Big Data Privacy Methods

Jayesh Surana[1], Akshay Khandelwal[2], Avani Kothari[3],
Himanshi Solanki[4], Meenal Sankhla[5]

[1]Assistant Professor (IT), *Sri Vaishnav Institute of Technology & Science , M.P, India*
[2] *Student, Sri Vaishnav Institute of Technology & Science, M.P, India*
[3] *Student, Sri Vaishnav Institute of Technology & Science, M.P, India*
[4] *Student, Sri Vaishnav Institute of Technology & Science, M.P, India*
[5]*Student, Sri Vaishnav Institute of Technology & Science, M.P, India*

***Abstract-* Big data is a term used for large and complex data sets that cannot be stored and processed using traditional data processing software. Since, big data require high computational power and storage, distributed system are used. Big data Analytics is a term used for deriving some meaningful and hidden data from the large data sets. The data sets are collected from social media, healthcare centers, data governance, institutions, etc. Thus, privacy and security of the data become the prime concern. This paper focus on the privacy and security concerns and the problems in the privacy of big data. The privacy in big data is divided into three stages-data generation, data storage and data processing. This paper also covers some traditional methods adopted for privacy in big data, the challenges faced by these techniques. The goal of this paper is to study the recent techniques adopted for privacy and draw their comparison in order to declare the most efficient technique among all of them.**

***Index Terms-* Big data Privacy and security Privacy preserving: k-anonymity: T-closeness, L-diversity, De-identification.**

I. INTRODUCTION

Big Data is used in many applications which use Predictive Intelligence that we humans exhibit in our everyday lives. The first and most prevalent example is online advertising, where predicting our intent when we search or read documents on the web. Companies use that data to provide focus add campaign and attract the target audience. For example if a user is surfing internet to buy a Camera, web companies can look at your search patterns and publish an ad for nearby Camera stores and/or discounts available on Cameras in Online e-commerce portals.

II. PRIVACY AND SECURITY CONCERNS IN BIG DATA

Privacy: Information privacy is the privilege to have some control over how the personal information is collected and used.

Security: It is the practice of defending information and information assets through the use of technology.

S. No.	Privacy	Security
1.	Privacy is the appropriate use of users information	Security is "confidentiality, Integrity, and availability" of data.
2.	Privacy is the ability to decide what information goes where.	Security offers ability to be confident that decisions are respected.
3.	The issue of privacy often applies to consumer rights to safeguard the data	Security may provide confidentiality. The overall goal of secured system is to protect an enterprise.

III. PRIVACY REQUIREMENTS IN BIG DATA

Due to the absence of standard security and privacy protection tools, many organizations decide not to use services of big data analytics. Developers should be able to verify that their applications conform to privacy agreements and that sensitive information is kept private regardless of changes in the applications and/or privacy regulations

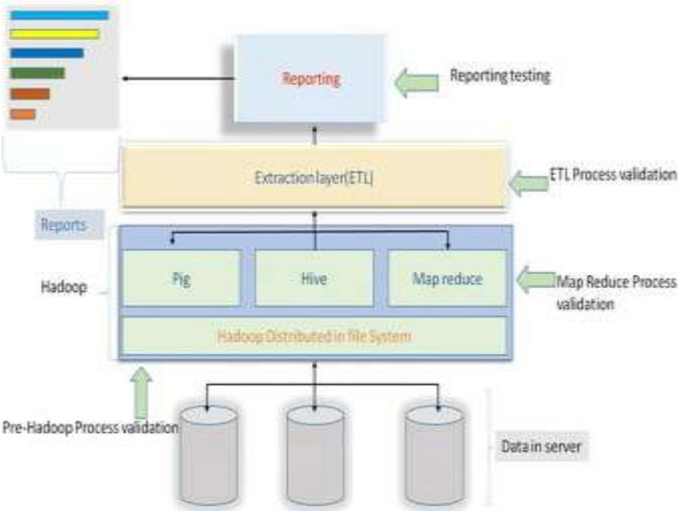


Fig. 1 Big data architecture and testing area new paradigms for privacy conformance testing to the four areas of the ETL (Extract, Transform, and Load) processes are shown here.

