Journal of Theoretical and Applied Information Technology

30th April 2022. Vol.100. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BIG DATA CHALLENGES: PRESERVING TECHNIQUES FOR PRIVACY VIOLATIONS

TAMER ABDEL LATIF ALI [1], MOHAMED HELMY KHAFAGY [2], MOHAMED HASSAN FARRAG[3]

[1]Department of Computer Science, College of Computing & Information Technology

Arab Academy for Science, Technology and Maritime Transport, Aswan, Egypt

[1,2]Department of Computer Science, Faculty of Computers & Information, Fayoum University, Egypt

[3]Department of Information Systems, Faculty of Computers & Information, Fayoum University, Egypt

E-mail: [1]Tamer2019@aast.edu, [1]Ta1284@fayoum.edu.eg, [2]Mhk00@fayoum.edu.eg, [3]Mohamed.farrag@fayoum.edu.eg

ABSTRACT

The endless and quick development of data has brought the consideration of analysts to utilize it within the most conspicuous way for decision-making in different rising applications. These huge data are greatly valuable and profitable for logical investigation, increment efficiency in trade, and make strides for humanity. It makes a difference from the open division to commerce exercises, healthcare to way better route, smart cities to national security. Despite the fact that there are more opportunities to work, the obstacles of handling these data have also grown. In this study, various aspects of big data, as well as their applications and limitations, are discussed. This paper focuses on privacy challenges, privacy violations, and privacy-preserving techniques in big data, smart cities, and IoT. This paper introduces a comparative study among the different privacy-preserving techniques showing their advantages and their drawbacks to propose a powerful and applicable privacy protection technique that securely guarantees data integrity when dealing with big datasets and sensitive information in important fields.

Keywords: *Big Data, Big Data Challenges, Security, Privacy, Privacy Violations, Privacy-Preserving Techniques, Data Integrity.*

1. INTRODUCTION

Big data bring numerous appealing openings and applications with numerous confrontations to handle. This paper introduces various challenges in big data in section II. Privacy violations will be discussed in section III. Section IV introduces a comparative study among different privacy-preserving techniques. Section IV was divided into three main parts. Part one discusses big data privacy-preserving techniques. While part two concerns smart cities' privacy-preserving techniques. In part three, IoT privacy-preserving techniques will be discussed. A conclusion, on the other hand, will be discussed in the conclusion section. Finally, the future work was outlined in the concluding section.

2. VARIOUS CHALLENGES IN BIG DATA

- Data storage: data comes from numerous sources by fundamental exercises, illustrations like weblogs, social systems, sensors, logical investigations, and tests that deliver an enormous sum of data. Depending on the weight of the data, this volume must be reduced and a network load balanced [1], [2], [3].

- Heterogeneity: The user's data are heterogeneous, while data investigation calculations anticipate homogeneous data for superior preparing and examination [3], [4].

- Inconsistency, deficiency, and quality of data: Enormous information gets data for the investigation, is coming from different sources with diverse unwavering quality levels, which may contain incorrect, dubious, and lost esteem information. Such information must be overseen sometime recently the investigation handle [3], [4].

- Timeliness: Each day, data is developing exponentially, which should be summarized, sifted, and put away. any case, for real-time applications, like extortion location, social systems, the insights of things, shrewdly transport frameworks, and biomedical, timeliness must be a top-most need.

Timeliness refers to the degree to which data is current with the world that it models [3], [4].

• Security: Big data storage necessitates massive storage systems such as servers, data warehouses, and even the cloud. It could be a concern to supply security to these storehouses as criminal bunches are focusing on these big data storehouses to pick up imperative and private information, which may incorporate customer's individual or monetary points of interest, employees' data, or company's mystery [3], [5], [6].

• Scalability: The volume of big data is the first impression, which poses the most pressing essential adaptive challenge [5], [6].

• Visualization: In big data, we consider the framework and consider the human viewpoint. We must guarantee that people can appropriately get it comes about and not get misplaced in huge data [5], [6].

• Fault tolerance: With modern innovations like cloud computing and big data, it is continuously required that in case the disappointment happens and the harm is done, at that point it must be inside a worthy limit instead of beginning which prepare from scratch [5], [6].

• Privacy: Protection of data is one of the preeminent concerns in enormous information. Patients' data must be kept private since there is a danger of improper use of personal information. This can get uncovered when coordination data from diverse sources. Protection of an individual implies having the right to choose the degree of interaction with the environment, and the sum of data permitted to be gotten to by the other party. Whereas within the security, it is sufficient to identify data as a "password" in expansion, security is between two trusted parties, whereas in protection, and the server provider (SP) may be a foe. Figure 1 below discusses the issues that illustrate the difference between privacy and security [5], [7], [8].

3. PRIVACY VIOLATIONS

There are many privacy violation situations. This paper summarizes them into four violations. The primary is tracking by the government, the moment is data collected by service providers, the third is re-identification assaults, and the fourth is data breaches. In tracking by government circumstance, data collected can be utilized to decide private client data. An example of tracking by government is PRISM where the US government collects data from major service providers for intelligence purposes. In data collection by benefit supplier circumstance, benefit suppliers can utilize followed data to create client profiles. Service providers moreover evaluate other secret data. Accidental sharing of google documents is an example of Information collection by service providers. Re-identification assaults circumstance happens when the namelessness of data is compromised through the method of re-identification. An assailant might have individual or monetary objectives for a re-identification assault. Confidential information about a governor was identified by linking medical insurance records and a voter registration database is a good example of the re-identification attacks. There are different causes of data breaches in a data breaches circumstance, extending from insider dangers to malware and misconfigured systems. An example of data breaches is Ashley Madison where a dating website was hacked, and confidential information was made public. .Personal details of almost157,000 customers of the UK's major telecom providers were leaked. Figure 2 below introduces privacy violations in society [7], [8].

