IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

An Exhaustive Review of the Privacy Preservation and Security Mechanisms in Big Data Life Cycle

Manjula GS[1*], T. Meyyappan[2]

[1,2]JP Morgan Services India Pvt., Ltd Bangalore – 560001,Karnataka (Industrialist)

[*]*Corresponding Author: gsmanjula@yahoo.co.in, Contact: 9845957942*

Available online at: www.ijcseonline.org

Abstract— As there is an exponential growth of data in every field of life, the assessment and extraction of data from the massive data sets has derived as a dreadful challenge in this golden era of big data. Conventional security methods cannot be adapted to big data due to its massive volume, and range. Undoubtedly, mining fruitful information from this massive data has been a universal interest for the organizations having large dataset. Big data life cycle includes three phases such as data generation, data storage, and data processing. In big data process, distributed systems are adapted since it needs large storage and high computational power. As many parties are engaged in these systems, the possibility of the violation in security concerns increases. In addition, since big data contains individual's personal information, privacy is the foremost security concern. The main objective of this paper is to present an exhaustive overview of the privacy preservation mechanisms in big data life cycle. The modern privacy-preserving methods such as the generalization are capable of effectively managing the privacy assaults on a sole data set, whereas the protection of privacy for multiple data sets continues to be hard. Therefore, with intention of conserving the secrecy of multiple data sets, it is desirable to initially anonymize whole data sets and thereafter encrypt them before amassing or exchanging them in cloud. The challenges in existing mechanisms and eventual research discussions relevant to privacy preservation in big data are mentioned. The security techniques to protect the data set from being accessed by illegal users are also discussed.

Keywords— Big Data, Conventional, Data Privacy, Mining, Security

I. INTRODUCTION

The quantum of the database is being increased due to the possibility of powerful and economical database systems. This fiery growth in data and databases has provoked a demand for new mechanisms and approaches. It can intelligently and spontaneously re-establish the processed data into fruitful information. Therefore, the big data mining has emerged as an important research area to transform abrupt quantities of data into meaningful knowledge. Today, big data mining is used in various fields, such as business, market analysis, forensic investigation, healthcare, criminology, and several other are fitting more attractive with time. There are a lot of concerns like data types, distinct sources of data, usability, capability and scalability of big data mining processes preserving of privacy, data security, etc., to be examined while scheming an effective big data mining technique All the ocean of data generated from various sources in distinct formats with very high speed is called as big data. Privacy preserving data mining (PPDM) for big data mainly focuses on two targets namely meeting privacy requirements and providing valid data mining results respectively [2].

Generally, privacy preservation is needed in personal information of the user (individual privacy) and information regarding the collective work of the users (collective privacy preservation). In individual privacy preservation, the fundamental target of data privacy is the preservation of personally trackable information. Information is considered personally trackable if it can be linked, directly/ indirectly, to an individual. Hence, while mining personal data, the attributes related with individuals are private and essentially be unveiled. Accordingly, miners are able to study from global models rather than from the aspects of an individual. Individual privacy preservation may not be sufficient. In collective privacy preservation, there is a requirement to preserve across learning sensitive knowledge depicting the collective activities of a group. The target here is security control systems while prevent unveiling of confidential information about individual persons at the time of furnishing cumulative information about groups as similar to statistical databases. Despite, unlike statistical databases, another target of collective privacy preservation is to protect crucial patterns which are predominant for strategic decisions. Thus, the goal of collective privacy preservation is to protect personally trackable information without the need for priori hypotheses and also some patterns and trends at the same time. As a field, it has admitted the new perceptions and methods like association rule learning and it has also been used in machine-learning algorithms like inductive-rule learning, in decision trees to the framework where abundant databases are involved [3]. In recent years, the growing need of the information management has led to an outburst interest in new big data mining mechanisms. Big data Mining is one of the phases in Knowledge Discovery in Databases process (KDD) process and both are entirely different from each other [4]. The KDD process implies the common procedure to resolve desirable knowledge from abundant data; while big data mining indicates the process of extracting interesting patterns from data depending on analysis. For processing the big data, one of the frameworks of Hadoop called Hadoop Distributed File system (HDFS) is used. Hadoop supports a distributed file structure and a frame-work for the investigating and transforming gigantic data sets using the Map Reduce archetype. Hadoop interfacing with HDFS supports the distributed storage of the distinct files available. Hadoop is intended to accomplish data partitioning and computation against several number of hosts, and executes computations in parallel close to their data [5]. This paper presents an exhaustive overview of the privacy preservation mechanisms in big data life cycle. This paper also provides the challenges in existing mechanisms and eventual research discussions relevant to privacy preservation in big data. Also, the security techniques to protect the data set from being accessed by illegal users are also discussed.