IV. BIG DATA PRIVACY IN DATA GENERATION PHASE

Data generation can be classified into active data generation and passive data generation.
The methods of privacy are as follows-

1. Access restriction.

If the data owner thinks that the data may uncover sensitive information which is not supposed to be shared, it refuses to provide such data.

2. Falsifying data

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

V. BIG DATA PRIVACY IN DATA STORAGE PHASE

Due to the numerous technologies available to handle the enormous amount of data, it is very challenging to store data privately.

Approaches to privacy preservation storage on cloud

Attribute based encryption: Access control is based on the identity of a user to have wholesome access over all resources.

Homomorphic encryption: Can be deployed in ABE (Attribute Based Encryption in which key or cipher text is dependent on attributes like residential address) scheme settings, updating cipher text receiver is possible.

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 two platforms.

VI. BIG DATA PRIVACY PRESERVING IN DATA 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 unauthorized disclosure. In the second phase, the aim is to extract meaningful information from the data without violating the privacy.

Traditional Methods of Privacy

De-identification is a traditional technique for privacy preserving data mining. In this method the data either go through the generalization or suppression method. In generalization the quasi-identifiers are replaced with more general but consistent values and in suppression some data is not revealed and is hidden by *. In order to prevent data from re-identification, the concepts of k-anonymity, l-diversity and t-closeness have been introduced.

Some common terms used in privacy fields of these methods-

1. Identifier Attribute- attributes that uniquely identify the individuals e.g.-PAN, no, name, social security no etc.

2. Quasi-Identifier- attribute whose value when taken together can uniquely identify an individual e.g. gender, age, date of birth etc.

3. Sensitive attribute- The information which is private and personal to the individual comes under sensitive data e.g. salary, disease etc.

4. Equivalence classes- are the group of all records that have same value on the quasi-identifiers.

- **K-anonymity**

The table is said to have k-anonymity if every record in the table is similar to at least k-1 other records with respect to every set of quasi-identifiers.

Non Anonymized table containing patient records

There are two regular technique for completing K-anonymity for some value of k

1 .Generalization-The attribute 'age' can be written in more general and broader form.eg age '19'can be written as <=20.

2. Suppression- In this method some values of the attribute are hidden using * symbol. For e.g. some values of zip code can be hidden using astrik. Thus, the 4-anonymous version table of table-1 can be written as-

Name	Age	Zip code	Disease
Priyash	29	47677	Heart Disease
Ram	24	47602	Heart Disease
Rohan	28	47905	Flu
Pranjal	27	47909	Cancer
Aayush	24	47607	Cancer

Table-1

Name	Age	Zip code	Disease
Priyash	20 <age<=30	476**	Heart Disease
Ram	20 <age<=30	476**	Heart Disease
Rohan	20 <age<=30	4790*	Flu
Pranjal	20 <age<=30	4790*	Cancer
Aayush	20 <age<=30	476**	Cancer

Table-2

Limitation-

1. K-anonymity is insufficient to prevent attribute disclosure.
2. It can suffer from Homogeneity attack and background knowledge attack.

- **L-diversity**

An equivalence class is said to have L-diversity if there are at least "well-represented" values for the sensitive attribute.

If one knows Peter salary is in range of 3k-5k then one can conclude that he has some stomach-related disease. Thus, leakage of sensitive information occur which give rise to more efficient method called t-closeness.