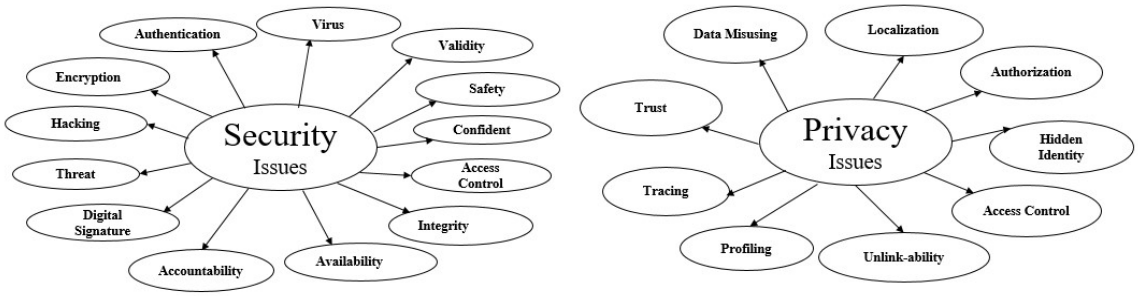

Figure 1: The difference between security and privacy [7]

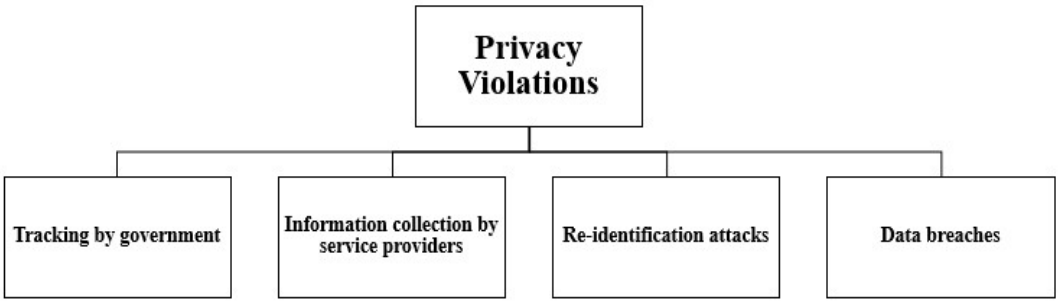

Figure 2: Privacy violation types of big data [7], [9], [10], [11]

4. PRESERVING TECHNIQUES OF PRIVACY

4.1. Preserving Techniques of Privacy in Big Data

4.1.1. Privacy-protection by slicing

In Slicing, the dataset partitioned horizontally and vertically. While in vertical Partitioning, the attributes are grouped into columns based on their relationships. Slicing can handle high-dimensional data according to attribute apportioning. But the horizontal partitioning has happened whereas gathering the tuples into diverse buckets. To break the relationship between distinct columns inside each bucket, values in each column are arbitrarily permuted. Using slicing technique preserves the associations within each column but breaks the association cross columns [12].

4.1.1.1. Vertical partitioning of the data

These algorithms gathering the exceedingly connected attributes are within the same column. There is no doubt that gathering the foremost connected attributes protects the data utility which means data should keep up privacy, and data is accommodating for analysis. Recognizable proof risk is provided by the association of uncorrelated attributes [12].

4.1.1.2. Horizontal partitioning of the dataset

Data partitioning of the dataset horizontally implies dividing the tuples into numerous buckets. This group of records is called a buckets tuple dividing technique which partitions the tuples into buckets and to check whether each bucket fulfills the l-diversity or not [12].

4.1.2. Privacy in the big data generation phase

4.1.2.1. Access restriction

Using anti-tracking expansions, advertisement/script blockers, encryption techniques, anti-malware, and anti-virus program makes us able to restrain the get to sensitive data [8].

4.1.2.2. Falsifying data

Using Socketpuppet, which employments deception to stow away an individual's online personality, or MaskMe, which implies making assumed names of their individual data such as credit number card or e-mail address can be defined as falsifying [8].

4.1.3. Privacy in the big data storage phase

4.1.3.1. Attribute-based encryption (ABE)

In this technique, the data owner establishes access policies, and data is encrypted in accordance with those policies. ABE is a cryptographic method that ensures end-to-end big data privacy in cloud storage systems. Only the users whose attributes comply with the data owner's access policies are allowed to decrypt the data [8].

4.1.3.2. Identity-based encryption (IBE)

IBE protects sender and receiver anonymity in a certificate-based public key infrastructure (PKI) by employing human identities as public keys, such as an IP address or email address [8].

4.1.3.3. Homomorphic encryption

Functions can be computed on encrypted data using this type of encryption. Anyone who acquires the Encryption of a message can likewise obtain the

Encryption of a function of that message by directly calculating on the Encryption [8].

4.1.3.4. Storage path encryption

This method entails breaking down large amounts of data into a series of sequential chunks, which are then saved on separate storage devices controlled by multiple cloud storage providers [8].

4.1.3.5. Usage of hybrid clouds

This entails utilizing both public and private clouds, as well as the inherent characteristics of public clouds, such as scalability and processing capacity, as well as providing security and prospective research opportunities for processing and storing large amounts of data from private clouds [8].

4.1.3.6. Big data integrity

- Proofs of Retrievability (POR): POR allows a cloud provider to demonstrate that a user may retrieve a whole set of data [8].
- Provable Data Processing (PDP): PDP is another technique used to assure the integrity and completeness of data [8].
- Public Auditing: Third-party data integrity verification is referred to as public auditing [8].
- Cloud Provider High Availability Redundant Independent Files (CPHARIF) Algorithm: It's a new strategy for boosting huge data replication availability. Data availability is closely related to data integrity and is just as vital as integrity. CPHARIF can increase availability and dependability while also lowering recovery costs [13].

4.1.4. Privacy in the big data processing phase using anonymization techniques

Figure 3 below shows the most important anonymization strategies.

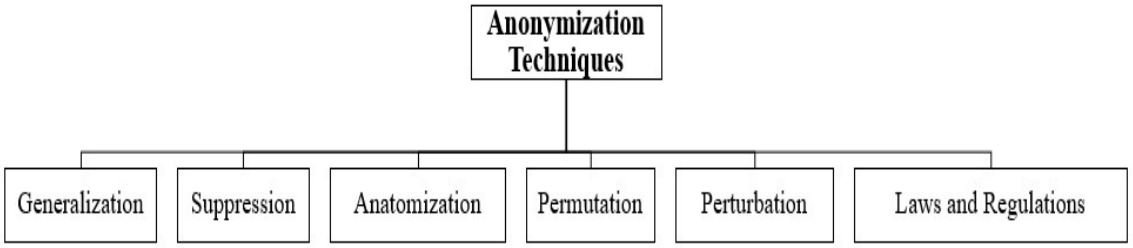

Figure 3: Different Anonymization Techniques [7], [8]

4.1.5. Cybersecurity measures to prevent data breaches

Espionage systems like as honeypots, preventive methods such as firewalls, encryption techniques, access records, and warning systems are examples of cybersecurity measures [7].