II.BIG DATA MINING AND ITS LIFE CYCLE

There is a efficient and effective mechanisms to process various dimensions of big data with respect to volume, velocity, speed and variety, arriving from different sources. Big data has multiple phases in life cycle, as shown in Fig. 1. Today, data are distributed and new technologies such as are cloud computing (Hadoop Map Reduce) developed to store and process large data warehouse.

Fig.1 Big data life cycle

Data can be generated from many distributed sources. In the past few years, there is a flare-up (90%) in the amount of data generated by humans and machines. Daily 2.5 quintillion bytes of data are generated on the web and 25TB of new data are generated on the facebook alone [6]. Due to the large, diverse and complex generation in data with a specific domain such as business, Internet, research, health care etc., the handling of data is hard for the traditional systems. Data storing and managing massive data sets takes place under data storage. Data storage has two elements namely hardware infrastructure and data management. Hardware infrastructure utilizes information and communications technology (ICT) assets for various distributed storage related tasks. Data management specifies the set of software positioned on top of hardware infrastructure to manage and query big data sets and also provides several interfaces to e communicate with and analyze stored data.

Data processing indicates primarily to the data collection, data transmission, pre-processing processes and thereby extracting useful information. Since data are from diverse sources i.e., sites that has text/ images/ videos, they are acquired with the help of dedicated data collection technology in the data collection phase. Then they are transmitted into a proper storage with the help of high speed transmission mechanism for its future use in various types of analytic applications in the data transmission phase. Finally, the meaningless and redundant parts of the data are removed in pre-processing phase to save the storage space. The massive data and discipline specific analytical methods are adapted by various applications to acquire fruitful information. The technical areas of classification under the budding data analytics are structured data analytics, text analytics, multimedia analytics, web analytics, network analytics, and mobile analytics.

III.SECURE MINING TECHNIQUES

A. Randomized Response (RR)

Randomized response (RR) technique can be used when sensitive questions are asked to respondents and they are hesitant to answer directly (Warner 1965). Examples of sensitive questions are alcohol consumption, sexual behavior and fraud. RR variables are misclassified absolute variables that are determined by knowing the conditional misclassification probabilities[7].The misclassification prevents the respondent’s privacy. The general scheme of RR is shown in Fig.2.

Fig.2 The general scheme of RR

In general, let $\dot{X}$ be the binary RR variable that models the latent status, X+ the binary variable that models the observed answer, and 'yes' is equal to 1and 'no' is equal to 2. Given the forced response design, the distribution of $\dot{X}$ is the 2-component finite mixture given by

$$IP(\dot{X} = \dot{x}) = \sum_{k=1}^{2} IP(\dot{X} = \dot{x}|X = k) IP\ (X = k) \quad (1)$$

Where $\dot{X} \in \{1,2\}$. The conditional probabilities $p_{jk} = IP(\dot{X} = j|X = k)\ for j, \in \{1,2\}$ are fixed by the forced response design and the known distribution of the sum of the two dice. The equation (1) represents that RR variables can be found as misclassified variables. The transition matrix of X that has the conditional misclassification probabilities p_{ik} for j, k∈{1,2} is given by

$$P_X = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} \quad (2)$$

Similar computations hold for more than two RR variables/ RR variables with more than two categories.

