

Privacy-Preserving Using Big Data Analytics

Dr. Geetha M¹, Bharath G N²

Associate Professor, Department of Master of Computer Applications, BIET, Davangere-577004.

bmgeethamallad@gmail.com, Davangere, Karnataka, India¹

P.G. Student, Department of Master of Computer Applications, BIET, Davangere-577004.

Bharathbharath4487@gmail.com, Davangere, Karnataka India²

ABSTRACT

In today's digital world, a huge quantity of information is gathered per second from internet shopping, social media, mobile apps, and numerous other venues. This large amount of information is called Big Data. Organizations benefit from this data analysis. and researchers to find useful patterns, make decisions, and improve services. But when looking at data, it's very important to protect users' privacy because personal and sensitive information can be shared if it's not done right.

The objective of the "Privacy-Preserving Using Big Data Analytics" project is to locate ways to analyze data without revealing any personal or sensitive information. This project uses two privacy-preserving methods: encryption and differential privacy. Differential privacy guarantees that no individual in the dataset may be identified by the analysis's findings. To do this, the method adds random noise to the data or the output. This way, the method keeps the data accurate while also protecting privacy. Encryption is another method utilized for this project. The method involves changing regular data into an encrypted format that is only readable by those who are permitted. The method helps keep data safe while it's being stored and sent over different networks in this way.

This project creates a user interface that lets people safely upload and analyze data. The API layer makes it possible for the interface and the server to talk to each other. The server processes data in an encrypted form. The models used in machine learning are used on encrypted or differentially private data to get useful information without giving away personal information. The project's objective is to create a large data analysis system that is dependable, safe, and aids in research and corporate decision-making without endangering user privacy. This system can be used in many areas where keeping data private is very important, like healthcare, banking, education, and government data analysis. In general, privacy-preserving technologies like encryption and differential privacy can make big analytics safer, more ethical, and more reliable for use in the future.

KEYWORDS : Data Privacy and Security, Differential Privacy Mechanism, Data Encryption for Protection, Secure Data Sharing and Collaboration, Federated Learning Support, Data Integrity and Authenticity, Transparency and Auditability, Bias Reduction and Fairness.

I INTRODUCTION

In the current technological era, researchers generate enormous volumes of data every second. industrial organizations, healthcare services, educational institutions, and many other domains. The great volume of data what we term big data. It is a tool which organizations use for decision making. Personal data, health care data, and financial data must be protected from access by unauthorized parties. This research paper which we have named "Privacy Preserving Using Big Data Analytics" looks at private data analysis.

Our goal in this research is to present a model which will allow researches or organizations to look at their data but which also guarantees privacy. In terms of differential privacy we add noise to the data which randomizes it and thus it is impossible to determine an individual which a certain set of data pertains to. Also we see in encryption which the data is put through a series of calculations which in turn codes the info.

Conclusion. The main goal of this paper is to show that when analyzing big data, there should be a method in place to ensure security, privacy, and data integrity. This approach would be very helpful in various fields, such as healthcare, banking, education, and government. It allows these sectors to analyze data without disclosing any private information.

II LITERATURE REVIEW

Recently, privacy issues related to Big data analysis has been observed due to the sudden increase in huge amounts of data which is very prevalent in fields such as healthcare, financial institutions, social media platforms, and Internet of Things(IoT) applications. Early research on this topic has focused on approaches such as k-anonymity, l-diversity, and t-closeness to preserve privacy. But these techniques have been noted to be vulnerable to attacks based on background knowledge and linkage.

Differential privacy is a method that is suitable for providing stronger privacy where noise is added in query answers to guarantee the privacy of individual entries, according to Cynthia Dwork. Another way to protect privacy is by using cryptographic approaches. One such approach is homomorphic encryption proposed by Craig Gentry where computations can be performed on encrypted information without decrypting it first. Another approach where collaboration is possible without compromising sensitive information is through the method of secure multi-party computation that relies on cryptographic approaches such as garbled circuits by Andrew Yao.

In conclusion, The current output of literature focuses on ensuring privacy while maximizing data usage, enhancing scalability, and adopting hybrid methodologies.

