ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Big Data Privacy Preservation and Security Protection

Dhanashri Varute[1], Dr. Arti Mohanpurkar[2]

Department of Computer Engineering, D.Y.Patil's School of Engineering and Technology, Lohegaon, India

HOD, Department of Computer Engineering, D.Y.Patil's School of Engineering and Technology, Lohegaon, India

ABSTRACT: Big data is a term utilized for huge informational indexes that have more changed and complex structure. These attributes normally correspond with extra troubles in storing, examining and applying further methods or extricating outcomes. Notwithstanding, there is a conspicuous inconsistency between the security and privacy of big data and across the utilization of big data. This paper concentrates on privacy and security worries in big data. There has been various security safeguarding systems created for security insurance at various stages (for instance, data storage, data generation, and data processing) of a big data life cycle. The objective of this paper is to give the privacy protection and security in big data.

KEYWORDS: Big data, Privacy and security, Fingerprint, Relational Database, Data Partitioning, Slicing, Generalization, SQL Injection.

I. INTRODUCTION

Today, the improvement of informatization and systems administration prompt hazardous development of data. As per measurements, 2 million clients are utilizing Google's web search tool in consistently, Facebook clients share 4 billion assets consistently, Twitter prepare 340 milliontweets consistently, At the same time, the vast measure of data are delivered ceaselessly in logical estimation, restorative administrations, fund, retailing. 8ZB data will be generated in 2015. The diary Science distributed a comparable issue about data handling. More exercises were completed in IT industry. The reuse of big data was centered on continuously, the potential estimation of big data are mining. At introduce, the improvement of big data still faces numerous issues, Security and privacy issues is one of the key issues that individuals recognized widely, currently, people's each word and activity on the web are recorded by businesses, including shopping propensities, companions contact circumstance, perusing propensities, seeking propensities, etc. Number of cases demonstrates that even after a substantial number of safe data is gathered, individual privacy will be uncovered. Truth be told, the security ramifications of big data are more widely, the danger individuals confronted, is not constrained to break of individual privacy, as other information, big data is confronting numerous security dangers amid capacity, handling, transmission, and so forth and it needs data security and privacy assurance. In any case, data security and privacy insurance in big data period is more troublesome than in the past (such as data security in distributed computing, and so on.). In the distributed computing, the specialist co-ops can control the capacity and operation of data. Nonetheless, clients still have some approach to ensure their own data, for example, data stockpiling and security figuring through specialized methods for cryptography, or operational natural security through trusted processing mode. And with regards to big data, many organizations is a maker of data, as well as store chiefs and clients of data, Therefore, only by simple specialized intends to restrict the organizations to utilize client's information, user privacy assurance is to a great degree troublesome. Right now, numerous associations have understood the security issues of big data, and make a move to focus on big data security issues .In 2012, the cloud security organization together (CSA) shaped a big data working gathering, went for discovering answers for data focus security and privacy issues. In this paper, in view of checking the exploration circumstance of big data, analyzes the security difficulties to big data, talks about the key innovation ofthe current big data security and privacy insurance.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Privacy and security concerns:

Privacy and security as far as big data is a critical issue. Big data security display is not recommended in case of complex applications because of which it gets crippled as a matter of course. Be that as it may, in its nonattendance, data can simply be traded off effectively. Accordingly, this area concentrates on the privacy and security issues.

Privacy- Data privacy is the benefit to have some control over how the individual data is gathered and utilized. Data privacy is the limit of an individual or gathering to prevent data about themselves from getting to be noticeably known to individuals other than those they give the data to. One genuine client privacy issue is the recognizable proof of individual data amid transmission over the Web.

Security is the act of protecting data and data resources using innovation, procedures and preparing from:-Unauthorized access, Disclosure, Disruption, Modification, Inspection, Recording, and Destruction. Privacy versus security Data privacy is centered around the utilization and administration of individual data—things like setting up approaches set up to guarantee that shoppers' close to home data is being gathered, shared and used in proper ways. Security focuses more on shielding data from malevolent assaults and the abuse of stolen data for benefit. While security is major for ensuring data, it's not adequate for tending to privacy.