B. Sequential Pattern Hiding (SPH)

Sharing of sequence data are increasing nowadays that enable mining in different applications and domains such as business, market analysis, forensic investigation, healthcare and criminology which are exposed to sensitive sequential patterns. This causes to infer about individuals/ leak sensitive information about organizations. This hazardous threat can be prevented by sequential pattern hiding method (SPH) which necessarily conceals sensitive patterns extracted from published data, while the data and the non-sensitive patterns are maintained as it is [8]. In SPH, the set of selected frequent non-sensitive patterns called side-effects are replaced in the place of sensitive patterns, thus retaining data utility with minimum distortion and side-effects and also avoids implausible symbol orderings existing in certain applications. SPH significantly permits more exact data analysis than the state-of-the-art method and it seems to be a challenging problem, as sequences have more complex syntactic than item sets, and calls for powerful solutions that offer high utility.

C.*Secure Multiparty Computation (SMC)*

Secure Multiparty Computation protects the sensitive data effectively by considering a set of parties who wish to mine their data in collaborate but does not want to reveal their datasets to each other. Thus, the distributed problem of PPDM is solved using cryptographic techniques by reducing it to the secure computation with respect to the distributed inputs. SMC produces the results of data mining with each party, having some of the private data join in a protocol, but not revealing data to other parties that are not already having them and thus not causing breaches of privacy. Therefore SMC is a computation protocol which simulates the input and output of the party and finally no party involved have knowledge about anything else except it's own inputs and the results. SMC focuses on three models of security namely the semi-honest model which assumes that each party follows the rule of the protocol, the malicious model which assumes that parties can cheat arbitrarily without compromising either security/ the results (exact results of the malicious party/ detect the malicious party) and an intermediate model which preserves privacy with non-colluding parties [9].

D. Cryptographic technique

The prime goal of cryptographic techniques is to implement access control to data stored in basically unrecognized repositories. Here, the authorized users are given access to the data they want at the same time the unauthorized users (outsiders/ malicious insiders) are ensured of non-accessing of sensitive data. The initial step in this process is to encrypt the data to hide it from unauthorized users. However, the problem of giving access to the authorized users is solved by group keying technique/ attribute-based encryption technique. The group keying technique/ broadcast encryption technique manages this through deliberate key management to ensure authorized users have the mandatory keys to decrypt the data. Next, the attribute-based encryption (ABE) depends on more powerful cryptographic techniques to automatically invoke the approval access [10].

IV.SUPPORTING AND PROVIDING PRIVACY PRESERVATION

PPDM secures data on performing data mining. Confidentiality is one of the crucial problems that emanates in any of the huge collection of data. The requirement for privacy is very much important due to the storage of important and confidential data that can be provoked by business concerns. For handling big data, the efficient methods that reduce the risk of mishandling of the data must be considered [11]. One of the novel methods which provide security to the data by partitioning and storing them separately is the Hadoop technology. Hadoop technology provides security by storing the data in distributed style in HDFS file system. In some other approach, there is some procedure of Authentication, Authorization and Accounting (AAA) is used for PPDM. Most of the techniques adopt the cryptographic technique by performing some plan of alteration on the original data without compromising mining for attaining privacy preservation. The techniques of privacy preservation is given below.

A. Hadoop Distributed File System (HDFS)

HDFS is an open source version of the implementation of the Google File System which stores, processes and analyzes both local and distributed data. HDFS performs data storage activity and it uses replication behaviour and scalability for fault tolerance and availability. HDFS consists of two important nodes namely data node and name node. The original file are split into many units and stored across the full cluster in the data nodes. And the name node has the namespace tree and the actual data node is blocked by the mapping of the namespace tree. Each Data Node is replicated into number of times and the Name Node retains the particulars of the replicated Data Nodes for the availability and fault tolerance reasons. Multiple application queries are executed in parallel by each of the data node [12].

