Journal of Advances and Scholarly Researches in Allied Education
Vol. 21, Issue No. 7, October-2024, ISSN 2230-7540

DOI: https://doi.org/10.29070/qnkj2h53

Empirical Assessment of Privacy-Preserving Techniques in Big Data Analytics

Farendrakumar Shrawan Ghodichor [1] * , Dr. Suraj Vishwanath Pote [2]

1. Professor, Department of Computer Science & Engineering, University of Technology, Jaipur, Rajasthan, India
farendrakumarghodichor@gmail.com ,
2. Professor, Department of Computer Science & Engineering, University of Technology, Jaipur, Rajasthan, India

Abstract: A privacy-preserving paradigm for large data data mining is proposed in this research. Two potential use cases have been defined according to the suggested architecture. The first is a collaborative filtering method that protects patients' privacy and is used to generate recommendations in a healthcare system where patients' data is randomly dispersed across different locations. When data is handled quickly and the information it contains can be used as a basis for decisions, the true benefit of big data may be realized in today's data-driven world. Using data mining techniques, valuable insights and patterns have been uncovered inside massive databases. Aside from the potential benefits to analytics, there are a number of security concerns that might arise from making all of this data available to data miners. This is because bad actors could potentially abuse this data. Therefore, it is important to strike a balance when it comes to the availability and security of data, as it is necessary to protect critical information without compromising application performance.

Keywords: privacy-preserving, data analytics, big data, multiple healthcare, filtering technique

X

INTRODUCTION

The proliferation of computer-based solutions across all industries has led to an explosion in both the quantity and diversity of available data. Reasonably priced computing, storage, and network connection have enabled the expansion. A lot of data, including personally identifiable information (such as a person's gender, zip code, health condition, caste, purchase history, religious beliefs, etc.), is preserved in the public domain. If the data holder wants to help improve companies, give value-added services to consumers, make predictions, forecasts, and recommendations, they may provide their data to an outside data analyst. The analyst will then analyze this information to uncover hidden patterns and obtain deeper insights. Among the most well-known uses of data analytics is in recommendation systems, which are used extensively by online marketplaces such as Amazon and Flipkart to propose goods to consumers according to their past purchases. Based on our interests, Facebook does propose friends, locations to visit, and movies to watch.

Disclosure of user activity data, however, might open the door to inference attacks, such as the possibility of gender identification using this data. In order to safeguard users' privacy, we have researched several methods that are already in use. There are benefits and drawbacks to each of these methods. The benefits and drawbacks of different approaches will be examined in this research, as well as the obstacles to privacy preservation research. The value of data and individual privacy are inherently incompatible. Privacy preservation in unstructured data with maximal data usefulness is handled by a data lake based

modernistic approach, which is also proposed in this work.

Big data refers to the massive assemblage of massive datasets that conventional computer methods are ill-equipped to handle. Accordingly, Big Data is massive amounts of data that combine several data types and are processed rapidly. Managing such massive amounts of data (at least 1 terabyte) has become an issue for many businesses as a result of the exponential expansion of data. Web servers, transactions, commercial sites, social networking sites, and other sources generate both structured and unstructured data, ushering in the big data age. To paraphrase Apache Hadoop (2010), "datasets which could not be captured, managed, and processed by general computers within an acceptable scope" [1] constitute large data. Processing or analyzing the vast quantities of data needed to get useful insights is the primary obstacle in big data analysis. One way to characterize big data is as seven V's problems.

The term "Big Data" describes a collection of data sets so large that traditional computer methods are inadequate for handling them. Big Data is more than just data; it incorporates data created by a wide range of devices, apps, and other technological infrastructures. Data captured from aircraft and helicopters that includes crew voices, various microphone recordings, etc., are just a few examples. In many fields, including medicine, academia, and business, publishing data is essential (Goswami 2017). Hadoop and other distributed programming environments are handling massive amounts of data as a consequence of problems with open platforms like social media and mobile devices. According to Mehta and Rao (2017), "Big Data" refers data sets that defy traditional processing methods due to their size or complexity methods to handle well. The three Vs of big data analytics are volume (a large quantity of data), variety (velocity (the pace of data generation and processing), and data types (structured, unstructured, and semi-structured)., also known as term of time or rush data). Subsequent research noted that the 3V's description was inadequate for explaining the current big data scenario, thus it was expanded to include two more V's: variability and veracity.