II. LITERATURE SURVEY

Big data [2, 7] particularly refers to data sets that are so huge or complex that conventional data preparing applications are not adequate. It's the substantial volume of data both structured and unstructured that immerses a business on an everyday premise. Because of late innovative advancement, the measure of data created by web, long range interpersonal communication destinations, sensor systems, social insurance applications, and numerous different organizations, is radically expanding step by step. All the huge measure of data delivered from different sources in various organizations with fast [3] is considered as big data. The term big data [4, 5] is characterized as "another era of advancements and models, intended to financially isolate an incentive from expansive volumes of a wide assortment of data, by empowering high-speed catch, disclosure and investigation". On the introduction of this definition, the properties of big data are reflected by 3V's, which are, volume, velocity and variety. Later investigations called attention to that the meaning of 3Vs is deficient to clarify the big data we confront now. Along these lines, veracity, value, variability, validity, venue, vagueness, and vocabulary were added to make some supplement clarification of big data [6]. A typical subject of big data is that the data are various, i.e., they may contain content, sound, picture, or video and so on. This contrasting characteristic of data is meant by assortment. So as to guarantee big data security, different components have been created as of late. These systems can be assembled in light of the phases of big data life cycle [1] i.e., data storage, processing, and generation. In data generation stage, for the assurance of security, get to limitation and additionally adulterating data strategies are utilized. The ways to deal with security assurance in data stockpiling stage are essentially in light of encryption systems. Encryption based methods can be additionally partitioned into Identity Based Encryption (IBE), Attribute Based Encryption (ABE) and capacity way encryption. Also, to ensure the touchy data, crossover mists are used where delicate data are put away in private cloud. The data handling stage joins Privacy Preserving Data Publishing (PPDP) and learning extraction from the data. In PPDP, anonymization strategies, for example, speculation and concealment are used to secure the protection of data. These instruments can be additionally isolated into bunching, characterization and affiliation control mining based strategies. While bunching and arrangement split the information data into different gatherings, affiliation govern mining based procedures locate the valuable connections and patterns in the information data [8]. To deal with assorted estimations of big data as far as volume, speed, and assortment, there is have to plan proficient and compelling structures to handle extensive measure of data touching base at fast from different sources. Big data needs to encounter numerous stages amid its life cycle. Starting at 2012, 2.5 quintillion bytes of data are made every day. The volumes of data are huge, the era speed of data is quick and the data/data space is worldwide [9]. Lightweight incremental calculations ought to be viewed as that are fit for accomplishing heartiness, high precision and least pre-handling inactivity. Like, if there should arise an occurrence of mining, lightweight component determination strategy by utilizing Swarm Search and Accelerated PSO can be utilized as a part of place of the customary characterization strategies [10]. Assist ahead, Internet of Things (IoT) would prompt association of everything that individuals think about on the planet because of which significantly more data

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

would be created than these days [11]. To be sure, IoT is one of the significant main thrusts for big data examination [9]. In the present computerized world, where loads of data is put away in big data's, the examination of the databases can give the chances to take care of big issues of society like healthcare and others. Smart energy big data analytics is also a very complex and challenging topic that share many common issues with the generic big data analytics. Smart energy big data involve extensively with physical processes where data intelligence can have a huge impact to the safe operation of the systems in real-time [12]. This can also be useful for marketing and other commercial companies to grow their business. As the database contains the personal information, it is vulnerable to provide the direct access to researchers and analysts. Since in this case, the privacy of individuals is leaked, it can cause threat and it is also illegal. The paper is based on research not ranging to a specific timeline. As the references suggest, research papers range from as old as 1998 to papers published in 2016. Also, the number of papers that were retrieved from the keyword based search can be verified from the presence of references based on the keywords. “Privacy and security concerns” section discusses of privacy and security concerns in big data and “Privacy requirements in big data” section covers the Privacy requirement in big data. “Big data privacy in data generation phase”, “Big data privacy in data storage phase” and “Big data privacy preserving in data processing” sections discusses about big data privacy in data generation, data storage, and data processing Phase. “Privacy Preserving Methods in Big Data” section covers the privacy preserving techniques using big data. “Recent Techniques of Privacy Preserving in Big Data” section presents some recent techniques of big data privacy and comparative study between these techniques.

A great deal of research has been done on watermarking or fingerprinting mixed media information. The whole action of copyright security of numeric social databases was started by Agrawal, Haas, and Kiernan, (2003). A watermarking plan which follows up on numeric qualities just and performs bit-level stamping is depicted. The basic thought of this plan is to embed particular mistake esteems in bit positions, qualities and the tuples in the connection which are pseudo-haphazardly recognized. This bit design constitutes the watermark. The secret key and the tuple's primaryl key are given as contribution to the pseudo–random succession generator. Despite the Least Significant Bit (LSB) based information concealing systems being productive, the loss of watermark without much harm to the information may come about because of moving the LSB by just a single position. Sion, Atallah, and Prabhakar, (2004) utilize marker tuples for apportioning which ends up noticeably helpless against addition or erasure assaults. The position of the marker tuples is not attractive to be adjusted. Shehab, Bertino, and Ghafoor, (2008) have exhibited a strong watermarking strategy for social information that inserts watermark bits in the information insights. A compelled advancement issue is defined for watermarking where the concealing capacity is boosted or limited in light of the bit to be implanted. Hereditary Algorithm and Pattern Search streamlining systems are connected and limitations are dealt with. A watermarking/fingerprinting framework for social databases is displayed in (Lafaye, Gross-Amblard, Constantin, and Guerrouani, 2008). The ease of use limitations are determined utilizing a worked in revelatory dialect. Whole number Linear calculation and plot secure matching calculation are the two watermarking techniques characterized and Tardos code is utilized for fingerprinting. It considers nearby and worldwide limitations on information while watermark addition is finished by altering the LSB of numerical esteems. Boneh and Shaw, (1998) and (Tardos, 2008; Skoric, Tatiana, Vladimirova, Celik, and Talstra, 2008) plans might be connected to get arrangement secure fingerprints. In (Boneh and Shaw, 1998), a paired randomized code is displayed by Boneh and Shaw where a somewhat randomized internal code is linked with external code. A checking suspicion is likewise proposed where the imprints at which their duplicates concur can't be modified by the colluders. A completely randomized paired fingerprinting code is portrayed by Tardos, (2008). This plan especially furnishes codes with short lengths. The proprietor require not know the quantity of clients ahead of time and another client can be included powerfully while utilizing Tardos code. In (Kamran, 2013), the plan created by M. Kamran et al., performs watermarking and addresses a few difficulties like strength against various assaults, learning conservation, and indicates ease of use imperatives once for all. All of multibit watermark created from date-time is embedded in each (numeric) ascribe of chose line to accomplish power even after assault on some chose piece of the dataset. Because of this 100 for every penny translating exactness is accomplished by the plan regardless of the possibility that just a single watermarked push stays in the database. Here, the creator guarantees that the deciphering exactness is free of the predetermined ease of use imperatives. Therefore, the proprietor needs to characterize the convenience limitations once for an application. A method for watermarking social databases is accordingly recommended which is strong and limits

