CSP 554 – Big Data Technologies

A Research Paper On

Privacy Protection in Big Data Environment: A Technological Perspective and Review

Done By:

Adarsh Mathad Vijayakumar
A20424847

List of Tables

List of Figures

Privacy Protection in Big Data Environment: A Technological Perspective and Review

Introduction

Due to the fast growth of emerging information technologies such as Internet of Things (IoT), cloud computing, Internet services, and social networking an increasing interest in big data security and privacy is aroused. The increasing amount of big data also increases the chance of breaching the privacy of individuals. An entire lifetime of big data can be broadly classified into three phases: big data generation; processing and analytics; storage and management. Since big data require high computational power and large storage, distributed systems are used. As multiple parties are involved in these systems, the risk of privacy violation is increased. The five salient features of big data: volume, variety, velocity, value, and veracity bring great challenges on protecting big data security and privacy during its whole lifetime. Big data normally contains valuable or sensitive information about user behaviors, preferences, interests, mobility, and so on. User privacy is easily leaked if the data cannot be protected well during its lifetime. Therefore, if we want to enjoy the convenience and benefits from big data, guarantee of its security and privacy becomes an essential task.

Entire Lifetime of Big Data

To handle different dimensions of big data in terms of volume, velocity and variety we need to design efficient and effective systems to process large amount of data arriving at very high speed from different sources. Big data has to go through multiple phases – like Data Generation, Data Storage and Data Processing - during its life cycle.

Figure a: Illustration of the stages of the big data lifetime

If we want to enjoy the convenience and benefits from big data, guarantee of its security and privacy becomes an essential task. We may get wrong information from big data once some insecure events occur in the above phases. Thus, security and privacy of big data in the above phases are very important. Herein, we regard these three phases as big data lifetime.

Privacy and Security Concerns in Big Data

Privacy and security are of the most important concern in the Big Data environment. A little compromise in this can lead to leak of the sensitive data or important data and misuse of it.

What's Privacy in big data environment? Information privacy is the privilege to have some control over how the personal information is collected and used. Information privacy is the ability of an individual or

a group or an organization to stop the information about themselves from being known to others except for the ones' they intend to share. One serious user privacy issue is the identification of personal information during transmission over the Internet.

What's Security in big data environment? Security is the practice of defending information and information assets through the use of technology from - Unauthorized access, Disclosure, Disruption, Modification, Inspection, Recording and Destruction.

Privacy vs Security - Data privacy is focused on the use and governance of individual data. Security concentrates more on protecting data from malicious attacks and the misuse of stolen data for profit. While security is fundamental for protecting data, it's not sufficient for addressing privacy. The following table shows the principle differences between Privacy and Security.

Privacy	Security
Privacy is the appropriate use of user's information	Security is the "confidentiality, integrity and availability" of data
Privacy is the ability to decide what information of an individual goes where	Security offers the ability to be confident that decisions are respected
The issue of privacy is one that often applies to a consumer's right to safeguard their information from any other parties	Security may provide for confidentiality. The overall goal of most security system is to protect an enterprise or agency
It is possible to have poor privacy and good security practices	However, it is difficult to have good privacy practices without a good data security program

Table 1: Privacy and Security

Why privacy is required in the big data environment?

The following environments may break the users' privacy in big data environment:

- During data transmission over the internet personal data or information's are shared with some external resources. Due to this type of incidents third party may conclude realities about the users.
- Sometimes personal data are gathered and exploit for the business purpose. For example, todays' online shopping is the hectic modern technology through this easily can be predicted the users' habits and lot of personal information.
- Data trickling occurs at the stage of storing and data processing phases. So privacy more important in the second and third phases of the big data lifetime.

Privacy preserving mechanisms is each stage of the big data lifetime

Big data privacy in data generation phase

Data generation can be classified into **Active Data Generation** and **Passive Data Generation**. By active data generation, we mean that the data owner will give the data to a third party, while passive data generation refers to the circumstances that the data are produced by data owner's online actions (e.g., browsing) and the data owner may not know about that the data are being gathered by a third party. The major challenge for data owner is that how can he protect his data from any third party who may be willing to collect them. The data owner wants to hide his personal and sensitive information as much as possible and is concerned about how much control he could have over the information. Minimization of the risk of privacy violation amid data generation can be done by either restricting the access or by falsifying data.

A. ACCESS RESTRICTION

If the data owner thinks that the data may reveal sensitive information which is not supposed to be shared, he can simply refuse to provide such data. For that, the data owner has to adopt effective access control methods so that the data can be prevented from being stolen by some third party. If the data owner is providing the data passively, some measures could be taken to ensure privacy, such as anti-tracking extensions, advertisement/script blockers and encryption tools. By using these tools, one can effectively limit the access to sensitive data. For the ease of use, most of these tools are designed as browser extensions.