B. The Security by Authentication, Authorization, and Accounting (AAA)

AAA is a structural framework that configures a set of three separate security functions in a rational manner

[13]. AAA provides a flexible way of performing the subsequent services:

Authentication—involves the method by considering identification of users, with login and password dialog, challenge and response, messaging, and, depending on the selected security protocol, encryption. Authentication identifies the user prior to accessing the network and network services. AAA authentication is configured by specifying a list of authentication schemes, and then implementing that list to several interfaces. The list of authentication schemes except default list describes the authentication types to be executed and the sequence of execution to be applied to a particular interface before performing any of the specified authentication schemes. If no other method list is defined, the default scheme list is automatically adapted to all interfaces. A defined scheme list disallows the default scheme list.

Authorization— involves the method by considering remote access control, along with one-time authorization/ authorization for particular service, individual user account list and profile, collaborators support, and IP and Telnet support. AAA authorization operates by gathering a set of attributes that illustrates what the user is authorized to execute. The assembled set of attributes is then compared to the database information for a particular user and the result is sent back to AAA that determines the substantial capabilities and restrictions of the particular user. The database can be located locally on the device/ access server/ it can be hosted on remote security servers (RADIUS or TACACS+). In remote security servers the users are authorized for specific rights with the help of attribute-value (AV) pairs. Like authentication, AAA authorization is configured by defining a list of authorization schemes, and then applying the defined list to various interfaces.

Accounting— involves the method for collecting and sending security server information used for auditing, client billing and reporting, namely user identities, start and stop times. Accounting enables users to track the services for accessing along with the consumed network resources. After activating AAA accounting, the network access server reports user functioning to the remote security server (RADIUS or TACACS+) which depends on implemented security method in the form of accounting records. Every accounting record comprises an accounting AV pairs and is kept on the access control server and can then be analyzed for client billing, auditing and network management. Like authentication and authorization, AAA accounting is configured by defining a list of accounting methods, and then applying the defined list to various interfaces [14].

C. Anonymizing Data Sets

In many data mining enterprises, access to huge amounts of personal data is necessary for inferences to be implicated. One way for preserving privacy in this section it to suppress a few sensitive data values [15] known as a k-anonymity model which was scheduled by Samarati and Sweeney. Assuming a table having n tuples and m attributes and z an integer with > 1, then the table is modified by suppressing the values of absolute cells in the table. The aim is to reduce the number of suppressed cells while validating that for each tuple in the modified table there are z-1 other tuples in the modified table identical to it. The solution for optimizing k-anonymized table for a given table instance can be found to be NP-hard even for binary attributes. There are yet O(z) approximation algorithm examined in for solving this problem and also proved to terminate.

D. Decision Tree Mining

The algorithm is designed is when two parties with database DB1 and DB2 intend to apply the decision tree algorithm on the joint database DB1 U DB2 without disclosing any unwanted information about their database. This technique is one of the semi honest adversary model attempts to minimize the number of bits communicated between the two parties using secure multi party computation. The conventional ID3 algorithm executes a decision tree mining by selecting the best attribute using information gain theory at each tree level to split it and hence the data is partition. The building of tree will end when the data in distinctively partitioned into a single class/ no attributes to split on. The best attribute is selected to minimize the entropy of the partitions and thereby maximizing the information gain. In the design of PPDM, the information gain for every attribute is calculated jointly over all the database cases without disclosing individual data. This solution reduces to privately computing x ln x in a protocol which receives x1 and x2 as input where x1 + x2 = x [16].

