Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

Big Data Security and Privacy

Muskan Parihar
M. Sc Data Science
Chandigarh University, Punjab

***Abstract*:- Earlier if we talk about 15-20 years back, data(traditional data) was limited because Social media, Online Transactions, E-Commerce, etc. was not in that much use and it was easy to store, process and protect the data due to its small volume and structured format, but day by day technology evolved following the world and new services get introduced due to which data generation increases which leads to the development of many techniques that can be used to store and process this amount of data. These technologies with their ability to extract information from large data sets for better decision-making process have created ways to maintain data, process data and new growth opportunities. But if data is not well protected from threats like phishing, hacking etc. all these processing becomes futile as if data falls in wrong hands, it could be misused. There are many ways to maintain data security and privacy but still it could be violated if not carried out properly. So while dealing with data, Security and Privacy becomes prime concern in order to protect it from attacks.**

Our purpose in this paper is to discuss the challenges faced while maintaining big data security and privacy and to explore some techniques that are used to deal with these challenges.

1. INTRODUCTION

We all are living in the era where data is generating rapidly, even the data generated in last 3-4 years is more than the data generated earlier in the whole century and it’s clear that future is filled with more data , So that means we are going to deal with a huge amount of data which keep on increasing.

This data is generating from various platforms like Social media, E- commerce websites, Stock market, data of particular organizations and list goes on. This data holds valuable information like how users share, view or engage with content in case of Social media; ratings and reviews of customers, preferences, shopping behavior, payment information, etc. in case of E-commerce; market data and trade related data in case of Stock market and so on. And our concern is to extract information from this much data for better decision making.

Big data is a technology that is used to analyze and extract the information from large and complex datasets such that traditional data-processing application software are incompetent to deal with them. Big data can be applied in various fields like E-commerce, banking, chemistry, data mining, cloud computing, finance, marketing, stocks, healthcare ,agriculture etc. It is a fast-growing field with exciting opportunities, that means with Big Data, there comes big opportunities but also big threats and challenges because of the volume and variety of data we are dealing with and Security and Privacy of data is one of them.

This security and privacy issue cannot be ignored and it is necessary to protect the sensitive personal information of customers from online criminals because data is an important asset to any organization and also for the millions of customers who trust such organization with their information.

2. CHALLENGES IN DATA PRIVACY AND SECURITY

Big Data is an emerging technology enabling organizations to have a greater insight into their huge amount of data for making reliable decisions for their business. The accumulation of data in big data systems also makes them a target for hackers. Organizations should be able to handle this data efficiently and must protect sensitive data so as to comply with the set of privacy laws. Securing and maintaining privacy of big data is difficult because of multiple reasons. Some are mentioned below:

1) **Complexity**: Earlier when big data didn’t come into existence, we used to store and process data in databases like Oracle and it was specific that how we need to deploy our database like for six to seven different servers and it was easy to protect them but with big data we deploy for example, hadoop system on thousands and thousands of servers, and because of this our infrastructure becomes much larger and it becomes difficult to protect.

2) With Big Data we get the value of our big data investments only by letting many analytics to have access of that data and if an outsider who could be some evil business rival or hacker, got the access then he/she would have many sensitive information and the problem is that for an outsider and because big data technologies do not provide an additional security layer to secure the data, gaining such access may not be difficult. Security of data often relies on perimeter security systems and if these systems are faulty then our big data becomes nothing but a low hanging fruit.

And it's not always an outsider, sometimes an organization may have corrupt IT specialist who can sell important information for their benefit as this Perimeter – based security are generally used for securing entry and exit points and what this IT specialist do inside the system remains a mystery and he/she can mine unprotected data.

3) **Velocity and Volume**: Data feeds of multiple types combined together come from different sources in real time with different security requirements. Like when dealing with high tech environments or retail or banking or simply while dealing with real world where decisions need to be made very quickly on data ,we are having real time security solution for alerting ,monitoring and blocking and sometimes these alerts leads to many false positives but in order to make decision quickly, we often ignore them which results to develop some loopholes.

4) **Environment**: If we look at big data environment it's not as simple as traditional database. It is difficult to manage as data warehouses store massive amounts of sensitive data such as personal data, financial transactions, etc. and it is required for Organizations and businesses to have a proper privacy and security infrastructure that enables only relevant data to be viewed by their employees.

3. METHODS USE TO DEAL WITHDATA PRIVACY AND SECURITY

There are many ways to deal with these challenges that come up with big data privacy and security. Companies often use different techniques like De-identification, Encryption etc. to deal with them so that no-one could extract sensitive information from the datasets.

1. **De-identification** is a traditional technique which is used to prevent someone's personal identity from being revealed [1]. This process is adopted as one of the main approach towards data privacy protection and it is commonly used in many fields of multimedia, communication, biometrics, big data etc.

