Journal of Positive School Psychology
2022, Vol. 6, No. 5, 9130–9138 

# Design and Verification of Hybrid model for Big Data Privacy-Preserving in D2D Communication Environment

Shelly Bhardwaj [1], Dr. Abhishek Kumar Mishra [2], Dr. Rahul Kumar Mishra [3]

*[1] Research Scholar, Department Of Computer Science & Engineering, IFTM University Moradabad, Uttar Pradesh, India.*
*[2] Associate Professor, Department Of Computer Science & Engineering, IFTM University Moradabad, Uttar Pradesh, India.*
*[3] Director School of Computer Science & Applications, IFTM University, Moradaba, Uttar Pradesh, India.*
*Email: [1] shellybhardwaj29@gmail.com, [2] abhimishra2@gmail.com, [3] rahulmishra@iftmuniversity.ac.in*

**Abstract**

Nowadays, a massive amount of datasets are being produced by numerous organizations which include but are not limited to private or governmental hospitals, educational institutions, the corporate world, and retail companies all around the world. There have been developed numerous methods in previous years for privacy preservation of the big datasets in the device-to-device (D2D) scenario. Nevertheless, such existing methods have diverse limitations in the modern world related to the secrecy and privacy of the big datasets because the datasets are increasing constantly and existing systems facing challenges in handling the gigantic amount of datasets on daily basis from hackers. In this article, the authors proposed a novel design and verification of a hybrid model for big data privacy-preserving in a D2D communication environment. The outcome of the suggested model demonstrates that the overall execution time and accuracy of this novel model are higher in comparison to the existing models. Further, our model is more robust against diverse assaults from hacker sides for securing the datasets and maintaining secrecy in the desired manner. Our proposed model offers accuracy values on selected datasets i.e. 500, 2000, 4000, 6000, 8000, and 10000, 92%, 92.5%, 93%, 96.5%, 97%, and 99.1%, respectively. In the future, this model can be updated according to diverse application and systems requirements for datasets privacy preserving more accurately for big data in a device to device communication environments.

**Keywords:** Big Data, Data Security, D2D, Hybrid Model, Privacy-Preserving.

## I. INTRODUCTION

The usual concept of dataset secrecy is the accessibility, privacy, as well as authenticity of a dataset. This is indeed a process of making certain that the data is protected against unwanted accessibility, guarantees whether it is trustworthy as well as correct, and makes certain that content becomes available anytime it gets needed. The strategy for dataset secrecy includes elements like obtaining the necessary dataset, safeguarding it, but also erasing any dataset which will never be needed anymore [1]–[5]. Contrarily, the right usage of data is to maintain confidentiality. In other terms, businesses, as well as dealers, must only utilize the datasets that have been given to people for a specific reason. For instance, any kind of corporation cannot sell or buy any customer' secured details regarding the identity to other parties if the customer purchases an item from them as well as gives them communication addresses, credit card details, and so on. Organizations must have a datasets protection

strategy with the express objective of protecting the confidentiality of their customers' sensitive datasets. Additionally, since the datasets are a resource for the business, firms must protect the overall confidentiality of the datasets. Nevertheless, no dataset protection strategy could defeat a company's determination to trade or collect customer details that have been confined to it [6]–[10]. The phrase "Big Data" encompasses a broad range amounts of digitized dataset that are gathered through various businesses including governmental agencies. There are quintillion gigabytes of datasets produced per day, or 95% of all the data packets in existence presently, has only been generated in the previous two decades. The pace, quantity, as well as range of Big Data, including larger-scaling cloud architectures, a range of dataset origins as well as platforms, and the streamed aspect of dataset capture, including higher-volume intra-cloud movement, exacerbate cybersecurity as well as confidentiality challenges. The usage of larger-scaling clouds architecture, with a variety of technology packages distributed over vast computing networking, also makes the enterprise more vulnerable to assault [11], [12]. Figure 1 illustrates the major big data applications.

