Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 01, January -2018

Data Privacy: Securing the Big Data

Rishabh Sinha[1]

[1]Research Scholar (MBA Tech), IT Department, SVKM's NMIMS MPSTME, Mumbai, India

Abstract- *Over thepast few years, the concern of data privacy and its security has been a major issue for most of the countries across the globe.The data breaches have had large implications to the society and one can clearly imagine the outcomes that it would have on the data integrity of the nation. Despite having strong data security mechanisms, there has always been incident of data breaches. The concern of data privacy is not a recent trend. There was a time when data was recorded and stored manually which lead to data duplication and often data losses. Such losses are also a form of data privacy breach. With everything going digital, large volumes of data can be stored in one go in multiple databases. When data is stored digitally, the risk of data duplication and redundancy can also be overcome. Storing data online has got several advantages over the manual methods of storing data.The data that is stored online can be accessed and replicated anytime and anywhere upon the ease of the users. Despite having several advantages, the online data is prone to severe disadvantages and threats as well. This paper discusses the advantages, disadvantages of storing data online. It also discusses the mechanisms that are required to store and secure the online data. The author also explains the various reasons breaching of data that is digitally stored.*

Keywords: *Data privacy; Data breach; Informationsecurity mechanisms; Big Data*

I. INTRODUCTION

Data breach and cybercrime are two words that are used interchangeably and are also closely related. But these two closely related terms are not synonymous. Data breach can be defined as an incident where "the name of an individual, or medical records or financial records (including the credit cards and debit cards) are potentially exposed to risk due to lack of secure mechanisms. The exposure of one' data can take place either through electronic means or through papers." Such loss of data can prove to be very damaging to an individual, irrespective of the means of data breach taking place through a hacker or any other method. On the other hand, cybercrime can be defined as "a criminal activity or a crime that uses Internet, a computer system or even a computer technology". In most of the cases, data breaches don't require the use of computers for achieving the goals. According to the International Organization for Standardization (ISO), data breach can be defined as "compromise of security that leads to the accident or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to protected data transmitted, stored or other processed".Humans produce a large volume and variety of data regarding individuals, things or the interactions that have occurred between them in the years passed by. This data can be replicated with ease at cheaper rates and are typically stored in online database, which enables easy searching of the same. An IBM report has estimated that nearly 2.5 Gigabytes (GB) of data is created every-day and the rate of data creation is continuously increasing. The continuous and daily emergence of data has led to the formation of a new concept related to data is known as Big-Data. Privacy is a term that can be used for both individuals as well as the society. In case of individuals, privacy can be defined as "the right of individuals for controlling his or her own data and to specify how these data are collected, used and shared".

II. CHARACTERISTICS OF BIG DATA

.

The characteristics of big data can be summarized as 7Vs. Out of these there are 3 major dimensions, namely- volume, velocity and variety. However, the other 'Vs' are useful for appreciating the real essence of big data and the effects. Each of these Vs is discussed below in brief.

i. <u>Data Volume:</u>

The big data is enormous and is continuously growing at very faster rates because of the fact that the Internet of Things has sensors across the globe and in all devices that are created every second. It is often estimated by Big Data specialists that the amount of data in 2020 would be nearly 50 times that was present a decade ago, in 2012. This enormous data is present in various forms such as images, videos, music and text data that's being uploaded by different users on various social media platforms.

ii. <u>Data Velocity:</u>

Velocity of data refers to the speed at which the data is being created, stored, analysed and visualized. With big data in the current scenario, data is produced in almost real-time. Due to the technological advancements, almost all the devices and machines, whether wireless or wired transmit as soon as the data is created. It is assumed that every minute around 100 hours of video is being uploaded on YouTube, more than 200 million emails are being sent, near about 20 million photos are being viewed, more than 3 lakh tweets are being tweeted and nearly 2.5 million searches is being performed on Google. This possesses are bigger challenge for organizations in order to cope with the enormous data generation speed and real-time usage.

iii. Data Variety:

With everything going digitized, the data so generated is highly varied. For extracting the meaningful information from the unstructured texts, images, audio and video from sensors into the IoT would require high computational and algorithmic power.

i. Veracity:

Veracity refers to the data that is verifiable and truthful in terms of originality. The huge potential of big data goes in vein if data analysis is carried out on inaccurate or an incomplete data, which are especially automated for decision-making or even providing data to unsupervised algorithms of machine learning. The inaccurate data would result in giving catastrophic results. As the data streams originate from diverse sources in various forms with variations in signal-to-noise ratios. These may be plentiful of accumulated errors, which are difficult to sort out when data reaches the Big Data analysis stage. Thus it is necessary to clean the data so that the veracity of final analysis isn't degraded.

v. Value:

For justifying the investments made on data collection, one should generate some value out of it, using the Big Data or traditional analytics tools. Various sites give recommendations to various sites based on the preferences of the users.

vi. Variability:

Variability refers to the data whose sense changes continuously. A unique decode challenge exists for limitless variability of big-data to correctly identify the sense of word by understanding the context.

vii. Visualization:

Visualization comes into play when the data is processed and presented in a readable and accessible manner. Techniques need to be adopted for representation of this enormous amount of data in an effective as well as efficient manner. One such method can be by converting the enormous data into graphical formats.

