ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.5

Controlling Privacy by using Methodology and Techniques of Big Data

Prof.Rajeswari.C[1], Gaurav Johri[2], Naman Singh[3]
Professor[1], Student[2, 3]
Department of School of Information Technology and Engineering[1], Department of Master of Computer Applications[2, 3]
VIT, University, Vellore, Tamilnadu, India

Abstract:
The major the privacy preservation mechanisms in big data are presenting the challenges for existing mechanisms. This paper also presents recent techniques of privacy preserving in big data when we discuss about the big data we may assume processing large files or data which is unable to be processed by common machine because its storage capacity is not so efficient. Using big data we can correlate huge amount of data which may have which may be having various hidden datasets which needs to be bring out but processing such large amount of datasets needs such a system which may be having massive storage capacities. This paper will throw out light on securities and privacy concerns that are prime factors faced by the enterprise in today's time. This paper focus on the importance of security covered by the present methods i.e. HybrExi, k-anonymitys, T-closenesss and L-diversities and its current demands in the modern world. The major role of my technology is to resolve the challenges. The goal of this paper is to provide a like hiding a needle in a haystack, identity based anonymization, differential privacy, privacy preserving big data publishing and fast anonymization of big data streams. This paper refer privacy and security aspects healthcare in big data. Comparative study between various recent techniques of big data privacy is also done as well.

Keywords: Big data, Privacies and securities, Privacy preserving's: k-anonymities: T- closeness, L-diversities, HybrEx's, PPDP, FADS

I. INTRODUCTION:

Big data mainly relates to large data sets that are so huge or so much expansive that present data processing techniques are not sufficient. It's the big volume of data—both structured and unstructured—that initiates a business on a daily basis. Due to many recent technological developments, the volume of data generated by internet, social sites, sensored networks, health specialist applications, and many other companies, is drastically increasing day by day. All the drastic measures of data produced from various sources in multiple formats with very high speed [3] is referred as big data.

The term big data can be called as "a new generation of technologies and architectures, designed to economically extract value from very huge volumes of a wide clusters of data, by enabling large-velocity captures, inventions and analysis". On the premise of this definition, the properties of big data are propagated by 3V's, which are, volume, velocity and variety. Recent studies points out that the definition of 3Vs is difficult to explain the big data we face now. Thus, veracity, validity, value, variability, venue, vocabulary, and vagueness are added to make some complement explanation of big data.

A well known idea of big data is that the data are heterogeneous, i.e., they may contain text, audio, image, or video etc. This varied quality of data is signified by variety. In order to ensure big data privacy, different mechanisms have been made in recent years.

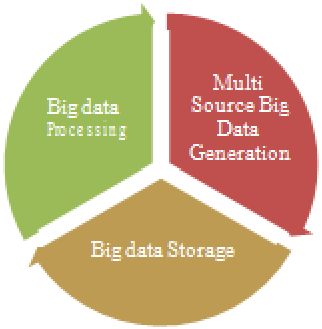

Figure.1.Bigdatalifecyclestagesofbigdatalifecycle,i.e.,datageneration,storage,andprocessingareshown

Privacy and security issues in big data

Privacy and security concerns: A Privacy and security concern in words of big data is an vital issue. Big data security model is not recommended in the event of complex applications due to which it gets unable by default. However, in its absence, data can always be compromised easily. As such, this section focuses on the privacy and security issues.

Privacy: Information privacy is the privilege to have some control of how the private information is collected and used. Information privacies are the capacities of an individuals or group to stop information about themselves from becoming known to people other than those they give the information to. One major user privacy issue is the knowing of personal information during transmission over the Internet

Security

Security is the practice of protecting informations and informations vital through the use of technology, processes and training from:-Unauthorizable access, Disclosure, Disruption, Modification, Inspection, Recording, and Destruction.