In addition to these tools, there are some alternative means, such as to use anti-malware and anti-virus software to protect the data stored digitally on their computer or laptop. These tools can help to protect user's personal data by limiting the access. Though there is no guarantee that one's sensitive data are completely protected from untrustworthy sources, making it a habit of clearing online traces/cookies/sessions of one's activity by using security tools can significantly reduce the risk.

B. FALSIFYING DATA

In some circumstances, it may not be possible to protect the personal data/sensitive data. In that case, data can be distorted using certain tools prior to the data gotten by some third party. If the data are distorted, then the true information cannot be easily revealed. The following techniques are utilized by the data owner to falsify the data:

- ***A tool Socketpuppet*** is used to hide online identity of individual by deception. Individual's true activities online are concealed by creating a false identity and pretending to be someone else. By using multiple Socketpuppets, the data belonging to one specific individual will be deemed as belonging to different individuals. In that way the data collector will not have enough knowledge to relate different socketpuppets to one individual. Hence, the user's true activities are unknown to others and the private information cannot be discovered easily.
- Certain ***security tools*** can be used to mask individual's identity, such as MaskMe. It allows users to create aliases of their personal information such as email address or credit card number. The data owner can use these masks whenever information is needed. This is especially useful when the data owner needs to provide the credit card details during online shopping.

Big data privacy in data storage phase

Storing high volume data is not a major challenge due to the advancement in data storage technologies. If the big data storage system is compromised, it can be exceptionally destructive as individual's personal information can be disclosed. Therefore, we need to ensure that the stored data are protected against such threats. In modern information systems, data centers play an important role of performing complex commutations and retrieving large amount of data. In distributed environment, an application may need several datasets from various data centers and therefore confront the challenge of privacy protection.

The conventional security mechanisms to protect data can be divided into four categories. They are file level data security schemes, database level data security schemes, media level security schemes and application level encryption schemes. Responding to the 3V's nature of the big data analytics, the storage infrastructure ought to be scalable. It should have the ability to be configured dynamically to accommodate various applications. One promising technology to address these requirements is storage virtualization, empowered by the emerging cloud computing paradigm. Storage virtualization is process in which numerous network storage devices are combined into what gives off an impression of being a single storage device. However, using a cloud service offered by cloud provider means that the organization's data will be outsourced to a third party such as cloud provider. This could affect the privacy of the data.

A. Approaches to privacy preservation storage on cloud

When data is stored on cloud, data security predominantly has three dimensions - confidentiality, integrity and availability. The first two are directly related to privacy of the data i.e., if data confidentiality or integrity is breached it will have a direct effect on users' privacy. Availability of information refers to ensuring that authorized parties are able to access the information when needed. A basic requirement for big data storage system is to protect the privacy of an individual. The approaches to safeguard the privacy of the user when data are stored on the cloud are as follows:

1) *Attribute Based Encryption (ABE)*: This is an encryption technique which ensures end to end big data privacy in cloud storage system. In this, access polices are defined by data owner and data are encrypted under those policies. The data can only be decrypted by the users whose attributes satisfy the access policies defined by the data owner. When dealing with big data one may often need to change data access policies as the data owner may have to share it with different organizations. This can be done in an easy and effective way as described next here. The data owner can send the queries to cloud to update the policy, and the cloud server can update the policy directly without decrypting the data.
2) *Identity Based Encryption (IBE)*: This is an alternative to the conventional Public Key Encryption which is proposed to simplify key management in a certificate-based public key infrastructure (PKI) by using human identities like email address or IP address as public keys. To preserve the anonymity of sender and receiver, the IBE scheme was proposed. By employing these primitives, the source and the destination of data can be protected privately.
3) *Homomorphic Encryption (HE)*: Homomorphic encryption has been recognized as one of the ideal approaches to securing and processing big data in remote servers including the cloud. Homomorphic Encryption (HE) is a method of secure computation which allows for calculations to be made on encrypted data without decrypting it and without giving away information about the operations being done.