4.1.6. Suggestions for restricting privacy violations

In [7], Shamsi et al. provide a model for minimising privacy violations, illustrating four forms of privacy violations as well as the involvement of various actors in reducing their impacts. The single-sided arrows denote a mechanism's function in preventing a certain form of privacy infringement. The double-sided arrows between separate entities, on the other hand, indicate stronger cooperation between them. Re-identification attacks, data breaches, government tracking, and information obtained by service providers are all covered in Figure 4. It also illustrates the responsibility for preserving privacy: computer scientists, governments, civil society, and organizations. It also contains methods of preserving privacy. They have enhanced anonymization techniques and scientific web technologies, enhanced Encryption and strong cyber defense, access mechanisms and intrusion detection and prevention, improving management strategy and organization issues, cyber policy and social ethics, research development, public awareness, and user-friendly laws [7].

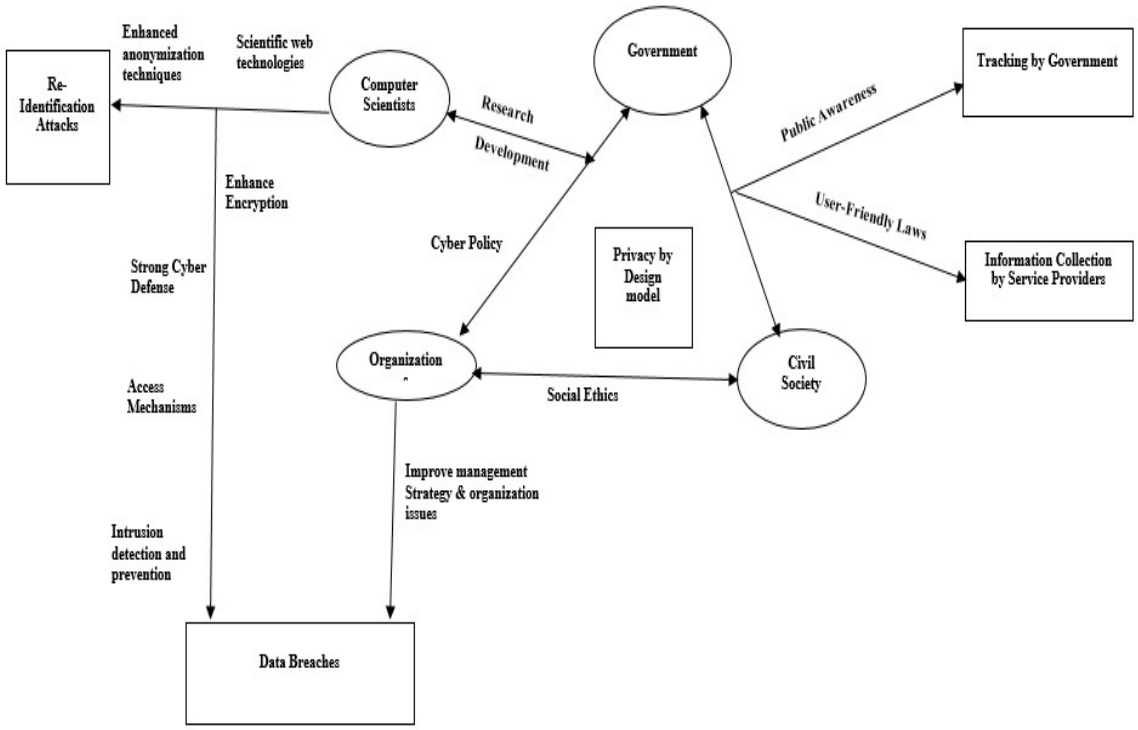

Figure 4: Model for improving privacy [7]

4.1.7. Proposed secured map reduce model (SMR)

Enhancing and optimization of Map-Reduce job is a very important thing while securing a Map-Reduce is a must. As the data passes through the phase called map-reduce, this new layer applies the privacy techniques to each individual piece of data. This privacy preserving approach must be a valuable encryption scheme, such that the overhead of additional algorithms does not interfere with the primary Big Data capability. When using this new proposed Secured Map Reduce (SMR) layer of Big Data, data can be safeguarded and secured. Between HDFS (Hadoop Distributed File System) and another layer known as Secured Map Reduce (SMR) Layer, this suggested solution offers a privacy layer. To increase the protection of the data, randomized algorithms and perturbation were applied. It begins with data gathered from weblog, streaming data, and social media. After that, the data is delivered to HDFS [14], [15], [16].

4.1.7.1. Encryption

The process of encoding or decrypting data refers to encryption. In this encryption method, there are two basic steps. Text data is converted to a number in the first level. The key-value pairs (KVP) model is then used to divide each word text into tokens, with each unique word being considered and the word's repetition in the given data being counted. This process also introduces high privacy by giving data. Each unique word is the key, and the quantity of repeats is important. To improve the degree of privacy, the randomization procedure in converted numerical data is performed at the second level. In the HybrEx model's vertical partitioning, the data is handled first in the private cloud during encryption and subsequently in the public cloud during decryption [14], [15], [17].

There are four stages to the encryption algorithm. The input stage, which contains the data file, is the first. The second phase is the mapper phase, which involves partitioning the data file, transforming words to numbers, and storing the results in a hash map and an SMR file. While the third phase is the reducer phase, in which combining the result is done. The fourth phase is the output phase, which contains the encrypted SMR file and the encrypted frequency file [14], [15].

4.1.7.2. Decryption

Decryption can be defined as the reverse process of Encryption. It refers to the process of decrypting data that has been encrypted. It is a method of transferring processed data to HDFS and then to the map-reduce layer. On the reconstruction step, decryption takes place in the map-reduce layer. This

decryption procedure is divided into two steps. The first stage is reverse randomization, which employs randomization to partially decipher the encrypted message. After reverse randomization, the second step is to convert numeric values to text data using the Key-Value Pairs (KVP) model. Because the order isn't indicated here, the data will be retrieved from the file where the order was written. A key is every token, and a value is the number of times it is repeated. Then, as Priyank Jain indicated in [15], it would correctly transform the number to text data.

The decryption algorithm also consists of four phases, like Encryption. The first is the input phase, which contains the encrypted file. The second phase is the mapper phase, in which the encrypted file is partitioned, mapper ids are read, and randomization is reversed using the SMR encrypted file. While in the reducer phase, generating the decrypted file is done after reading the hash map. The output phase, which contains the decrypted file of words and the decrypted file of frequencies, is the final phase [14], [15].

