Enhance Privacy in Big Data and Cloud via Diff-Anonym Algorithm

Imad Ali Hassoon
Faculty of Automatic and Computers
University "Politehnica" of Bucharest
Bucharest, Romania
emadali1980@yahoo.com

Nicolae Tapus
Faculty of Automatic and Computers
University "Politehnica" of Bucharest
Bucharest, Romania
ntapus@cs.pub.ro

Anwar Chitheer Jasim
Faculty of Automatic and Computers
University "Politehnica" of Bucharest
Bucharest, Romania
anwaralbazooni@gmail.com

***Abstract*— the main issue with big data in cloud is the processed or used always need to be by third party. It is very important for the owners of data or clients to trust and to have the guarantee of privacy for the information stored in cloud or analyzed as big data. The privacy models studied in previous research showed that privacy infringement for big data happened because of limitation, privacy guarantee rate or dissemination of accurate data which is obtainable in the data set. In addition, there are various privacy models. In order to determine the best and the most appropriate model to be applied in the future, which also guarantees big data privacy, it is necessary to invest in research and study. In the next part, we surfed some of the privacy models in order to determine the advantages and disadvantages of each model in privacy assurance for big data in cloud. The present study also proposes combined Diff-Anonym algorithm (K-anonymity and differential models) to provide data anonymity with guarantee to keep balance between ambiguity of private data and clarity of general data.**

Big Data; cloud; privacy models; K-anonymity; differential Privacy.

I. INTRODUCTION

Big Data and cloud are currently important in the territory of computer science. These two topics have drawn the attention of researchers as well as designers and businessmen. The aim of the researchers is to identify the best methodology of extracting information through enormous information of various types of data. Organizations need to improve their frameworks for various applications. This enthusiasm from various groups will decide the development in research in enormous information applications. Big data is specific and it differs depending on the activity domain:

- Healthcare
- Science
- Education
- Governance
- Sporting events
- Banking processes.

Almost all of these fields need some data to be public, while the rest must be privacy-protected to support customer confidence and companies during the use of big data.

In another words, big data is changing the very way of business, from social insurance to retail and farming. The rate at which information is collected on every conceivable activity means that there are increasing opportunities to fine-tune procedures and operations to squeeze out every last exclusion of efficiency. Of course, in business once an item has been developed or manufactured it should be sold and distributed [9]. The amount of client information, including you and me, effectively assembled by big retailers lets them know who will need to purchase what, where and when. Also, the cloud changes the ways data is stored and gives new views on information management, while applications are becoming easier to be used by anyone.

II. RELATED WORKS AND BACKGROUND

a) related works

Big data have the same details and properties as any other data, as well as the issue of privacy maintenance, especially when these data are stored in cloud or used by third party. Many related works have reached good results in solving the same problem as the present study. In [5] a highly scalable approach with of two-phased TDS is proposed to obtain data anonymization over MapReduce in cloud. They implement their approach by splitting data into two phases: one phase for the original data sets to be split in smaller groups of data; another phase to achieve the anonymization goal in parallel with these groups of data. Soria-Comas, Jordi, et al. [15] analyzed the synergy between k-anonymity and differential privacy. The results showed that k-anonymity can be employed to improve differential privacy and increase data protection guarantee. Thus, the authors proposed to combine these two methods to provide privacy for microaggregation data. In [2] the authors examined privacy models and focused on disclosure risk limitation. Also, the study evaluated two privacy methods (k-anonymity and differential privacy) in terms of computational cost, composability and link-ability when dealing with big data. The research considers that Anonymization is the best tool that guarantees privacy for big data and reduce the risk of disclosure. In [3], the authors reviewed the improvements of the differential privacy along with other methods. The study focused on selected epsilon for a better trade-off between privacy and utility of the datasets and proposed to achieve in future work further results of data privacy by differential methods that are better in combination with other methods or with cryptography.

b) background of privacy methods

1. **K-anonymity**.

The aim of k-anonymity is to cover data sets by restricting the intruders and not allowing disclosure of private data. Its purpose is to block any cases of individuals' identity disclosure. The objective of k-anonymity is to create a collection of quasi-

identifiers in the anonymized data to indicate at least k-tuples, which are purported equality tuples.

K-anonymity was proposed in 2002 by Sweeney [22], and was further developed in 2008 by Lodha and Thomas [19].

Definition 1*: Let RT (A1, . . . , An) be a table and QIRT be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in RT [QIRT] appears with at least k occurrences in RT [QIRT] (Sweeney, 2002)* [1], [2], [22].

