DOI: http://dx.doi.org/10.26483/ijarcs.v11i2.6522

ISSN No. 0976-5697

Volume 11, No. 2, March-April 2020

International Journal of Advanced Research in Computer Science

SURVEY REPORT

Available Online at www.ijarcs.info

A SURVEY ON DATA PRIVACY THREATS AND PRESERVATION TECHNIQUES

Lalita Takle
Electronics Engineering
D.J Sanghvi College of Engineering, Vile Pae
Mumbai, India

Mihir Sircar
Computer Engineering
Atharva College of Engineering, Malad
Mumbai, India

Advait Tare
Computer Engineering
Atharva College of Engineering, Malad
Mumbai, India

Abstract: Sublime measures of Data are being produced by online business, different applications, banks, schools, and so forth by virtue of advanced innovation. Pretty much all industries are endeavoring to adjust to this gigantic data. Big Data phenomenon has begun to gain noteworthiness. However, this massive amount of data may also lead to many privacy issues, which make Big Data Protection a major concern for any organization. Privacy Preservation methods are becoming progressively significant in order to tackle such security issues. Therefore, the paper investigates different security dangers and, in this manner, expresses the techniques for their avoidance. A general point of view for privacy preservation has been recommended.

Keywords: Big Data, Data Mining, Privacy Risks, Privacy Protection, Information Security.

I. INTRODUCTION

With the increasing amount of modernization and the use of technology, a huge amount of data is created every minute. Every aspect in our day-to-day life is connected with the internet in some way or the other, be it education, banking, appliances, vehicles, e-commerce and so on. The amount and variety of data is increasing exponentially as a result of numerous computer applications in almost every field. This enormous data is termed as Big Data and its proper analysis and processing is of utmost importance. Big Data when processed aptly has proven to be quite significant in boosting the growth of various sectors. As stated in [1],recommendation systems that are commonly seen by web based premises such as Amazon, Flipkart, to advise goods to consumers according to their buying habits are some of the most notable implementations of Big Data Analysis. Facebook suggests 'people you may know', spots to visit as well as film recommendations based on our interests[1] The vast amount of information, that additionally involves personal and confidential information such as identity, postal address, sickness, religion, shopping cart and so on, can be accessed by anyone. Information collectors may share such information with data analysts who then study this data and distinguish essential data that may help to improve organizations, offer incentives to clients, foresee and advise. Discharging user activity data, however, can lead to statistical assaults, such as recognizing user identity according to his/her online activity. We have considered different strategies for preserving privacy that are being used to safeguard against potential threats. This paper examines each of these strategies and likewise gives a statistical data to underline on security concerns.

So as to serve our potential readers with various degrees of need, we have organized the paper as follows. We examine the Definition and Features of Big Data in Section 2. The Privacy Threats in Big Data are introduced in Section 3. In Section 4, we have examined the different Privacy Preservation Methods. One can allude to the Analysis and Statistics in Section 5 and 6 individually. Finally, we outline the paper in Section 7.

II. DEFINITION AND FEATURES OF BIG DATA

According to [2] in resources, Big Data is an area that discusses ways of analyzing, intentionally extracting data from or in any case handling extremely large or complex information sets that can be managed using conventional application software for data processing. Its complexities include the accumulation, storage, investigation of data, search, transmission, representation, demand, update, data protection and source of information. Big data is related to five essential concepts: volume, velocity, variety, veracity and value. However, we cannot essentially watch and monitor what happens at the moment when we handle Big Data. Big Data also contains data of sizes beyond the traditional programming limit to process within a fair amount of time and value. The present use of the term Big Data generally refers to the use of precious statistics examining user behavior or other propelled methods of data analysis which are used to obtain an incentive from knowledge and occasionally to a particular dataset.