4.1.8. The SMR performance measures

- Running time: Because the proposed method's performance can be assessed in milliseconds, it can be efficiently cleaned. It also allows for greater scalability [15].
- CPU utilization: CPU utilization can be increased by using the SMR technique. While additional resources such as memory space and input-output devices can be used to boost CPU performance [15].
- Memory usage: The data structure and the amount of data are the most important things on which the memory usage depends [15].

When increasing the number of cores 40–160, the parallel execution of Hadoop SMR layer operations reduces the time because the nodes act at the same speed as the data itself, as illustrated in Figure 5[15].

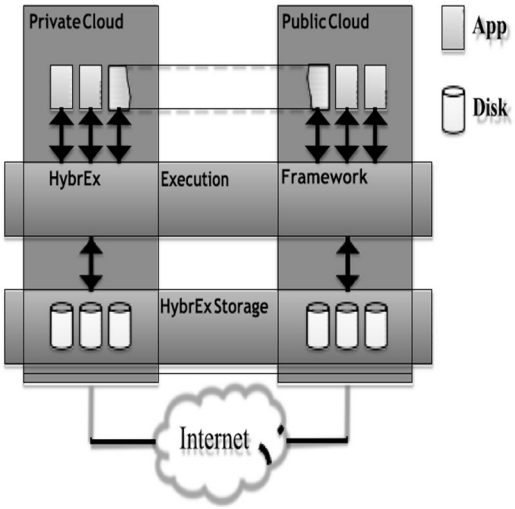

Figure 5: The time taken in relation to the number of cores and records [15]

4.1.9. Privacy protection for cloud Big data using a hash function

Figure 6 and figure 7 below illustrate privacy protection for cloud tenant's big data using the hash function. Encryption using the hash function will be illustrated in the pre-processed phase of the big data, as figure 6 shows. The decryption will be illustrated in the reconstruction phase of the big data, as shown in figure 7 [17].

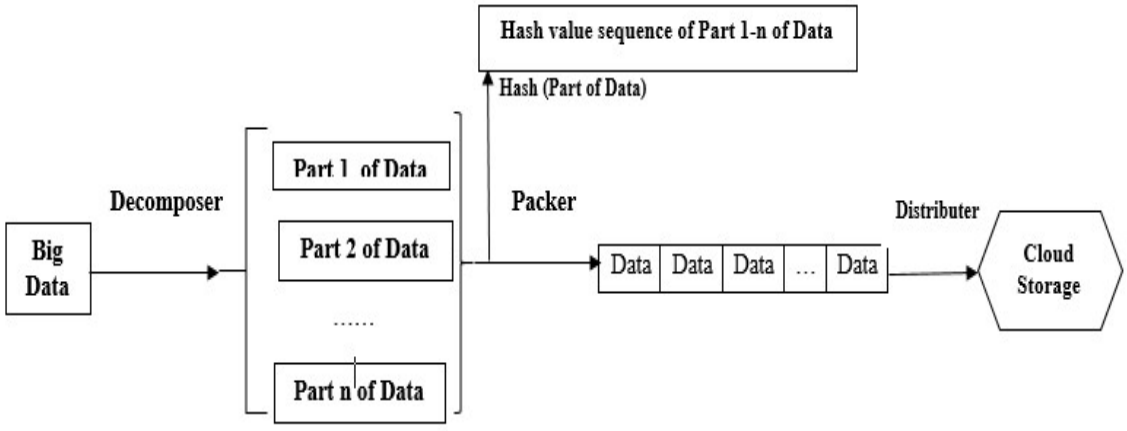

Figure 6: Pre- processed of the Big Data [17]

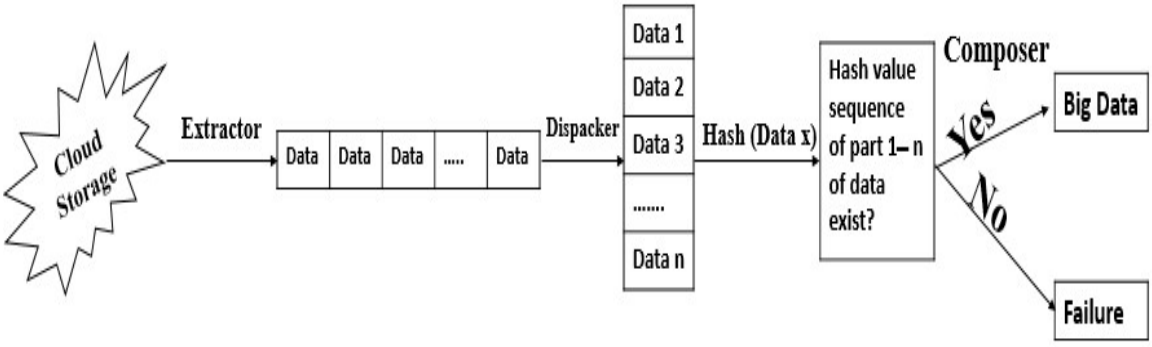

Figure 7: Reconstruction of the Big Data[17]

4.2. Preserving Privacy of Smart Cities

4.2.1. Foggy dummies

This approach is utilized in fog computing, and the fundamental idea is to create extremely intelligent dummies to preserve the user's privacy. Adnan Ahmed Abi Sen uses this technique to swap queries between fogs before sending them to the server provider (SP) and then exchanging the responses. This will be accomplished by fogs cooperating to exchange data before sending it to the server provider [18].

Foggy Dummies Advantages [18]: -

- Increasing the level of privacy
- There is no overhead.
- There are no network overheads.
- Server provider cannot detect this dummy.
- Increasing the cache hit ratio and decreasing SP connections
- No loss of accuracy

4.2.2. Blind third party (BTP) technique

The essential point is why we must rely on a third-party (TP) to keep the user safe from SP. That is, we are transferring the problem from one server to another. This strategy is dependent on Fog's role as a broker between the user and the SP in each area [18].

The distinction is that we employ the following measures to prevent the Fog from viewing the user data: - The user uses the SP public key to encrypt his query (location, data, and the new key UK). Then, in the same cell, he transmits his question to the Fog. After then, Fog will act as an anonymizer, hiding the user's ID and resending the user's query to SP, which will not be able to detect the user's UID and will simply answer the inquiry and encrypt the replies by the UK. Finally, SP transmits the result to Fog, who is unable to read it and just sends it back to the user [18].

