Dr Vikas Lamba, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 10, Issue 9, (Series-V) September 2020, pp. 29-34

RESEARCH ARTICLE OPEN ACCESS

Study and analysis of privacy preservation and security concerns in big data

Tanveer Ahmad Dar[1], Dr Vikas Lamba[2], Zubair Ahmad Mir[3]
Department of Computer Science and Engineering, Vivekananda Global University

ABSTRACT:
Big data is a series of huge and complex data units that cannot be stored and processed by the conventional records processing systems. Hence big data requires high computational strength and storage and usually on distributed system Big data are used. Big data analytics method reading invisible records styles from the bigger statistics units. The information units are accrued from diverse resources i.e., Social media, business quarter, healthcare, information governance, various institutions, etc. So, privacy and security is foremost subject in huge facts. This paper especially makes a speciality of l-anonymity strategies maintain the privacy of facts. This research ambitions to focus on 3 important anonymization techniques used in a scientific subject specifically, K-anonymity, l-range, and T-closeness.
Keywords: Big data, Privacy, security, Data Anonymization, l-diversity, t-closeness

---

Date of Submission: 07-09-2020 Date of Acceptance: 22-09-2020

I. INTRODUCTION

Big data is related as the big series of statistics. So, this information cannot be processed with the aid of traditional technologies. Big data is differentiated from the traditional technology in three approaches i. e., Volume, Speed and variety.

Volume: represents large quantity of information. Information generated from multiple resources such as Facebook, Twitter, Instagram, Business zone, hospitals, Testing Labs, institutions etc.

Speed: represents the pace of information

Variety: of data is coming from distinctive sources in many forms such as images, audio, video, files, emails, economic transactions, simulations, 3D models, etc. In this paper we file numerous techniques that can help privacy in big data.

The privacy and security in big data becoming more challengeable concern. In complex applications, no model is suggested yet for the security of big data due to which it gets disabled by default. However, without which data can always be easily compromised. This part focuses on the privacy and security concern of big data.

Privacy is concerned with the privilege to have control over the data collected and used by the person. Data privacy is the capacity of single person or group of persons to stop data about themselves from becoming known to the person other than they gave privilege to. The identification of personal data during Internet transmission Internet is one of the serious privacy issue[1].

Security is concerned to the execution of protecting data and its benefits through the use of technology by the processes and training from unauthorized access, leak Privacy vs. safety Data private-ness is centred on the use and governance of person data—things like putting up insurance policies in area to make sure that consumers' private statistics is being collected, shared and utilized in excellent ways. Security concentrates greater on defending data from malicious assaults and the misuse of stolen information for profit [2]. While security is imperative for defending data, it's now not adequate for addressing privacy.

Privacy requirements in big data

Big data analytics draw in various organizations; a hefty portion of them figure out not to utilize these offerings because of the absence of standard safety and privacy protection tools. These sections analyse viable strategies to upgrade large data platforms with the assist of privacy protection capabilities. The foundations and improvement strategies of a framework that supports

- The specification of Privacy insurance policies managing the get right of entry to to statistics saved into goal massive facts platforms,
- The technology of productive enforcement video display units for these policies, and
- The integration of the generated video display units into the goal analytics platforms. Enforcement strategies proposed for usual DBMSs show up insufficient for the massive records context due to the strict execution

requirements wanted to deal with giant statistics volumes, the heterogeneity of the data, and the velocity at which records should be analysed

BIGDATA SECURITY AND PRIVACY CHALLENGES

Security center of attention on shield the enterprise. Data privacy focuses on person user's information. There are broadly speaking three targets of security are secrecy, reliability and accessibility. As indicated by way of the article with the aid of Cloud Security Alliance (CSA) [3], there are especially many difficulties in the discipline of Big Data security and protection as referenced beneath:

1) Secured calculations in dispersed programming systems
2) Security great practices for non-relational facts stores
3) End-point enter approval and sifting
4) Real-time protection observing
5) Privacy-safeguarding records mining and examination
6) Cryptographically upheld facts pushed security
7) Granular get right of entry to control
8) Secure records stockpiling and exchanges logs
9) Granular
10) Data provenance

PRIVACY PROTECTION IN BIGDATA

Privacy is principal challenge in big data so we want environment friendly privacy renovation methods. Privacy without delay related to customers. Privacy normally focuses on user's person data as a substitute than complete series of data. The privacy preservation methods can be used defend the individuals sensitive information. Privacy is vital in three tiers i.e. information generation, data storage, information processing. In this paper focusing on statistics of Anonymization, and l-diversity in large data storage.