III EXISTING SYSTEM

1. Traditional Big Data Analytics Systems : In traditional big data systems, collected information from several sources is stored and processed in a centralised server or cloud platform. Large-scale analysis is possible with these platforms, but when private data—such as financial transactions, health records, and personal information—is stored simultaneously, there are significant security and privacy risks.

2. Rule-Based and Access Control Systems : Older data analytics systems used rule-based security or access control lists to manage who could view or modify data. They relied on fixed policies like “who can access what data” However, these systems weren't smart enough to handle large-scale or sensitive information dynamically.

3. Differential Privacy-Based Analytics Systems : A modern era privacy concept called Differential Privacy (DP) was created to protect individual-level information in data analysis. It introduces random sound to the dataset or query output so that the contribution of any single person's data cannot be identified.

4. Federated Learning and Distributed Privacy Systems : Without gathering raw data in one location, Model training is made possible via federated learning across several devices. To protect user data on their local devices, only the model updates are sent to a central server.

5. Hybrid Privacy-Preserving Frameworks : Many modern systems use a hybrid approach by combining Differential Privacy, Encryption, and sometimes Federated Learning to enhance security. This layered protection makes it harder for attackers to breach data at any single point.

6. Real-Time Monitoring and Secure Interface : Administrator Interface: Modern systems include real-time dashboards that allow administrators to keep an eye on system performance and detect suspicious activity.

IV PROPOSED SYSTEM

The present solution is that which we put forth an advanced system which uses many privacy technologies in big data analytics.

We put forward the idea of privacy and also efficient results. As was the case with past methods that used a single technology for issues of data accuracy and privacy, what we have is a very different approach which uses many cryptographic techniques which we apply from the time of data collection through to its analysis.

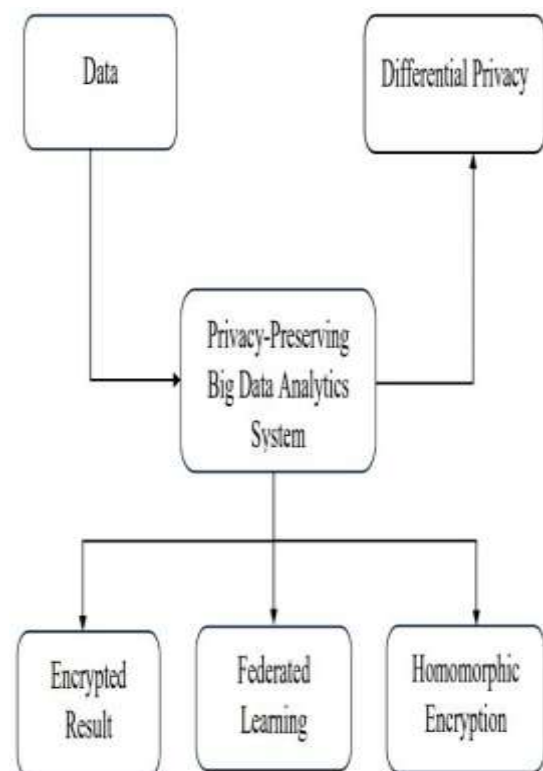
Homomorphic encryption is used in cases of very private info which has to be processed as a whole in encrypted form. The machine does not get to see any unencrypted info at all.

We use this system for compliance with privacy regulations also it facilitates secure internal info transfer and at the same time it instills trust in systems which handle large scale data. Also it is a feature in many fields like health, banking, schools etc. due to present day privacy solutions and engineering improvements.

Advantages of Proposed System

- 1. Enhanced Data Privacy :** The system protects sensitive information by applying privacy-preserving techniques such as data anonymization, differential privacy, encryption, or secure computation, lowering the possibility of data breaches and unauthorized access.
- 2. Secure Data Processing:** It enables analytics on encrypted or protected data without exposing raw information, ensuring confidentiality during computation and storage.
- 3. Improved Trust and Compliance:** By safeguarding personal data, The system assists businesses in adhering to data protection laws. (such as GDPR and other privacy laws) and builds user trust.
- 4. Maintains Data Utility:** The proposed approach balances privacy and analytical accuracy, ensuring that useful insights can still be extracted from large datasets without compromising sensitive information.
- 5. Big Data Scalability :** The system is made to manage high-volume, high-velocity, and high-variety datasets efficiently, making it suitable for real-world big data environments.
- 6. Protection Against Re-identification Attacks :** Advanced privacy techniques reduce the chances of linking anonymized data back to individuals.
- 7. Supports Distributed Environments :** The system can work in cloud, edge, or federated environments where data remains decentralized while still enabling collaborative analytics.
- 8. Decreased Chance of Data Abuse :** Since raw Sensitive information is not immediately visible, but insider threats and third-party misuse is significantly minimized.