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

bending. Distinctive methodologies and sorts of watermarking calculations for numeric social databases are recommended in (Khataeimaragheh, and Rashidi, 2010; Huang, Yue, Chen, He, Chen, 2009; Manjula, Settipalli, 2010; Hamadou, Sun, Shah, Gao, 2011). In (Manjula, Settipalli, 2010), a watermarking procedure versatile to added substance assault is proposed. A novel model is appeared in (Kamran, and Farooq, 2013) to consequently characterize client requirements for any information mining dataset to accomplish data safeguarding. A mystery gathering parameter is utilized for gathering the important lines and the nearby and worldwide convenience limitations are connected. The arrangement capability of the elements and a few different attributes are protected with the end goal that the mining of the datasets is not influenced. Watermark security is likewise guaranteed by checking the framework against the assaults. Some current overviews (Halder, Pal, and Cortesi, 2010; Mohanpurkar and Joshi, 2011) exhibit a survey of a few watermarking methods which apply diverse ways to deal with accomplish possession insurance. A couple of systems have endeavored fingerprinting of social databases and a not very many have guaranteed for conspiracy secure fingerprinting. Imprints are embedded in same places of various duplicates and furthermore perform bit-level stamping which have an indistinguishable detriments from said before in this segment. At the point when concentrate is on copyright security of Electronic Medical Records (EMRs) (Kamran, and Farooq, 2012) at that point the blunder inclusion ought not damage the convenience imperatives as it might bring about misdiagnosis of the malady. This misdiagnosis may prompt immense money related misfortune because of unseemly treatment and in the most pessimistic scenario, death toll as well. Thus data conservation by entirely holding fast to the predetermined ease of use requirements is basic on account of EMR copyright assurance. Appropriately, a novel plan proposed in (Waghmode, and Mohanpurkar, 2014) and (Mohanpurkar, and Joshi, 2015b) manages numeric databases without essential key and grouping measurements assume an imperative part. The procedure utilized as a part of (Gursale, and Mohanpurkar, 2014) and (Mohanpurkar, and Joshi, 2015a) is the place the essential key in the social databases shapes an essential info. In the strategy in exhibit paper essential key is of awesome significance. The impact of applying the novel unique mark addition procedure on (Mohanpurkar, and Joshi, 2015b) is expounded in (Mohanpurkar, and Joshi, 2015c). Agreement evasion is made conceivable in the proposed framework due to a remarkable unique mark installing strategy connected which is clarified in detail in additionally segments. The proposed framework is in this way found to give multi-crease benefits.

III. PROBLEM STATEMENT

To design and implement a system that can provide privacy and security to overall system in HDFS environment. With the data privacy system prevents the data from various attacks for sensitive information breaches and with the security system provides copyright protection along with traitor identification (if any).