4) *Storage Path Encryption:* In the proposed scheme, the big data are first separated into many sequenced parts and then each part is stored on a different storage media owned by different cloud storage providers. To access the data, different parts are first collected together from different data centers and then restored into original form before it is presented to the data owner. A trapdoor function has been incorporated in this scheme. It is a function which is easy to compute in one way and difficult to compute in the opposite direction without some additional information. The owner of the big data will keep the storage index information.
5) *Proxy Re-Encryption (PRE)*: Proxy re-encryption technique with respect to secure cloud data and its application, is a method to keep sensitive user data confidential against untrusted servers. Cryptographic methods are used to provide security and access control in clouds. Data over the network needs to be encrypted. Let's consider a simple situation to understand PRE – proxy re-encryption is generally used when one party, say Bob, wants to reveal the contents of messages sent to him and encrypted with his public key to a third party, Chris, without revealing his private key to Chris. Bob does not want the proxy to be able to read the contents of his messages. Bob could designate a semi-trusted proxy server to re-encrypt one of his messages that is to be sent to Chris. This generates a new key that Chris can use to decrypt the message. Now if Bob sends Chris a message that was encrypted under Bob's key, the proxy will alter the message, allowing Chris to decrypt it. This method allows for many applications such as e-mail forwarding etc...

Comparison of the Encryption Schemes:

Encryption Scheme	Features	Limitations
Identity Based Encryption (IBE)	• Access control is based on the identity of a user • Complete access over all resources	• Less time efficiency in large environments. • Granular access control is hard to implement • Data to be processed must be downloaded and decrypted • Changing ciphertext receiver is not possible
Attribute Based Encryption (ABE)	• Access control is based on user's attribute • More secure and flexible as granular access control is possible	• Data to be processed must be downloaded and decrypted • Updating ciphertext receiver is not possible
Homomorphic Encryption (HE)	• Computations are preferred on the encrypted data • Very much secure	• Computational overhead is very high
Proxy Re-Encryption (PRE)	• Updating Ciphertext receiver is possible • Can be deployed in IBE or ABE scheme settings	• Data to be processed must be downloaded and decrypted • Computational overhead

Table 2: Comparison of Encryption Schemes

B. Integrity verification of big data storage

When cloud computing is used for big data storage, data owner loses control over data. The outsourced data are at risk as cloud server may not be fully trusted. The data owner needs to be strongly convinced that the cloud is storing data properly according to the service level contract. One way to ensure privacy to the cloud user is, to provide the system with the mechanism to let data owner verify that his data stored on the cloud is intact. The integrity of data storage in traditional systems can be verified through number of ways i.e., Reed-Solomon code, checksums, trapdoor hash functions, message authentication code (MAC), and digital signatures etc. Therefore, data integrity verification is of critical importance. It is highly prescribed that the integrity verification should be conducted regularly to provide highest level of data protection.

Big data privacy preserving in data processing

Big data processing paradigm categorizes systems into batch, stream, graph, and machine learning processing. In data processing stage privacy protection can be divided into two parts. First part to concentrate on protection of data from unwanted leakage during processing stage since the collected information's have more sensitive data. Second part mainly focusing on the extracting meaningful information from the data without losing privacy.

Techniques for Protecting Privacy in Big Data

Privacy in big data has raised serious concerns bringing into notice the need for efficient privacy preservation methods. *Here we'll see the three privacy preservation methods*: **data anonymization, differential privacy and notice & consent**.

1. Data Anonymization

Data anonymization is the process of changing data that will be used or published in a way that prevents the identification of key information. It is also sometimes referred as data de-identification. In this method key pieces of confidential data are obscured in a way that maintains data privacy. Organizations release data publicly by anonymizing it. Anonymization in this case generally refers to hiding identifier attributes (attributes that uniquely identify individuals) like full name, license number etc. The main problem with data anonymization is that data may look anonymous but re-identification can be done easily by linking it to other external data. The attributes like gender, date of birth, zip code that can be combined with external data to re-identify individuals are called quasi identifier attributes. *There are three privacy preserving methods of Data Anonymization*, namely

a. K - Anonymity
b. L - Diversity
c. T - Closeness

There are some common terms used in privacy field of this method:

- **Identifier attributes** include information that uniquely and directly distinguishes individual such as full name driver license, social security numbers.
- **Quasi-identifier attributes** a set of information such as birth date, gender, age, zipcode. That can be combined with other external data in order to re-identify individuals.
- **Sensitive attributes** are private and personal information.
- **Intensive attributes** are the general and the innocuous information
- **Equivalence classes** are sets of all records that consist of the same value on the quasi-identifiers.

a. K – Anonymity

A dataset is called k-anonymized if for any tuple with given attributes in the dataset there are at least k-1 other records that match those attributes. In the context of *k*-anonymization problems, a database is a table which consists of *n* rows and *m* columns, where each row of the table represents a record relating to a particular individual from a populace and the entries in the different rows need not be unique. The values in the different columns are the values of attributes connected with the members of the population. Table-3 is a non-anonymized database comprising of the patient records of some fictitious hospital.