By employing public and private keys to apply the TTP idea without the necessity for third-party trust and information sharing, BTP offers a good answer to this challenge. However, there was a drawback to this approach: BTP might get information on the user without the user's knowledge on a permanent basis by sending inquiries to the SP on his own without the user's knowledge. All of this is done to assure the success of the BTP method and to overcome its drawbacks, thereby raising the level of privacy and security of user data for Internet of Things apps and allowing for a higher number of these applications to be supported without fear of constraints [18].

4.2.3. Double foggy cache technique

The basic idea behind this methodology is to use classic cooperative methods to overcome the problem of peer trust. Meanwhile, use SP to preserve your privacy. This method, in particular, can be viewed as a step forward in the work. To accomplish this, we proposed inserting two caches in the Fog that would operate as brokers between peers. The first is for questions, while the second is for responses [18].

The double foggy cache technique advantages:

- Enhancing the performance by decreasing the server Provider connections.
- There is no need for a user to rely on another to keep him.
- The server provider is unable to acquire any information about the user's actions.
- Increasing of the cache hit-ratio

4.2.4. Privacy Rule Ontology

This proposed ontology consists of three categories of privacy. The first category is the content privacy, which includes information about the privacy of sensed data produced by IoT devices. The context privacy class, on the other hand, depicts information concerning the privacy of non-sense data from IoT devices. While the Privacy Rules category covers the use of privacy-preserving rules by IoT devices, such as data swapping, random noise, and data micro aggregation in our case [19].

Gheisari et al. in [19] proposed that this ontology uses three techniques to preserve privacy in the smart city: -

- Micro Aggregation.
- Swapping.
- Random Noise.

The proposed privacy-preserving procedure is a flowchart. Each IoT device initially communicates its ID to the server, which comprises ontology. The server subtracts one from the duration of its privacy rule. If the lifespan is positive after decreasing, the server returns the current privacy mechanism for the device to use.

However, if the privacy rule's lifespan reaches zero, it will change the device's owner at random, mislead enemies, and alert the device. The server then selects another appropriate privacy rule for the device to use and returns its equal number, which is chosen at random in this scenario. Future research is needed to determine the optimal next privacy-preserving regulation for each type of device. The IoT device then delivers its processed data to the ontology server at the network's edge, after applying the chosen new privacy approach. The data will then be sent to a cloud computing environment for additional processing [19].

4.3. Privacy Preserving Technique in IoT

4.3.1. Preserving proposed solutions in IoT

For the IoT environment, Dabbagh et al. [20] suggested a novel authentication mechanism. To successfully detect emulation attacks, this suggested system employs device fingerprinting approaches as well as transfer learning. While Addo et al. [21] described the visual privacy gaps in collecting and processing emotion data for tailored advertising and provided a reference framework for maintaining end-user privacy throughout the emotion analytics lifecycle. While Otgonbayar et al. suggested an anonymization approach for releasing IoT data streams, Otgonbayar et al. Under a time-based sliding window, this method anonymizes tuples with comparable descriptions in a single cluster. Seliem et al. [23] summarised the recommended solutions introduced to address the privacy challenges in the IoT environment as the following:

a. Authentication and Authorization.
b. Plug-in architecture and Edge Computing.
c. Denaturing and Data Anonymizing Techniques.
d. Data Summarization and Digital Forgetting.

4.3.2. Blind third party (BTP)

This technique addresses the disadvantage of exposing data in the TTP approach. By removing the anomaly in the standard TTP methodology (Trust in TP), our method makes it useable. The user would gain all of the benefits of utilizing the TP without disclosing any data using this method. TP is called BTP for the same reason. Also, even if there are no other beers in the vicinity, the user can utilize this method to speed up the answer [24].

Advantages of the BTP Approach [24]

- The following points summarizes the advantages of BTP: -
- Enhancing privacy and there is no track for the user
- There is no outer attack for user's data
- Releasing the overload from the user
- Changing the identities of users
- There is no additional load on the link itself.
- BTP can Support IoT applications with a lot of data output from sensors.
- TTP and encryption techniques are integrated.

Disadvantages of The BTP Approach [24]

- The following is the summary of BTP drawbacks: -
- The risk of BTP and SP colluding to violate users' privacy.
- Encryption may create device overload for some users.
- Consuming more power in Encryption process by user's devices.

4.3.3. Blind peer approach

Yamin et al. [24] demonstrate that this technique overcomes the major flaw of earlier systems in which BTP interacts with SP to compromise the user's privacy. A user's query would be replaced with another peer in the same area, and then encrypted using PK of SP (SPPK), leaving the other peer with no choice but to pass the query to the service provider, who would decrypt it. The new idea behind the BLP technique is to rely on

collaboration among many peers rather than dealing with a single TP. The encryption of the answer can be done using the Session Key and then it can be forwarded to the peer while sending the answer to the main user since the answer can't be read. Then this approach would repeat the same steps with peers' other queries.

Advantages of Blind Peers:

All of the benefits of the P2PCache and BTP approaches are preserved, but all of the disadvantages are eliminated. This technique removes the availability of collusion between the service providers and BTP. There would be no additional pressure on any peer, and the TP would not bottleneck because the user would be dealing with a different peer each time [24].

Disadvantages of Blind Peers:

There would be collusion between BTP and SP. While a suitable peer may be accessible in many other cooperative strategies, it may not be in this approach [24].

4.3.4. Integrated blind parties (IBPs)

By integrating the BLP and BTP, IBP's method doubles the degree of privacy. When a peer isn't present in the area, the user can only rely on the BLP in this technique. In this scenario, the peer can employ the BTP technique rather than the user. This strategy can also be applied to a set of seven approaches. In the event of a resource shortage, the user will be able to swap the query with another peer without encrypting the data. We could combine Cache with Bloom filter to improve performance in the same way that the DCA approach suggests. The next section will summarize the most important evaluation measures [24].

4.3.5. Evaluation Measures

4.3.5.1. Average time vs. queries number

BTP and DCA were extremely similar. Because Encryption affects the time of each query, BTP took the shortest time. In terms of the amount of time spent, it was still better than DCA. The same step happened with P2PCache and BLP, P2PCache, on the other hand, was superior than BLP because they both employed the same method, and BLP added encryption as an extra precaution to improve privacy [24].