E. Association Rule Mining

The association rule mining in privacy preserving is described for a horizontally partitioned data set where the Transactions are distributed across n sites. Let L = $\{l, l_2,....l_n\}$1be an item set and Z = $\{Z_1, Z_2,....Z_n\}$ Be the transactions set where each $Z_i \subseteq L$. A transaction Z_i contains an item set $X \subseteq L$ only if $X \subseteq Z_i$. An association rule connotation is of the form $X \Rightarrow Y$ $(X \cap Y = 0)$ with support δ and confidence ζ i f δ percentage of the transactions in Z contains $X \cup Y$ and ζ percentage of transactions that include X also includesY. The global δ count of an item set is the sum of all local δ counts. The global ζ of a rule can be expressed in terms of the global δ:

$$\delta_g(X) = \sum_{i=1}^{n} \delta_i(X)$$

$$\zeta(X \to Y) = \delta_g(X \cup Y)/\delta_g(X)$$

The aim of the association rule mining is to discover all rules with global δ and global ζ higher than the user specified minimum δ and ζ [17]. The steps viz., candidate set generation, local pruning, itemset exchange and support count are utilized in this algorithm. Candidate set generation use the Apiori algorithm to get the candidate k item sets. For each X in the local candidate set, local pruning scan the local database to compute the support of X Item set exchange computes a secure union of the large item sets over all sites. Support Count computes a secure sum of the local supports to get the global support.

F. EM Clustering

The algorithm groups' data into clusters based on the attributes values. EM algorithm is the familiar one in clustering which works on both discrete and continuous attributes. The algorithm for privacy preservation is described for a horizontally partitioned data set. Assuming the data is 1-D with only one attribute x and are partitioned across c sites. Each site has data items ($n = \sum_{l=1}^{c} n_l$). Let $y_{ij}^{(t)}$ denote the cluster membership for the ith cluster for the jth data point at the (t)th EM round. In the B step, the values of mean for cluster i (μ_i), variance for cluster i (σ_i^2) and estimate of proportion of items i (π_i) are calculated using the following equations (5-7):

$$\sum_{j=1}^{n} y_{ij}^{(t)} x_j = \sum_{j=1}^{c} \quad (5)$$

$$\sum_{j=1}^{n} y_{ij}^{(t)} = \sum \quad (6)$$

$$\sum_{j=1}^{n} y_{ij}^{(t)} (x_j - \mu_i^{(t+1)})^2 = \sum_{l=1}^{c} \sum_{j=1}^{n_l} y_{ijl}^{(t)} (x_j - \mu_i^{(t+1)})^2 \quad (7)$$

The second part of the sum in all these equations is local to every site. It is convenient to find that sharing this value does not disclose x_i to the other sites without necessarily sharing , and the inner sum values. In the C step, the y values can be partitioned and computed locally given the global μ_i , σ_i^2 and π_i without involving any data sharing across sites [18].

G. Frequency Mining

In frequency mining, the n customers $C_1, C_2,....C_n$ are assumed and each customer has a Boolean value f_i . The algorithm finds out the total number of 1s and 0s without knowing the customer values. In other words, the problem computes the sum $\sum_{i=1}^{n} f_i$ without disclosing each f_i . Here, each customer is allowed to send only one communication to the miner without any further interaction and there is no communication between the customers. Because of these restrictions, the secure sum protocol cannot be used. The technique [8] uses the additively homomorphic property of a variant of the ElGamal encryption which is described as below:

Let D be a group in which discrete logarithm is hard and let d be a generator in D. Each customer C_i has two pairs of private/public key pair $(a_i, A_i = d^{a_i})$ and $(b_i, B_i = d^{b_i})$. The sum $A = \sum_{i=1}^{n} A_i$ and $B = \sum_{i=1}^{n} B_i$, along with D and the generator d is known to everyone. Each customer sends to the miner the two values $l_i = d^{f_i} . A^{b_i}$ and $m_i = B^{a_i}$. The miner computes $t = \prod_{i=1}^{n} \frac{l_i}{m_i}$. The value of f for which $d^g = t$ represents the sum $\sum_{i=1}^{n} f_i$. Since 0 < f < n, this is easy to find by encryption and comparison. All the keys are assumed to be properly distributed properly. When the protocol begins, the frequency mining protocol prevents each user's privacy against the miner and up to (n-2) corrupted users [19].