The 3V's, i.e. Volume, Velocity and Variety are inherent to the Big Data, whereas the other 4V's, i.e. Variability, Veracity, Value and Visualization reflect the gigantic complexity of big data regarding the ones who are going to process, analyse and gain benefit from it.

III. INFORMATION SECURITY IN BIG DATA

Information security often refers to the mechanisms that are used for protection of data. Information security can be summarized as the study and practices that are adopted for protection of data available in all forms and is stored in either an IT system or reduction in the paper/physical mediums. Further, information security can also be conceptualized as protection of data from all kinds of threats that might be perpetrated by the means of malicious outsiders or individuals that have the legitimate access to the IT systems and the data. The practice of data protection includes terms such as

confidentiality, integrity and availability.

i. Confidentiality:

It means protecting of data in all forms, from any unauthorized access throughout the entire lifecycle of the data (starting from data creation to data destruction). Unauthorized access of data includes the access by individuals that are not affiliated with the underlying organizations that store the data. It further includes individuals within the organization who have purposefully exceeded their authority of accessing information/ the term confidentiality is often implicated whenever an organization suffers data breach.

ii. Integrity:

Integrity refers to ensuring the data within the IT systems are accurate. The data in IT systems can be either recorded or reproduced on any physical media. The IT system creators and managers need to implement controls that are within the systems for ensuring that users enter and process the data correctly and are conflicting data elements are identified and resolved. Integrity further requires that only the authorized users have got the ability to change, move or delete certain types of data files. When data has integrity, it is considered to be an accurate one and hence can be relied upon for decisionmaking.

iii. Availability:

Availability refers to ensuring of data to be readily available whenever an IT system needs it. Data availability can be ensured by the stakeholders in numerous ways, including designing systems that can be redundant and resistant to any sort of foreign or malicious attacks. It also includes backing-up of data on a regular basis.

IV. PRIVACY CONCERNS IN BIG DATA

There are three major concerns about big data. These include the following-

i. Volume:

Volume refers to the amount of data that is stored, collected and used by various and happens to be one of the common concerns in context to the big-data. The concept of volume includes two aspects- the first one being the number of records that are being included in big data collection, and whether contribution is from institutions, agencies and other organizations. The question that is being asked in this aspect is **"how much information does one have?"**The second aspect of volume is regarding the number of data elements collected by an individual. The question being asked in this aspect is **"how much does one know about me?"**

ii. Sensitivity:

The sensitivity of data in big data collections is a major concern. There are two different concerns in this context.
The first one being that data sets may be included of different categories of data in any number of possible combinations. Some common types of data categories are identifiable, de-identified, anonymous and aggregate data.
The second sensitivity issue is related to the first one but it includes a nuance with reference to the identifiable data in collection of big data. Some of the data elements such as Social Security Number or SSN and name are the most sensitive data types because they can identify the individuals.

iii. Access:

Personal data is being collected and stored in one or the forms. This information pertaining to personal data is obtained through personal mobile devices and the Internet. This collected data are almost designed to be accessed by multiple entities from different locations serving different purposes.
The access concerns of personal data falls under the following two categories-

I. Protecting the data from external actors without any legitimate access to the data *and*
II. Protecting the data from internal actors such as individuals that have legitimate access to the system and intentionally exceed their scope of the pre-approved authority and can access the data from any unapproved devices or make mistakes and data is disclosed accidentally.

Most of the companies and organizations understand the important connection that exists between transparency, privacy and trust. It is a challenging task to be transparent in this data- intensive era. With the evolution of Internet of Things, a large number of connected devices don't provider a user interface for presenting the information to the consumers regarding data collection. The devices are increasing numerously, adding to the information mountain regarding the companies present to the consumer in terms of privacy policies. Companies providing connected devices should recognize that providing transparency should require creative thinking. Visual and auditory cues along with immersive apps and websites should be employed for describing to consumers in a meaningful and relatively simple way and the nature of information being collected.