To apply k-anonymity if k=3 or k=N, then at least three rows or N rows should have equivalence class. Generalization after achieving that is applied to the operations; first is done on the dataset and after the k-anonymization technique is applied [2], [7].

2. Differential-privacy

Differential-privacy is one of the privacy models used for anonymization, which provides privacy safeguards more than other models, such as k-anonymity, T-Closeness or L-diversity [10]. Differential-privacy implies publishing the results of a query with some noise added to the results of the query. In this case, the attacker can't guess the results of the query because it contains noise with 100% guarantee. Differential-privacy has several drawbacks. The first major impediment is that differential privacy fails to give assurances with dataset linkage and attribute in data. Usually, this model is preferred in cases where the result of congruence queries is few and with low sensitivity. This makes differential-privacy the best in restricted classes of queries [2]. This model was initially proposed by Cynthia Dwork in 2008 [18], [20]. The assumption is the anonymization of the component set between the queries of the client and the response of the database. In 2007 Talwar and McSherry proposed a tool which guarantees to add noise in output results to include quality on functions which were not robust [21]. Then in 2008 Dwork suggested to use for privacy a mechanism that includes some multi sided noise on queries of individuals [18], [20]. The matrix system on differential privacy proposed in 2010 by Li et al. responds to predicated multiple queries that could increase precision [16]. In 2010, Friedman and Schuster analyzed the architecture for differential privacy tools [17]. Muralidhar and Sarathy (2011) evaluated the performance and privacy of Laplacian noise addition for numeric data to achieve the differential privacy for numeric data [14], [8]. Microsoft Company discussed in 2012 in a report entitled “Differential Privacy for Everyone” that differential privacy could be a vital part in data privacy of individuals by embedding multi layers of noise and original data to protect from analysts and attackers. When sending a query for data by analysts or attackers, the privacy guard responds with datasets that include layers of noise to protect the data. If the results contain a privacy breach, the mechanism will append noise on the query depending on the differential privacy mechanism. So, the result will contain loud reaction and distorting of personal data in small values which will help protect the analysis of researchers without affecting the privacy of the individual. Finally, differential privacy as one of the most important models to provide privacy, aims to split data to small parts while adding noise to the queries to guarantee that it won’t affect the analysis of the researcher nor questions the privacy of the individual [12]. Over the previous years, many ways to add noise on data to protect individual privacy have been proposes in numerous research on differential privacy.

III. PURPOSE OF THIS RESEARCH

As we mentioned, the importance of providing privacy for big data in cloud resides in the fact that this data are usually processed or used by third party. Therefore, in this case the problem is the selection of the privacy methods that can support high safety for data that will be treated or used by third party, as well as during sending or receiving from other parties. In order to reach this purpose, the present study proposes to combine multi methods and we present an overview of the methods of privacy (K-anonymity, Differential-privacy) and we explain how to combine these two methods to form a new algorithm called “Diff-Anonym”.

IV. PROPOSED DESIGN PROCESS

The proposed design process includes a suggested combination of the two models of privacy and implementation of the algorithm Diff-Anonym to provide privacy for big data or cloud.

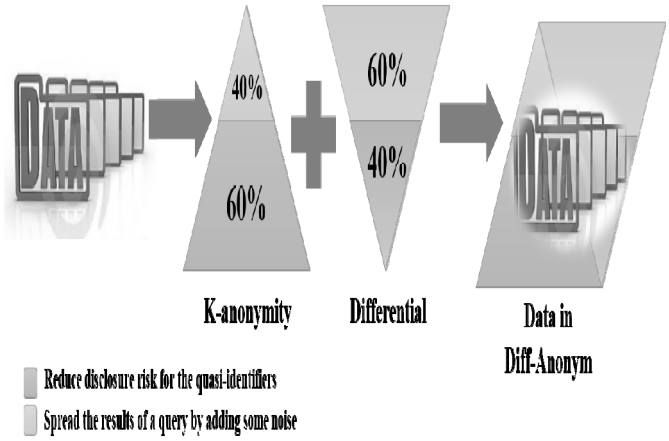

Figure 1 Design process of Diff-Anonym

The proposed framework gives privacy to data loaded in it. It is equipped for handling big data. The effectiveness of the framework proposed lies in providing anonymity without influencing the speed of operations. It can be scaled from mini data to any size of data. Execution of processes doesn't relate on value or size of data.

The following algorithm explains the sequence of processes:

Algorithm.1 Diff-Anonym

Input: Data set from any size of data to include privacy.

Output: Data set with privacy models (k-anonymous - differential).

Step 1: Upload data into framework.

Step 2: Select fields of attributes to arrange in new temporary tables.

Step 3: Detect quasi identifier in temporary tables.