H. Naïve Bayes Classifier

Naïve Bayes classifier simplifies the learning assignment by assuming the attributes are independent in the given class. They find their application in practical applications like text classification and medical diagnosis. Naïve Bayes classification can be defined as follows. Let $X_1, X, \ldots. X_n$ be n attributes and U be the class attribute. Let each attribute X_i have a domain $\{x_i^1, x_i^2, \ldots, x_i^d\}$ and class attribute U has a domain $\{u^1, u^2, \ldots, u^d\}$. The data point for the classifier looks like $(x_{j1}, x_{j2}, \ldots. x_{jn,} u_j)$. Given a new instance $(x_{j1}, x_{j2}, \ldots. x_{jn,})$, the most likely class can be found using the equation:

$$u = \text{argmax } P(u^l) \prod, \qquad (8)$$

This can be written is terms on number of occurrence € as:

$$u = \text{argmax } €(\iota, \qquad (9)$$

The privacy preserving naïve bayes learner accurately learns the Naïve Bayes classifier, but the miner learns nothing about each user's sensitive data except the knowledge derived from the classifier itself. To learn the classifier, all the miner need to learn and the pair are denoted by a Boolean value, the frequency mining technique is used to compute the Naïve Bayes model with the privacy constraints [20].

V. PRIVACY PRESERVATION FOR BIG DATA

Privacy in big data is a serious concern which brings into notice the demand for efficient privacy preservation techniques [21]. In this section, the three privacy preservation methods namely data anonymization, notice and consent and differential privacy are discussed. Also the application of these methods to big data and their limitations when applying these methods to big data are also discussed.

A. Data anonymization

Data anonymization is the method of changing data that will be used/ published in a way that protects the recognition of key information referred to as data re-identification. In this method crucial pieces of confidential data are concealed in a method that maintains privacy of the data. Organizations delivers data publically by anonymizing it. Here, Anonymization denotes hiding identifier attributes namely full name, license number, voter id etc. The main issue with data anonymization is that data looks anonymous but re-identification can be executed conveniently by associating it to other external data [22] called quasi identifier attributes. Re-identification of anonymous medical records can be executed using external voter list data and the attributes like gender, date of birth, zip code can be combined with external voter list data to re-identify individual users. As the size and variety of big data increases, the chance of re-identification also increases. This is the limitation of data anonymization in big data privacy. The following three methods (K-Anonymity, L-Diversity and T – Closeness) show the privacy preservation based on data anonymization.

K-Anonymity

A dataset is said to be k-anonymized if for a given tuple with its own attributes in the dataset there are at least k-1 other records that match those attributes [23]. K-anonymity is achieved by applying suppression and generalization. In suppression, quasi identifiers are replaced/ obscured by some constant values while in generalization, quasi identifiers are replaced by more general values. As an example, Table 1 shows the voter id and name as the identifier attributes and age, date of birth (DOB), and city as quasi identifiers with income as a sensitive attribute. Table 2 shows the suppression of age attribute and Table 3 shows 2-anonymized version of table 2. K-anonymous data is sensitive to unsorted matching, temporal and complementary release attacks.

Table 1: Base Dataset

Age	Sex	City	Income
25	M	Delhi	80,000
25	M	Gurgaon	45,000
25	M	Gurgaon	18,000
27	M	Delhi	32,000
27	F	Delhi	25,000
27	F	Delhi	34,000
33	F	Gurgaon	12,000
33	F	Gurgaon	26,000
33	M	Delhi	20,000

Table 2: Anonymous Dataset

Age	Sex	City	Income
25	M	Delhi	80,000
25	M	Gurgaon	45,000

25	M	Gurgaon	18,000
27	M	Delhi	32,000
27	F	Delhi	25,000
27	F	Delhi	34,000
33	F	Gurgaon	12,000
33	F	Gurgaon	26,000
33	M	Delhi	20,000