4.3.5.2. Response time

BTP had the fastest response time for the user's request, but it was very close to DCA's performance because using one party as a server in BTP with encryption is clearly better than using two caches, especially when considering whether or not a query is replied. Because it employed encryption and needed searching for and connecting with an available peer, BLP was the weakest performer. This slows down response times while still giving the strongest data privacy protection. Because all of the situations transmitted only one query to SP, with no further dummy or requests for extra results, BTP/BLP, along with P2PCache and DCA, had the lowest cost of number of inquiries. Encryption had no effect on the number of queries, but it did reduce the size, which can be viewed as a step closer to perfect privacy and security [24].

4.3.5.3. Entropy

Because the user didn't send any personal information, BLA had the highest Entropy (E = 1). Furthermore, due of the Trust Factor, BLA supersedes the set of seven techniques [24].

4.3.5.4. Ubiquity

The BLA technique delivers maximum Ubiquity (U = 2) while maintaining maximum privacy. This is due to the fact that each user in BLA sends a large number of requests to various users throughout the cell [24].

4.3.5.5. Cache hit ratio

BLA can be utilized with the caching strategy to improve performance vs. privacy, as we stated earlier. This can happen if you utilize strategy number two with Bloom Filter and only store real user requests in the cache [24].

4.3.6. IoT applications of blind approach

The following are the most famous application in the internet of things that used a blind approach [24]:

- In smart transportation systems, a blind method can be employed, which is one of the most critical components in preserving smart city privacy.
- A blind approach can be used in protecting smart home privacy.
- In underwater wireless sensor networks, a blind technique can be utilized to prevent eavesdropping and malevolent nodes from getting data from other sources.
- A blind approach can be used to protect medical data efficiently

4.3.7. Enhanced AES encryption technique

The phases of the enhanced AES encryption method are depicted in Figure 8. The enhancement is accomplished by realizing both correct and broken-down ciphertexts, data adjustment analysis, and secret key recovery enhancements that are ready to be integrated into physically unclonable

functions (PUFs) that enable randomized secure storage and secret key recovery. The steps of the improved AES algorithm are shown in figure 8 below [25].

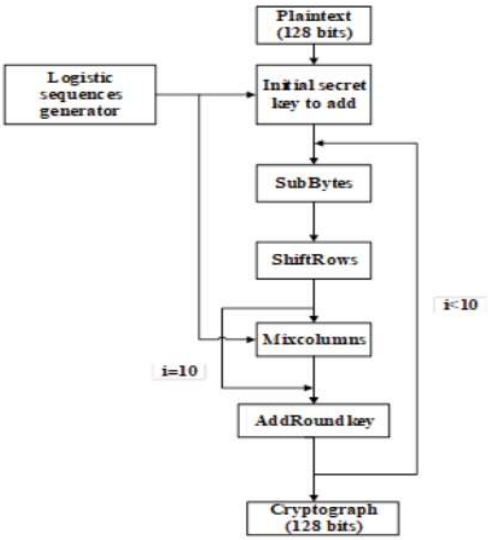

Figure 8: Flowchart for an Enhanced AES algorithm [25]

Anajemba et al. in [25] proposed that the algorithm is divided into three main steps. The first phase is AES encryption, which includes defining the plain text and the key, conducting the pre-round transformation using the plain text, taking the key length into account to confirm the rounds of transformation, and ciphertext implementation. The second is AES decryption, which reverses the steps of the symmetric encryption technique AES encryption function. While the third is a key generation, which contains generating the key, the first four words are created from the key, which is used to calculate the required set of words, receiving the next word, and repeating the previous steps until all the words are made from the keys [25].

4.3.8. Enhanced BTP

In this enhanced approach, there is a new factor added to the old BTP, which is a unique token. This new technique consists of seven factors. A unique token is defined when the user sends a hidden code within a query to the service provider (SP) while SP returns the previous query token. Then SP will store the token for each ID generated by the third party, so the previous one cannot be used in a later query. When the third party inquiries from SP, a change will occur on the user's token, and the user will discover unauthorized access to his data by third party, so the proposed technique will be a powerful guarantee that there is no breakthrough. [26].

4.3.9. Light weight cryptography techniques(LWCT)

Based on an oil spill detection application, LWCT is utilized to secure a data transmission framework for the internet of things. Through locative and boundary value aggregation, this strategy eliminates duplicate data transmission. The suggested method protects data transfer by combining known lightweight cryptographic techniques with simple ID-based authentication. Future enhancements could include adding intelligence to sensor nodes to make decisions about oil spills [27].

4.3.10. Block Nested Loop (BNL) Skyline Algorithms

This method is used to determine which encryption algorithm is best for ensuring data protection and privacy issue. The author of the Skyline algorithm considers two primary parameters: the rate of variation and the number of dimensions. Figure 9 below shows the steps of applying the BNL skyline technique [28].

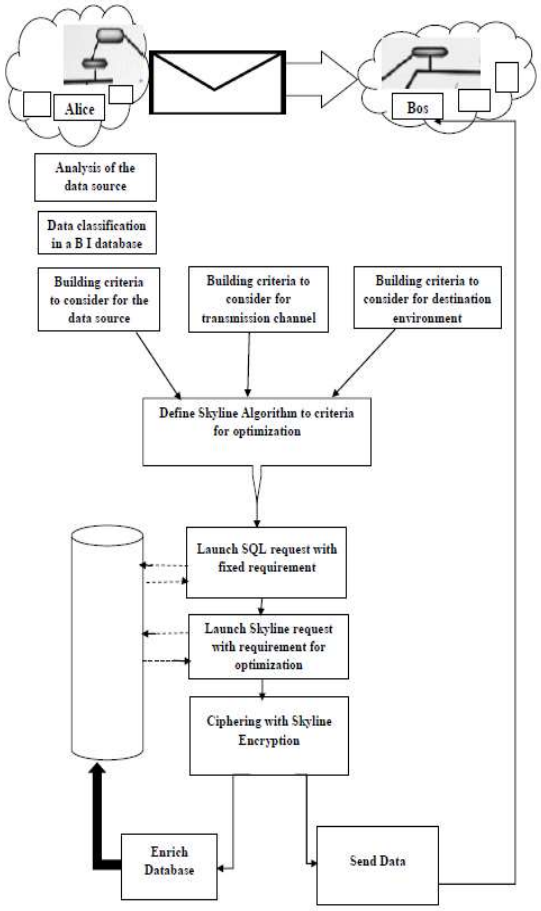

Figure 9: BNL Skyline algorithm for enhancing encryption [28]

5. RESULTS AND EVALUATION

As seen from results, improved AES, LWCT and SMR have the highest level of privacy but improved AES has high processing time and low data integrity. While LWCT has a medium level for processing time and also a medium level of data integrity. While has low processing time but also low data integrity level.