Big Data can be depicted by the following attributes as per the cited works in [3] :

i. **Volume** - Big data volume is the amount of data generated and processed. The size of the data decides the importance and possible insight and whether or not it can be classified as Big Data [3].

ii. **Variety** - Variety in Big Data refers to the type and quality of knowledge that allows individuals processing it to use the subsequent understanding in a viable manner. Big Data draws from text, pictures, sound and video; and finishes missing parts by integrating information [3].

iii. **Velocity -** Variety in Big Data refers to the type and quality of knowledge that allows individuals processing it to use the subsequent understanding in a viable manner. Big Data draws from text, pictures, sound and video; and finishes missing parts by integrating information [4].

iv. **Veracity** - For Big Data, it is the all-inclusive term that alludes to the quality of information and the esteem for information. The quality of the information collected will fluctuate enormously, affecting the exact analysis [4].

v. **Value** - The majority of Data having no Value is of nothing more than a bad memory to the organization, except if it is transformed into something helpful. Information in itself is of no utilization or significance. Rather, it should be changed over into something important to extract Information. Consequently, one can express that Value is the most significant V among the 5V's.

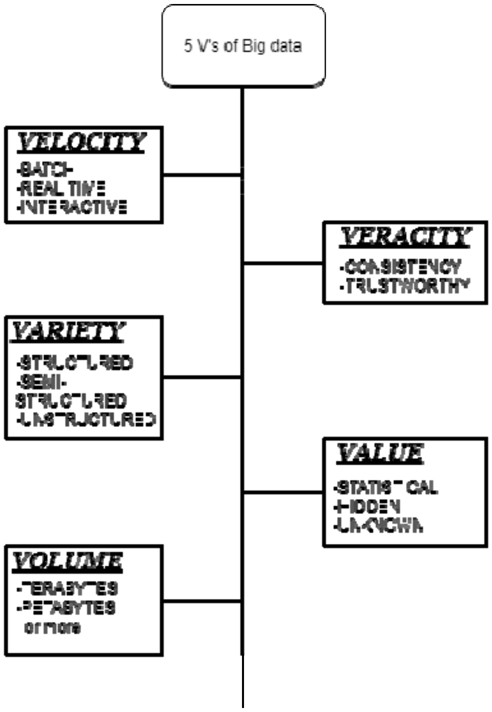


Figure 1. 5V's of Big Data

III. Risks In Privacy Of Big Data

In accordance with the work cited in [5], Privacy is an individual or group's right to keep their lives and personal affairs out of public view, or to monitor the flow of knowledge about themselves. In the event that the information is made public, there is a danger to the privacy of individuals, as the information is in control of the data holder. Having said that, this turns out to be favorable for the advertising agencies and marketing associations as they get a simple method to track you. Likewise, this furnishes them with much approved data in a more affordable manner. Likewise, this furnishes them with much approved data in a more affordable manner.

Through Big Data, administrations build up a profile of their buyers with much extraordinary and precise data. Making online activity a standard, they could pass judgment on personal likes and interests, for example, contribution in governmental issues, travelling preferences, social propensities and different things of different individuals. This may assist them with increasing individual data about the people belonging to a particular nation as well [6].

However, most associations guarantee their conduct as a stage to improve the client's online experience. In spite of the fact that it's very evident that such tracking could be utilized in a negative manner. For example, this could lead the insurance agencies to scrutinize the clients about inclusion dependent on these big data profiles. However, this practice is as of now been begun. Yet, this issue would never be tackled by limiting big data collection. At this period of Big Data and technological progression one can't deny the reality whether it appears to be helpful or not. Therefore, the real and valid methods for Big Data storage ought to be created so as to ensure the security and privacy which could prompt a protected and gainful practice. For example, the identification of malicious activities through authentic big data collection could be made a lot simpler [7].

The disclosure of data gathered and the reason for which it would be utilized could wipe out numerous privacy issues. Therefore, big data handlers should open such data to explain big data protection and security challenges.

Notwithstanding, the most significant component for the clients is to know how the information is been gathered, who can get to it and how the access is taken. Likewise, for the associations it is important to clarify the security technique they use for keeping up the client's gathered information. Through this the ventures could guarantee their client's trust.