Big data analysis has difficulties due to its bulk, pace, diversity, unpredictability, honesty, and ambiguity. Media files such as text, music, images, videos, etc., are examples of information gathered coming from a wide range of places. Linking structured, semi-structured, and unstructured data sets is what data integration is all about. This method, however, is ineffective against linkage assaults. For ways to share data while protecting users' privacy, many anonymization methods are being considered, such as k-anonymity. Below is the outline for the remainder of the article. In Section 2, the characteristics of big data are outlined.. The third section provides a synopsis of the issues surrounding data privacy. Methods for Privacy-Preserving Data Mining are Surveyed in Section 4, and Methods for Privacy-Preserving Data Publishing are Surveyed in Part 5. Section 7 concludes the essay, whilst Section 6 offers a summary of the review.

LITERATURE REVIEW

Venish, Luan et.al. (2023) Many sectors, including healthcare, banking, marketing, and more, have been profoundly affected by the fast development of big data analytics. Yet, serious worries over data privacy have surfaced in relation to the gathering, storing, and processing of massive datasets. The extraction of insights from large databases while protecting people' privacy is a pressing issue. While significant data analysis is still possible, privacy-preserving strategies in big data analytics seek to safeguard sensitive

information during data collection, storage, and processing. Examining the merits, shortcomings, and potential uses of alternative privacy-preserving methods, this study delves into topics including differential privacy, anonymization, and encryption. The advantages and disadvantages of these technologies for improving data privacy without sacrificing analytical skills are also discussed in the paper, as are the difficulties brought about by the sheer volume and complexity of big data. With an eye toward the future, this article surveys the current landscape of privacy-preserving techniques and investigates potential avenues for incorporating privacy protection into big data analytics.

Shah, Samarth et.al. (2024) Opportunities for insights via big data analytics have been generated by the exponential expansion of data in recent years. But there are serious privacy problems that arise from this data explosion, particularly when personal details are at stake. To guarantee the ethical use of data while keeping analytical effectiveness, privacy-preserving approaches have become an important focus of study. Modern methods for protecting personal information in big data analytics are discussed in this article. Methods including homomorphic encryption, safe multi-party computing, differential privacy, and data anonymization are tested for their performance and domain-specific application. Despite ongoing obstacles such as re-identification attacks, anonymization approaches continue to strive for the removal of personally identifying information while maintaining analytical value. Differential privacy strikes a balance between privacy and accuracy by introducing calibrated noise to data. With secure multi-party computing, remote systems may benefit from collaborative analytics without compromising data privacy. Homomorphic encryption guarantees security during analysis by enabling calculations on encrypted data. Dealing with computational overheads, scalability, and legal compliance are essential components of integrating privacy-preserving approaches into big data operations. Combining several privacy-preserving strategies may improve overall resilience and overcome individual constraints, as shown in this work. Such methods are becoming more important due to the evolution of privacy legislation such as GDPR, and comparable frameworks.

Baig, Hidayath. (2020) The amount from embedded devices' data (e.g., sensor readings, automobiles, and cellphones) is growing at an exponential rate in today's technology-driven society. Massive data expansion has resulted from people's involvement with the Internet, particularly in the realms of e-commerce, e-governance, and social media. While there has been some early success from a technology standpoint in handling such massive amounts of data, the potential advantages of this data explosion are substantial and helpful. Storage, data interchange, curation, transportation, analysis, visualization, analytics, privacy, and information in general are just a few of the data-related issues that come along with the advantages. Given the widespread agreement that data privacy is paramount, protecting information derived by data analytics ought to be an equally pressing concern. Therefore, protecting the confidentiality of data produced by big data analytics is an essential subject that this study aims to investigate. In this work, we first examine the many privacy-threatening phenomena that exist in the context of big data platforms, such as disclosure, monitoring, identification, discrimination, etc., and then we examine the several privacy-preserving approaches that are now accessible. We would also want to talk about what we can do next to make big data analytics more private by pointing out their shortcomings.