IV. SYSTEM ARCHITECTURE

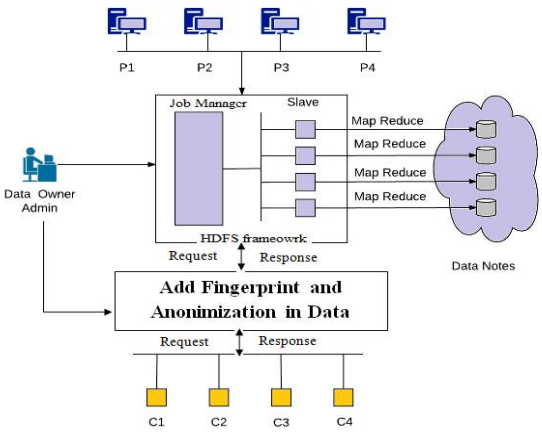

Fig.1: System Architecture

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Distributed computing is creating as a viable model for engaging speedy time-to-respond in due order regarding present day vast scale data concentrated applications. The benefits of this model join capable asset usage, improved execution, and ease of use through customized resource booking, bit, and data organization. Continuously, the MapReduce structure is used for acknowledging distributed computing bases, which modifies the application change prepare for exceedingly versatile figuring bases. Arranging a MapReduce setup includes numerous execution essential diagram decisions, for instance, center point handle compel and limit confine, choice of record structure, plan and distributing of data, and decision of framework topology, to name a couple. Likewise, a typical setup may incorporate tuning of numerous parameters to separate perfect execution. Aside from some site-specific encounters, for e.g. Google's MapReduce system this setup space is by and large unexplored. Regardless, assessing how applications would perform on specific MapReduce setups is fundamental, especially to optimize existing setups and building new ones. In this work, we get a diversion approach to manage examine the impact of diagram choices in MapReduce setups. We are stressed with how decisions over gathering plot, run-time parameters, multi-inhabitance and application design impact application execution. We build up an exact test framework, MRPerf, to totally get the diverse diagram parameters of a MapReduce setup. MRPerf can gauge the impact of various parts on application execution, and also get the capricious coordinated efforts between the components. We foresee that MRPerf will be used by masters and experts to perceive how their MapReduce applications will bear on a particular setup, and how they can improve their applications and stages. The general target is to empower MapReduce sending by methods for usage of MR-Perf as a feedback instrument that gives exact parameter tuning, instead of the surviving erroneous experimentation approach.
The proposed framework is composed and executed to safeguard the privacy of an individual data and give security to that data in distributed database framework utilizing cutting calculation. Cutting calculation segments the data both on a level plane and vertically and safeguards preferable data utility over speculation and can be utilized for participation revelation assurance utilizing HDFS structure. This framework additionally utilizes fingerprint as security approach. At the point when data proprietor give the data in the meantime we will include a few fillers into the touchy characteristics from genuine data and make it complex which can't use for produce Foundation Information (BK) too sham copyright moreover.

Algorithm 1: Slicing algorithm

Input: Data set with D, providers n, with C
Output: Slice view (T*) with provider
Step 1: read data from (D up to null)
Step 2: for each (attributes in table)
For each (tupels in tables)
Step 3: set quasi identifier (QIfr) and sensitive attributes (SA)
Step 4: Apply generalization technique it will classify the tupples in QIfr groups
Step 5: Apply anonymization on relative information attributes
Step 6: While **(**verify data-privacy(D, n, C) = 0) do
if (Di → D) verified with QIfr then
add Di up to when K-anonimity
else early stop
Bucket(i1) → D;
Step 7: permute the data with (I=(I(null-1)))
Step 8: Apply Pruning on(D)
Step 9: Apply step 1,2,3 on Bucket(i1)
Step 10: if (C fails with (D)&& (p#1))
Bucket(i2) → Bucket(i1(j))
Step 11: Display all (Bucket (i2)!=null)
Step 12: end while
Step 13: end for

Fig 2: Slicing Algorithm

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Algorithm 2: Data Partitioning Algorithm

The dataset DB is a database relation as DB= (P_k, A_0,….,A_{n-1}), where P_k is the primary key and A0, An-1 are n attributes. The dataset DB is divided in m non-overlapping partitions by using hashing technique stated in Equation (1) as {P0,...., Pm-1} such that for any of two partitions $P_i \cap P_j = \{\}$, where $i \neq j$. The data partitioning algorithm (Kamran, 2013) as shown in fig. 6.6 is applied to partition the dataset into groups. For each tuple t ϵ DB, the MAC is computed to assign tuples to the respective partitions by using hash function H as:

$$\text{Partition (r)} = H(K_s \| H(t.P_k \| K_s)) \% m \ldots\ldots\ldots\ldots\ldots (1)$$

Here, '$t.P_k$' is the primary key of tuple t, 'H()' is secure hash function and '||' is the concatenation operator and K_s is the owner's secret key. The hashing technique is used for dividing the dataset into m partitions. Hash method for partitioning maps data to partitions based on a hashing algorithm chosen applies to the partitioning key that is identified. The hashing algorithm evenly distributes rows among partitions, giving partitions approximately the same size. Hash partitioning is the ideal method for distributing data evenly across devices. Hash partitioning is also an easy-to-use alternative to range partitioning, especially when the data to be partitioned is not historical or has no obvious partitioning key.

Input: Dataset DB, Seceret key K_s, Number of Partitions m
Output: Partitioned data $P_0,\ldots,P_{m-1}$
Steps:
1. For each tuple t ϵ DB.
2. Do partition r using equation (1)
3. Insert r into P_r
4. end for
5. return $P_0,\ldots,P_{m-1}$

Fig 3: Data Partitioning Algorithm

Algorithm 3: Threshold Computation Algorithm

The threshold value is computed using Equation (2) where μ is the mean and σ is the standard deviation of values of attributes A in DB.

$$T = f * \mu + \sigma \ldots\ldots\ldots (2)$$

The confidence factor '*f*' which can have the value between 0 and 1which is kept secret because of which it becomes difficult to guess the selected tuples. For each attribute the threshold is calculated. The attributes of a tuple having value above its corresponding threshold computed is selected for marking and union of such selected tuples is taken into consideration to form DB_1.