SYSTEM ARCHITECTURE



V EXPERIMENTAL RESULTS

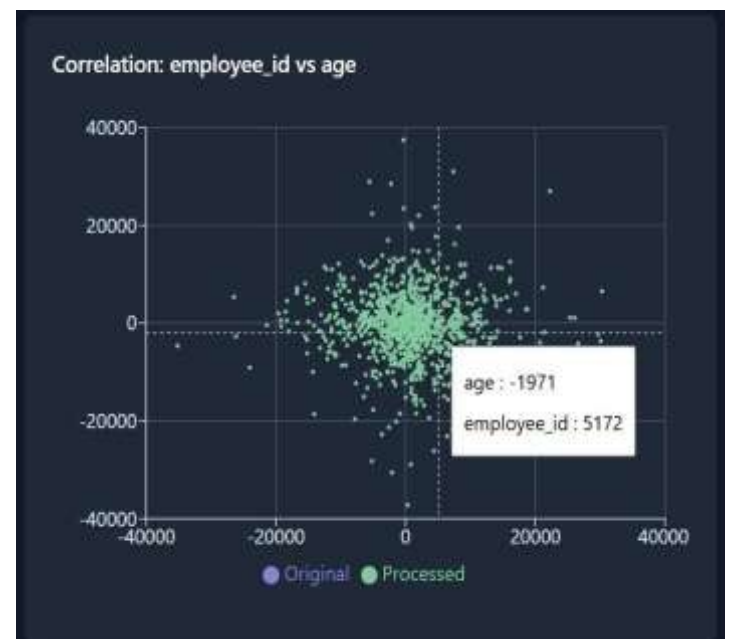
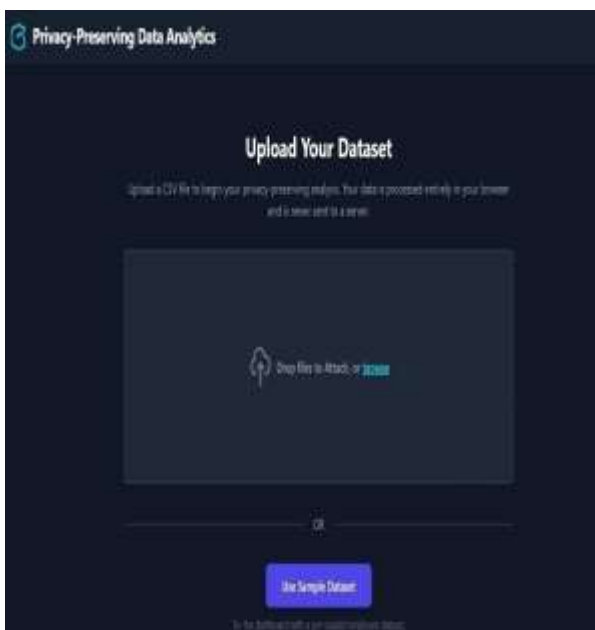
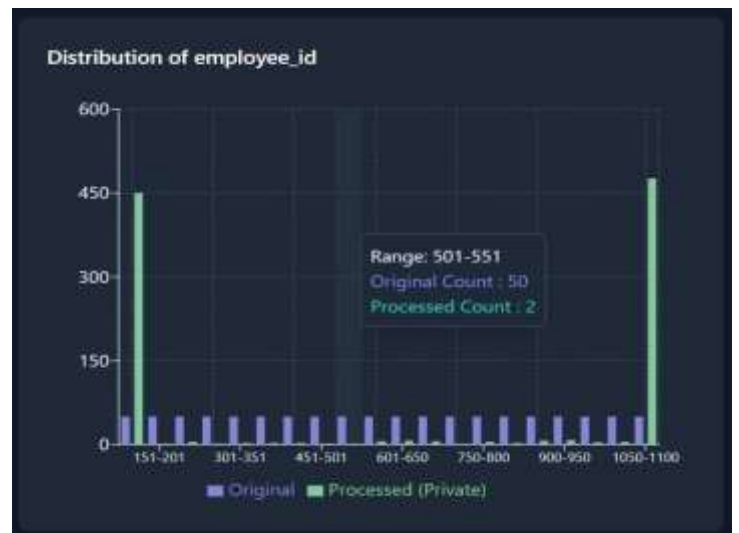
The present study reports that we have in fact determined the put forth model to be very effective as it relates to data utility and privacy.