Shekhawat, Hema et.al. (2019) Big Data refers to massive datasets that are too big to be handled by conventional computing methods. There are a variety of forms that make up big data, including structured

and unstructured data. Analyzing large data using the methods now used by typical RDBMS is inefficient in terms of reaction time. Due to several characteristics, such as quantity, diversity, worth, unpredictability, visibility, speed, and authenticity, big data presents numerous obstacles. Because big data consists of massive datasets that cannot be stored on a single system, a new distributed platform is necessary to evaluate and store these datasets for use in future predictions and decisions. To handle and store this massive quantity of data on commodity technology, Hadoop offers a framework for distributed data processing. With Hadoop on the cloud, it's possible to process distributed queries across many datasets. The cloud is where big data is kept so that massive datasets may be processed and analyzed. Using privacy-preserving methods in untrusted cloud servers has become more difficult as a result of this paradigm shift. The security of the private data was guaranteed using a privacy-preserving technique using cryptography. In order to ensure the privacy and security of data, three methods are presented in this study. The article explains homomorphic encryption, attribute-based encryption, and order-preserving encryption methods. These methods work best in the cloud to protect sensitive information, and they're ideal for big data since they keep massive datasets efficient and scalable, which is crucial for making decisions.

Tran, Yen et.al. (2019) We provide a thorough overview of big data analytics that protect user privacy in this article. We provide well-crafted taxonomies that categorize this difficult area of study in great depth and provide systematic viewpoints. We shed light on new research on hot subjects in the area. We also highlight potential areas for further study with respect to future big data analytics that ensure user privacy. In order to meet the many different privacy-related circumstances that may arise in reality, this survey might be a valuable reference for developing new privacy-preserving strategies.

RESEARCH METHODOLOGY

The EMS-PPCF system model. Assume that A1, A2,..., Ap are the entities who want to work together in order to formulate the suggestions. In a situation where DAz is an element of Az, the whole dataset D is randomly distributed among several parties (p), as shown by D = DA1 || DA2 || DA3...DAp. The item set IAz belongs to DAz and is unique to each Az party. In order to determine how similar their item sets are to one another; they exchange them in the protected form. The rating matrix for party Az is represented by the item vectors in the item set IAz. Everyone involved with Presumably, EMS-PPCF knows the user ID and the item label the public. Various parties are able to stay in sync with this. Using the encrypted form, we just exchange the ratings. Each party uses these item similarities as an offline model ModelAz to deliver the user a forecast shows that when one party, A1, gets a query qa, about item i, as submitted by user a, it uses its offline model, ModelA1, to construct the prediction Preda,i and transmits it to user a. All parties A2...Ap are collaborating with party A1, which is called the Master Party (MP).

DATA ANALYSIS

A dataset with numerical ratings has been used for recommendation generating purposes due to the lack of a dataset based on cumulative binary ratings ClaMPP operates by using three datasets that are available to the public: Movielens Public (MLP), Movielens 20M (MLM), and Jester experimentally examined to assess the quality of its predictions. These datasets' attributes are shown in Table 4.1. Following the steps outlined Binary format is applied to these datasets. Each dataset is divided into a training dataset that makes up 70% and a testing dataset that makes up 30%. To evaluate the accuracy of the predictions, a 10-

fold cross-validation procedure is used. There are two measures taken into account: the F-measure and the Classification Accuracy (CA) (F1). Resolving the proportion of correctly predicted items to the total number of forecasts made is how CA is calculated. For every user u, if we have a set of x recommendations (Pu), then CA is:

$$CA = \sum_{u \in U} \frac{\#(l \in P_u | p_{u,l} = r_{u,l})}{x}$$

where pu,l is prediction generated for user u, for item l and ru,l is an original prediction for user u of item l.

F1 is computed as:

$$F1 = \frac{2\psi\omega}{\psi + \omega}$$

where ψ stands for accuracy, which is the ratio of relevant suggestions to all recommendations, and ϋ for recall, which is the ratio of relevant suggestions to all relevant items. ψ and ω are calculated as:

$$\psi = \frac{1}{U} \sum_{u \in U} \frac{\#(l \in P_u | p_{u,l} = r_{u,l})}{x}$$

$$\omega = \frac{1}{U} \sum_{u \in U} \frac{\#(l \in P_u | p_{u,l} = r_{u,l})}{(l \in P_u | r_{u,l} = 1) + (l \in P_u | r_{u,l} = 0)}$$

We first analyze our experimental data to make sure that ADD-collaboration does, in fact, increase the prediction quality of all involved participants. Figure 1 shows the results of calculating Given different datasets and parties (t), the CA and F1 values. There are seven different groups to which the suggestion data is randomly assigned. As indicated in Table 4.2, CA is determined when one party makes suggestions for t = 1. We take into account the average of each party's CA values while making forecasts. At time t = 3, CA is shown by the collaborative suggestion generation that takes place between three participants. We have explored every conceivable permutation of three individuals (t=3) from a pool of seven for the purpose of developing joint suggestions. They are evaluated based on the mean of their individual CA scores.