Name	Age	Gender	State of domicile	Religion	Disease
Ramya	29	Female	Tamil Nadu	Hindu	Cancer
Yamini	24	Female	Andhra Pradesh	Hindu	Viral infection
Salini	28	Female	Tamil Nadu	Muslim	TB
Sunny	27	Male	Karnataka	Parsi	No illness
Joshna	24	Female	Andhra Pradesh	Christian	Heart-related
Badri	23	Male	Karnataka	Buddhist	TB
Ramu	19	Male	Andhra Pradesh	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
John	17	Male	Andhra Pradesh	Christian	Heart-related
Jhonny	19	Male	Andhra Pradesh	Christian	Viral infection

Table-3: Non-anonymized database

There are two regular techniques for accomplishing k-anonymity for some value of k.

a). Suppression:

In this method, certain values of the attributes are supplanted by an asterisk '*'. All or some of the values of a column may be replaced by '*'.

b). Generalization:

In this method, individual values of attributes are replaced with a broader category. For instance, the value '15' of the attribute 'Age' may be supplanted by '≤20', the value '23' by '20 < age ≤ 30', etc.

We shall see how this is applied to the non-anonymized database in the table-3. The following table-4 illustrates the usage of Suppression and Generalization methods of K-Anonymity.

Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Andhra Pradesh	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Andhra Pradesh	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Andhra Pradesh	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Andhra Pradesh	*	Heart-related
*	Age ≤ 20	Male	Andhra Pradesh	*	Viral infection

Table-4: 2-Anonymized table of the table-3 database.

In the above 2-Anonymized table we have, replaced all the values in the 'Name' attribute and each of the values in the 'Religion' attribute by a '*'. It is inferred that the values of the 'Age' attribute are generalized which is a type of anonymization technique as discussed above. Table-4 has 2-anonymity with respect to the attributes 'Age', 'Gender' and 'State of domicile' since for any blend of these attributes found in any row of the table there are always no less than two rows with those exact attributes.

K-anonymous data can still be vulnerable to attacks like unsorted matching attack, temporal attack, and complementary release attack. Therefore, we move towards L-diversity method of data anonymization.

b. L – Diversity

L-diversity technique of data anonymization tries to bring diversity in the sensitive attribute of data. It ensures that each equivalence class of quasi identifiers has at least L different values of sensitive attribute. It is a form of group-based anonymization that is utilized to safeguard privacy in data sets by reducing the granularity of data representation. This decrease is a trade-off that results outcomes in some loss of viability of data management or mining algorithms for gaining some privacy.

The *l*-diversity model is an extension of the *k*-anonymity model which diminishes the granularity of data representation, utilizing methods including generalization and suppression, in a way that any given record maps onto at least *k* different records in the data. Let's see this with an example.

Age	Sex	City	Income
24	M	San Jose	60,000
24	M	Denver	75,000
24	M	Denver	29,000
24	M	San Jose	100,000
26	F	San Jose	29,000
26	F	San Jose	84,000
26	M	San Jose	91,500
26	F	San Jose	88,000
32	M	Denver	45,000
32	F	San Jose	34,000
32	F	Denver	34,000
32	M	Denver	50,000

Table-5

Age	Sex	City	Income
24	Person	ncr	60,000
24	Person	ncr	75,000
24	Person	ncr	29,000
24	Person	ncr	100,000
26	Person	ncr	29,000
26	Person	ncr	84,000
26	Person	ncr	91,500
26	Person	ncr	88,000
32	Person	ncr	45,000
32	Person	ncr	34,000
32	Person	ncr	34,000
32	Person	ncr	50,000

Table -6

Table-5: base dataset. **Table-6:** Anonymized Dataset (Using Generalization), 3-Diverse Dataset.

In Table-5 'Income' is a sensitive attribute. For data to be L-diverse there should be at least L different values of income associated with each equivalence class. Table-6 shows 3- diverse version of table-5 since each equivalence class has at least 3 different values for sensitive attribute 'Income'.

The problem with this method is that it depends upon the range of sensitive attribute. If we want to make data L diverse whereas sensitive attribute has less than L different values, fictitious data is to be inserted. This fictitious data will enhance the security but may result in problems during analysis. Also, L-diversity method is prone to skewness and similarity attack and thus can't prevent attribute disclosure

c. T – Closeness

T-closeness is an advancement of l-diversity group-based anonymization that is used to preserve privacy in data sets by decreasing the granularity of a data representation. This reduction is a trade-off that results in some loss of adequacy of data management or mining algorithms in order to gain some privacy. The t-closeness model extends the l-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute.

