Nanotechnology Perceptions
ISSN 1660-6795
www.nano-*ntp.com*

Privacy Preserving Analytics: Privacy-By-Design For Big Data Analytics

Bheemisetty Venkata Sriram Pavan Kumar

Assistant Professor, Department of Computer Science and Information Technology, KL University, Green Fields, Vaddeswaram, Guntur Dt. Andhra Pradesh, India.

Massive data are being generated exponentially every day. Therefore, analytics over data is inevitable nowadays to gain meaningful insights. Big Data Analytics (BDA) is powerful in critical applications while making effective decisions. Since the data to be processed are enormous in local systems, it is getting stored and processed in cloud platforms. Most of the clouds are public which are third-party resources. Security and privacy are the greatest concerns when it comes to the cloud. In the Big Data era, secure and private BDA has acquired the center of attention. This survey analyzes the various security and privacy solutions for BDA in the cloud environment from the combined perspectives of the three main subsystems: secure access control, secure data storage, and secure and private learning. Various techniques are studied and presented in each subsystem. The primary aim of this paper is to provide an overview of secure and private BDA in the cloud. Research challenges and future research directions in this area have also been presented.

Keywords: analyze, data, future, information, privacy, private, resources, security, system.

Introduction

With the popularity of multimedia applications and social networks, various multimedia data (i.e., texts, images, and videos) on the internet have shown exponential growth.. By regarding the storage cost and the computation efficiency, it is becoming more and more popular for data owners to employ cloud services. However, the data owners afraid of cloud services to reveal their private information, such as location and financial status. Moreover, the data analysis (such as feature extraction, retrieval, model construction, etc.) may easily leak important private information. For example, recent study in machine learning have demonstrated that sensitive data can be recovered from models. In this case, both cybersecurity and knowledge discovery are extremely important for analyzing big data. This special issue collects some recent studies on current machine learning techniques as well as privacy-preserving data analysis.

Wen et al. proposed a new clustering method considering both the local structure and the global structure for conducting nonlinear clustering. Specifically, the proposed method learns a robust spectral representation of the original data in the kernel space, and then introduces both the technique of feature selection and the method of adaptive graph learning into the proposed model. Furthermore, the proposed model utilizes low-rank constraint to make the adaptive graph to achieve the purpose of one-step clustering.

Advanced analytics**,** machine learning and other data science techniques are powerful tools for transformation. However, because "big data**"** entails large and complex data sets, the privacy risks associated with such endeavors are incredibly high.

Not only are organizations legally obligated to protect personal identifiable information (PII) from external threats, how companies use such data is also coming under increased scrutiny. As a result, preserving privacy of users has become a key requirement for many web-scale analytics and reporting applications.

Organizations looking to enable the sharing, processing or analysis of personal data without compromising privacy are increasingly adopting privacy preserving data analytics (PPDA) strategies**.** Rather than a specific tool or technology, PPDA represents a privacy-first approach to delivering data analytics.
Though PPDA first and foremost requires an effective, mathematically robust definition of privacy, it also relies on a combination of data protection systems and technologies - most of which result in data anonymization - to secure data. The following is an overview of some of those approaches.

K-anonymity, L diversity and T closeness
Often referred to as the power of 'hiding in the crowd,' the concept of k-anonymity revolves around the idea that by combining sets of data with similar attributes, it will inherently obscure identifying information about any one of the individuals contributing to that data. In other words, if an individuals' data is pooled in a larger group, any information in the group could correspond to any single member.

This process is known as 're-identification,' or the practice of tracing data's origins back to the individual it is connected to in the real world.

Though k-anonymization is suitable for data with low dimensionality, as it's difficult to group data with high dimensionality such as time series data, it may not work for every data science project, especially considering the cost of data minimization. In addition, it also falls short when it comes to protecting against homogeneous pattern attacks and background knowledge attacks.

To help overcome the limitations of k-anonymization, two extensions have been developed: l-diversity and t-closeness. L-diversity works by increasing the entropy and diversity in sensitive attributes, thereby further reducing the granularity of data. T closeness, on the other hand, builds on l-diversity by requiring that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table.

Randomization
Randomization is the process of adding "noise" to the data to hide the actual values of the individual records. Even though the data is masked, aggregate behavior of the data distribution can be reconstructed by subtracting out the noise from the data.

The benefits of randomization are that it can be applied during data collection and preprocessing and there is no anonymization overhead. One of the downsides of randomization is that it cannot be applied to large datasets due to time complexity and data utility.