V. TYPES OF ONLINE & EXTERNAL SECURITY THREATS

The different types of online security threats are-

i. Virus Threats:

Threat is a computer virus program that is written to alter the operation of a computer without any permission or knowledge of the user. A virus replicates and executes itself and brings damage to the host computer in the process.

ii. Spyware Threats:

This is a serious computer threat and is a program that can monitor ones online activities or even install programs without any consent for any profit or to capture the personal information.

iii. Hackers:

Hackers are the programmers who can victimize others for their own gain by breaking into computer systems for the purpose of stealing, changing or destroying the information of the host computer. This is usually considered as an act/form of cyber –terrorism.

iv. Phishing Threats:

This type of threat includes the attempt of stealing sensitive financial or personal information through the means of fraudulent emails or instant messages. When the host computer is connected to the Internet, it is subjected to attacks through the network communications. Some of the common phishing attacks include-

- *Bonk: an attack on the Microsoft TCP/IP stack that can crash the attacked host computer*
- *RDS_Shell: a method for exploiting the Remote Data Services component of the Microsoft Data Access Components that lets a remote attacker run the commands with system privileges*
- *Win Nuke: an exploit that can use the NetBIOS for the purpose of crashing the older Windows computers*

v. Viral Web Sites:

Users can be forced to visit certain websites, often by the means of email message. These sites might be the source of viruses or Trojans. Such sites are known as viral web sites and are often made to look like the some well- known web sites and also can have similar web addresses to the imitating sites. Users are forced into the trap often by visiting the sites, downloading and running virus or Trojan, which can infect the host system and ultimately becoming the subject of hacker attacks.

vi. Spyware, Adware and Advertising Trojans:

These are often installed with the other programs without any consent of the users. They can record the behaviours of the users on the Internet, displaying the targeted ads to the user and lead to downloading other malicious software on the computers. These are often included in the programs, so that one can download freely from the Internet or sometimes along with CDs that are given free with magazines. Spywares usually don't carry viruses, but they can use the system resources and thereby slowing down the Internet connection with the displaying of ads. If the spyware contains any bug or fault it can make the computer unusable, and marking the main concern as privacy. There are some spyware that can download more serious threats on to the computer, such as the Trojan Horses.

vii. Unsecured Wireless Access Points:

In case of a wireless access point hasn't been secured then anyone having wireless devices such as laptop, PDA etc. will be able to connect to it, then it can it can access the Internet and all the computers on the wireless network.

viii. Bluesnarfing:

Bluesnarfing is the act of stealing personal data, especially calendar and contact information through a Bluetooth enabled device.

ix. Social Engineering:

Tricking the computer users to reveal the computer security or any private information such as passwords, email addresses by exploiting the natural tendency of an individual to trust.

VI. TECHNOLOGIES FOR PRIVACY PROTECTION

In a big data environment, the technology for privacy protection is mainly studied form the following perspectives-

- User privacy protection
- Data content verifiable *and*
- Access control

Following are the technologies that are used for data privacy protection-

i. Anonymity Data ProtectionTechnology:

In a big data environment, anonymity protection is necessary for protecting the data. Considering the case of social media, anonymity protection can be classified into user identity anonymity, attributes anonymity and relationships anonymity. User information and user attributes must be hidden when it is published and the relationship anonymity should hide the relationship between users and data upon their release.

ii. Data Watermarking Technology:

Data watermarking refers to the identification information that is embedded imperceptibly within the data carrier and these doesn't affect the method of its usage. This is usually used for copyright protection of multimedia data and also there exists a watermarking scheme for databases as well as the text files. The characteristics of randomness and dynamic data make the watermarking methods very different on the marked database, document and the multimedia files.

iii. Data Provence Technology:

Due to the diversification of data sources, it becomes necessary to record the origin and dissemination process for providing additional support to the dissemination process in terms of decision making and decisions. Data provence method is a labelled method and through the label, one can know which data in the table acts as a source. One can easily check the correctness of the result or update the data with a minimum price.

VII. PRIVACY AND CONFIDENTIALITY

i. Privacy:

Privacy is the state of an individual's data which is free from public interruption and intrusion. A data can be said to be in the privacy state, when it is apart from public attention and observations. A virtual boundary can be drawn around the data for ensuring the information access from the others usage. A data is said to be in the privacy mode, when an individual doesn't wish to disclose his/her information in front of the others.

ii. Confidentiality:

Confidentiality refers to the state of data when it is either intended or expected from someone to keep it a secret. A confidential data can be entrusted when an individual doesn't want to disclose his information/data to any unauthorized person.

VIII. DIFFERENCES BETWEEN PRIVACY AND CONFIDENTIALITY

Privacy and confidentiality are two interchangeable terms that are used in data protection. Privacy is considered to be more personal or private, whereas confidentiality is considered to be more of professionalism.