As rightly cited in [8], Big Data’s most formidable security challenges :

i. **Inaccurate Data**: Cyber criminals will generate data and 'dump' it into one's data lake in order to intentionally compromise the efficiency of big data research. For instance, cyber criminals can infiltrate your network and make your sensors show false results. If a manufacturing company uses such data, it may result in failure in the production process. In this way, one would lose the chance to repair issues unless one suffers significant harm. With fraud prevention, these problems could be overcome.

ii. **Involvement of non-native mappers**: Upon gathering the big data, parallel processing takes place. MapReduce framework is one of the approaches used here. When a mapper processes this data and assigns it to other storage resources, it is split between various bulks. In the event that an outsider approaches one’s mappers' code, it is possible for them to fabricate the settings of the current mappers or include 'non-native’ ones. The issue here is that, getting access might not be too difficult because large data systems do not necessarily have an additional data protection layer.

iii. **Risk of sensitive data mining**: Typically, Big data security uses perimeter-based safety which means both 'entry and exit points' are protected. Even then systems succumb to infiltrations. Such failures may often cause rival companies to mine unprotected data. If this information is related to the launch of a new product / service, financial operations of the company or to personal user data, then the company can incur huge losses. Data can be made secure by adding additional perimeters or via anonymization.

iv. **Information Provenance**: The provenance of information makes matters even more complicated. Since its role is to document and store the origin of a particular data, we can only imagine how big a metadata set can be. To store such tons of information and all its corresponding manipulations is indeed a complicated task. Thus, we can say Information Provenance is an unavoidable concern in Big Data

v. **Inadequacy in security strength of NoSQL**: NoSQL databases are now a prominent big data science phenomenon and it is its popularity which is causing problems. Logically, NoSQL databases are constantly being developed with new features, even then, security is abused and ignored. With the advancement in databases, it is expected that security too would be upgraded. Yet, at this point too, it is most frequently overlooked.

vi. **Failure in conducting audits**: Safety audits in Big Data allow businesses to become aware of their safety vulnerabilities. While this suggestion is advisable to follow periodically, it is seldom implemented in practice. Big data research already possesses numerous problems and issues, audits will only aid in further increasing these complications. Furthermore, these audits are impractical due to the lack of time, money, trained staff or clear understanding about business security requirements.

IV. Strategies In Privacy Preservation Classification

There are numerous strategies for privacy security of data mining; our privacy protecting classification techniques are dependent on the following aspects, for example, data distribution, data distortion, data mining algorithms and privacy protection [9]. Detailed explanation of which is provided below:

i. **Distribution of data**: At present, some algorithms perform data protection on centralized data while some implement the same on distributed data. Distributed data consists of vertical partitioned information. Databases record data as horizontally or vertically where vertically partitioned data comprises of attribute values[9].

ii. **Algorithms for data mining**: An algorithm in data mining is a collection of algorithms and calculations that construct a data model. The algorithm analyzes the data you provide in order to construct a model, searching for specific patterns or trends. The algorithm uses the study results through a series of iterations to determine the best parameters for the development of the mining model. Such parameters are then applied to derive operable trends and accurate statistics in the entire data collection[9].

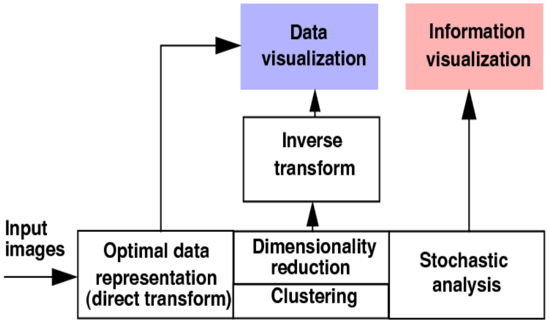


Figure 2. Block diagram of Data Mining tools

iii. **Algorithms for data mining**: An algorithm in data mining is a collection of algorithms and calculations that construct a data model. The algorithm analyzes the data you provide in order to construct a model, searching for specific patterns or trends. The algorithm uses the study results through a series of iterations to determine the best parameters for the development of the mining model. Such parameters are then applied to derive operable trends and accurate statistics in the entire data collection[9].