Privacy* vs. *security

Data privacy is concerned on the use and governing of private data—things like setting up policies in place to ensure that consumer' private informations is being gathered, shared and utilized in appropriate ways. Security focusses more on protecting data from malicious attacks and the misuse of stolen data for profit While security is fundamental for protecting data, it's not sufficient for addressing privacy. Table 1 focuses on additional difference between privacy and security. Privacy requirements in big data Big data analytics draw in various organizations; a hefty portion of them decide not to utilize these services because of the absence of standard security and privacy protection tools. These sections analyse possible strategies to upgrade big data platforms with the focuses on additional difference between privacy and security help of privacy protection capabilities. The foundations and development strategies of a framework that supports:

1. The specification of privacy policies managing the access to data stored into target big data platforms,
2. The generation of productive enforcement monitors for these policies, and
3. The integration of the generated monitors into the target analytics platforms. Enforcement techniques proposed for traditional use

Table.1. Difference between privacy and security

	Privacy	**Security**
1	Privacy is the appropriate use of user's information	ecurity is the "confidentiality, integrity and availability" of data
2	rivacy is the ability to decide what information of an individual goes where	ecurity offers the ability to be confident that decisions are respected
3	he issue of privacy is one that often applies to a consumer's right to safeguard their information from any other parties	ecurity may provide for confidentiality. The overall goal of most security system is to protect an enterprise or agency

Focuses on additional difference between privacy and security Help of privacy protection capabilities. The foundations and development strategies of a framework that supports:

1. The specification of privacy policies managing the access to data stored into target big data platforms,

.

2. The generation of productive enforcement monitors for these policies, and The integration of the generated monitors into the target analytics platforms

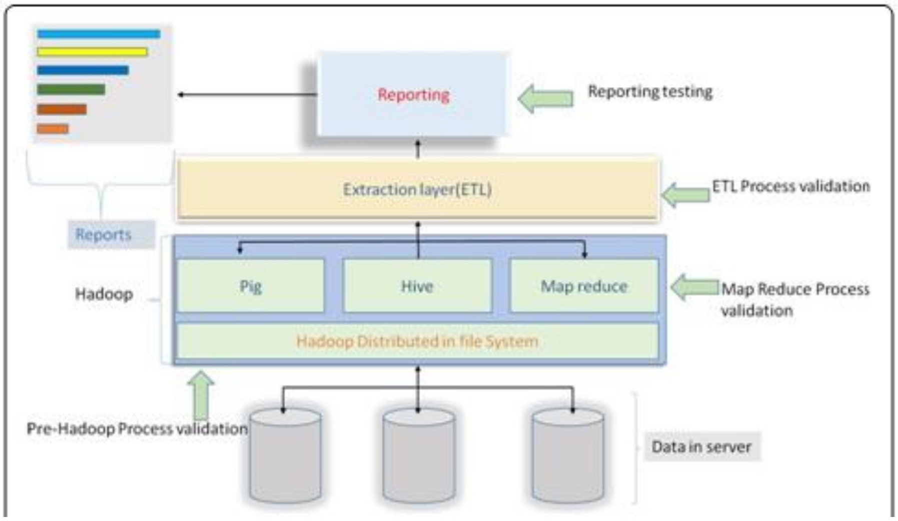

Figure.2. Big data architecture and testing are new paradigms for privacy on formance testing to the four areas of the ETL (Extract, Transform, and Load) processes are shown here

1. *Map-reduce process validation* this process changes big data assets to effectively react to a query. Privacy terms can tell the minimum number of returned records required to cover individual values, in addition to constraints on data sharing between various processes.

2. *ETL process validation* Similar to step (2), warehousing rationale should be confirmed at this step for compliance with privacy terms. Some data values may be aggregated anonymously or excluded in the warehouse if that indicates high probability of identifying individuals.

3. *Reports testing* reports are another form of questions, conceivably with higher visibility and wider audience. Privacy terms that characterize ‘purpose’ are fundamental to check that sensitive data is not reported with the exception of specified uses.

Big data privacy in data generation phase

Data generation can be classified into active data generation and passive data generation. By active data generation, we mean that the data owner will give the data to a third party , while passive data generation refers to the circumstances that the data are produced by data owner’s online actions (e.g., browsing) and the data owner may not know about that the data are being gathered by a third party. Minimization of the risk of privacy violation amid data generation by either restricting the access or by falsifying data.

1. ***Access restriction*****:** If the data owner thinks that the data may uncover sensitive information which is not supposed to be shared, it refuse to provide such data. If the data owner is giving the data passively, a few measures could be taken to ensure privacy, such as anti-tracking extensions, advertisement or script blockers and encryption tools.