Table 3: Anonymized Dataset

Age	Sex	City	Income
25	M	Delhi	80,000
25	M	Gurgaon	45,000
25	M	Gurgaon	18,000
27	M	Delhi	32,000
27	F	Delhi	25,000
27	F	Delhi	34,000
33	F	Gurgaon	12,000
33	F	Gurgaon	26,000
33	M	Delhi	20,000

L-Diversity

L-diversity technique attempts to bring dissimilarity in the sensitive attribute of data and it ensures that each equivalence class of quasi identifiers has atleast L different values of sensitive attribute [24]. The technique depends upon the range of sensitive attribute. If there is a need to make data L diverse whereas sensitive attribute has less than L different values, fictitious data is to be added to enhance the security but result in problems during analysis. Also L-diversity method is liable to skewness and similarity attack and thus can't protect attribute revealing .

T – Closeness

A class has t-closeness if the distance between the distribution of a sensitive attribute in the class and whole table is not greater than a threshold T. A table has t-closeness if all equivalence classes has t-closeness [25]. The merit of t-closeness is that it protect attribute revealing.

B. Notice and consent

Notice and Consent is the most common privacy preservation method for web services [26]. When the user accesses a new application/ service, a notice specifying privacy concerns is displayed every time. The consumer must authorize the notice before using the application/ service. This method permits the user to assure his privacy rights thereby putting the burden on the individual user. This method bears numerous challenges when applying it to big data. In most of the problems, uses of big data are unknown at the time of specifying notice and consent. This needs the notice to alter every time big data is used for a distinct purpose. Also big data is collected and processed quickly that it builds burden on consumers to consent the notice.

C. Differential privacy

Differential Privacy method enables to extract useful answers from databases containing personal information while offering strong individual privacy preservation [27]. It aims to reduce the chances of individual identification while querying the data. The differential privacy method is shown in fig. 3.

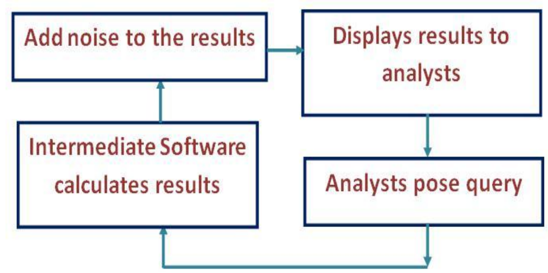

Fig. 3. Differential privacy process

Unlike data anonymization, data is not altered in differential privacy and individual users do not have direct access to the database. The results are calculated by an interface adds suitable inaccuracies. Unlike data anonymization, the original data set is not altered and suppression or generalization is not needed, distortion is added to the results (extract hidden value) based on the data type, type of questions etc., in the case of differential privacy.

VI.CONCLUSION

As there is an exponential growth in data everyday and it is impossible to conceive the next generation applications without producing and executing data driven processes. In this paper, a comprehensive survey on the privacy issues is conducted when dealing with big data. The privacy challenges in big data life cycle are discussed in the context of big data applications. The security techniques to protect the data set from being accessed by illegal users are also discussed.

Although several research works have been performed to preserve the user's privacy from data generation to data processing, several open issues and challenges still prevails.

REFERENCES

[1] NasrinIrshadHussain, BharadwajChoudhury, SandipRakshit, ―A Novel Method for Preserving Privacy in Big-Data Mining, International Journal of Computer Applications (0975 – 8887), Vol. 103, No. 16, pp. 22-25, October 2014.

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers, "Big data: The next frontier for innovation, competition, and productivity," Mickensy Global Institute , pp. 1–137, Jun. 2011.

[3] B. Matturdi, X. Zhou, S, Li, and F. Lin, "Big data security and privacy: A review," China Communications, vol. 11, no. 14, pp. 135–145, Apr. 2014.

[4] P.Samarati, "Protecting respondent's privacy in micro data release," In IEEE Transaction on knowledge and Data Engineering, pp.010-027, 2001

[5] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: Privacy and data mining," in IEEE Access, vol. 2, pp. 1149–1176, Oct. 2014.

