many transformation techniques which are generalization, suppression and anatomization. Generalization is attained if a parent domain value is swapped with its child domain values within its domain taxonomy tree for preserving big data privacy. In Suppression, the original values are masked into data records with one pre-given symbol, thus all the sensitive information of the attribute is hidden. Anatomization divides the sensitive values and attribute values without modifying the values of attributes, and then these are placed into different locations for implementing privacy preservation.

Data Update

For efficiency, once it is update occurs all the data sets are not acceptable to be anonymized Hence, the best option is to just anonymize the updated part and regulate the data sets which are already anonymized. For data sets to be anonymized, level of anonymization defines the degree of privacy preservation.

Anonymous Data Management

In the cloud computing, the expenses of resources of computation and storage depends on the demand of the user that how much resources are used by him. In the feature of payment for each-usage, instead of re-computing the data sets over and over again, it is helpful to store some definite part of the intermediate data sets. A huge amount of independently anonymous data sets is retained which are not safe because of the disclosure of privacy problems. As the PSI module gives privacy specifications which are flexible, different anonymous data sets can be obtained by anonymizing an input raw data set.

Therefore, resulting in recovery of privacy-sensitive information can be done from multiple anonymous data sets. For referring this information disclosure issue in multiple data sets, a technique which includes encryption for confirming the preservation of privacy. Actually, all the anonymous data sets can be encrypted and sent to particular users. But, it will be very costly for encryption of all the intermediate data sets as many data users access or process these anonymized input data sets repeatedly. Hence, for saving the expense of privacy preservation only some part of intermediate data sets is encrypted and yet the preservation of privacy can be ensured.

V. CONCLUSION

Privacy of Big data has turned to be a main problem as it is directly associated with the customers. Now the privacy in big data analytics must be necessarily assured by an organization. Instead of focusing on collection of data, uses of data should be focused in Privacy measures. These Privacy measures must be changed according to unexpected uses of big data and its size. Methods such as anonymization have limited potential when applied to big data. In Notice and consent technique also does not guarantee the customer for ensuring privacy. In Differential privacy, privacy of big data is possible but there is an issue with this technique that is before using the differential privacy model the query must be known to the analyst. After changing and applying to big data and devoid of actually altering the data privacy may be ensured.

REFERENCES

[1] X. Zhang, C. Liu, S. Nepal, C. Yang, J. Chen, "Privacy Preservation over Big Data in Cloud Systems," Security, Privacy and Trust in Cloud Systems, pp 239-257, Springer.

[2] M. V. Dijk, A. Juels, "On the impossibility of cryptography alone for privacy-preserving cloud computing," Proceedings of the 5th USENIX conference on Hot topics in security, August 10, 2010, pp.1-8.

[3] J. Sedayao, “Enhancing cloud security using data anonymization”, White Paper, Intel Coporation.

[4] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, pp. 557–570, 2002.

[5] N. Li, T. Li, S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity**,** " IEEE 23rd International Conference on Data Engineering, 2007, pp. 106 - 115.

[6] F. H. Cate, V. M. Schönberger, "Notice and Consent in a World of Big Data," Microsoft Global Privacy Summit Summary Report and Outcomes, Nov 2012.

[7] J. Salido, "Differential privacy for everyone," White Paper, Microsoft Coproration, 2012.

[8] S. Sagiroglu and D. Sinanc, “Big Data: A Review,” Proc. International Conference on Collaboration Technologies and Systems, 2013, pp. 42-47

[9] Y. Demchenko, P. Grzsso, C. De Laat, P. Membrey, “Addressing Big Data Issues in Scientific Data Infrastructure,” Proc. International Conference on Collaboration Technologies and Systems, 2013, pp. 48-55.

[10] Top Ten Big Data Security and Privacy Challenges, Technical report, Cloud Security Alliance, November 2012

[11] S. H. Kim, N. U. Kim, T. M. Chung, “Attribute Relationship Evaluation Methodology for Big Data Security,” Proc. International Conference on IT Convergence and Security (ICITCS), 2013, pp. 1-4.

[12] S.H. Kim, J. H. Eom, T. M. Chung, “Big Data Security Hardening Methodology Using Attributes Relationship,” Proc. International Conference on Information Science and Applications (ICISA), 2013, pp. 1-2.

[13] Big Data Privacy Preservation, Ericsson Labs, http://labs.ericsson.com/blog/privacy-preservation-in-big-data-analytics

[14] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and cell suppression," Technical report, SRI International, 1998.

[15] O. Heffetz and K. Ligett, “Privacy and data-based research,” NBER Working Paper, September 2013