Data distribution

Data distribution is a data protection technique whereby the data is distributed across many sites. In addition to making sensitive data more difficult to access, it also creates a backup and restoration system because if one component of the system is breached or goes down, the rest remain. However, data distribution can increase costs and complexity.

Distribution is typically accomplished in one or two ways:

- **Horizontal distribution of data** - data is distributed across many sites with the same attributes.
- **Vertical distribution of data** - data is distributed across different sites under custodian of different organizations.

Cryptographic techniques

Cryptographic techniques make information readable by the sender and receiver, but unintelligible to anyone else. For example, encryption, one of the most popular cryptographic techniques, obfuscates plaintext data by transforming into an unreadable, encoded format known as ciphertext. Only those with a digital key can access or read the encrypted information.

Multidimensional Sensitivity Based Anonymization

Multidimensional Sensitivity Based Anonymization (MDSBA) is an anonymization algorithm designed to be applied to large data sets with reduced loss of information and predefined quasi identifiers.

MDSBA builds on two preexisting anonymization algorithms, top–down specialization (TDS) and by bottom–up generalization (BUG)**,** both of which require continuous iterations with conditional statements, which can result in multiple times of heavy scan for the whole data records, data loss, scalability issues and high computation costs.

In a nutshell, MDSBA works by parallelizing data for big data frameworks and reducing the computation overhead of data iteration by providing pre-calculated k-anonymity parameters and pre-determined attributes for anonymization. In addition to protecting the privacy of data, MDSBA provides a fine-grained access control for multi-level of user's permissions**.**

Techniques and Challenges for Preserving Privacy in Big Data Analytics

Posted July 11, 2024, Last Revised July 30, 2025

In the era of big data, organizations are able to collect, store, and analyze vast amounts of information to gain valuable insights. However, this also raises significant privacy and security concerns, especially as data often contains sensitive information about individuals. Ensuring privacy in big data analytics is a complex task that requires advanced techniques and careful consideration of various challenges. This blog post discusses methods for preserving privacy in big data analytics, such as differential privacy, homomorphic encryption, and federated learning, and explores the associated challenges.

The Importance of Privacy in Big Data Analytics

With the increasing amount of data being generated and analyzed, protecting individuals' privacy has become paramount. Privacy-preserving techniques ensure that sensitive information remains confidential and that the insights derived from data analytics do not compromise personal privacy. Failure to address privacy concerns can lead to legal ramifications, loss of customer trust, and significant reputational damage.

Techniques for Preserving Privacy

1. **Differential Privacy**

Differential privacy is a mathematical framework that ensures the privacy of individual data points while allowing useful insights to be extracted from the data. It provides a quantifiable measure of privacy and guarantees that the inclusion or exclusion of a single data point does not significantly affect the overall analysis.

Working of Differential Privacy

1- Noise Addition: Random noise is added to the data or the results of queries, making it difficult to infer the presence or absence of any individual data point.
2- Privacy Budget: A privacy budget, denoted by ε (epsilon), controls the amount of noise added. A smaller ε provides stronger privacy but may reduce the accuracy of the analysis.

Applications

- Census Data: Differential privacy is used by statistical agencies, such as the US Census Bureau, to protect the privacy of respondents while publishing aggregate statistics.
- Machine Learning: Implementing differential privacy in machine learning models ensures that training data cannot be reverse-engineered from the model.

2. **Homomorphic Encryption**

Homomorphic encryption allows computations to be performed on encrypted data without decrypting it. This ensures that sensitive data remains secure even while being processed, providing strong privacy guarantees.

How Homomorphic Encryption Works-

- Encryption: Data is encrypted using a homomorphic encryption scheme before being sent to a third party for processing.
- **Computation on Encrypted Data:** The third party performs the required computations on the encrypted data.
- **Decryption:** The results of the computations, still encrypted, are sent back to the data owner, who decrypts them to obtain the final results.

Applications

- Secure Data Outsourcing: Organizations can outsource data processing to cloud providers without exposing sensitive information.
Privacy-preserving Data Analysis: Researchers can perform data analysis on encrypted datasets without accessing the raw data.

3. **Federated Learning**

Federated learning is a distributed machine learning approach that enables model training across multiple devices or servers holding local data samples, without exchanging the data itself. This technique enhances privacy by keeping raw data on local devices and only sharing model updates.

How Federated Learning Works

- Local Training: Each device trains a local model on its own data.
- Aggregation: The local models are sent to a central server, which aggregates them to create a global model.
Model Update: The global model is sent back to the devices, where it is refined with further local training.

Applications

- Mobile Devices: Federated learning is used in mobile applications, such as predictive text and personalized recommendations, without sending user data to centralized servers.
- Healthcare: Hospitals can collaborate on machine learning models for medical research without sharing patient data.