iv. **Protection of Privacy**: There is a need to carefully adjust data to achieve a high-data utility in order to ensure security. The following are some explanations for these. (1) Adjust data according to the dynamic algorithm strategy, and only alter the values chosen, but not all the values that make data loss the least of all. (2) Advances in encryption facilitates secure multiparty communication. In the case where each site only knows its input and output, the measurements are safeguarded. (3) The technique of data reconstruction may replicate original distribution of data from arbitrary data [9].

V. Privacy Protection Approaches

Numerous Privacy preserving systems were developed, yet a large portion of them depend on anonymization of information. The list of privacy preservation methods is given underneath.

A. Anonymization

Data anonymization is the process of erasing or encrypting the identifiers linking individuals to stored data to protect privacy or sensitive Information [10]. For example: you may use an anonymization method that preserves the data but keeps the source anonymous to run personally identifiable information (PII) such as names, social security numbers and addresses. According to sitography stated in [11], different methodologies to anonymize data are given below :

1) Data anonymization techniques

Screening the data is an easy way to protect the security of your personal information while using data for predictive modelling or compilation. Scrubbing basically deletes personal data such as name, address and date of birth. Nevertheless, it might be possible to use cross referencing this with public data or other sources for filling out the "missing holes" in the scrubbed dataset. The classic example was that a MIT student found an individual with scrubbed health records by cross-referencing them to voting records. A further common method used for anonymizing sensitive information is tokenization which includes replacing personal information such as a name with a token, for example, using a numerical representation of that name. The token may nevertheless be used as a reference to the original data.

2) Extensive data anonymization methods

Differential privacy and k-anonymity are more advanced approaches that help solve the de-anonymization of data.

a) Differential privacy

In order to mask personal identifiable information, differential privacy utilizes mathematical mechanisms to add random noise to the original dataset while allowing the same query across the original dataset to be returned with probabilistic effect to similar search results. An analogy tries to cover a panda with the head of a horse; only making enough camouflage to make sure it isn't a panda. When asked, the counts of toys that belongs to the secret panda are revived without a single panda toy being seen. For instance, with iOS 10, Apple began using differential data privacy to expose user behaviour and activity patterns without identifying individual users. This allows Apple to examine transactions, history of web surfing and health details while protecting its privacy

b) K-anonymity

In addition, K-anonymity adds data. The method is to look for a certain k number of persons containing the same identifiable combination of attributes, so that a person in this group is hidden. Identifiable information as the age can be generalized to replace the age with an approximation of less than the age of 25 years or more than the age of 50 years. However, failure to randomize sensitive data to mask means that k-anonymity may be vulnerable to hacking

B. Randomization

The technique of randomization uses methods of distorting data to add some noise to the original data. The recovery of individual values is longer by adding noise to the original data but the only thing that can be recovered is the aggregated distribution cab[12]. This method provides a proper equilibrium between the protection of privacy and the discovery of knowledge. [13] allows for some kind of disturbance. The data records are complemented by additional perturbation, randomized noise.[13]. It is also possible to get the data distributions from the random records. In order to disturb records, multiplying disturbance, random projections of random rotation techniques or random rotating techniques are used. This method includes random disturbance and a randomized reaction scheme based on noise. This results in an effective and high-quality method of information.

C. Data Distribution

The information is conveyed on different websites in Horizontal as well as Vertical manner. (Fig3)

1) Horizontal distribution

a) Support a group of analysts.
b) Has access to total of atomic information for all clients.
c) Analysts compose ad-hoc queries.
d) Queries take seconds to minutes to obtain results.
e) Is for the most part cluster stacked.
f) Stability of stack generally doesn't affect the core product, so having a day down time won't cause issues to the end user.