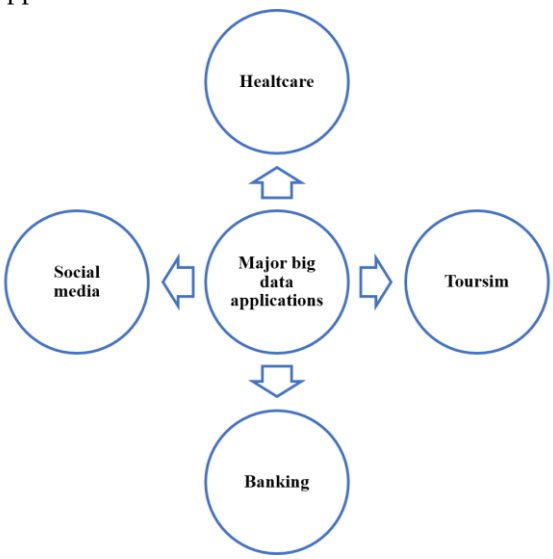


***Figure 1:*** *Illustrates the major big data applications.*

The 3 primary categories of the big dataset are organized, semi-structured, as well as unorganized. Relevant codecs including user-described ranges are preserved within structural metadata. Such datasets are created without the involvement of any people, whether manually or automatically. The structural dataset is generally processed using query techniques. Unstructured datasets are fundamentally different in terms of creation, storing, administration, as well as retrieving processes from the regular dataset, which are well-structured as well as kept in hierarchical networks so that entries may be quickly searched utilizing simple searching techniques. It's indeed difficult to handle effectively because of the absence of predictability, style, as well as pattern [13].

These days, businesses use Hadoop to handle unorganized datasets quickly and effectively utilizing the grouping method. Similar to XML, a semi-organized dataset does not fundamentally have a certain amount or format. Every day, unorganized dataset resources including social media, gadgets, blogs, including digitized images produce over 2.8 quintillion diverse bytes of datasets. Unstructured datasets are unquestionably growing quickly. Unstructured datasets are more difficult to sort through and analyze than organized datasets. Corporations are learning how to collect data that can be transformed into linkages as well as trends since it is so useful. Lack of data derived through unorganized datasets may cause chances to be missed. The easiest way to evaluate unorganized information is using cutting-edge analytical software. Companies may create strategies to lessen deception, stop criminality, find wastage, as well as uncover terrorist attacks utilizing such technologies. It is indeed very difficult to find trends that lead to significant findings because of the variety, speed, as well as amount of the datasets pouring throughout systems [14].

## II. LITERATURE REVIEW

In [15], D. Wu et al. discussed a novel scalable as well as privacy-preserving bigger dataset aggregation mechanism. Investigators are paying close focus on the gathering, preservation, transfer, as well as evaluation of

enormous sensor datasets since larger-scale Wireless Sensing Networks (WSNs), have lately constituted an essential component of Bigger Data. An energy-constrained massive sensor dataset gathering approach called Scalable as well as Privacy-preserving Bigger Datasets Aggregation (Sca-PBDA) was suggested throughout this research to address the confidentiality needs of larger-scale WSNs. In [16], P. Jain et al. discussed another review article regarding the secrecy and privacy of the larger datasets. Big datasets refer to exceedingly massive data collections with more intricate as well as diversified structures. Such traits typically result in increased challenges while preserving, analyzing, and using supplementary methods or retrieving findings. The phrase big datasets analytics refers to the overall practice of analyzing enormous volumes of complicated datasets to uncover buried trends or locate concealed relationships. Nevertheless, there seems to be a clear inconsistency between the growing usage of big datasets as well as their safety as well as confidentiality. To distinguish between confidentiality as well as protection and to address confidentiality needs in big datasets, this study concentrates on confidentiality as well as protection issues in big datasets.

In [17], W. Haoxiang et al. discussed the effective privacy preservation technique for big datasets to handle accurate dataset mining procedures. Physical settings, as well as everyday life, are permeated by computing technology, which also generates an enormous quantity of datasets that may be analyzed. Nevertheless, there's still increasing worry because if the acquired datasets aren't properly cleaned before being submitted for examination, possibly confidential information could become accessible. While there exist several privacy-preserving techniques, none of them are effective, adaptable, and therefore have no issues with confidentiality or the value of the datasets. To preserve the massive dataset's privacy, this study proposes a novel nonreversible perturbation method which is to be called PABIDOT, which is effective yet highly scalable compared to previous approaches.