Going forward from this report the introduced model does enable the feature of data sets upload via a user's interface which in turn will run only on that user's machine.

Once the info is uploaded to the platform users also have the option to set their privacy settings which include allocation of privacy budget and which also sees to the management of sensitive attributes. These settings in turn which play a role in the application of privacy protection methods like the use of noise and masking which in turn protect sensitive information.

Upon analysis I see that the means, minimums and maximums are similar for both data sets. The unprocessed and processed data sets have the same descriptive statistics. There are no differences, between them. The differences are very small and not significant.

Conclusion: Based on the analysis, there is no meaningful relationship between Employee ID and age. This holds true for both the original (raw) data and the filtered dataset, indicating that Employee ID does not influence or relate to an employee's age in any way.



VI CONCLUSION

Big in the field is a Privacy Preservation System which is a result of the integration of many data science, cryptography and privacy engineering methods which we have put together to form a strong base for doable data analytics. We see in practice that via Differential Privacy, Homomorphic Encryption, and Federated Learning we achieve privacy at every step of the analysis.

The system is put in place for privacy and security of data as well as compliance which includes global standards like GDPR and HIPAA. We have developed a web portal interface which allows our staff to easily analyze results without at the same time putting any personal data at risk.

Present study reports on the growth of privacy preserving data mining in today which we live in where we see large scale data creation in seconds. We put forth a model which is very much a success in it's identification and handling of sensitive information in health care, banking and academic fields.

In total this research did what it set out to do which was to present a model for security, intelligence and privacy. This research also sets a base for future studies in privacy preserving computations.

VII FUTURE ENHANCEMENT

1. Create adaptive privacy systems that would enable dynamic control of the privacy system depending on the data being used and the purpose of its use.
2. Use Federated Learning algorithms that make learning possible without transferring data from one party to another.
3. Use encryption techniques and computations for minimizing costs and providing faster processing.
4. Extend the proposed mechanism to conduct real-time privacy preserving data analytics.

5. Make models for machine learning that consider privacy without compromising accuracy.

VIII REFERENCES

1. Adam Smith, Kobbi Nissim, Frank McSherry, and Cynthia Dwork (2006). Noise to Sensitivity Calibration in Private Data Analysis (Differential Privacy). Journal of Cryptography Theory.
2. Shokri, R. and V. Shmatikov (2015). Deep Learning while protecting privacy. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS).
3. Craig Gentry (2009). Ideal lattices are used for fully homomorphic encryption. ACM Symposium on Computer Theory (STOC).
4. McMahan, B. et al. (2017). Learning Deep Networks with Efficient Communication from Decentralized Data. The 20th International Conference on Artificial Intelligence and Statistics (AISTATS).
5. Google AI Blog (2019). Differential Privacy: Protecting Data in the Machine Learning Era.
6. Microsoft Research (2020). Federated Learning and Privacy-Preserving AI: Techniques and Applications. Technical Report, Microsoft Research.
7. IBM Research (2021). Using Homomorphic Encryption to Provide Secure Data Analytics. IBM Technical Paper on Privacy Technologies.
8. Abadi, M. et al. (2016). Deep learning's differential privacy. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security.
9. Bonawitz, K. et al. (2019). Federated Learning: Obstacles, Approaches, and Prospects. Journal of Machine Learning Systems.
10. PySyft Documentation (2023). OpenMined Library for Privacy-Preserving Machine Learning.