2) Vertical distribution

a) Supports a huge number of clients.
b) Has access to just their domain of information.
c) Queries are pre-canned, or have constrained scope for adjusting.
d) Is sort of real time.
e) Stability of stack impacts the core product, as it's a client facing feature.

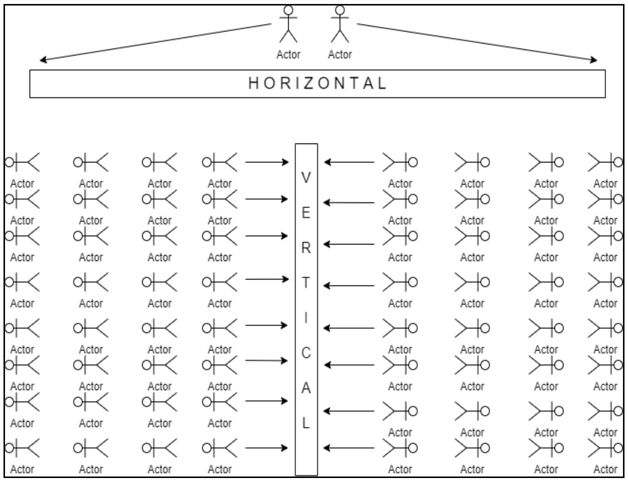


Figure 3. Horizontal and Vertical Distribution

D. Cryptographic techniques

In the presence of an enemy, cryptographic methods are used to maintain data confidentiality and integrity. Different cryptographic methods like symmetrical key encryption or public key cryptography may be used for the transport and storage of data, depending on the security needs and threats involved[14]. Furthermore, homomorphic encryption enables various calculations on encrypted data without the need for computation to decrypt data. From the point of view of privacy, this technology is useful to shield your sensitive information from being leaked from transport and storage servers[14] . Various cryptographic techniques are given as stated by source links [15].

1) Symmetric key encryption

It is often known as hidden key encryption. This is an encryption technique with a hidden key exchanged by all parties. The information is encoded and decoded by a common key. The sender encrypts data with the key. The receiver would use the common key to decode the message when sent. Today, all protected live traffic such as live phone discussions; video sharing, etc. are encoded using symmetrical encryption algorithms[15].

2) A symmetric key encryption

Different keys are used to encrypt and decrypt a message in asymmetric key encryption. The most useful asymmetric key algorithms are those where no key can be released while the other is secure. This public-key-private-key system, also known as public-key encryption, has many major benefits. The need to pass hidden keys to other users is avoided. So, for both authentication and encryption, the algorithm can be used. In the processing of big database blocks, asymmetric cypher algorithms are based on strong mathematical operations[15].

3) Monoalphabetic figures

The character of the plain text is continuously changed in the ciphertext in monoalphabetic substitution irrespective of its role in the text. When the algorithm, for example, says letter A is modified to letter D in the plaintext, in other words, there is an on-going interaction between letters in the plaintext and the ciphertext. For example: the lowercases are used to display the plaintext; the uppercases are used to display the ciphertext. The cipher is possibly monoalphabetic, because both 1's are encrypted as 0's [15].

4) Caeser ciphers

Caesar Cipher is a special case of substituting techniques where an alphabet three places down a line is substituted for each alphabet of a letter. Caesar cipher is only vulnerable to an attack by statistical ciphertext. To interact with his soldiers, Jules Caesar used this additive cipher. Thus, additive ciphers are called as Caesar cipher. For his correspondence, Caesar cipher used a 3 key. For instance, let "hello" message be encrypted with a 15 key. All characters are then modified from 15 in the ciphertext. Then the answer is hello = WTAAD[15].