In another study [18], M. A. P. Chamikara et al. discussed another scheme for big datasets using the distributing machine learning scheme. Edge computing as well as dispersed machine learning have developed to a point where they could completely transform a given business. Distributed gadgets, like IoT (Internet of Things), frequently generate a lot of datasets, leading to bigger data that could be crucial in revealing underlying trends as well as other breakthroughs in many industries, including medical, finance, and law enforcement. Datasets from industries like healthcare as well as finance may include highly sensitive information that, if improperly cleaned, might become available. However, this anticipated outline has miscellaneous restrictions in the modern world. In [19], Y. Zhao discussed another review study on diverse privacy-preserving schemes for providing secrecy as well as maintaining the confidentiality of the bigger data effectively. Presently data mining relies heavily on the grouping approach, which has been quite successful in addressing a variety of industry issues, including community analytics, picture recovery, tailored recommendations, activities forecasting, and so on. The classic grouping approach, as well as the newly developed numerous grouping technique, are both initially reviewed in this study. This enormous expense of processing as well as memory prevents the present approaches from performing well when grouping is done on the huge datasets, despite their improved effectiveness on certain smaller or specific samples. That problem could be successfully solved with the help of cloud technology, however, it poses a danger to the confidentiality of any consumer or business. In [20] C. Eyupoglu et al. discussed another effective large dataset anonymization set of rules rooted in the chaos as well as the perturbation methods. Big datasets had also recently drawn a growing amount of attention. Big data's growth creates significant challenges for the confidentiality control methods that are

essential for exchanging as well as analyzing information. The biggest problem in anonymity preservation is safeguarding critical data about particular people whilst keeping the usefulness of overall released data collection. Throughout this case, information anonymization techniques are used to safeguard records from identity theft as well as linkage assaults.

## III.METHODOLOGY

### *3.1 Design:*

The data variety, as well as volume, is increasing rapidly in the modern era. There are diverse sources of big data nowadays because of the accessibility of computing devices in multiple areas such as hospitals, educational institutions, large corporations, and many more. Most of the datasets which are to be transmitted over the server are unstructured which requires proper organization of easy accessibility along with required secrecy against diverse hackers in a device-to-device (D2D) communication environment. For maintaining big data privacy and secrecy, the authors designed a novel hybrid prototypical for big dataset privacy-preserving in a D2D communication environment.