Input: Partitioned data $P_0,\ldots,P_{m-1}$
Output: Dataset DB_1
Steps:
1. for i=0 to m-1 do
2. for each A(attribute) ϵ P_i do
3. Compute μ(mean) and σ(standard deviation) on Attributes
4. Calculate T by using equation (2)
5. End for
6. End for
7. Return $DB_1 \leftarrow R_{>T}$

Fig 4: Threshold Computation Algorithm

Algorithm 4: Hash-Value Computation Algorithm

Tuples having even hash values are selected from the data set DB_1. The MD5 hash function is applied to the data set DB_1 using Equation (3) and dataset DB_2 is obtained. This step further reduces the size of the dataset selected for marking and so leads to reduced distortion.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Even value (r) = H (K_s|| r.P_k) % 2 ……. (3)

Input: Dataset DB_1
Output: Dataset DB_2
Steps:

1. for each r ϵ DB_1 do
2. Calculate even value r using equation (3)
3. If r==0 then
4. Insert r into DB_2
5. Else do not consider tuple for fingerprinting
6. End if
7. End for
8. Return DB_2

Fig 5: Hash-Value Computation Algorithm

Algorithm 5: Fingerprint Insertion Algorithm

The tardos generated fingerprint bits b1,b2,.., bn are inserted in the each partition P_i. By using algorithm given in fig. 6 the fingerprint is inserted into a dataset. Fingerprint embedding function (F^I) is used to convert dataset DB into fingerprinted database DB_f.

$$F^I : (DB, f_c) \rightarrow DB_f \text{ ……(4)}$$

Fingerprinting function F^I for i^{th} row and j^{th} column is

$$F^I = \eta_{ij} + v_{ij} \text{ ……(5)}$$

Where, ij is one instance of modification which is performed on the tuple i^{th} row in jth column, computed according to b bit.
If b bit is equal to one, then η_{ij}, subject to constraints Uc ,on data value v_{ij} of an attribute as

$$\eta_{ij} = \alpha\% \text{ of vij,with } \alpha > 0 \text{ …….(6)}$$

and if b bit is equal to zero, then

$$\eta_{ij} = \alpha\% \text{ of vij,with } \alpha < 0 \text{ …….(7)}$$

where, α is fixed for every fingerprinted tuple and is defined by data owner/admin ad used in fingerprint decoding phase. Usability constraints are defined in terms of mean and standard deviation of attributes of DB. These two measures remain approximately the same before and after fingerprinting the data.
Data manipulation vector (Δ) is used to keep the record of transformation from DB to DB_f by adding the parameter η in every encoding step. The output data set DB_f is then made available for buyer. Algorithm in fig 6 represents steps involved in embedding fingerprint into dataset.

Input: Data set DB, Data set DB_2, f_c, Secret key K_s, Primary key P_k, BuyerID
Output: Fingerprinted Data set DB_f
Steps:

1. DB_f = DB
2. DB_2 = Get EvenHashValue Dataset(DB_1,K_s)
3. for each tuple r in DB_2 do
4. select attribute pseudo randomly using secret key, primary key and buyer id
5. if bit b == 1 then
6. compute η from equation (6)
7. else
8. compute η from equation (7)
9. end if
10. DB(temp) (η+ DB(temp))
11. end for
12. insert η into Δ
13. return DB_f

Fig.6: Fingerprint Insertion Algorithm

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Algorithm 6: Generation of Fingerprint Bits

Fingerprint is generated by using Trados code. Fingerprinting is used to trace the guilty users who redistribute an unauthorized data. Fingerprints should be inserted into dataset such that location of it should not be revealed to traitor. A Collusion Attack, where more than two pirates compare their copies and build their own copy as they want is important and needed to be handled. For example, Consider 3 fingerprint values m1 = 0110, m2 = 1000, and m3 =1110 used for 3 buyers b1, b2 and b3 respectively. Buyer b1 and b2 compare their respective copies and modify it to build a copy where it di□ers. So, the fabricated fingerprint is 1110 which is actually the fingerprint for buyer b3. So it can happen that buyer b1 and b2 actually collude to pirate the database but user b3 is identified as victim after detection. To avoid accusation of such innocent users collusion-secure codes have been designed like Boneh and shaw scheme code[1998], Tardos Scheme of B.Skoric, S. Katzenbeisser, and M. Celik [2008]. Out of these Tardos scheme is better because-

1. Codewords have small length as compared to other Schemes.
2. Number of users need not be pre-defined.
3. Implementation is e□cient and simple.

Following are the steps to generate Tardos Fingerprinting code:

Two Phase Process:

1. First distributor choose the codewords distribution.
2. Fingerprinting matrix BT is built. The rows of the matrix BT are the codewords embedded into the copies that the users receive.