- ➢ Mathematically,

According to [15],

C (ciphertext) = (P+K) mod 26 (K= key)

P (plaintext) = (C-K) mod 26

To encrypt message “hello”

Plaintext ‘h’= 7 encryption (7+15) mod 26 C= 22=> W

Plaintext ‘e’= 4 encryption (4+15) mod 26 C= 19=> T

To decrypt “WTAAD”

Ciphertext W=22 decryption (22-15) mod 26 P= 7=>h

Ciphertext T=19 decryption (19-15) mod 26 p= 4=>

5) Multidimensional sensitivity based anonymization (MDSBA)

Bottom up Generalization and Top down Generalization are the standard anonymization approaches which have been used for well- standardized data logs. However, it is very difficult to apply the same to large sets of data, which leads to scalability and loss of information. Anonymization based upon multidimensional sensitivity is an enhanced form of anonymization which is more powerful than the traditional method of anonymization. Anonymization-based multidimensional sensitivity is a stronger anonymization strategy, which can be implemented with reduced data loss and predefined quasi-identifiers in large sets of data. Apache MAP REDUCE was used for handling large data sets as part of this process. The data will be broken down to blocks of sixty-four MB or 128 MB of each and spread over various nodes without considering the details inside the blocks in the traditional Hadoop Distributed Files System.

The content, as part of the Multidimensional Sensitivity Based Anonymization, is part of different packages depending on the probability distribution of semi-identifiers by using Apache Pig's scripting channels. Multidimensional Sensitivity Based Anonymization also uses bottom up generalization of some class value characteristics where class represents a sensitive attribute[16]. As contrasted with traditional block approaches, data distribution was successful. Four semi-identifiers using Apache Pig were used to anonymize data. Since data is divided vertically into several classes, the context information attack can be covered if the bag contains only a few attributes. It is also difficult to map the information from external sources so that precise information can be revealed to any person. The implementation using Apache Pig has been achieved in this way. Apache Pig is a language to write; thus, the efforts to improve are less. Nonetheless, as compared to Map Reduce jobs, Apache Pig's code effectiveness is comparatively smaller since any Apache Pig script must eventually be converted to a Map Reduce job. Multidimensional Sensitivity Based Anonymization is ideally suited to big data but only if the data is at rest. Multidimensional Sensitivity Based Anonymization for data streaming cannot be implemented [17].

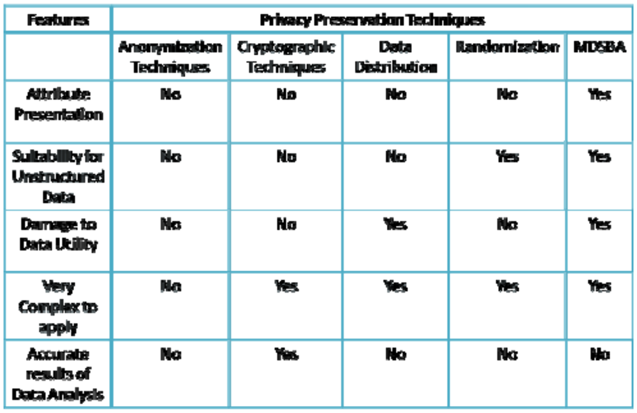
VI. ANALYSIS

The above Strategies for Protection of Privacy are analyzed below in the Table 1. Perhaps the most ideal approaches to ensure against big data security dangers is to comprehend the risks and execute measures to diminish probable occurrences.

- Remove unnecessary data: Many organizations reserve the entirety of their information, yet a portion of this can be casted off once the association recognizes which data is the most valuable.