While there are some drawbacks summarized in a long time, complexity, and data integrity violations. This paper summarized the Evaluation of the previous techniques as shown in Table 1 below. Table 1 summarizes all advantages and drawbacks of each technique. All techniques have a low level of deception for a bad user.

As seen from results, improved AES, LWCT and SMR have the highest level of privacy but improved AES has high processing time and low data integrity. While LWCT has a medium level for processing time and also a medium level of data integrity. While has low processing time but also low data integrity level.

While there are some drawbacks summarized in a long time, complexity, and data integrity violations. This paper summarized the Evaluation of the previous techniques as shown in Table 1 below. Table 1 summarizes all advantages and drawbacks of each technique. All techniques have a low level of deception for a bad user

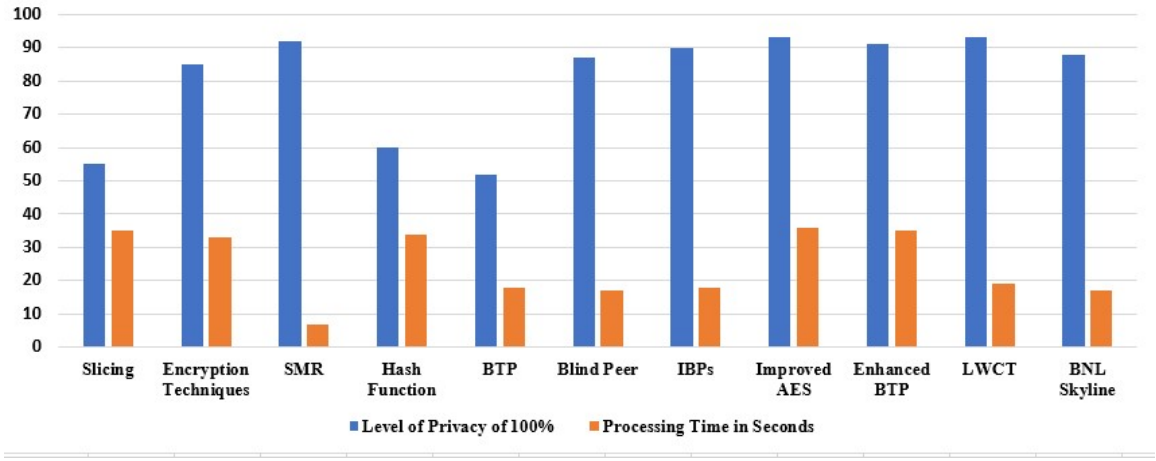

Figure 10: Comparison of All Privacy Techniques According to Privacy Level and Processing Time in Seconds

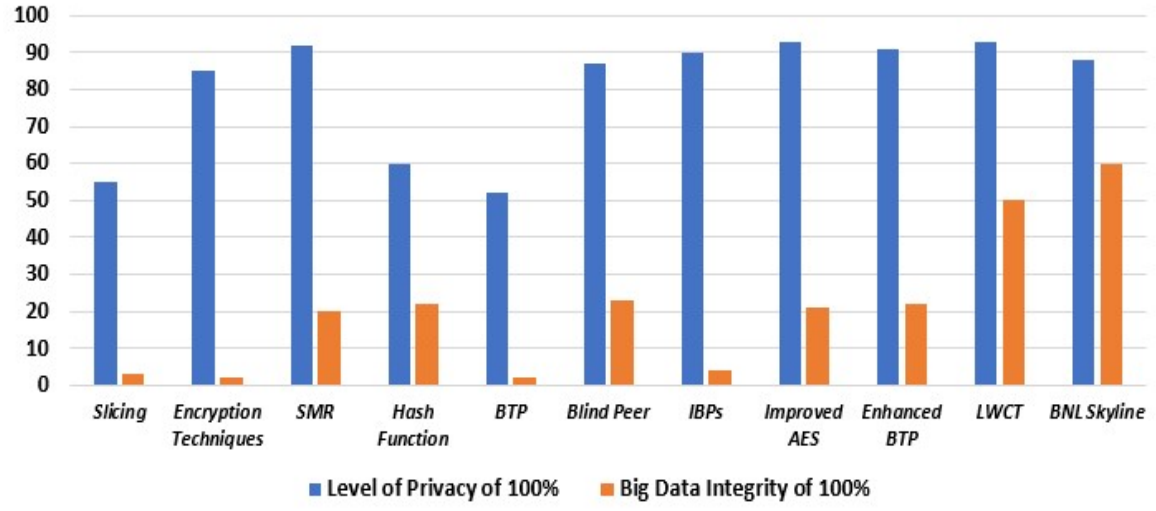

Figure 11: Comparison of All Privacy Techniques According to Privacy Level and Data Integrity

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Privacy-Preserving Techniques Evaluations

Technique	Level of Privacy	Big Data Integrity	Processing Time	Deception Level
Slicing	Medium	Lost	High	Lost
Encryption Techniques	High	Lost	High	Lost
SMR	High	Low	Low	Low
Hash Function	Medium	Low	High	Lost
BTP	Medium	Lost	Medium	Lost
Blind Peer	High	Low	Medium	Lost
IBPs	High	Lost	Medium	Lost
Enhanced AES	High	Low	High	Lost
Enhanced BTP	High	Low	High	Low
LWCT	High	Medium	Medium	Lost
BNL Skyline	High	Medium	Medium	Lost

6. CONCLUSION

This paper listed the most important big data challenges and focused on privacy challenges. It summaries privacy violation situations and their examples in real life. It also provides a list of the most efficient and popular techniques used to protect data privacy in different fields. It divided privacy-preserving techniques into three categories. The first was about preserving big data privacy. The second category was about smart cities' privacy and how to preserve it. While the third category was about how to preserve privacy in the IoT environment. As concluded from this paper, these techniques have their advantages that are summarized in protecting privacy in big data, the internet of things, and smart cities. Another advantage is decryption simplicity, especially when dealing with small datasets. This study introduces a powerful evaluation and comparison between different privacy-preserving techniques in different fields such as big data, smart cities and IoT. This evaluation is based on the most significant factors for the data owners. The evaluation factors are level of privacy, big data integrity, processing time and deception level for unauthorized users. This study does not take into consideration the different attribute types such as image and polynomial. This is what the author tends to cover in the future work.