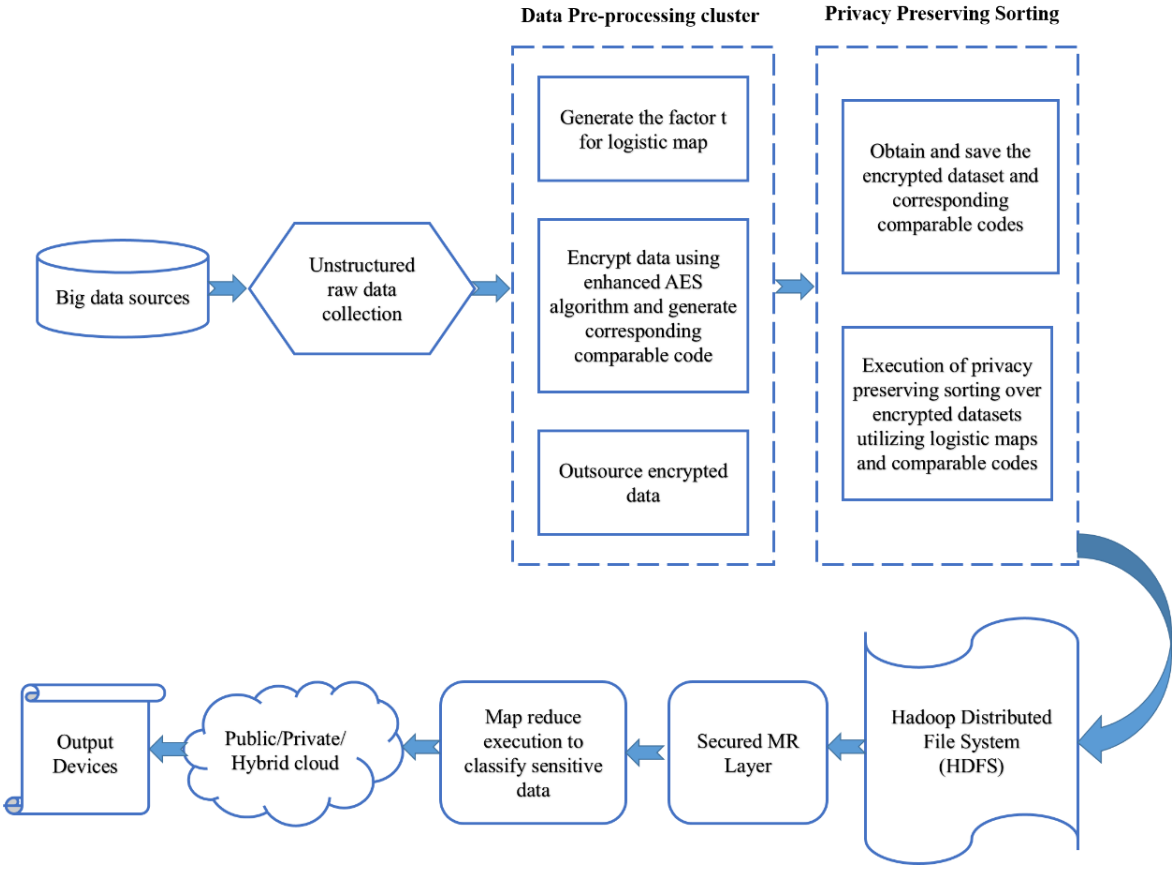


***Figure 2:*** *Demonstrates the design of a suggested hybrid model for big data privacy-preserving in D2D communication.*

Figure 2 illustrates the design of the suggested hybrid model for big datasets privacy-preserving within D2D communication. The working procedure of our suggested model is described as follows. Initially, the datasets produced by numerous computing gadgets such as smartphones, laptops, tablets, surveillance cameras, and many others are acquired and stored within a database. From this entire dataset, the unstructured raw data is collected and stored separately and the cleaning procedure of the data is done accurately for safeguarding the dataset's confidentiality as well as secrecy. In the subsequent step, the data pre-processing is accomplished. In this pre-processing step, first, originate the t factor for

the logistic map and then after the encryption procedure is done using the AES (Advanced Encryption Standard) algorithm and produce the corresponding comparable codes efficiently. Once the datasets are to be encrypted properly, then this dataset is outsourced for further processing i.e. privacy preserving sorting. In this segment, first, acquire and save these encrypted datasets along with the corresponding comparable codes in real-time. Later, the execution of privacy-preserving sorting over the encrypted datasets utilizing the logistic maps as well as comparable codes is accomplished. Once the privacy-preserving sorting is completed, the entire dataset is translated to the Hadoop-Distributing Files-System (HDFS). Just after the HDFS, we installed MR (Map-Reduce) layer for the dataset's secrecy as well as confidentiality against the hacker in real-time D2D communication. Later, the map-reduce execution procedure is done for classifying the sensitive datasets for additional datasets secrecy. After sensitive and other data segregation, all the datasets are to be translated over the public, private or hybrid cloud according to the data category in real-time to translate this cleaned and secured data to diverse devices.

### *3.2 Instrument:*

Our proposed model is designed and verified by using the most popular and recognized software platform Hadoop which is based on MapReduce. We utilized a personal computing system which includes the resulting system measures: 8GB RAM, SSD 1TB, OS (Operating System) 64 bits, processor i7 Intel along with Window 11. MapReduce is gaining more popularity these days as its open access platform, easy interface, and very minimal computational complexities. Furthermore, this toolbox helps researchers in effectively handling the critical datasets with required secrecy and privacy. This toolbox is one of the most pragmatic programming prototypical inside the Hadoop context which is utilized for accessing the bigger datasets inside the HDFS. This is a key element, integrated into the operation of the Hadoop framework.