Phase 1:

In this phase the distributor picks a random variable for the code construction. The distributor does it in the following way:

1. The distributor picks some suitable parameter t_p where $t_p=1/(300c)$
2. The distributor computes the parameter

$$t^1_p = \arcsin(\text{sqrt}(t_p)) \quad \ldots\ldots\ldots\ldots(8)$$

For all $1 \le i \le N$ the distributor picks r_i uniformly at random from the interval $[t^1_p, \pi/2 - t^1_p]$. The distributor puts $p_i = \sin^2 r_i$, where $1 \le i \le N$ and thus $t_p \le p_i \le 1 - t_p$.

Phase 2:

Fingerprinting matrix B_T is built. The entries of B_T are such that $P[B_T(j,i)=1] = p_i$ and $P[B_T(j,i)=0] = 1 - p_i$.
Where, j denotes the number of user j = 1, ..., n and i denotes the code length i = 1,... l.

Fig. 7: Tardos Code Generation

V. RESULT AND DISCUSSION

Table 1 shows the output of system after execution of slicing algorithms. Slicing will provide the better security than existing approach.

ID	Name	Gender	Zip Code	Qualifi-cation	Age
128	*****	female	***** 5	*****	[26-50]
129	*****	male	***** 8	*****	[26-50]
130	*****	female	***** 5	*****	[26-50]
131	*****	male	***** 6	*****	[0-25]

Table 1: Slicing result

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

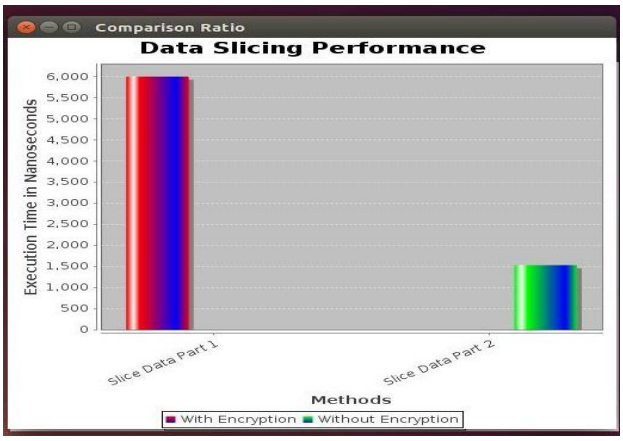

Fig.8: Data Slicing Performance

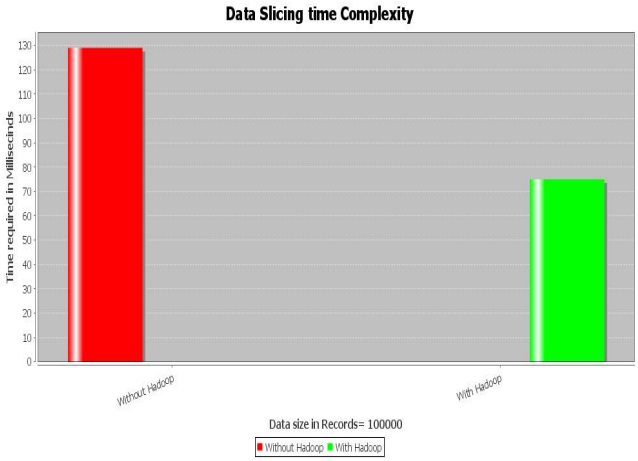

Fig.9: Data Slicing Performance on Hadoop

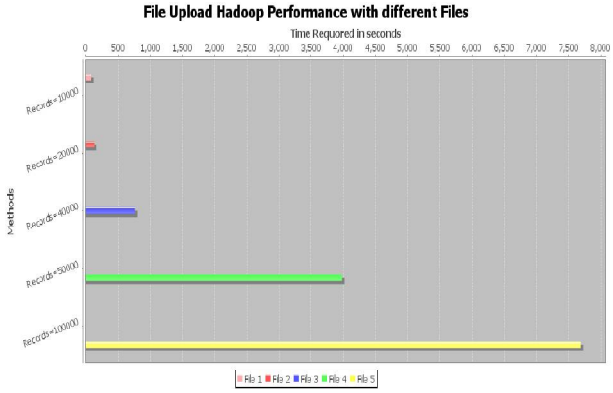

Fig.10: File Uploading Time Performance on Hadoop

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Attribute	Before Fingerprinting Mean	After Fingerprinting Mean	Difference
First	68.8076	68.8097	0.0021
Second	70.2106	70.2189	0.0083
Third	67.5071	67.5127	0.0056
Fourth	69.8714	69.8898	0.0184
Fifth	68.4989	68.5051	0.0062

Table 2: Mean of Attribute values before and after fingerprint insertion

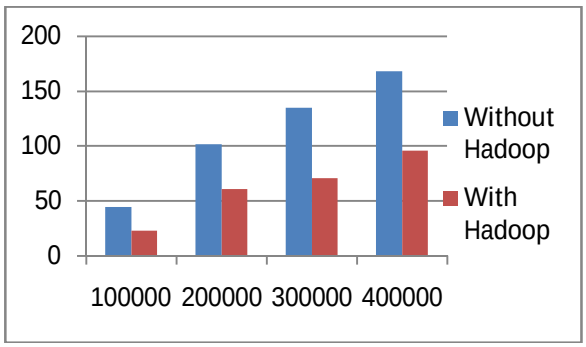

Fig .11 : System performance measure proposed vs existing

Fig.11 shows overall system performance using Hadoop and using normal system.