So in today's digital rich environment, by what means does a data holder, such as a health care, public intervention, or monetary association, share person-specific archives in such a way that the released statistics remain almost useful but the individuality of the persons who are the subjects of the data cannot be determined? One way to accomplish this is to have the released information adhere to k-anonymity,

a.**Big data privacy in data storage phase**

Data stores huge quantity of data. There are precise strategies to keep privacy in storage.

Classifications of attributes are

Security consist of especially three dimensions i. e. Confidentiality, integrity, availability [4]. Cryptographic encryption mechanisms are public key encryption, identification particularly based encryption, attribute primarily based encryption etc. In public key encryption approves an data sender to scramble the facts below the general populace key of recipient, the recipient decrypts the statistics underneath personal key recipient. So, there can be leakage of data. This cryptographic mechanism does now no longer fulfil every one of the stipulations of consumers in the state of affairs of correct sized facts stockpiling

In ordinary encryption mechanisms can't acclaim the anonymity of cipher textual content receiver /sender. So, absolutely everyone can without difficulty acquiring a cipher textual content (e.g. cloud server), if any one is aware of the public key of the discern message, that is scrambled below the proprietor of the discern content material .So, outsider can besides a lot of a stretch receives the undeniable text[5].

b. **Data Anonymization**

Data Anonymization capability via putting off the private important points to keep the privacy of customers [6,7]. It is additionally referred to as de-identification. Whenever businesses releasing the records publicly through anonymizing it. Anonymization alludes to concealing the identifier characteristics (the characteristics that specially apprehend the columns) like Aadhar range ,bank a/c variety ,full name, licence number, voter identification etc. This nameless information hyperlinks to exterior records [8]. With the quit purpose to hold records from re-identification, the thoughts of k-anonymity, diversity range and t-closeness have been presented.

Key attributes	**Quasi Identifiers**	**Sensitive attributes**	**Non-sensitive attribute**
Based on the attributes uniquely identifies tuples for example: Social security number, pan number, Aadhar number, voter id, driving license number etc.	An arrangement of traits that can be conceivably connected with outside data to re-distinguish entities. for example: phone no., DoB, gender etc.	some of the attributes contains sensitive value with respect to data owner. for example: Income and health issues etc	disclosing the non-sensitive attributes will not break the secrecy of user

i.k-anonymity: It can be used prevent record linkage. Release of data is said to have the k-anonymity [2, 9] property if the information for each person contained in the release cannot be perceived from at least k-1 individuals whose information show up in the release. Therefore, k-anonymity provides privacy protection by guaranteeing that each record relates to at least k individuals even if the released records are directly linked (or matched) to external information. To preserve the privacy, the following Anonymization techniques are applied to the data [10, 11, 12]. Suppression and Generalization

Suppression: quasi identifiers are supplanted or darkened by some steady qualities like 0, *, @ and so on. for example: some values like license number, Aadhar number can be invisible using asterisk.

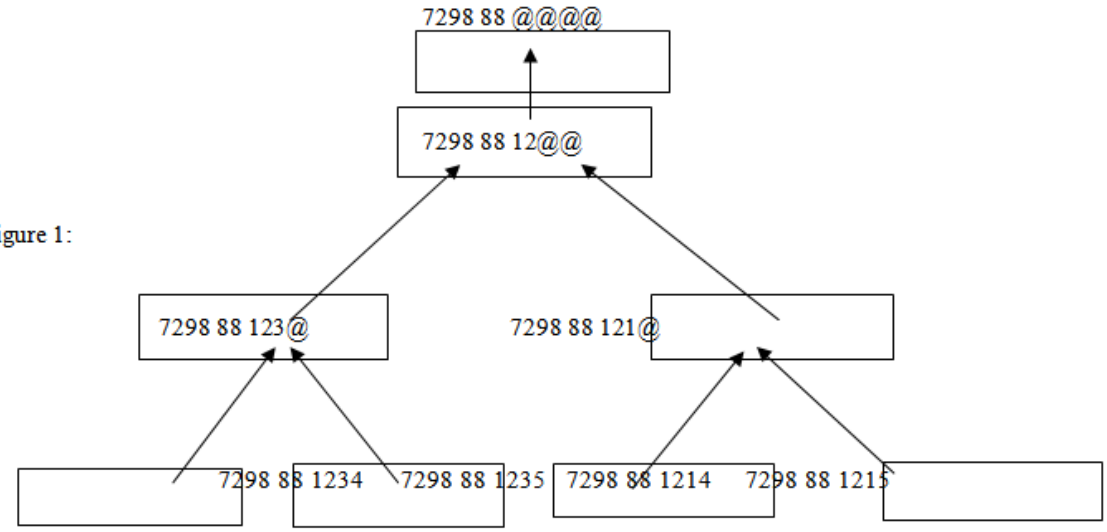
Generalization: Some values are replaced by parent values. The trait age can be written in more broad shape. for example: The attribute age can be written in more general form. like age is equal to 19 (age=19). The age attribute can be generalized to age is less than or equal to 20 (age<=20).The below table1 shows the patient record. Whereas pin code, Age, district are Non-Sensitive attributes. Disease is sensitive attribute.