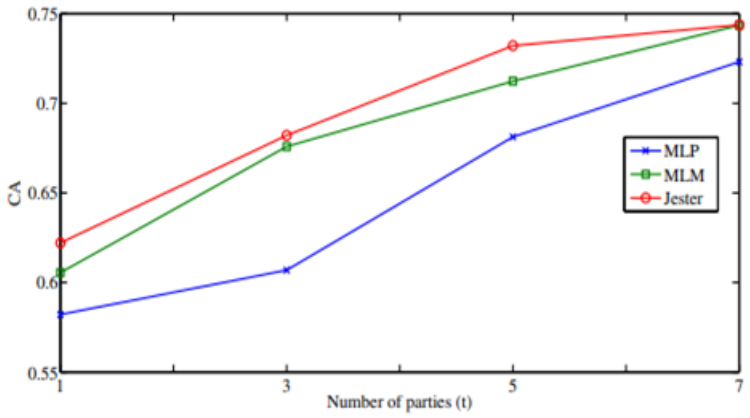
Table 1: CA values for t = 1

S.no	Party	CA		
		MLP	MLM	Jester
1	1	0.5911	0.6122	0.6534
2	2	0.586	0.5987	0.6211
3	3	0.5521	0.5993	0.5989
4	4	0.6129	0.6212	0.6434
5	5	0.5744	0.6116	0.6061
6	6	0.5779	0.5829	0.6326
7	7	0.5801	0.6129	0.5991
Average		0.5821	0.6055	0.6221

Performance

When new data is added into the off-line model, it needs to be updated frequently. Because of this, its calculation time shouldn't be too high, as this might impede the updating process. While there is a comparable technique that only works with two parties, our suggested solution is designed for many parties. As a result, we have only looked at two parties in our experimental comparison of the off-line computing costs of ClaMPP with another analogous technique. When t=2, ClaMPP's off-line calculation cost is O(m)HPE.enc, Mul, and Dec. Find the overall running time of cryptographic functions using the benchmarks offered at CRYPTO++ tool kit. On that basis, displays the calculation time required to generate models using various techniques. Compared to the other comparable approach, ClaMPP's model generation time is lower.

Collaboration improves prediction quality, as shown in Figure 2, where the CA and F1 values rise as the number of cooperating parties (t) grows. When computing results on integrated data, there must be little to no loss of accuracy as a consequence of privacy protections. During an on-line period, ClaMPP does not execute any operations, which impacts the quality of the predictions. Using the PPCPP technique to calculate likelihood probabilities in the offline phase has no effect on the output accuracy, and using homomorphic encryption-based procedures to compute prior probabilities has the same effect. The same MLP dataset is used for all experimental trials to confirm that ClaMPP does not lose accuracy. Figure 3 displays the results from the original combined data set with and without ClaMPP, demonstrating that the use of privacy measures does not compromise accuracy.