- ***Falsifying data:*** In some circumstances, it is unrealistic to counteract access of sensitive data. In that case, data can be distorted using certain tools prior to the data gotten by some third party A tool Socket puppet is utilized to hide online identity of individual by deception.

- Certain security tools can be used to mask individual’s identity, such as Mask Me. This is especially useful when the data owner needs to give the credit card details amid online shopping.

Big data privacy in data storage phase

Storing high volume data is not a major challenge due to the advancement in data storage technologies, for example, the boom in cloud computing if the big data storage system is compromised; it can be exceptionally destructive as individuals’ personal information can be disclosed. In distributed environment, an application may need several datasets from various data centres and therefore confront the challenge of privacy protection. The conventional security mechanisms to protect data can be divided into four categories. They are file level data security schemes, database level data security schemes, media level security schemes and application level encryption schemes . Responding to the 3V’s nature of the big data analytics, the storage infrastructure ought to be scalable. Approaches to privacy preservation storage on cloud when data are stored on cloud, data security predominantly has three dimensions, confidentiality, integrity and availability. The first two are directly related to privacy of the data i.e., if data confidentiality or integrity is breached it will have a direct effect on users privacy. Availability of information refers to ensuring that authorized parties are able to access the information when needed. The approaches to safeguard the privacy of the user when data are stored on the cloud are as follows:

- *Attribute based encryption* Access control is based on the identity of a user complete access over all resources.

- *Homomorphism encryption* Can be deployed in IBE or ABE scheme settings updating cipher text receiver is possible.

- *Storage path encryption* It secures storage of big data on clouds.

- *Usage of Hybrid clouds* Hybrid cloud is a cloud computing environment which utilizes a blend of on-premises, private cloud and third-party, public cloud services with organization between the two platforms.

Integrity verification of big data storage

At the point when cloud computing is used for big data storage, data owner loses control over data. The outsourced data are at risk as cloud server may not be completely trusted. The data owner should be firmly convinced that the cloud is storing data properly according to the service level contract. To ensure privacy to the cloud user is to provide the system with the mechanism to allow data owner verify that his data stored on the cloud is intact.

Big data privacy preserving in data processing

Big data processing paradigm categorizes systems into batch, stream, graph, and machine learning processing . For privacy protection in data processing part, division can be done into two phases. In the first phase, the goal is to safeguard information from unsolicited disclosure since the collected data might contain sensitive information of the data owner. In the second phase, the aim is to extract meaningful information from the data without violating the privacy.

Privacy preserving methods in big data

Few traditional methods for privacy preserving in big data is described in brief here. These methods being used traditionally provide privacy to a certain amount but their demerits led to the advent of newer methods.

De- identification

De-identification [1] is a traditional technique for privacy-preserving data mining, where in order to protect individual privacy, data should be first sanitized with generalization (replacing quasi-identifiers with less particular but semantically consistent values) and suppression (not releasing some values at all) before the release for data mining. Mitigate the threats from re-identification; the concepts of k-anonymity , l-diversity and t-closeness have been introduced to enhance traditional privacy-preserving data mining..

- Privacy-preserving big data analytics is still challenging due to either the issues of flexibility along with effectiveness or the de-identification risks.

- De-identification is more feasible for privacy-preserving big data analytics if develop efficient privacy-preserving algorithms to help mitigate the risk of re-identification.

There are three -privacy-preserving methods of De-identification, namely, K-anonymity, L-diversity and T-closeness. There are some common terms used in the privacy field of these methods:

- *Identifier attributes* include information that uniquely and directly distinguish individuals such as full name, driver license, social security number.

- *Quasi-identifier attributes* means a set of information, for example, gender, age, date of birth, zip code. That can be combined with other external data in order to re-identify individuals.
- *Sensitive attributes* are private and personal information. Examples include, sickness, salary, etc.
- *Insensitive attributes* are the general and the innocuous information.
- *Equivalence classes* are sets of all records that consist of the same values on the quasi-identifiers.

K- anonymity

A release of data is said to have the *k*-anonymity 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. In the context of *k*-anonymization problems, a database is a table which consists of *n* rows and *m* columns, where each row of the table represents a record relating to a particular individual from a populace and the entries in the different rows need not be unique. Table 2 is a non-anonymized database comprising of the patient records of some fictitious hospital in Hyderabad. There are six attributes along with ten records in this data. There are two regular techniques for accomplishing *k*-anonymity for some value of *k*.