Table1: shows the patient record

Non-Sensitive			Sensitive	
S.No.	Phone	Age	District	Disease
1	7298 88 1234	30	Anantnag	BP Influenzas tumour
2	7298 88 1235	30	Kulgam	tumour
3	7298 88 1236	25	Pulwama	
4	7298 88 1237	40	Kulgam	
5	7298 88 1238	40	Anantnag	Influenzas tumour
6	7298 88 1239	30	Shopian	BP
7	7298 88 1212	40	Kulgam	BP
8	7298 88 1213	40	Anantnag	
9	7298 88 1214	45	Kulgam	Gout
10	7298 88 1215	25	Kulgam	Gout
11	7298 88 1216	35	Kulgam	Influenzas
12	7298 88 1217	42	Kulgam	Influenzas

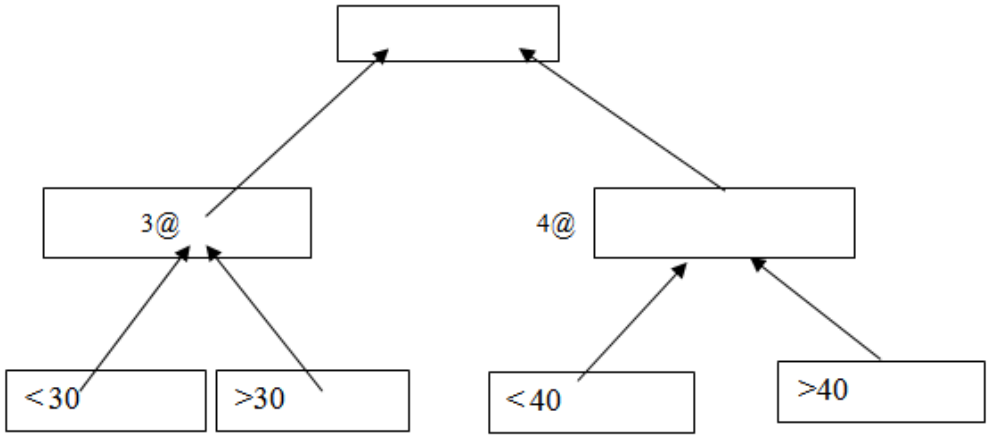
This paper signifies generalization to include suppression by striking on each value of generalization hierarchy a new symbol, atop the old symbol. The new symbol is the attribute's suppressed value. The height of each value generalization hierarchy is thereby incremented by one. No other changes are necessary to incorporate suppression. Figure 1 and Figure 2 provides examples of domain and value generalization hierarchies expanded to include the suppressed maximal element (@@@@).

Figure 1:



@@

Figure 2:



In Table2: Phone no. and age value are generalized including suppression.

Non-Sensitive			Sensitive	
S.No.Phone		Age	District	Disease
1 2 3 4	7298 88 12@@ 7298 88 12@@ 7298 88 12@@ 7298 88 12@@	3@ 3@ 3@ 4@	Anantnag Kulgam Pulwama Kulgam	BP Influenzas tumour tumor
5 6 7 8	7298 88 12@@ 7298 88 12@@ 7298 88 12@@ 7298 88 12@@	4@ 3@ 4@ 4@	Anantnag Shopian Kulgam Anantnag	Influenzas tumour BP BP
9 10 11 12	7298 88 12@@ 7298 88 12@@ 7298 88 12@@ 7298 88 12@@	4@ 3@ 3@ 4@	Kulgam Kulgam Kulgam Kulgam	Gout Gout Influenzas Influenzas

ii. L-diversity

It is a form of group-based anonymization that is utilized to safeguard privacy in data sets by reducing the granularity of data representation. This decrease is a trade-off that results outcomes in some loss of viability of data management or mining algorithms for gaining some privacy. The l-diversity model (Distinct, Entropy, Recursive) [2, 8, 9] is an extension of the k-anonymity model which diminishes the granularity of data representation utilizing methods including generalization and suppression in a way that any given record maps onto at least k different records in the data. The l-diversity model handles a few of the weaknesses in the k-anonymity model in which protected identities to the level of k-individuals is not equal to protecting the corresponding sensitive values that were generalized or suppressed, particularly when the sensitive values in a group exhibit homogeneity. The l-diversity model includes the promotion of intra-group diversity for sensitive values in the anonymization mechanism. The problem with this method is that it depends upon the range of sensitive attribute. If want to make data L-diverse though sensitive attribute has not as much as different values, fictitious data to be inserted. This fictitious data will improve the security but may result in problems amid analysis. Also L-diversity method is subject to skewness and similarity attack [8] and thus can't prevent attribute disclosure.