- Holding onto pointless information expands risk, so organizations ought to set apart what is required to draw new experiences and expel the rest

- Strengthen back-door security: Just as the home security, the back-door of technology systems is once in a while also protected as front access zones. In organizations, this implies areas where data is stored however isn't a part of a 'live' system. For instance, data could be reproduced in test situations or disaster recovery platforms, which are less likely to have comprehensive protection



Features	Privacy Preservation Techniques				
	Anonymization Techniques	Cryptographic Techniques	Data Distribution	Randomization	MDSBA
Attribute Presentation	No	No	No	No	Yes
Suitability for Unstructured Data	No	No	No	Yes	Yes
Damage to Data Utility	No	No	Yes	No	Yes
Very Complex to apply	No	Yes	Yes	Yes	Yes
Accurate results of Data Analysis	No	Yes	No	No	No

Figure 4. Analysis

VII. Results And Discussion

As stated in survey cited in [18],57% of the world's population had access to the Internet in January 2019. However, it does come at a price to be linked to the world's largest knowledge base. Whilst some see privacy loss as a necessary evil, others see it as an injustice they seek to avoid, or at least mitigate, through various strategies. Worldwide, as compared to a year earlier, 53% of internet users are more concerned with their privacy [18]. As the Internet is fast and the state of cyber-crime is changing, it is difficult for policymakers and technology developers and individual users to keep up with the increasingly innovative ways of online communication.

81% of internet users in the United States thought their data was highly or very vulnerable to hacker data[18]. The most common concern among internet users is the exposure of sensitive personal data such as credit card data or social security data that could lead to robbed ID and financial loss. 22% of internet users in the USA reported during a March 2019 survey that storing confidential information internet was not safe enough for them, and 40% said they have been misused [18]. The presence of many governments in their own country's online activities has become a further source of concern for internet users. These concerns in the United States were further strengthened as the 2013 leak of sensitive information on the United States National Security Agency (NSA) and the global surveillance of foreign nationals and US citizens by its international allies.

Several regular web customers and IT experts argue that government should not have access to encrypted information systems, and multiple internet users have employed encryption methods to cover up their data from the government. As per [18],in terms of administration and protection of personal information, only 24% of global online customers have substantial trust in their governments. In comparison, national security authorities legitimize their ability to access personal data from users in view of terrorism incidents from previous years, and force device vendors to work on codes allowing access to those data during suspicious behavior. Given this situation, it is hardly surprising that 66% of internet users said they were more concerned about their privacy online than of their own government in the February 2019 survey. Overall, 49% of respondents in Europe knew very much or a little bit about their internal data security and data privacy laws but just 29% of North American respondents said the same thing[18].

The malicious use or release of personal information aimed to humiliate, threaten or otherwise harm someone's reputation is another growing cause of concern in the online medium. A survey carried out in [18], December 2018 found that 21% of the victims of online abuse faced online harassment because of politics, and 20% reported online hate because of gender. With attacks including threats to physical safety and sexual abuse online, it is no wonder that most internet users prefer to continue to protect private information. In spite of these concerns for online privacy, a significant number of online users are willing to consider certain privacy risks for comfort. Despite these online privacy concerns, a large share of online users is willing to accept certain privacy risks in favour of convenience.

VIII. Conclusion

Big Data requires additional safety requirements when collecting, processing, evaluating as well as transmitting data. Throughout this survey, we have relatively analyzed studies on Big Data, Safety and Security [19]. At this point no solid answer has been developed for unstructured data. Conventional data mining algorithms may be used for classification as well as clustering concerns, however they may not be used for privacy preservation, notably while handling data associated to a particular person [17] .

A solid demand on the part of the governments of all nations for law enforcement to guarantee privacy of an individual is observed. One of the severe protection hazards is the smartphone. Generally, users do not even read the user agreement before downloading any app which leads to the leakage of a lot of personal data through a number of applications operating on our smartphone without us being aware of. Subsequently, there seems to be a pressing need to make individuals learn about the different security flaws that can add to the discharge of personal data [17]. A lot of challenges are coming up with the rapid development of IoT and Big Data; the volume of content is massive, yet the reliability is poor and the information differs from a number of data sources inherently possessing a sizeable number of types as well as forms of representation [20]. Also, the information is diverse, as structured, quasi-organized, and sometimes totally unstructured. This poses new challenges to privacy and opens up research issues.