To understand this technique, consider that we are having a medical data consist of certain attributes like Patient's name, Aadhaar no., Date of birth, gender, pin code, Visit date, Diagnosis, Medication, Total charge. So before releasing this data to third party we will de-identify the data by removing patient's name and Aadhaar no. and in this way sensitive information will be hidden from the third party. But it's not guaranteed that data is well protected here as if third party is having another election dataset of the people living in that area having attributes like Name, Aadhaar no, date of birth, gender, pin code, Party affiliation, and date voted. Even after removing patient's name and Aadhaar no from medical data, there are few attributes that are common in both datasets and i.e. date of birth, gender and pin code and this third party use this common attributes which are said to be quasi- identifiers as a source to connect both dataset and extract sensitive information like patient's name and disease diagnosed and this type of attack is known as Re-identification attack and an actual survey similar to this scenario was conducted in US which has uniquely identified the data of 87% population using this quasi-identifiers.

Quasi identifier is the minimal set of attributes which is used to uniquely identify the data that means only those attributes will be taken as quasi identifier which will be enough to extract hidden information from another dataset. And this Re- identification and linking attacks take place mainly because of quasi identifier. So in order to prevent our data from such attacks, we should mask the values of quasi identifier using suppression and generalization based anonymization methods.

In Suppression, data do not get revealed, it is hidden .We suppress quasi identifier using special symbol like asterisk '*' .For example we are having a data in which a person's age is 36 so to suppress age we can write age =3* which means that age lies in the range.

The quasi-identifiers are replaced with more general but consistent values in Generalization. For example age 36 can be replaced by age < 40.

But due to this reverse process of data Re-identification which can be used to identify individuals' data such that the sensitive information may be linked back to an individual using de-identified or anonymized data. The notions of k- anonymity, l-diversity, and t-closeness were proposed to prevent data from being re-identified.

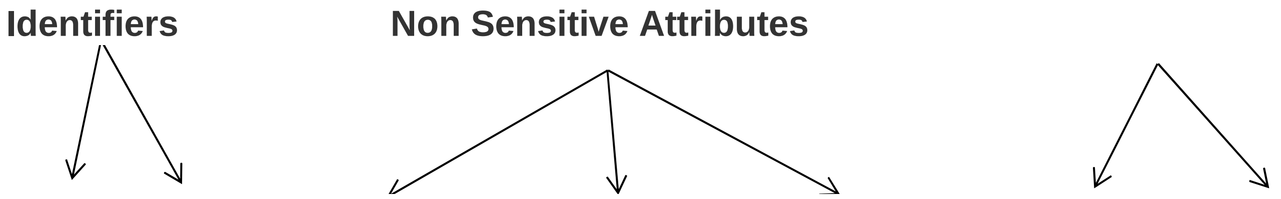
K-anonymity

Concept of *K*-anonymity was introduced to address the risk of re- identification of anonymised data. Latanya Sweeney introduced the k- anonymity privacy model in her paper "Protecting privacy when revealing information: k-anonymity and its implementation by generalisation and suppression" in 1998.

A dataset can have the *k*-anonymity property [1,2] if we coarsen quasi identifier attributes such that every tuple in the table shares its quasi identifier value with at least k-1 other values present in the table or simply we can say that in each release of data and at least k-1 individuals cannot perceive the information for each person in it. To understand *k*-anonymization problems, let's consider a table of a medical data having *10* rows and *7* columns, where each row represents a record relating to a certain individual, and the entries in the various rows do not need to be unique. The values in the columns are the values of the table's members' characteristics

Let us first have the complete table of a medical consist of 'Name', 'Adhaar No.', 'Pin Code', 'Age', 'Nationality' and 'Disease' and after removing key identifiers that are 'Name' and 'Adhaar No' ,we have another table which consists of remaining non sensitive and sensitive attributes.

Here's the Complete Table



Name	Adhaar No	Pin Code	Age	Nationality	Disease	Gender
Priyash	0000 0000 0000 0000	452006	29	Indian	Sugar	M
Ram	0000 0000 0000 0000	452002	24	Indian	Sugar	M
Rohan	0000 0000 0000 0000	452006	40	Indian	Flu	M
Pranjal	0000 0000 0000 0000	451010	43	Indian	Cancer	M
Aayush	0000 0000 0000 0000	451010	45	Indian	Blood Pressur e	M
Peter	0000 0000 0000 0000	452002	25	Indian	Sugar	M
Reyansh	0000 0000 0000 0000	451010	48	Indian	Sugar	M
Virat	0000 0000 0000 0000	452002	42	Indian	Cancer	M
Jai	0000 0000 0000 0000	452002	38	Indian	Cancer	M

Here's the table after removing identifiers

Pin Code	Age	Nationalit y	Diseas e
452006	29	Indian	Sugar
452002	24	Indian	Sugar
452006	40	Indian	Flu
451010	43	Indian	Cancer
451010	45	Indian	Blood Pressure
452002	25	Indian	Sugar
451010	48	Indian	Sugar
452002	42	Indian	Cancer
452002	38	Indian	Cancer

In k- Anonymitiy table will be gone through different de-identification techniques like suppression and generalization, So after applying these techniques we will have 3-Anonymous Table i.e. a table in which every row is repeating k-1 times i.e. 2 times in this case except that row and this trend can be observed in the whole table as

PIN CODE	AGE	NATIONALITY	DISEASE
452***	>35	*	Cancer
452***	>35	*	Cancer
452***	>35	*	Flu
451***	40-50	*	Sugar
451***	40-50	*	Cancer
451***	40-50	*	Blood Pressure
452***	<30	*	Sugar
452***	<30	*	Sugar
452***	<30	*	Sugar

Now due to this 3 Anonymous table, when medical data goes to third party, it would be somewhat difficult to determine the identity of the individuals in that data set.