#	Privacy	Confidentiality
1	The state of being secluded is known as privacy	It refers to the situation when its expected form someone, that he will not divulge the information to any other person
2	It is the right to be let alone	An agreement between the persons standing in fiduciary in order to maintain the secrecy of sensitive information and the documents
3	Limits the access of the public	Prevents unauthorized access to the information as well as documents
4	It applies to individuals	It applies to information
5	It is the personal obligatory or choice of an individual	It is an obligatory when there exists a professional and legal information

Table 1: Differences between Privacy and Confidentiality

IX. TYPES OF NON-TECHNICAL SECURITY THREATS

i. Insider:

An insider can be defined someone who has got the legitimate access to the network. As the information is accessed by the insiders, it can be easily stolen, copied, deleted or changed. The insider threats can be at times damaging regardless whether they have occurred due to user carelessness or malicious attempts.

ii. Poor Passwords:

Keeping the strong user passwords is critical in terms of data protection. It is important for users to have the access for most of the sensitive information. The modern password cracking programs can easily crack the weak passwords. These types of passwords include the common words or word groups that can be found in a dictionary. Due to the mentioned reason, user passwords are considered to be much weaker than the randomly-generated passwords. The user-generated passwords can follow a predictable pattern or association with something in the life of user and thus become more vulnerable to crack the passwords. On the contrary the randomly generated passwords are difficult to remember as well as guess and also they can't be cracked easily.

iii. Physical Security:

Physical security is essential for preventing unauthorized access to the sensitive data as well as protecting the organization's personnel and resources. An effective physical security system is an integral part of a comprehensive security program. Physical safety measures includes securing the access to dedicated computers, server rooms, routers and all those areas that are responsible for processing or storing sensitive data.

iv. Insufficient Backup and Recovery:

Lack of robust data backup and recovery and solution puts the organization's data at risk and also undermines the effectiveness of related IT operations. Data and system recovering capabilities allow the organization in reducing the risk of damages that are associated with data breaches. Thus, it becomes essential to conduct regular backups of critical data and store the backed-up data in a safe and secured manner.

v. Improper Destruction:

Paper documents including reports and catalogues might contain some sensitive data. Until and unless these documents are properly destroyed by the means of shredding or incinerating, these documents can be salvaged and misused. Discarded electronic devices including systems and external memory sources that were once used for processing and storing sensitive data are also vulnerable, till the time all the stored data is erased properly. Data breaches might take place if recovery tools are used for extraction of an improperly erased or an overwritten data.

vi. Social Media:

Use of organization' devices and network resources for accessing the social media websites also poses a high level of data security. Social networking sites are often targeted by malware, which represents a high degree of spam and are often used for gaining information for the purpose of identity theft.

vii. Social Engineering:

Breaking into a network doesn't require any technical skills. Accessing the sensitive information can be gained by manipulation of legitimate users and their trust. Caution should be advised whenever communicating to any account or network information, Socially engineered attacks are the means for some hackers for gaining passwords, access codes, IP addresses, router or server names and any other type of information that can be exploited for breaking into a particular network

X. CONCLUSION

The data privacy concern is not a new issue. It has been the major concern of data analysts over the past many years. When data was being stored manually, the theft of data was a concern, and now with everything going digitized, the foreign and internals threats as discussed in the paper have become a major concern. Despite using strong data encryption mechanisms, there are cases of data breaches, which in turn have resulted in loss of severe and important data. Every day large amount of data is generated by users in diverse forms and these needs to be secured using proper and strong encrypting mechanisms. Crucial information of any nation such as the army and artillery details, if it gets leaked, then it can become a threat for the nation and make the nation vulnerable to any kind of attacks. In case of a private data breach, there can be psychological extortion to the user, wherein one can torture and demand for some of the most crucial information.

Thus it becomes essentially important for IT data systems to secure data from both internal as well as external threats.

REFERENCES

[1]Dr. PuneetGoswami, Ms. Suman Madan: "A Survey on Big Data & Privacy Preserving Publishing Techniques" at Advances in Computer Sciences and Technology (Vol. 10, Number 3, 2017) pp.395-408

[2]Ira S. Rubinstein: "Big Data: The End of Privacy or a New Beginning?" at International Data Privacy Law (Vol. 3 No.2, 2013)

[3] Julio Moreno et. Al: "Main Issues in Big Data Security" at future internet (2016)

[4] Elisa Bertino: "Data Security-Challenges and Research Opportunities"

[5] "Data Security: Top Threats to Data Protection" by Privacy Technical Assistance Center (December, 2011)

[6] Lei Xu et. al: "Information Security in Big Data: Privacy and Data Mining" at IEEE (Volume 2, 2014)

[7] AbidMehmoodet. al: "Protection of Big Data Privacy" at IEEE (2016)

[8] Gang Zeng: "Research on Privacy Protection in Big Data Environment" at International Journal of Engineering Research and Applications (Vol.3, Issue 5, May 2015)

[9] Ateeq Ahmad: "Types of Security Threats and It's Prevention" at International Journal Computer Technology and Applications (Vol 3(2))