### *3.3 Data Collection:*

Due to the increasing amount of dataset which is collected, saved, but also altered, conventional safety, as well as confidentiality solutions, seem inadequate to keep up with the transformations that Big Datasets has introduced towards the electronic realm. Access management restrictions, routers, as well as intrusion monitoring programs for networking protection could potentially be breached, allowing for the re-identification as well as linking of previously anonymized datasets to a particular person for malicious objectives. To deal with particular problems, a range of additional rules have indeed been suggested. Nevertheless, there have been numerous situations where earlier outlined guidelines could outcome in confidentiality infractions, including the preservation of electronic mails datasets for a particular time (for instance up to the 2 decades). Big Datasets have introduced obstacles to personal confidentiality, such as extrapolation as well as accumulation, that also make it feasible to re-identifying people after signifiers are eliminated from the selected essential dataset. Table 1 illustrates the randomized dataset's correlation coefficients.

***Table 1:*** *Illustrates the randomized dataset's correlation coefficients.*

| **Sl. No.** | **The ID of the used datasets** | **The scale of the dataset** | **Avg. value of correlation coefficient** |
|---|---|---|---|
| **1** | 1 | 500 | -0.145 |
| **2** | 2 | 1000 | -0.105 |
| **3** | 3 | 2000 | -0.085 |
| **4** | 4 | 4000 | -0.055 |
| **5** | 5 | 7000 | -0.064 |
| **6** | 6 | 12000 | -0.075 |

### *3.4 Pseudo Code:*

This safety pyramid, a long-standing conundrum, holds that when safety controls are tightened, the usability, as well as functioning of systems, are harmed. One should suggest a

reasonable strategy since, for instance, if a rule limits companies' accessibility to uncooked datasets analytics as well as modification, firms won't be allowed to expand existing operations. In short, this entire Big Dataset environment has to be reevaluated as well as closely studied within the context of safety as well as protection concerns, spanning architecture and administration to confidence rules, consistency, and datasets quality.

Algorithm 1:

Step 1: Initiate

Step 2: t = randm (from 0, $F_{min}$({$Fd_1$, $Fd_2$, $Fd_3$, $Fd_4$…….. $Fd_m$})/(1.$\mu^m$))

Step 3: for $Fd_i \subset$ {$Fd_1$, $Fd_2$, $Fd_3$, $Fd_4$…….. $Fd_m$} do

Step 4: Generate dataset pair $Fdg_i$, for $Fd_i$,

Step 5: $Fdg_i$.e = Encr ($Fd_i$, Fk); // Originate encrypted datasets

Step 6: $Fdg_i$.c = FL (t/$Fd_i$, Fn); // Originate comparable codes

Step 7: end for

Step 8: terminate

## IV. RESULTS AND DISCUSSION

Hadoop is an open-source distributed processing system that uses the MapReduce paradigm to process huge datasets and is extensively used for data processing by large corporations such as Google, Linked In, Facebook, and Yahoo. However, because this framework was not designed to operate in an untrustworthy environment, the required security safeguards were not included. Many Big Dataset technologies, including Hadoop, as well as Twitter Storm, the Pig, but also Hive, as well as MapReduce, and many more, lack adequate security protection, making infrastructure a secrecy issue for Big Datasets administration as well as analytics. Conversely, despite its safekeeping flaws, the Hadoop platform drew a lot of attention and was chosen as one of the key Big Data platforms, necessitating the need to figure out how to add the required security safeguards, especially because hackers frequently target data stored in the cloud. We'll go through Hadoop's two key security flaws here. (1-Cloud Data Access, 2-HDFS secrecy) as well as momentarily cover the strategies which may be utilized to ensure data security and privacy while creating a Hadoop system.