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 t. A table is said to have t-closeness if all equivalence classes have t-closeness. The main advantage of t-closeness is that it prevents attribute disclosure.

Data anonymization can be applied to big data, but the problem lies in the fact that as size and variety of data increases, the chances of re-identification also increase. Thus, anonymization has a limited potential in the field of big data privacy.

Data Anonymization Method	Limitations	Computational Complexity
K - Anonymity	• It can suffer from Homogeneity attack and background knowledge attack • K -Anonymity is insufficient to prevent attribute disclosure	$O(k.\log k)$
L - Diversity	• L – Diversity is difficult to achieve • L – Diversity is insufficient to prevent attribute disclosure	$O(n\hat{}2/k)$
T - Closeness	• T- Closeness does not deal with identity disclosure	$2^{O(n)O(m)}$

Table 7: Comparison of Data Anonymization methods

2. Notice and Consent

The most common privacy preservation method for web services is notice and consent. Every time an individual access a new application or service, a notice stating privacy concerns is displayed. The consumer needs to consent the notice before using the service. This method empowers an individual to ensure his privacy rights. It puts the burden of privacy preservation on the individual.

When applied to big data, this method poses numerous challenges. In most of the cases, use of big data are unexpected or unknown at the time when notice and consent is given. This requires the notice to change every time big data is used for a different purpose. Also, big data is collected and processed so rapidly that it creates burden on consumers to consent the notice. A method by which notice and consent can be modified for big data is the use of third parties offering a choice of different privacy profiles.

3. Differential Privacy

Differential Privacy is a method enabling analysts to extract useful answers from databases containing personal information while offering strong individual privacy protections. It aims to minimize the chances of individual identification while querying the data. The method of differential privacy is shown in figure b.

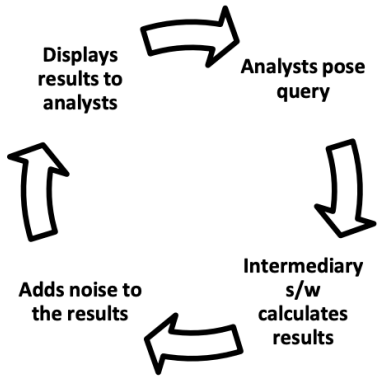

Figure b: Differential Privacy Process

As opposed to anonymization, data is not modified in differential privacy. Users don't have direct access to the database. There is an interface that calculates the results and adds desired inaccuracies. It acts as a firewall. These inaccuracies are large enough that they protect privacy, but small enough that the answers provided to analysts and researchers are still useful.

The advantages of Differential Privacy over anonymization are:

- The original data set is not modified at all. There is no need for suppression or generalization.
- Distortion is added to the results by mathematical calculations based on the type of data, type of questions etc.
- The distortion is added in such a way that value hidden is useful to analysts.

Conclusion and Future Work

Big data privacy has become an important issue since it is directly related to customers. It is now essential for an organization to promise privacy in big data analytics. Privacy measures should now focus on the uses of data rather than collection of data. They should be modified with respect to the size and unexpected uses of big data. In this paper, I have investigated the privacy challenges in big data by first identifying big data privacy requirements and also the privacy and security concerns in the big data environment. Privacy challenges in each phase of big data life cycle are presented along with the advantages and disadvantages of using current methods. This paper also presents traditional as well as recent techniques of privacy preserving in big data. Techniques like anonymization have limited potential when applied to big data. Notice and consent method also burdens the customer for ensuring privacy. Differential privacy may be seen as a viable solution for big data privacy.

As a future direction, perspectives are needed to achieve effective solutions to the scalability problem of privacy and security in the era of big data, especially to the problem of reconciling security and privacy models by exploiting the map reduce framework. In area of healthcare services, social media and web usage as well, more efficient privacy techniques need to be developed. Differential privacy is one such sphere which has got much of hidden potential to be utilized further. One problem with this method is that analyst should know the query before using the differential privacy model. When modified and applied to big data, it may ensure privacy without actually modifying the data.

References

[1] https://pdfs.semanticscholar.org/6f31/8a5d53e919574960eb6e84030c37f18ee2d1.pdf
[2] https://www.researchgate.net/publication/301716085_Protection_of_Big_Data_Privacy
[3] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0059-y#Sec40
[4] https://www.ijedr.org/papers/IJEDR1702165.pdf
[5] https://ieeexplore.ieee.org/document/8250688