iii.T-closeness

It is a further improvement of l-diversity group based anonymization that is used to preserve privacy in data sets by decreasing the granularity of a data representation. This reduction is a trade-off that results in some loss of adequacy of data management or mining algorithms in order to gain some privacy. The t-closeness model (Equal/Hierarchical distance) [2, 13] extends the l-diversity model by treating the values of an attribute distinctly by considering the distribution of data values for that attribute. An equivalence class is said to have t-closeness if the distance between the conveyance of a sensitive attribute in this class and the distribution of the attribute in the whole table is less than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness. The main advantage of t-closeness is that it intercepts attribute disclosure. The problem lies in t-closeness is that as size and variety of data increases, the odds of re-identification too increases.

Privacy-preserving big data publishing

The publication and dissemination of raw data are crucial components in commercial, academic, and medical applications with an increasing number of open platforms, such as social networks and mobile devices from which data might be gathered, the volume of such data has also increased over time [14]. Privacy-preserving models broadly fall into two different settings, which are referred to as input and output privacy. In input privacy, the primary concern is publishing anonymized data with models such as k-anonymity and l-diversity. In output privacy, generally interest is in problems such as association rule hiding and query auditing where the output of different data mining algorithms is perturbed or audited in order to preserve privacy. Much of the work in privacy has been focused on the quality of privacy preservation (vulnerability quantification) and the utility of the published data. The solution is to just divide the data into smaller parts (fragments) and anonymize each part independently [15]. Despite the fact that k-anonymity can prevent identity attacks, it fails to protect from attribute disclosure attacks because of the lack of diversity in the sensitive attribute within the equivalence class. The l-diversity model mandates that each equivalence class must have at least l well-represented sensitive values. It is common for large data sets to be processed with distributed platforms such as the MapReduce framework [16, 17] in order to distribute a costly process among multiple nodes and accomplish considerable performance improvement. Therefore, in order to resolve the inefficiency, improvements of privacy models are introduced.

II. CONCLUSION

Big data is analysed for bits of knowledge that leads to better decisions and strategic moves for overpowering businesses. Yet only a small percentage of data is actually analysed. In this paper, we have investigated the privacy challenges in big data by first identifying big data privacy requirements and then discussing whether existing privacy preserving techniques are enough for big data processing.

REFERENCES

[1]. Machanavajjhala, A., Kifer, D., Gehrke, J.andVenkitasubramaniam, M., 2007. *l*-diversity:Privacy beyond *k*-anonymity.*ACM Transactions onKnowledge Discovery from Data (TKDD)*, 1(1), p. 3.

[2]. Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on.IEEE, 2007.

[3]. A Cloud Securty Alliance Collaborative research, “Expanded Top Ten Big Data Security and Privacy challenges” , April 2013.
[4]. Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage Kaitai Liang, Willy Susilo, Senior Member, IEEE, and Joseph K. Liu 2015.
[5]. Privacy Preservation in the Age of Big Data : A Survey John S. Davis II, Osonde A. Osoba
[6]. Protection of Big Data Privacy IEEE access January 2016
[7]. Privacy Preservation in Big DataAugust 2014
[8]. L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, pp. 557–570, 2002.
[9]. Ton A, Saravanan M. Ericsson research. [Online]. http://www.ericsson.com/researchblog/data-knowledge/big-data-privacy-reservation/2015.
[10]. Qin Y, et al. When things matter: a survey on data-centric internet of things. J Netw Comp Appl. 2016; 64:137–53.
[11]. Fong S, Wong R, Vasilakos AV. Accelerated PSO swarm search feature selection for data stream mining big data. In: IEEE transactions on services computing, vol. 9, no. 1. 2016.
[12]. Middleton P, Kjeldsen P, Tully J. Forecast: the internet of things, worldwide. Stamford: Gartner; 2013.
[13]. Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-9804, SRI Computer Science Laboratory; 1998.
[14]. Feng Z, et al. TRAC: Truthful auction for location-aware collaborative sensing in mobile crowd sourcing INFOCOM. Piscataway: IEEE; 2014. p. 1231–39.
[15]. HessamZakerdah CC, Aggarwal KB. Privacy-preserving big data publishing. La Jolla: ACM; 2015.
[16]. Dean J, Ghemawat S. Map reduce: simplied data processing on large clusters. OSDI; 2004.
[17]. Lammel R. Google’s MapReduce programming model-revisited. Sci Comput Progr. 2008;70(1):1–30.