(a) CA Values

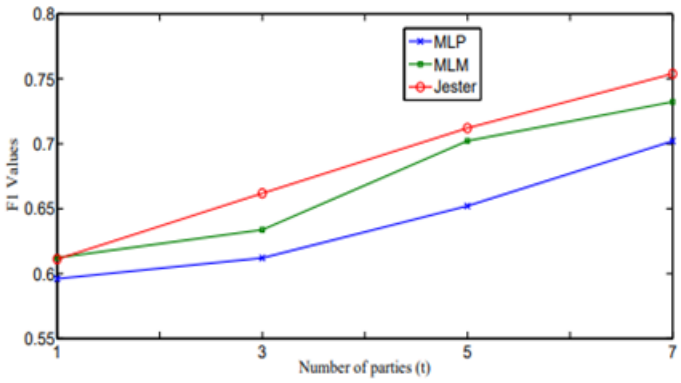


Figure 2: Results for the MLP, MLM, and Jester datasets in terms of CA and F1

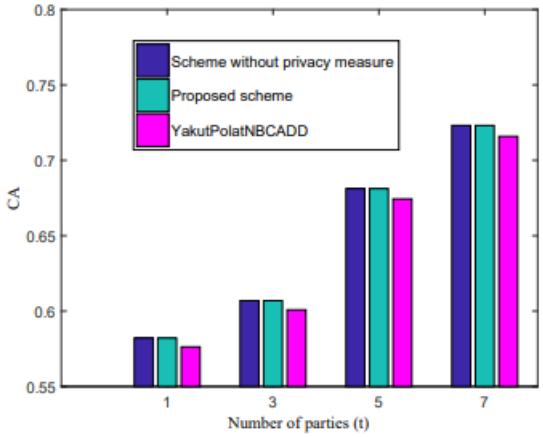


Figure 3: Comparing CA values for MLP with and without privacy parameters

It may become difficult to synchronize the actions carried out by several parties and operate individual cryptographic primitives over diverse network domains, which is a concern with the suggested system. The user-item interaction is seen in both the MLP and MLM datasets. There is only one interaction type in the Jester dataset as well: user-joke. As previously mentioned by Pham et al., datasets also represent the homogenous information network.

CONCLUSION

The rapid development of IT has had a profound impact on the healthcare sector. A lot of people who are sick or injured look for information online, including symptoms, causes, diagnosis, treatment options, doctors, and hospitals. With effective data collection, mining, and analysis, patient-oriented decision-making may boost the healthcare recommender system's efficiency. This is assuming that the data is spread out throughout various geographical locations. For healthcare suggestion generation on ADD among several parties, this paper proposes a PPCF method that employs the use of homomorphic encryption, masking, and randomization. Three methods have been proposed in our research. Predictions are generated in Protocol III, item vector lengths are computed in Protocol II using the homomorphic encryption

methodology, and item similarity is produced in Protocol I using the PPSDPC method. While being computationally efficient, the proposed technique safeguards the patient's anonymity. A multi-party privacy-preserving naive Bayesian classification system that operates on the cloud algorithm called ClaMPP is suggested for use in an e-commerce recommender system. We have put out four procedures. Secure calculation of conditional probabilities using the PPCPP method is carried out by Protocols I and II. Homomorphic encryption is used by Protocol III for prior probability calculations, while online prediction generation is carried out by Protocol IV. Multiple parties working together increase the accuracy of prediction generation and ensure the security of the proposed plan. Furthermore, the suggested method's accuracy loss as a result of privacy protections is almost nonexistent.

References

1. Venish, Luan & Blake, Harrison & Bernardi, G.. (2023). Privacy-Preserving Techniques in Big Data Analytics: Challenges and Opportunities.
2. Shah, Samarth & Khan, Shakeb. (2024). Privacy-Preserving Techniques in Big Data Analytics. 1. 521-541.
3. Baig, Hidayath. (2020). Privacy-Preserving in Big Data Analytics: State of the Art.
4. Shekhawat, Hema & Sharma, Samiksha & Koli, Reetika. (2019). Privacy-Preserving Techniques for Big Data Analysis in Cloud. 1-6. 10.1109/ICACCP.2019.8882922.
5. Tran, Yen & Hu, Jiankun. (2019). Privacy-preserving big data analytics a comprehensive survey. Journal of Parallel and Distributed Computing. 134. 10.1016/j.jpdc.2019.08.007.
6. Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2017). A survey on homomorphic encryption schemes: Theory and implementation. ACM Computing Surveys (CSUR), 49(4), 1–35.
7. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. IEEE Symposium on Security and Privacy (SP), 3–18.
8. Kairouz, P., McMahan, H. B., Avent, B., et al. (2019). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1–210.
9. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1–19
10. Gentry, C., Halevi, S., & Vaikuntanathan, V. (2015). Homomorphic encryption from learning with errors: Concept and efficiency. Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS), 97–106.
11. Dwork, C., & Roth, A. (2016). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407
12. Zhu, T., Li, G., & Zheng, K. (2020). Privacy-preserving data mining techniques: Trends and challenges. IEEE Transactions on Knowledge and Data Engineering, 32(1), 141–156

13. Chamikara, M. A. P., Bertok, P., Khalil, I., Liu, D., & Camtepe, S. (2018). Efficient privacy-preserving protocol for big data storage and querying. Future Generation Computer Systems, 83, 151–160

14. Zhang, Y., Yang, Q., & Chen, T. (2021). Privacy-preserving machine learning with homomorphic encryption and secure computation. ACM Computing Surveys (CSUR), 54(4), 1–36

15. Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. (2019). Tools for privacy-preserving distributed data mining. Journal of Knowledge and Information Systems, 59(2), 287–314