7. FUTURE WORK

In the future, a powerful and applicable privacy protection technique will be proposed. This technique will cover all attribute types. It may be a hybrid of the previous techniques or a newly proposed one that will securely guarantee data integrity when dealing with big datasets and sensitive information in important fields.

REFERENCES:

[1] E. Sarhan, A. Ghalwash, and M. Khafagy, "Queue weighting load-balancing technique for database replication in dynamic content web sites," *Proceedings of the 9th WSEAS International Conference on Applied Computer Science*, pp. 50-55, 2009.

[2] M. H. Mohamed and M. H. Khafagy, "Hash semi cascade join for joining multi-way map reduce," *Proceedings of 2015 SAI Intelligent Systems Conference*, pp. 355-361, 2015.

[3] P. V. Desai, "A survey on big data applications and challenges," *Proceedings of the Second International Conference on Inventive Communication and Computational Technologies (ICICCT),* IEEE, pp. 737-740, 2018.

[4] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

challenges: greening big data," *IEEE Systems Journal*, Vol. 10, No. 3, September 2016.

[5] S. Boubiche, D. E. Boubiche, A. Bilami, and H. Toral-Cruz, "Big data challenges and data aggregation strategies in wireless sensor networks," special section on real-time edge analytics for big data in the internet of things, *IEEE Open Access Journal*, Vol. 6, pp. 20558-20571, 2018.

[6] M. M. Shendi, H. M. Elkadi, and M. H. Khafagy, "A study on the big data log analysis: goals, challenges, issues, and tools," *International Journal of Artificial Intelligence and Soft Computing*, Vol. 7, No. 2, pp. 5-12, 2019.

[7] J. A. Shamsi and M. A. Khojaye, "Understanding privacy violations in big data systems," *IT Professional*, Vol. 20, No. 3, pp. 73-81, 2018.

[8] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo "Protection of big data privacy," *IEEE access*, Vol. 4, pp. 821-1834, 2016.

[9] P. A. Laplante, "Who's afraid of big data?" *IT Professional*, Vol. 15, No. 5, pp. 6–7, 2013.

[10] M. De Goede, "The politics of privacy in the age of preemptive security," *International Political Sociology*, Vol. 8, No. 1, pp. 100–104, 2014.

[11] D. Barth-Jones, "The 're-identification' of governor william weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now," *SSRN Electronic Journal*, 2012. [Online]. Available at SSRN: https://ssrn.com/abstract=2076397.

[12] K. Vani and B. Srinivas, "Enhanced slicing for privacy-preserving data publishing," *The International Journal of Engineering and Science (IJES)*, Vol. 2, No. 10, pp. 1-4, 2013.

[13] M. R. Kaseb, M. H. Khafagy, I. A. Ali, and E. M. Saad, "An improved technique for increasing availability in big data replication," *Future Generation Computer Systems*, Vol. 91, pp. 493-505, 2018.

[14] H. Mahmoud, A. Hegazy, and M. H. Khafagy, "An approach for big data security based on hadoop distributed file system," *2018 International Conference on Innovative Trends in Computer Engineering (ITCE)*, IEEE, pp. 109-114, 2018.

[15] P. Jain, M. Gyanchandani and N. Khare, "Enhanced secured map-reduce layer for big data privacy and security," *Journal of Big Data*, Vol. 6, No. 1, pp. 1-17, 2019.

[16] M. S. Shanoda, S. A. Senbel, and M. H. Khafagy, "JOMR: Multi-join optimizer technique to enhance map-reduce job," *2014 9th International Conference on Informatics and Systems*, pp. PDC-80-PDC-87, 2014.

[17] H. Cheng, W. Wang, and C. Rong, "Privacy protection beyond encryption for cloud big data," *Proceedings of 2nd International Conference on Information Technology and Electronic Commerce*, IEEE, pp. 188-191, 2014.

[18] A. A. Abi Sen, F. A. Eassa, and K. Jambi, "Preserving privacy of smart cities based on the fog computing," *International Conference on Smart Cities, Infrastructure, Technologies and Applications*, Springer, Cham, pp. 185-191, 2017.

[19] M. Gheisari, G. Wang, and S. Chen, "An edge computing-enhanced internet of things framework for privacy-preserving in smart city," *Computers and Electrical Engineering*, Vol. 81, 106504, 2020.

[20] Y. Sharaf-Dabbagh and W. Saad, "On the authentication of devices in the internet of things," *2016 IEEE 17th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, IEEE, pp. 1-3, 2016.

[21] I. D. Addo, P. Madiraju, S. I. Ahamed, and W. C. Chu, "Privacy preservation in affect-driven personalization," *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC),* IEEE, Vol. 2, pp. 400–405, 2016.

[22] A. Otgonbayar, Z. Pervez, and K. Dahal, "Toward anonymizing iot data streams via partitioning," *2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pp. 331–336, 2016.

[23] M. Seliem, K. Elgazzar, and K. Khalil, "Towards privacy-preserving iot environments: a survey," *Communications and Mobile Computing*, Vol. 2018, pp. 1-15, 2018.

[24] M. Yamin, Y. Alsaawy, A. B. Alkhodre and A. A. A. Sen, "An innovative method for preserving privacy in internet of things," *Journal of Sensors*, Vol. 19, No. 9, pp. 3355, 2019. [Online]. Available: https://doi.org/10.3390/s19153355.

[25] J. H. Anajemba, C. Iwendi, M. Mittal, and T. Yue, "Improved advanced encryption standard with a privacy database structure for iot nodes," *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 201-206, 2020.

[26] A. A. A. Sen, F. A. Eassa, K. Jambi, N. M. Bahbouh, S. S. Albouq, and A. Alshanqiti, "Enhanced- blind approach for privacy protection of iot," *2020 IEEE 7th International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, pp. 240-243, 2020.

[27] H. V. Abhijith and H. S. Rameshbabu, "Secure data transmission framework for internet of things based on oil spill detection application," *International Journal of Advanced Computer Science and Applications (IJACSA),* Vol. 12, No. 5, pp. 189-195, 2021.

[28] S. Trichni, F. Omary, and M. Bougrine, "New smart encryption approach based on multidimensional analysis tools," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 12, No. 5, pp. 666-675, 2021.