Step 4: Split tables into mini tables.

Step 5: Apply k-anonymity to mini temporary tables.

Step 6: Detect equal attributes in results.

Step 7: Spread the results of apply k-anonymity.

Step 8: Add noise to the data which already have equal attributes

in results.

Step 9: Re-combine the results in big data set.

V. EXAMPLE ABOUT DIFF-ANONYM

To explain the processes of the algorithm proposed, we suppose random data to give a brief example of the expected results. Examine the in-team-football dataset to be used for healthcare.

Sq_no	*ID_no*	*Name_ Player*	*Age_ player*	*Gender_ player*	*Position_ Player*	*Status_ Player*
1	121010	Hanna	22	F	Physician assistant	Influenza
2	121011	Sami	31	M	Striker	Ready
3	121012	Lyla	23	F	Physician assistant	Malaria
4	121013	Dieaa	33	M	Defender	Infected knee
5	121014	Samir	36	M	Defender	Ready
6	121015	Safe	28	F	Doctor	Influenza
7	121016	Jasmine	29	M	Defender	Hepatic
8	121017	Oday	37	M	Striker	Infected knee
9	121018	Younis	41	M	Goalkeeping	Muscle spasm
10	121019	Mustafa	48	M	Coach assistant	Influenza
11	121020	Jasim	46	M	Coach	Ready
12	121021	Karim	44	M	Goalkeeping coach	Ready

Table No.1 The original dataset

The ordinary routine of distributing individual particular information, is directly hiding the fields of the table which have explicit identifiers. The explicit identifiers in Table No.1 are the ID_no. and name_Player [2]. The result of the dataset after hiding these explicit identifiers is given in Table No.2.

Sq_no	*Age_player*	*Gender_player*	*Position_Player*	*Status_Player*
1	22	F	Physician assistant	Influenza
2	31	M	Striker	Ready
3	23	F	Physician assistant	Malaria
4	33	M	Defender	Infected knee
5	36	M	Defender	Ready
6	28	F	Doctor	Influenza
7	29	M	Defender	Hepatic
8	37	M	Striker	Infected knee
9	41	M	Goalkeeping	Muscle spasm
10	48	M	Coach assistant	Influenza
11	46	M	Coach	Ready
12	44	M	Goalkeeping coach	Ready

Table No.2 Result after hiding the explicit identifiers

In steps 3, and 4 we detect the quasi identifiers in the temporary tables and split the tables into mini tables. To apply this in table No.2, we detected the quasi identifiers (Age_ Player and Position_Player) and we implement step 4. The result is presented in table No.2A

Sq_no	*Age_player*	*Position_Player*
1	22	Physician assistant
2	31	Striker
3	23	Physician assistant
4	33	Defender
5	36	Defender
6	28	Doctor
7	29	Defender
8	37	Striker
9	41	Goalkeeping
10	48	Coach assistant
11	46	Coach
12	44	Goalkeeping coach

Table No.2A Dataset after split of the quasi identifier

Sq_no	*Gender_player*	*Status_Player*
1	F	Influenza
2	M	Ready
3	F	Malaria
4	M	Infected knee
5	M	Ready
6	F	Influenza
7	M	Hepatic
8	M	Infected knee
9	M	Muscle spasm
10	M	Influenza
11	M	Ready
12	M	Ready

Table No.2B Dataset after split of the normal data

In step 5 to apply k-anonymity we can suppose k=2, as comparability ought to have no less than two columns [2]. Table 2A includes the fields of quasi identifiers which are processed. After achieving the results, the operations are applied, first done on the dataset. After that the k-anonymization technique can be applied and the results can be seen in Table No.3.

Sq_no	*Age_player*	*Position_Player*
1	20-25	Professional
3	20-25	Professional
6	25-30	Professional
7	25-30	Player
2	30-35	Player
4	30-35	Player
5	35-40	Player
8	35-40	Player
9	40-45	Player
12	40-45	Professional
10	45-50	Professional
11	45-50	Professional

Table No.3 Results after k-anonymities application on table with k = 2

In step 6 we find in table No.3 the tuples with a similar equivalence in tuples that include the same value of the quasi-identifiers (tuples [1,3], [2,4], [10,11]). Table No.3 includes three equivalence tuples. One of the suppositions with this model of k-anonymity is that the data owner acknowledges the quasi-identifiers and that requires taking full care of the privacy data which needs to be hidden in the published information. The drawback of this model is that it can suffer from two kinds of attacks. The homogeneity assault is the first kind of attack on k-anonymity. This attack often succeeds when lacking diversity in the sensitive attribute. Background attack is another kind of attack on k-anonymity and it occurs when the striker possesses some background about the personal data of individuals. Thus, the ability of the k-anonymity approach to prevent disclosure is quite low taking into account the drawbacks above. In this case, attackers could be able to have additional information about an individual and guess more knowledge about any data in the published tables when the attributes disclosure is not protected [2].