This name knob (for instance, a master node), dataset nodes, as well as secondary names nodes are the three primary components of Hadoop's distributed file system. To maintain availability and quick reaction times, HDFS produces several clones of each block of data. However, HDFS has certain authentication difficulties, for which Kerberos (authentication protocol) has been recommended as a solution for enabling the nodes to confirm their individuality to others. Additional trouble which HDFS confronts is the unavailability of the identification nodes (for instance, the masters naming nodes), for which the utilization of second naming nodes (for instance the slave naming nodes) is to be recommended, which may be retrieved if the master name node fails. If the criterion specified on Naming-Nodes Securities-Enhance (NNSE) grips, the administrator grants access to the slave node. The Bull's eye Algorithm is used for data monitoring to maintain the safety of duplicated datasets as well as to ensure that admittance is only allowed to authorized individuals.

Whenever a person engages in questionable conduct, the computer keeps a list of phrases connected with that behavior as well as creates a lower crucial record of that data. From there, a highly crucial record is created by comparing the regularity of something like the lower crucial records to see regardless of whether it has exceeded its upper threshold. The final stage of this self-assuring process stops suspicious individuals from using the connection. A completely automated surveillance framework that gathers as well as categorizes channel reports, acts as a filter for individuals which contextualizes as well as corresponds to the information to produce the required facts and figures but instead makes precise forecasts regarding connectivity behavior, as well as occurrences, is required to identify abnormalities inside the information flow brought on by the diverse disposition of Big Dataset.