VI. CONCLUSION AND FUTURE ENHANCEMENT

Data is anonymized by removing the personal details to preserve the privacy of users. It indicates that it would not be possible to identify an individual only from the anonymized data. However, due to the availability of huge volumes of data and powerful data analytic tools, the existing anonymization techniques are becoming increasingly ineffective. In big data scenarios, anonymization needs to be more than just masking or generalizing certain fields. One needs to carefully analyze if the anonymized data are vulnerable to any attacks. For that, different attack models and information loss metric for big data anonymization is studied. Moreover, a Fingerprinting approach is used which can provide security against the ownership theft as well as it also helps trace the traitor if any unauthorized copy is found. In case of relational databases, insertion of fingerprint bit can change numeric data to some extent. This change in numeric data may not lead to loss in knowledge from database. The proposed insertion algorithm is independent primary key and it can efficiently trace the traitor.

Future Enhancement

The proposed architecture can be developed on hadoop base system with cube materialization and mapreduce with multi cloud environment. Also subset of holistic measures that are partially algebraic and propose the technique of value partitioning to make them easy to compute in parallel can be identified. Design algorithms that partition the cube lattice into batch areas to effectively exploit both the parallel processing power of MapReduce and the pruning power of cube materialization algorithms. Further the ability to surface interesting cube groups as part of the cube computation process can be demonstrated. Experiments over real and synthetic data show that MR-Cube algorithm efficiently distributes the computation workload across the machines and is able to complete cubing tasks at a scale where prior algorithms fail.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

REFERENCES

[1] 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.
[2] Abadi DJ, Carney D, Cetintemel U, Cherniack M, Convey C, Lee S, Stone-braker M, Tatbul N, Zdonik SB. "Aurora: a new model and architecture for data stream management". VLDB J. 2003;12(2):120–39.
[3] 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.
[4] Big data at the speed of business, [online]. http://www-01.ibm.com/soft-ware/data/bigdata/2012.
[5] 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.
[6] Gantz J, Reinsel D. "Extracting value from chaos". In: Proc on IDC IView. 2011. p. 1–12.
[7] Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV. "Big data analytics: a survey". J Big Data Springer Open J. 2015.
[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] Agrawal, R., Haas, P., & Kiernan, J. (2003). "Watermarking Relational Data: Framework, Algorithms and Analysis". The VLDB Journal, 12(2), 157–169. doi:10.1007/s00778-003-0097-x
[14] Blayer, O., & Tassa, T. (2008). "Improved versions of Tardos' Fingerprinting scheme. Designs, Codes and Cryptography", 48(1), 79–103. http://www.openu.ac.il/home/tamirtassa/ Publications/tardos.pdf doi:10.1007/ s10623-008-9200-z
[15] Boneh, D., & Shaw, J. (1998). "Collusion – Secure Fingerprinting for Digital Data". IEEE Transactions on Information Theory, 44(5), 1897 – 1905.
[16] Gursale, N., & Mohanpurkar, A. (2014). "A Robust, Distortion Minimization Fingerprinting Technique for Relational Database". International Journal on Recent and Innovation Trends in Computing and Communication. 2(6), 1737 – 1741.
[17] Gursale, N., & Mohanpurkar, A. (2014). "A Robust, Distortion Minimization Fingerprinting Technique for Relational Database". International Journal on Recent and Innovation Trends in Computing and Communication. 2(6).1737 – 1741.
[18] Halder, R., Pal, S., & Cortesi, A. (2010). "Watermarking Techniques for Relational Databases: Survey, Classification and Comparison". Journal of Universal Computer Science, 16(21), 3164–3190.
[19] Hamadou, A., Sun, X., Shah, S. A., & Gao, L. (2011). "A Hybrid Watermarking Scheme for Relational Databases Copyright Protection and Tamper Proofing". International Journal of Advancements in Computing Technology, 3(8), 18–28. doi:10.4156/ijact.vol3.issue8.3
[20] Huang, K., & Yue, M., Chen, He, Y., Chen, X. (2009). "A Cluster-Based Watermark Technique for Relational Database". Proceedings of the First International Workshop on Database Technology and Applications (pp. 107110). doi:10.1109/DBTA.2009.38
[21] Kamran, M., & Farooq, M. (2012). "An Information-Preserving Watermarking Scheme for Right Protection of EMR Systems". IEEE Transactions on Knowledge and Data Engineering, 24(11), 1950–1962. doi:10.1109/ TKDE.2011.223
[22] Kamran, M., & Farooq, M. (2013). "A Formal Usability Constraints Model for Watermarking of Outsourced Datasets". IEEE Transactions On Information Forensics And Security, 8(6), 1061–1072. doi:10.1109/ TIFS.2013.2259234
[23] Kamran, M., Suhail, S., & Farooq, M. (2013). "A Robust, Distortion Minimizing Technique for Watermarking relational Databases Using Once-for-all Usability Constraints". IEEE Transactions on Knowledge and Data Engineering, 25(12), 2694–2707. doi:10.1109/TKDE.2012.227 Khataeimaragheh,
[24] H., & Rashidi, H. (2010). "A Novel Watermarking Scheme For Detecting And Recovering Distortions In Database Tables". International Journal of Database Management Systems, 2(3), 1–11. doi:10.5121/ ijdms.2010.2301
[25] Lafaye, J., Gross-Amblard, D., Constantin, C., & Guerrouani, M. (2008). "WATERMILL: An Optimized Fingerprinting System for Databases under Constraints". IEEE Transactions on Knowledge and Data Engineering, 20(4), 532–546. doi:10.1109/TKDE.2007.190713
[26] Li, Y., Swarup, V., & Jajodia,S. (2005). Fingerprinting Relational Database: Schemes and specialities. IEEE Transactions on Dependable and Secure Computing, 2(1), 34-45.
[27] Mahmoud E. F., Horng, S., Lai, J. L., Run R. S., Chen, R. J., & Khan, M. K. (2012). "A Blind reversible method for watermarking relational databases based on a time-stamping protocol". Expert Systems with Applications, 39, 3185-3196.
[28] Manjula, R., Settipalli, N., (2010) "A new Relational Watermarking Scheme Resilient to Additive Attacks". International Journal of Computer Applications, 10(5),1-7.
[29] Mohanpurkar, A., & Joshi, M. (2011). "Applying Watermarking For Copyright Protection, Traitor Identification And Joint ownership: A Review". Presented atInternational IEEE Conference WICT 2011 (pp. 1018-1023). doi:10.1109/WICT.2011.6141387
[30] Mohanpurkar, A., & Joshi, M. (2015a). "A Fingerprinting Technique for Numeric Relational Databases with Distortion Minimization". Proceedings of theInternational Conference on Computing Communication Control and Automation. IEEE. , 655-660. doi:10.1109/ICCUBEA.2015.134