Based on the above results of step 6, the k-anonymity can't provide full anonymity of the data. In the following steps, we attempt to solve the issues in order to provide high privacy for data. In steps 7 and 8 we will apply differential to add noise for the equivalence classes from the table No.3A:

Sq_no	*Age_player*	*Position_Player*
1	20-25	Professional
3	>20	P####

6	25-30	Professional
7	25-30	Player
2	30+	Player
4	30-35	P****
5	35-40	Player
8	35-40	Player
9	40-45	Player
12	40-45	Professional
10	-50	**********
11	4**	**********

Table No.3A. Results of differential application

In table No.3A, the results include different kinds of anonymity with noise which guarantees the data privacy. In step 9 we re-arrange the results in normal way of data. Table 4 shows the end results after applying the Diff-Anonym algorithm.

Sq_no	*Age_player*	*Gender_player*	*Position_Player*	*Status_Player*
1	20-25	F	Professional	Influenza
2	30+	M	Player	ready
3	>20	F	P####	Malaria
4	30-35	M	P****	infected knee
5	35-40	M	Player	ready
6	25-30	F	Professional	Influenza
7	25-30	M	Player	Hepatic
8	35-40	M	Player	infected knee
9	40-45	M	Player	Muscle spasm
10	-50	M	**********	Influenza
11	4**	M	**********	ready
12	40-45	M	Professional	ready

Table No.4 Results of Diff-Anonym algorithm application

Table No.4 includes all data after application of the Diff-Anonym algorithm. The data has now become ready to be stored in the Big Data location or published as anonymous data. In general, if the results after step 9 contain privacy breaches, the last steps need to be repeated to append noise on the query depending on differential privacy mechanism. So, the result will contain loud reaction and distorting of personal data in small values, in which case can be used by researchers for analysis without affecting the privacy of the individual.

VI. Discussion

The methods k-anonymous, t-Closeness and L-diversity [2], [10], [11] are the most common privacy protection methods. The methods t-Closeness and L-diversity deal with enhanced ability of preventing privacy leaks, and have been widely used in research. The k-anonymity method provides privacy by limited disclosure of data. The differential privacy supporters view k-anonymity as an old-fashioned privacy method that offers only poor disclosure limitation guarantees, while differential privacy detractors criticize the limited utility of differentially private outputs and especially of not having access to the data set [13]. Thus, the present study focuses on disclosure limitations of big data in cloud, for which we selected to combine k-anonymous method and differential privacy.

In our simulation, we presented a small example and we can observe the differences of the results in table No.4 compared with table No.1. These results prove that we can implement Diff-Anonym algorithm to provide high privacy for big data. Also, we achieved our goal of reducing the possibilities of guess privacy in the case of attack on the data. The advantages of our proposal are increased guarantee by differential privacy and limited identity disclosure of individuals by K-anonymity method. The combination of these two methods could contribute to provide high privacy for big data.

VII. Conclusion and future work

The variety of privacy models and whichever provides a guarantee to maintain big data privacy requires research and study to determine the best and the most appropriate one to be applied in the future. In this paper, we reviewed models of privacy and focused on two models to be implemented in the system for privacy management in big data.

Based on our review on privacy models, k-anonymity has especially capacity to ensures against identity disclosure, as indicated above, but it is not all in all secured against attribute disclosure: In case the attacker has knowledge about the individuals, and the data are similar in several rows with the same attribute, often k-anonymity is not able to provide privacy for the target individual [2], [6], [7]. In addition, to solve the drawbacks of k-anonymity we propose the solution with differential privacy model as it could guarantee true privacy for data and individuals. Other strategies, for example, anonymization is prone to different attacks. With differential privacy model, almost any quasi identifier attack could be prevented because it is based on different data set generated each time with different layers of noise in the original data. Regardless of the attacker or analyst having additional information about the target data set [2], [4], the layers of noise should prevent them from being able to interpret the data. This way, the data privacy is guaranteed. Privacy infringement for big data happens because of limitation, privacy guarantee rate or dissemination of accurate data which is obtainable in the data set. In addition, we proposed in this paper a new Diff-Anonym algorithm combining k-anonymity and differential models to provide anonymity on data with guarantee to keep balance between the ambiguity of private data and clarity of general data. In future works, we will implement our proposal with two models with improved algorithm to manage privacy of big data, which will guarantee data safety from attribute attack and ensure identity disclosure prevention.