Limitations:

1. L-diversity is difficult to achieve
2. L-diversity is insufficient to prevent attribute disclosure.

Age	Salary	Disease
29	3k	Gastric ulcer
22	4k	gastritics
23	5k	Stomach cancer
52	6k	Flu
43	11k	pneumonia
36	8k	bronchitis

Table-3

Age	Salary	Disease
2* 2* 2*	3k 4k 5k	Gastric ulcer Gastritics Stomach cancer
>=40 >=40 >=40	6k 11k 8k	Flu Pneumonia bronchitis

Table-4(diverse version of table-3)

- **T-closeness**

It is a further improvement of l-diversity group based on annoymization that is used to preserve privacy in data sets by decreasing the granularity of a data representation

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold.

Limitation-

It does not deal with identity disclosure.

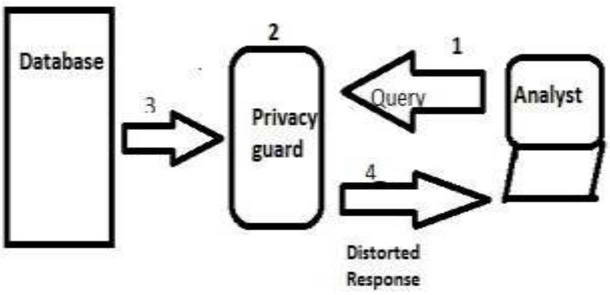
Comparison of the methods-

Sno	Methods name	Computaional Complexity
1.	K-anonymity	O(klogk)
2.	L-diversity	O(n^2/k)
3.	T-closeeness	$2^{O(n)O(m)}$

VII. RECENT TECHNIQUES IN BIG DATA PRIVACY

Differential privacy

Differential privacy is one of the effective methods to deal with privacy threats .In this model the personal information is not revealed or modified by the analyst to use. The analyst does not have any access to the information directly instead an intermediary is used which serve as a privacy guard. The privacy guard takes up the queries from the analyst and gives the result with little distortion. When the privacy risk is low we can think of the distortion as inaccuracies that are small enough that it does not affect the quality of the answer but it is large enough to protect the individual privacy.



Steps in differential privacy are as follows-

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

Step 2- The privacy guard takes the queries and procees it using the databse.
Step 3-The privacy guard then gets the answer from the database.
Step 4- The privacy guard add little distortion to the data, then finally provide to the data analyst.

Advantages

1. The original data is not modified or revealed to the end user.
2. There is no need of generalization and suppression techniques.
3. The response is distorted based on the level of risk without affecting the quality of response.
4. The distortion is added in such a way that value hidden is useful to analyst.

VIII.CHALLENGES TO SECURITY AND PRIVACY IN BIG DATA

- Increased Potential for Large-scale Theft or Breach of Data
- Increased Potential for Large-scale Theft or Breach of Data
- Long Term Availability of Sensitive Datasets
- Data Quality/Integrity and Provenance Issues
- Unwanted Data Correlation and Inferences
- Algorithmic Accountability

IX. MOST SIGNIFICANT PRIVACY RISKS

1. Privacy breaches and embarrassments
2. Anonymization could become impossible
3. Data masking could be defeated to reveal personal info.
4. Unethical actions based on interpretations.
5. Big data analytics are not 100% accurate.
6. Discrimination.
7. Few legal protections exists for individuals
8. Big data will probably exist forever.

X CONCLUSION

Big data is large amount of data which is unorganized and unstructured. Big data privacy is very important issue in while organizing big data. In this paper we have studied different methods of securing big data in three phases i.e. data generation, data storage and data processing and drawn a conclusion of best method among them. So, different methods of privacy preserving methods may be studied and implemented in future. There is the lot of future scope in the privacy of big data methods.

REFERENCES

[1] 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

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

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

[4] Li N, et al. t-Closeness: privacy beyond *k*-anonymity and *L*-diversity. In: Data engineering (ICDE) IEEE 23rd international conference; 2007.

[5] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-diversity: privacy beyond *k-anonymity*. In: Proc. 22nd international conference data engineering (ICDE); 2006

[6] http://download.microsoft.com/…/Differential_Privacy_for_Everyone.pdf.

[7] Big Data:Opportunities and Privacy Challenges By Hervais Simo .