Table.2. A Non-anonymized database consisting of the patient records

Name	Age	Gender	State of domicile	Religion	Disease
Ramya	29	Female	Tamil Nadu	Hindu	Cancer
Yamini	24	Female	Andhra Pradesh	Hindu	Viral infection
Salini	28	Female	Tamil Nadu	Muslim	TB
Sunny	27	Male	Karnataka	Parsi	No illness
Joshna	24	Female	Andhra Pradesh	Christian	Heart-related
Badri	23	Male	Karnataka	Buddhist	TB

It is a non-anonymized database comprising of the patient records of some fictitious hospital in Hyderabad

1. ***Suppression*** In this method, certain values of the attributes are supplanted by an asterisk '*'. All or some of the values of a column may be replaced by '*'. In the anonymized Table 3, replaced all the values in the 'Name' attribute and each of the values in the 'Religion' attribute by a '*'.

2. ***Generalization*** In this method, individual values of attributes are replaced with a broader category. For instance, the value '19' of the attribute 'Age' may be supplanted by ' ≤20', the value '23' by '20 < age ≤ 30', etc.

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, and Recursive) 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

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) extends the *l*-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute. Comparative analysis of de‑identification privacy methods advanced data analytics can extricate valuable information from big data but at the same time it poses a big risk to the users' privacy. There have been numerous proposed approaches to preserve privacy before, during, and after analytics process on the big data. This paper discusses three privacy methods such as K-anonymity, L-diversity, and T-closeness..

HybrEx

Hybrid execution model is a model for confidentiality and privacy in cloud computing. It executes public clouds only for operations which are safe while integrating an organization's private cloud, i.e., it utilizes public clouds only for non-sensitive data and computation of an organization classified as public, whereas for an organization's sensitive, private, data and computation, the model utilizes their private cloud. It considers data sensitivity before a job's execution. It provides integration with safety.

Recent techniques of privacy preserving in big data

Differential privacy

Differential Privacy is a technology that provides researchers and database analysts a facility to obtain the useful information from the databases that contain personal information of people without revealing the personal identities of the individuals. This is done by introducing a minimum distraction in the information provided by the database system. In mid-90s when the Commonwealth of Massachusetts Group Insurance Commis-sion (GIC) released the anonymous health record of its clients for research to benefit the society [16]. GIC hides some information like name, street address etc. so as to protect their privacy. Latanya Sweeney (then a PhD student in MIT) using the publicly available voter database and database released by GIC, successfully identified the health record by just comparing and co-relating them. Thus hiding some information cannot assures the protection of individual identity. Differential Privacy (DP) deals to provide the solution to this problem as shown Fig. 4. In DP analyst are not provided the direct access to the database containing personal information. An intermediary piece of software is introduced between the database and the analyst to protect the privacy. This intermediary software is also called as the privacy guard.

Step 1 The analyst can make a query to the database through this intermediary privacy guard.

Step 2 The privacy guard takes the query from the analyst and evaluates this query and other earlier queries for the privacy risk. After evaluation of privacy risk.

Step 3 The privacy guard then gets the answer from the database.

Step 4 Add some distortion to it according to the evaluated privacy risk and finally provide it to the analyst.

The amount of distortion added to the pure data is proportional to the evaluated privacy risk. If the privacy risk is low, distortion added is small enough so that it do not affect the quality of answer, but large enough that they protect the individual privacy of database. But if the privacy risk is high then more distortion is added.