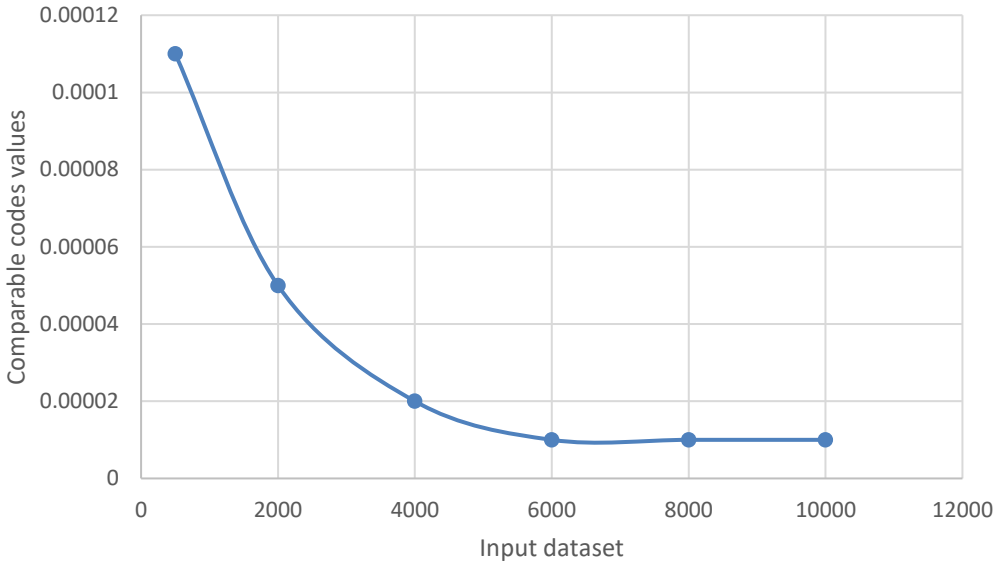


***Figure 3:*** *Illustrates the relationship between the input datasets as well as comparable codes.*

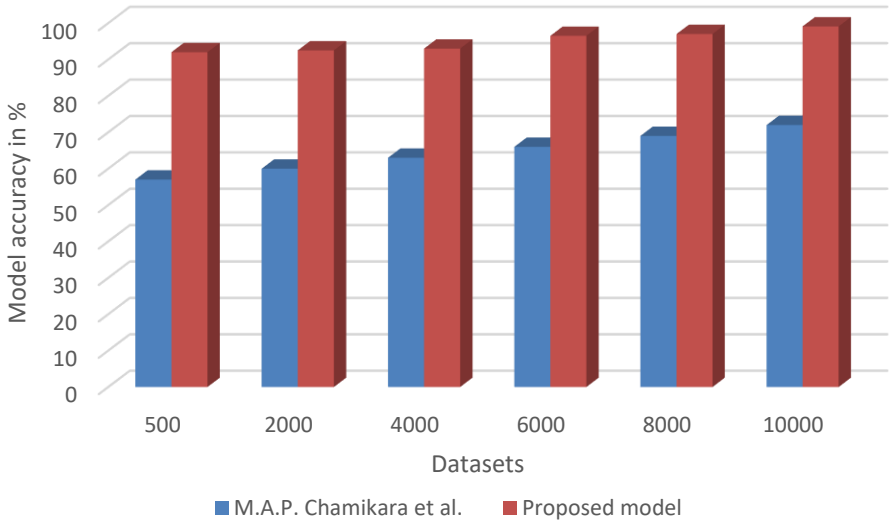


***Figure 4:*** *Illustrates the accuracy of the proposed model and the existing model in percentage.*

Figure 3 illustrates the relationship between the input datasets as well as comparable codes. The values of the comparable codes on chosen datasets i.e. 500, 2000, 4000, 6000, 8000, and 10000 are obtained at 0.00011, 0.00005, 0.00002, 0.00001, 0.00001, and 0.00001, respectively which are pragmatic and highly considerable in case of large datasets translation in the D2D communication environment. Figure 4 illustrates the accuracy of the proposed model and the existing model in percentage. For performance evaluation, we measured the accuracy of our model and compared it with the existing M.A.P. Chamikara et al. [21] model. From Figure 4, it is clear that our model accuracy is higher when compared with the existing model in the real-time execution of the model. Table 2 illustrates the dataset preprocessing time estimates (in ms). The dataset preprocessing time estimates (in ms) for the existing M.A.P. Chamikara [21] et al. model and proposed model on the scale of the dataset i.e. 500, 1000, 2000, 4000, 7000, 12000 are 0.98ms, 0.99ms, 1ms, 1.2ms, 1.4ms, and 1.8ms as well as 0.60ms, 0.67ms, 0.72ms, 0.74ms, 0.77ms, 0.82ms, respectively.

***Table 2:*** *Illustrates the dataset preprocessing time estimates (in ms).*

| Sl. No. | The scale of the dataset | M.A.P. Chamikara [21] et al. model | Proposed model |
|---|---|---|---|
| **1** | 500 | 0.98 | 0.60 |
| **2** | 1000 | 0.99 | 0.67 |
| **3** | 2000 | 1 | 0.72 |
| **4** | 4000 | 1.2 | 0.74 |
| **5** | 7000 | 1.4 | 0.77 |
| **6** | 12000 | 1.8 | 0.82 |

## V. CONCLUSION

Regarding big data, the question of privacy and security is paramount. This is because the big dataset safekeeping pattern is not indorsed for complicated applications and is disabled through evasion. The dataset, on the other side, could always be negotiated without it. Privacy and security are the primary concerns in this area. Permission to gather and utilize personal information is referred to as privacy. As a result, individuals and groups can prevent the disclosure of their personal information to persons other than those who provide it. During Internet transmission, personal information might be identified, which poses a major privacy risk to users. Datasets confidentiality is apprehensive along with the use as well as control of individual datasets, namely the launching procedures to assure that consumer individual datasets are collected, communal, as well as used in the most suitable manner possible. Big Data is enabled by the ability to receive a stream of data from a diversity of sources and in a variety of formats: structural data or non-structural data, for instance. As a consequence, this is more essential than ever to ensure that the dataset is accurate and reliable so that it can be used effectively. The estimated dataset preprocessing times (in ms) for the conventional M.A.P. Chamikara et al. approach as well as the suggested model is 0.98ms, 0.99ms, 1ms, 1.2ms, 1.4ms, and 1.8ms, as well as 0.60ms, 0.67ms, 0.72ms, 0.74ms, 0.77ms, and 082ms, correspondingly. In the past, integrity has been described as the ability to maintain data's consistency, correctness, and trustworthiness during its lifespan, it safeguards data from illegal modifications.