Big data safety and reliability are the main problems that need to be addressed further in the future [19]. Different strategies, innovations and approaches need to be developed in order to facilitate human-computer interactions or current systems need to be optimized for accurate performance [19]. It is expected that this study will hopefully help in understanding Big Data, its environment and build better frameworks, mechanisms and remedies for now as well as in the coming days.

IX. References

[1] D. Yang, B. Qu, and P. Cudré-Mauroux, “Privacy-Preserving Social Media Data Publishing for Personalized Ranking-Based Recommendation,” IEEE Trans. Knowl. Data Eng., 2019, doi: 10.1109/TKDE.2018.2840974.

[2] Zhang Luke, “Why all businesses should pay attention to Big Data,” Techpoint, 2019. .

[3] R. Kitchin and G. McArdle, “What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets,” Big Data Soc., vol. 3, no. 1, p. 2053951716631130, Feb. 2016, doi: 10.1177/2053951716631130.

[4] P. Michele, “Machine Learning 101-An Introduction,” towardsdatascience, 2018. .

[5] A. L. Pepper and B. F. Thau, “Privacy Versus Security in the Workplace,” Priv. Data Secur. Law J., no. February 2008, pp. 92–100, 2008.

[6] L. Kammourieh et al., “Group Privacy in the Age of Big Data,” 2017, pp. 37–66.

[7] K. Spiller et al., “Data privacy: Users’ thoughts on quantified self personal data,” in Self-Tracking: Empirical and Philosophical Investigations, 2017.

[8] M. S. Merkow, Secure, Resilient, and Agile Software Development. 2019.

[9] X. Qi and M. Zong, “An Overview of Privacy Preserving Data Mining,” Procedia Environ. Sci., 2012, doi: 10.1016/j.proenv.2012.01.432.

[10] V. S. Iyengar, “Transforming data to satisfy privacy constraints,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, doi: 10.1145/775088.775089.

[11] M. Rebecca and I. Raja, “Data Privacy and Data Anonymization Techniques,” datasciencedojo, 2017. .

[12] C. C. Aggarwal and P. S. Yu, “A General Survey of Privacy-Preserving Data Mining Models and Algorithms,” 2008.

[13] M. Haina, “A Survey on Privacy Preservation Used in Data Mining Techniques,” Int. J. Comput. Sci. Inf. Technol., vol. 6, no. 3, pp. 2339–2341, 2015.

[14] R. Gaire et al., “Crowdsensing and Privacy in Smart City Applications,” in Smart Cities Cybersecurity and Privacy, 2019.

[15] R. S. Kartha and V. Paul, “A New Cryptosystem Based On Polyalphabetic Substitution Scheme With Multiple Number Of Cipher,” 6th IRF Int. Conf., vol. 2, no. 8, pp. 40–44, 2014.

[16] K. Wang, P. S. Yu, and S. Chakraborty, “Bottom-up generalization: A data mining solution to privacy protection,” in Proceedings - Fourth IEEE International Conference on Data Mining, ICDM 2004, 2004, doi: 10.1109/icdm.2004.10110.

[17] P. Ram Mohan Rao, S. Murali Krishna, and A. P. Siva Kumar, “Privacy preservation techniques in big data analytics: a survey,” J. Big Data, vol. 5, no. 1, p. 33, 2018, doi: 10.1186/s40537-018-0141-8.

[18] J. Clement, “Online privacy - Statistics & Facts,” Statista, 2019. .

[19] D. Sinanc, R. Terzi, and S. Sagiroglu, A survey on security and privacy issues in big data. 2015.

[20] F. Chen, P. Deng, J. Wan, D. Zhang, A. V Vasilakos, and X. Rong, “Data Mining for the Internet of Things: Literature Review and Challenges,” Int. J. Distrib. Sens. Networks, vol. 11, no. 8, p. 431047, Aug. 2015, doi: 10.1155/2015/431047.