Challenges in Preserving Privacy

1. **Balancing Privacy and Utility**

Ensuring privacy often involves adding noise or limiting data sharing, which can reduce the accuracy and utility of the analysis. Finding the right balance between privacy and utility is a significant challenge.

2. **Scalability**

   Implementing privacy-preserving techniques at scale can be computationally expensive and complex, especially for large datasets and high-dimensional data.

3. **Compliance with Regulations**

Organizations must navigate various data privacy regulations, such as GDPR, HIPAA, and CCPA, which impose strict requirements on data handling and protection. Ensuring compliance while performing big data analytics adds an additional layer of complexity.

4. **Data Integrity and Quality**

Adding noise or encrypting data can impact data integrity and quality. Ensuring that privacy-preserving techniques do not significantly degrade the quality of data analysis is crucial.

5. **Technological Complexity**

Advanced privacy-preserving techniques like homomorphic encryption and federated learning require specialized knowledge and expertise, which may not be readily available in all organizations.

Best Practices for Privacy-preserving Big Data Analytics

1. **Adopt Privacy by Design**

Integrate privacy-preserving techniques into the design and development of data analytics systems from the outset, rather than as an afterthought.

2. **Use Privacy-enhancing Technologies**

Leverage advanced technologies like differential privacy, homomorphic encryption, and federated learning to protect sensitive data.

3. **Implement Strong Access Controls**

Ensure that access to sensitive data is restricted to authorized personnel only, and use robust authentication and authorization mechanisms.

4. **Regularly Audit and Monitor**

Conduct regular audits and monitoring of data processing activities to ensure compliance with privacy policies and regulations.

5. **Educate and Train Staff**

Provide training and education to staff on the importance of data privacy and the use of privacy-preserving techniques.

With the significant improvements in mobile digital devices and wireless networking technologies, we have witnessed the explosion of multimedia data. Because it is dynamic, vast in volume, and heterogeneous, this data not only evokes various novel data-driven services and applications, but also brings considerable security threats. In this article, the authors focus on privacy leakage issues in multimedia systems and study how to maximize the total privacy weights and upgrade the security level given predefined time and resource constraints. To this end, they propose a selective privacy-preserving method that adaptively allocates encryption resources according to the privacy weight and execution time of each data package. That is, it selects the encryption method with the appropriate complexity and security level for each multimedia data package. It first divides the data randomly into two parts, then performs XOR operations and generates cipher keys in different cloud storages to prevent users' original information from being attacked by untrusted cloud operators. Extensive simulation results have demonstrated the advantages and superiority of the proposed method over previous schemes. This article is part of a special issue on cybersecurity.

The availability of an increasing amount of user generated data is transformative to our society. We enjoy the benefits of analyzing big data for public interest, such as disease outbreak detection and traffic control, as well as for commercial interests, such as smart grid and product recommendation. However, the large collection of user generated data contains unique patterns and can be used to re-identify individuals, which has been exemplified by the AOL search log release incident. In this paper, we propose a practical framework for data analytics, while providing differential privacy guarantees to individual data contributors. Our framework generates differentially private aggregates which can be used to perform data mining and recommendation tasks. To alleviate the high perturbation errors introduced by the differential privacy mechanism, we present two methods with different sampling techniques to draw a subset of individual data for analysis. Empirical studies with real-world data sets show that our solutions enable accurate data analytics on a small fraction of the input data, reducing user privacy risk and data storage requirement without compromising the analysis results.

Conclusion

Preserving privacy in big data analytics is a complex but essential task. Techniques such as differential privacy, homomorphic encryption, and federated learning offer powerful tools for protecting sensitive information while enabling valuable data insights. However,

implementing these techniques requires careful consideration of challenges such as balancing privacy and utility, scalability, regulatory compliance, data integrity, and technological complexity. By adopting best practices and leveraging advanced privacy-preserving technologies, organizations can build robust data analytics systems that protect individual privacy and comply with regulatory requirements, while still harnessing the power of big data. As the field of data privacy continues to evolve, staying informed about the latest developments and methodologies will be crucial for maintaining trust and safeguarding sensitive information in the digital age.