We shift to L-diversity for data anonymization because K-anonymous data may still be vulnerable to attacks like homegeniety.

L- diversity

The l-diversity model can be thought of as an updated version of the k- anonymity model. It decreases the granularity with which data is represented. It also employs generalization and suppression techniques to ensure that any given record appears at least k times in the data. The l- diversity model addresses some of the sort-comings of the k-anonymity model like in case of homogeneity attacks and by additionally maintaining the diversity of sensitive fields of a dataset.

T- closeness

It's a more advanced version of L-diversity, and it's used to maintain data privacy by preserving the granularity of a data representation.

The dataset is likely to have t-closeness if the distance between the distribution of a sensitive attribute and the distribution of the attribute in the dataset is less than a threshold [6].

t- Closeness in addition to generalisation and suppression, helps one to use various anonymization strategies. For instance, instead of suppressing the entire record, some sensitive features of the record can be hidden; one advantage is that the number of records in the anonymized table is more accurate, which has numerous uses [6].

2. Encryption

The fundamental idea behind the concept of encryption is that whenever we transfer some encrypted information then our computer convert that data into a cipher text i.e. the result of simple text after encryption or simply defined as an unreadable output of an encryption algorithm which can only be put back into a readable form after decoding it. This concept is actually in use since long time even before computer was invented to send secret messages.

Even if you use applications like whatsapp, you can see that chat or whatever the information shared is encrypted i.e. no third party can have that information in between both ends and in this way data can be protected from hackers to hack data as there is complex Encryption algorithm behind these data which makes data difficult to decrypt. These algorithms ensure confidentiality of data. They also play an important role in keeping an eye on key security initiatives which includes authentication which verifies the origin of message , integrity which ensures that the content of message have not changed and non-repudiation that assures that the sender cannot deny the messages he had sent.

Applications of using Encryption:

1) Encryption Protects Privacy: Encryption is a technique for safeguarding confidential data, such as personal information. This helps to ensure anonymity and privacy and also reducing opportunities for surveillance by criminals. It also ensures the security of communication between client apps and servers.

2) Data is quite unsafe when travelling from one location to another i.e. the path through which data travels and Encryption works during this time only, ensuring no matter where data has been kept or how it isused.

So in short Encryption is also a solution for privacy and security issue in big data. It’s just that it works with the help of complex algorithm like Homorphic Encryption algorithm, Verifiable computation algorithm (outsource computing), Message digest algorithm, Key rotation algorithm, DES Algorithm and Rijndael Encryption Algorithm which also make it difficult for criminals to decode the data.

Other than these techniques De-identification and Encryption, we also have many other techniques that are working effectively in this field like data cryptography, differential privacy, end point filtration etc. Every technique has their own methodology to deal with security and privacy of data but at last each technique’s goal is to prevent data from attacks and misuse.

4. CONCLUSION

In this paper we come to know that no matter how advanced big data technology is, the very first priority is securing and maintaining the privacy of big data in order to protect data from malicious attacks, ensuring safety of data to stop it from falling into wrong hands. And there are many techniques which can be used to protect big data from such harms like we have gone through De-Identification, Encryption. There are other technologies also like Data Cryptography, End point filtration etc. which are used to deal with Data security and privacy using different ways and algorithms. But even after having so many techniques to secure the data, there are few shortcomings of these techniques too as the amount, source, type and speed at which the data is generating, it is quite difficult to protect it from attacks .So we need more advance techniques to deal with Big Data Privacy and Security to put an end to such frauds.

ACKNOWLEDGMENT

I would like to express my gratitude to my supervisor Dr. Bhagat Singh for his suggestions to improve the manuscript.

REFERENCES

[1] Jayesh Surana, Akshay Khandelwal, Avani Kothari, Himanshi Solanki, Meenal Sankhla, Big Data Privacy Methods 2017 IJEDR, Volume 5, Issue 2, ISSN: 2321-9939.

[2] Priyank Jain, Manasi Gyanchandani & Nilay Khare, Big data privacy: a technological perspective and review, 2016.

[3] Alex Bekker, Buried under big data: security issues, challenges, concerns, Head of Data Analytics Department, Science Soft 2018.

[4] M. Manikandakumar (Thiagarajar College of Engineering, India) and E. Ramanujam (Thiagarajar College of Engineering, India), Security and Privacy Challenges in Big Data Environment, 2018.

[5] N Li, S Venkata subramanian, T Li t-closeness: Privacy beyond k- anonymity and I-diversity 2007.