[6] Chen, M.; Mao, S.; Liu, Y. Big data: A survey. Mob. Netw. Appl, Vol.19, pp. 171–209, 2014.

[7] Wenliang Du and Zhijun Zhan, "Using Randomized Response Tech-niques for Privacy-Preserving Data Mining," SIGKDD '03, August 24-27, 2003, Washington, DC, USA.

[8] ArisGkoulalas-Divanis, &GrigoriosLoukides, "Revisiting Sequential Pattern Hiding to Enhance Utility", ACM, August 2011.

[9] Yehuda Lindell, Benny Pinkas, ―Secure Multiparty Computation for Privacy-Preserving Data Mining, The Journal of Privacy and Confiden-tiality, Vol. 1, no. 1,, pp. 1-39, 2009.

[10] B. Pinkas, ―Cryptographic techniques for privacy-preserving data mining, SIGKDD Explore, Vol. 4, no. 2, pp. 12-19, 2002.

[11] Pingshui WANG, Survey on Privacy Preserving Data Mining, International Journal of Digital Content

[12] 20 essential Hadoop tools for crunching Big Data [Online] availa-ble:http://bigdata-madesimple.com/20-essential-hadoop-tools-for-crunching-big-data/

[13] Akhil Mittal, "Trustworthiness of Big Data," International Journal of Computer Applications (0975 – 8887), Vol. 80, no.9, October, 2013.

[14] Boris Glavic,"Big Data Provenance: Challenges and Implications for Benchmarking", Specifying Big Data Benchmarks,Vol. 8163, Springer Berlin Heidelberg, 2014.

[15] MogreNeha V., &PatilSulbha,"Slicing: An Approach for Privacy Preservation in High-Dimensional Data Using Anonymization Technique,"IRAJ International Conference, Pune 2013

[16] K.Anbazhagan, Dr.R.Sugumar, M.Mahendran, R.Natarajan, "An Efficient Approach for Statistical Anonymization Techniques for Privacy Preserving Data Mining,"International Journal of Advanced Re-search in Computer and Communication Engineering, Vol. 1, no. 7, pp. 482-485, September 2012.

[17] Y.H.Wu. C.Chiang and A.L.P.Chen. "Hiding Sensitive Association Rules with Limited Side Effects", IEEE Transaction on Knowledge and Data Engineering, Vol.19, no.1, pp 29-42, 2007

[18] GarimaSehgal, Dr. KanwalGarg "Comparison of Various Clustering Algorithms" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5, no. 3, pp. 3074-3076, 2014.

[19] M.-Y. Lin, P.-Y. Lee, and S.-C. "Hsueh. Apriori-based frequent itemset mining algorithms on MapReduce,"In Proc. ICUIMC, pages 26–30.ACM, 2012.

[20] Singh, Pravesh Kumar, and MohdShahid Husain. "Books Reviews using Naıve Bayes and Clustering lassifier." Second International Conference on Emerging Research in Computing, Information, Communication and Applications " (ERCICA-2014), pp. 886-891, 2014.

[21] J. Domingo-Ferrer, D. Sanchez, and J. Soria-Comas. "Database anonymization: privacy models, data tility, and microaggregation-based inter-model connections,". Morgan & Claypool, 2016

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

[23] Khaled El Emam,&Fida Kamal Dankar, Protecting Privacy Using k-Anonymity, J Am Med Inform Assoc. Vol. 15, no. 5, pp. 627–637, 2008.

[24] 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.

[25] Meraj Fatima , Dr. T.K.ShaikShavali , G. Kumar, "Strict Privacy with Enhanced Utility Preservation by T-Closeness Through Microaggregation,"Interrnational Journal of Advanced Technology and Innovative Research, Vol.08, no.15, pp. 3026-3035, October-2016.

[26] 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.

[27] [27]Friedman A, Schuster A. Data mining with differential privacy. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA, pp.25–28, July 2010.