References

[1] M. Barbaro and T. Zeller. A face is exposed for aol searcher no. 4417749. The New York Times, Aug. 2006. Google Scholar

[2] J. Bennett and S. Lanning. The netflix prize. In Proceedings of KDD cup and workshop, volume 2007, page 35, 2007.Google Scholar

[3]
B. Berjani and T. Strufe. A recommendation system for spots in location-based online social networks. In Proceedings of the 4th Workshop on Social Network Systems, SNS '11, pages 4:1--4:6, New York, NY, USA, 2011. ACM. Digital Library Google Scholar

[4]
A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In Proceedings of the 40th annual ACM symposium on Theory of computing, pages 609--618, New York, 2008. ACM. Digital Library Google Scholar

[5]
T.-H. H. Chan, E. Shi, and D. Song. Private and continual release of statistics. ACM Trans. Inf. Syst. Secur., 14(3):26:1--26:24, Nov. 2011. Digital Library Google Scholar

[6]
K. Chaudhuri and N. Mishra. When random sampling preserves privacy. In Proceedings of the 26th annual international conference on Advances in Cryptology, CRYPTO'06, pages 198--213, Berlin, Heidelberg, 2006. Springer-Verlag. Digital Library Google Scholar

[7]
R. Chen, G. Acs, and C. Castelluccia. Differentially private sequential data publication via variable-length n-grams. In Proceedings of the 2012 ACM conference on Computer and communications security, CCS '12, pages 638--649, 2012. Digital Library Google Scholar

[8]
G. Cormode, C. Procopiuc, D. Srivastava, and T. T. L. Tran. Differentially private summaries for sparse data. In Proceedings of the 15th International Conference on Database Theory, ICDT '12, pages 299--311, New York, NY, USA, 2012. ACM. Digital Library Google Scholar

[9]
Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports, Mar. Google Scholar

[10]
C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, Automata, Languages and Programming, volume 4052 of Lecture Notes in Computer Science, pages 1--12. Springer Berlin Heidelberg, 2006. Digital Library Google Scholar

[11]
C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: privacy via distributed noise generation. In Proceedings of the 24th annual international conference on The Theory and Applications of Cryptographic Techniques, EUROCRYPT'06, pages 486--503, Berlin, Heidelberg, 2006. Springer-Verlag. Digital Library Google Scholar

[12]
C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In In Proceedings of the 3rd Theory of Cryptography Conference, pages 265--284, Heidelberg, 2006. Springer-Verlag. Digital Library Google Scholar

[13]
L. Fan and L. Xiong. An adaptive approach to real-time aggregate monitoring with differential privacy. Knowledge and Data Engineering, IEEE Transactions on, 26(9):2094--2106, Sept 2014. Google Scholar

[14]
Y. Hong, J. Vaidya, H. Lu, and M. Wu. Differentially private search log sanitization with optimal output utility. In Proceedings of the 15th International Conference on Extending Database Technology, EDBT '12, pages 50--61, New York, NY, USA, 2012. ACM. Digital Library Google Scholar

[15]
G. Kellaris and S. Papadopoulos. Practical differential privacy via grouping and smoothing. In Proceedings of the 39th international conference on Very Large Data Bases, PVLDB'13, pages 301--312, 2013. Digital Library Google Scholar

[16]
A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In Proceedings of the 18th international conference on World wide web, WWW '09, pages 171--180, 2009. Digital Library Google Scholar

[17]
N. Li, W. Qardaji, and D. Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, ASIACCS '12, pages 32--33, 2012. Digital Library Google Scholar

[18]
X. Long, L. Jin, and J. Joshi. Towards understanding traveler behavior in location-based social networks. In Global Communications Conference (GLOBECOM), 2013 IEEE, 2013. Crossref Google Scholar

[19]

A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, pages 277--286, 2008. Digital Library Google Scholar

[20]
F. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. volume 53, pages 89--97, 2010. Digital Library Google Scholar

[21]
K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In Proceedings of the thirty-ninth annual ACM symposium on Theory of computing, STOC '07, pages 75--84, New York, NY, USA, 2007. ACM. Digital Library Google Scholar

[22]
D. Proserpio, S. Goldberg, and F. McSherry. Calibrating data to sensitivity in private data analysis: A platform for differentially-private analysis of weighted datasets. Proc. VLDB Endow., 7(8):637--648, Apr. 2014. Digital Library Google Scholar

[23]
V. Rastogi and S. Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pages 735--746, 2010. Digital Library Google Scholar

[24]
G. Yuan, Z. Zhang, M. Winslett, X. Xiao, Y. Yang, and Z. Hao. Low-rank mechanism: optimizing batch queries under differential privacy. Proc. VLDB Endow., 5(11):1352--1363, July 2012. Digital Library Google Scholar

[25]
C. Zeng, J. F. Naughton, and J.-Y. Cai. On differentially private frequent itemset mining. Proc. VLDB Endow., 6(1):25--36, Nov. 2012. Digital Library Google Scholar