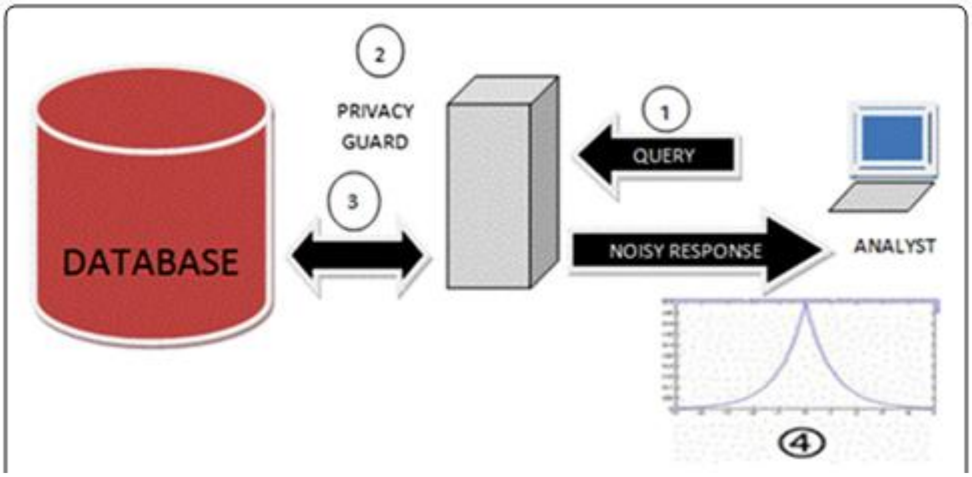

Figure.4. Differential privacy bigdata differential privacy (DP) as a solution to privacy-preserving in big data is shown

II. IDENTITY BASED ANONYMIZATION

These techniques encountered issues when successfully combined anonymization, privacy protection, and big data techniques to analyse usage data while protecting the identities of users. Intel Human Factors Engineering team wanted to use web page access logs and big data tools to enhance convenience of Intel's heavily used internal web portal.

III. REFERENCES

[1]. Abadi DJ, Carney D, Cetintemel U, Cherniack M, Convey C, Lee S, Stone-braker M, Tatbul N, ZdonikSB. Aurora: a new model and architecture for data stream management. VLDB J. 2003; 12(2):120–39.

[2]. Kolomvatsos K, Anagnostopoulos C, Hadjiefthymiades S. An efficient time optimized scheme for progressive analytics in big data. Big Data Res. 2015; 2(4):155–65.

[3]. Big data at the speed of business, [online]. http://www-01.ibm.com/soft-ware/data/bigdata/2012.

[4]. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A. Big data: the next frontier for innovation, competition, and productivity. New York: Mickensy Global Institute; 2011. p. 1–137.

[5]. Gantz J, Reinsel D. Extracting value from chaos. In: Proc on IDC IView. 2011. p. 1–12.

[6]. Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analytics: a survey. J Big Data Springer Open J. 2015.

[7]. Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. In: IEEE translations and content mining are permitted for academic research. 2016.

[8]. Jain P, Pathak N, Tapashetti P, Umesh AS. Privacy preserving processing of data decision tree based on sample selection and singular value decomposition. In: 39th international conference on information assurance and security (lAS). 2013.

[9]. Qin Y, et al. When things matter: a survey on data-centric internet of things. J Netw Comp Appl. 2016;64:137–53.

[10]. 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.

[11]. Middleton P, Kjeldsen P, Tully J. Forecast: the internet of things, worldwide. Stamford: Gartner; 2013.

[12]. Hu J, Vasilakos AV. Energy Big data analytics and security: challenges and opportunities. IEEE Trans Smart Grid. 2016; 7(5):2423–36.

[13]. Porambage P, et al. The quest for privacy in the internet of things. IEEE Cloud Comp. 2016;3(2):36–45.

[14]. Jing Q, et al. Security of the internet of things: perspectives and challenges. WirelNetw. 2014;20(8):2481–501.

[15]. Han J, Ishii M, Makino H. A hadoop performance model for multi-rack clusters. In: IEEE 5th international conference on computer science and information technology (CSIT). 2013. p. 265–74.

[16]. Gudipati M, Rao S, Mohan ND, Gajja NK. Big data: testing approach to overcome quality challenges. Data Eng. 2012:23–31.

[17]. Xu L, Jiang C, Wang J, Yuan J, Ren Y. Information security in big data: privacy and data mining. IEEE Access.
2014; 2: 1149–76.

[18]. Liu S. Exploring the future of computing. IT Prof. 2011;15(1):2–3.

[19]. Sokolova M, Matwin S. Personal privacy protection in time of big data. Berlin: Springer; 2015.

[20]. Cheng H, Rong C, Hwang K, Wang W, Li Y. Secure big data storage and sharing scheme for cloud tenants. China Commun. 2015; 12(6):106–15.