References

[1] Jordi Soria-Comas, Josep Domingo-Ferrer, (2016),"Big data privacy: challenges to privacy principles and models", Data Sci. Eng. (2016) 1(1):21–28, DOI 10.1007/s41019-015-0001-x

[2] Nancy Victor* and Daphne Lopez, (2016), “Privacy models for big data: a survey”, January 2016, DOI: 10.1504/IJBDI.2016.073904.

[3] X.Yao, X.Zhou, J.Ma, (2016), Differential Privacy of Big Data: An Overview, 2016 IEE .HPSC-IDS.2016.9.

[4] T.Feng, Y.Guo, Y.Chen (2016), “A Differential Private Collaborative Filtering Framework Based on Privacy-Relevance of Topics”, 2016 IEEE. (ISCC), 978-1-5090-0679-3/16.

[5] Zorige Priyanka, K Nagaraju, Dr. Y Venkateswarlu, (2014),”Data Anonymization Using Map Reduce on Cloud based A Scalable Two-Phase Top-Down Specialization”, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, December 2014.

[6] Kavitha Sl Yamini S2 Raja Vadhana p3 (2015), “An Evaluation on Big Data Generalization Using k-Anonymity Algorithm on Cloud”, IEEE Sponsored 9th International Conference on Intelligent Systems and Control.

[7] Anirban Basu, Toru Nakamura, Seira Hidano, Shinsaku Kiyomoto(2015), “k-anonymity: risks and the reality”, IEEE DOI 10.1109/Trustcom-BigDataS.

[8] Zakerzadeh, H. and Osborn, S.L. (2011). “Faanst: fast anonymizing algorithm for numerical streaming data”. Data Privacy Management and Autonomous Spontaneous Security, pp.36–50, Springer, Berlin, Heidelberg 2012).

[9] D.Wang, Zh.Han, (2015), “Sublinear Algorithms for Big Data Applications”, ISBN 978-3-319-20448-2 (eBook).

[10] Sánchez D., Domingo-Ferrer J., Martínez S. (2014), “Improving the Utility of Differential Privacy via Univariate Microaggregation”, Domingo-Ferrer J. (eds) Privacy in Statistical Databases. PSD 2014. Lecture Notes in Computer Science, vol 8744. Springer, Cham.

[11] Jordi Soria-Comas, Josep Domingo-Ferrer, D.Sánchez and S.Martínez, (2015), "t-Closeness through micro aggregation: strict privacy with enhanced utility preservation", IEEE Transactions on Knowledge and Data Engineering, Vol.27, no.11, pp. 3098-3110, Oct 2015, ISSN: 1041-4347.

[12] Xiao, Y., Xiong, L., Yuan, C. (2012). “Differentially private data release through multidimensional partitioning”. Proceedings of the 7th VLDB conference on Secure data management (SDM’10), pp. 150-168 (2010).

[13] Adeel Anjum, Adnan Anjum (2014), “Differentially Private K-anonymity”, 12th International Conference on Frontiers of Information Technology, IEEE DOI 10.1109/FIT.2014.37.

[14] Sarathy, R. and Muralidhar, K. (2011) ‘Evaluating Laplace noise addition to satisfy differential privacy for numeric data’, Transactions on Data Privacy, Vol. 4, No. 1, pp.1–17.

[15] Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., & Martínez, S. (2014). Enhancing data utility in differential privacy via microaggregation-based k-anonymity. The VLDB Journal, 23(5), 771-794.

[16] Li, C. et al. (2010) ‘Optimizing linear counting queries under differential privacy’, Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, ACM.

[17] Friedman, A. and Schuster, A. (2010) ‘Data mining with differential privacy’, in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp.493–502.

[18] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum and S. Vadhan (2009). “On the complexity of differentially private data release: efficient algorithms and hardness results”. In: Proc. of the 41st Annual Symposium on the Theory of Computing-STOC 2009, pp. 381-390, 2009.

[19] Lodha, S. and Thomas, D. (2008) “Probabilistic anonymity”, in Privacy, Security, and Trust in KDD, pp.56–79, Springer, Berlin, Heidelberg.

[20] C. Dwork, (2008) “Differential privacy: a survey of results”, in Theory and Applications of Models of Computation, pp.1–19, Springer, Berlin, Heidelberg.

[21] McSherry, F. and Talwar, K. (2007) ‘Mechanism design via differential privacy’, in 48th Annual IEEE Symposium on Foundations of Computer Science, 2007, FOCS’07, IEEE, pp.94–103.

[22] Sweeney L .(2002) “k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No.5 , pp.557–570.