## REFERENCES


1. R. Beakta, “Big Data And Hadoop: A Review Paper,” Spl. Issue, 2015.
2. N. Elgendy and A. Elragal, “Big data analytics: A literature review paper,” 2014. doi: 10.1007/978-3-319-08976-8_16.
3. E. Al Nuaimi, H. Al Neyadi, N. Mohamed, and J. Al-Jaroodi, “Applications of big data to smart cities,” J. Internet Serv. Appl., 2015, doi: 10.1186/s13174-015-0041-5.
4. “A Review Paper on Ranking of Product on Big Data,” Int. J. Innov. Eng. Technol., 2017, doi: 10.21172/ijiet .81.033.
5. J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, and X. Zhou, “Big data challenge: A data management perspective,” Front. Comput. Sci., 2013, doi: 10.1007/s11704-013-3903-7.
6. M. Martis, N. V. Pai, R. S. Pragathi, S. Rakshatha, and S. Dixit, “Comprehensive survey on hadoop security,” 2019. doi: 10.1007/978-981-13-6001-5_17.
7. S. Sinha, S. Gupta, and A. Kumar, “Emerging Data Security Solutions in HADOOP based Systems: Vulnerabilities and Their Countermeasures,” 2019. doi: 10.1109/ICCCIS48478.2019.8974535.
8. N. Sirisha, K. V.D. Kiran, and R. Karthik, “Hadoop security challenges and its solution using KNOX,” Indones. J. Electr. Eng. Comput. Sci., 2018, doi: 10.11591/ijeecs.v12.i1.pp107-116.
9. M. M. Shetty and D. H. Manjaiah, “Data security in Hadoop distributed file system,” 2017. doi: 10.1109/ICETT.2016.7873697.
10. H. Chen and Z. Fu, “Hadoop-Based Healthcare Information System Design and Wireless Security Communication

Implementation," Mob. Inf. Syst., 2015, doi: 10.1155/2015/852173.
11. A. A. Bakar, R. Ramli, and F. A. Rahim, "Efficient Cryptographic-Based Technique for Privacy Preservation in Industries Practising Big Data," Adv. Sci. Lett., 2018, doi: 10.1166/asl.2018.11799.
12. A. Elsir, O. Elsier, A. Abdurrahman, and A. Mubarakali, "Privacy preservation in big data with data scalability and efficiency using efficient and secure data balanced scheduling algorithm," J. Sci. Ind. Res. (India)., 2019.
13. S. Kim, H. Lee, and Y. D. Chung, "Privacy-preserving data cube for electronic medical records: An experimental evaluation," Int. J. Med. Inform., 2017, doi: 10.1016/j. ijmedinf.2016 .09.008.
14. K. Sujatha and V. Udayarani, "Deep restricted and additive homomorphic ElGamal privacy preservations over big healthcare data," Int. J. Intell. Comput. Cybern., 2022, doi: 10.1108/IJICC-05-2021-0094.
15. D. Wu, B. Yang, and R. Wang, "Scalable privacy-preserving big data aggregation mechanism," Digit. Commun. Networks, 2016, doi: 10.1016/j.dcan.2016.07.001.
16. P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: a technological perspective and review," J. Big Data, 2016, doi: 10.1186/s40537-016-0059-y.
17. W. Haoxiang and S. S, "Big Data Analysis and Perturbation using Data Mining Algorithm," J. Soft Comput. Paradig., 2021, doi: 10.36548/jscp .2021. 1.003.
18. M. A. P. Chamikara, P. Bertok, I. Khalil, D. Liu, and S. Camtepe, "Privacy preserving distributed machine learning with federated learning," Comput. Commun., 2021, doi: 10.1016/j.comcom .2021.02.014.
19. Y. Zhao, S. K. Tarus, L. T. Yang, J. Sun, Y. Ge, and J. Wang, "Privacy-preserving clustering for big data in cyber-physical-social systems: Survey and perspectives," Inf. Sci. (Ny)., 2020, doi: 10.1016/j.ins. 2019.10.019.
20. C. Eyupoglu, M. A. Aydin, A. H. Zaim, and A. Sertbas, "An efficient big data anonymization algorithm based on chaos and perturbation techniques," Entropy, 2018, doi: 10.3390/e20050373.
21. M. A. P. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil, "Efficient privacy preservation of big data for accurate data mining," Inf. Sci. (Ny)., 2020, doi: 10.1016/j.ins.2019.05.053.