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

[31] Mohanpurkar, A., & Joshi, M. (2015a). "A Fingerprinting Technique for Numeric Relational Databases with Distortion Minimization".International Conference on Computing Communication Control and Automation., IEEE DOI , 655-660. doi:10.1109/ICCUBEA.2015.134
[32] Mohanpurkar, A., & Joshi, M. (2015b). "Fingerprinting Numeric Databases with Information Preservation and Collusion Avoidance". International Journal of Computer Applications, 130(5), 13-18.
[33] Mohanpurkar, A., & Joshi, M. (2015b). "Fingerprinting Numeric Databases with Information Preservation and Collusion Avoidance". International Journal of Computer Applications (0975 – 8887). 130(5).13-18.
[34] Mohanpurkar, A., & Joshi, M. (2015c). "The Effect of the Novel Anti-Collusion Fingerprinting Scheme on the Knowledge from Numeric Databases". International Journal of Scientific and Engineering Research, 6(12), 334 – 338.
[35] Mohanpurkar, A., & Joshi, M. (2015c). "The Effect of the Novel Anti-Collusion Fingerprinting Scheme on the Knowledge from Numeric Databases". International Journal of Scientific and Engineering Research. 6(12). ISSN 2229-5518. 334 – 338.
[36] Shehab, M., Bertino, E., & Ghafoor, A. (2008). "Watermarking Relational Databases Using Optimization Based Techniques". IEEE Transactions on Knowledge and Data Engineering, 20(1), 116–129. doi:10.1109/ TKDE.2007.190668
[37] Sion, R. Atallah, M., & Prabhakar, S. (2004). "Rights Protection for Relational Data". IEEE Transaction on Knowledge and Data Engineering, 16(12), 1509-1525.
[38] Skoric, B., Tatiana, U., Vladimirova, T. U., Celik, M., & Talstra, J. C. (2008). "Tardos Fingerprinting is Better Than we Thought". IEEE Transactions on Information Theory, 54(8), 3663–3676. doi:10.1109/TIT.2008.926307
[39] Tardos, G. (2008). "Optimal Probabilistic Fingerprint Codes", Journal of ACM, 55(2). Article, 10, 10–24.
[40] Uzun, E., Stephenson, B. (2008). "Security of Relational Databases in Business Outsourcing". HP Laboratories HPL-2008-168, 1-21.
[41] Varsha Waghmode, V., & Mohanpurkar. (2014). "A. Collusion Avoidance in Fingerprinting Outsourced Relational Databases with Knowledge Preservation". International Journal on Recent and Innovation Trends in Computing and Communication., 2(5), 1332–1337.