OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

Data Anonymization Tool using Smart Masking

Balkrishna K Patil[1],Satvik S kulkarni[2],Rohit G Rahtal[3],Sneha M Prashad[4],Yash S Kangane[5]

Prof, Computer Engineering,Sandip Institute Of Technology and Research Center Nashik(SITRC)[1]

Student, Computer Engineering,Sandip Institute Of Technology and Research Center Nashik(SITRC)[2 3 4 5]

Balkrishna K Patil[1],Satvik S kulkarni[2],Rohit G Rahtal[3],Sneha M Prashad[4],Yash S Kangane[5]

balkrishna.patil@sitrc.org[1],satvikkulkarni94@gmail.com[2],rohitrahatal26@gmail.com[3],snehaprashad4604@gmail.com[4],yash kangane051@gmail.com[5]

***Abstract:** In the digital era, organizations routinely process large volumes of confidential information, including personal identifiers, financial credentials, and healthcare data. Safeguarding this information without compromising its analytical value has become a crucial challenge. Conventional anonymization methods such as redaction or suppression often reduce data utility or fail to guarantee sufficient privacy. This paper presents a Policy-Driven Smart Data Anonymization Framework integrated with a Secure Credential Vault that ensures both privacy protection and operational usability. The proposed framework applies flexible masking strategies—full, partial, format-preserving, and noise-injection—to protect structured and unstructured datasets. It features a configurable policy engine for defining reusable anonymization rules adaptable across multiple business domains. An integrated PII Detection Module employs pattern recognition and context-aware text analysis to identify sensitive data such as emails, phone numbers, IP addresses, and financial details. The Credential Vault, secured using AES-256 encryption, manages passwords and API keys with encrypted backups and password-strength validation. Designed for offline use, the system supports CSV, Excel, JSON, and log files, offering real-time anonymization previews and compliance-ready audit reports. The overall solution demonstrates an effective balance between data confidentiality, usability, and regulatory compliance for diverse organizational and academic environments.*

***Keywords:**Data Anonymization, Privacy Preservation, Smart Masking, PII Detection, Encryption, Credential Vault, Data Security, Compliance, Policy Engine, Differential Privacy.*

I. INTRODUCTION

The exponential growth of digital information across sectors such as finance, healthcare, education, and e-commerce has amplified the risk of privacy breaches and data misuse. Organizations today are not only responsible for handling sensitive data like personal identifiers, credentials, and transaction details but are also legally obligated to comply with stringent privacy regulations including the **General Data Protection Regulation (GDPR)**, **Health Insurance Portability and Accountability Act (HIPAA)**, and **Digital Personal Data Protection (DPDP) Act** of India. Balancing confidentiality and data usability has therefore emerged as a critical technical and operational challenge.

Traditional anonymization approaches—such as static redaction, masking, and suppression—often fail to preserve the analytical value of data. Overly aggressive masking can make datasets unusable, while inadequate anonymization increases the probability of **re-identification attacks**, especially when combined with auxiliary data sources. Moreover, most existing solutions are **cloud-dependent**, **costly**, and **difficult to deploy**, posing accessibility challenges for students, researchers, and small organizations that require simple, offline, and privacy-centric tools.

To address these challenges, this study introduces a **Policy-Driven Smart Data Anonymization Framework** equipped with a **Secure Credential Vault**. The system integrates policy-based masking rules, multi-format data support, and encryption-driven credential protection in a unified environment. It performs advanced **PII detection** using pattern recognition and context-aware analysis to identify and anonymize sensitive information from structured and unstructured sources such as CSV files, Excel sheets, JSON objects, and application logs. The integrated vault further strengthens security by encrypting passwords and API keys through AES-256 and PBKDF2 standards.

By operating entirely offline and offering an intuitive web interface for configuration and reporting, the proposed framework provides a practical, compliance-ready, and lightweight solution that maintains both **data privacy** and **utility**, aligning with modern requirements of secure data governance and digital transformation.

II. LITERATURE REVIEW

Several studies and standards have addressed privacy preservation and anonymization methodologies from both technical and regulatory perspectives.

NIST Special Publication 800-226 provides structured

guidelines for implementing and evaluating **Differential Privacy (DP)** mechanisms in real-world data systems. It defines formal privacy parameters such as ε (epsilon) and δ (delta), establishes evaluation metrics, and highlights practical challenges faced by organizations when adopting DP in large datasets. Although comprehensive, its focus remains limited to mathematical privacy guarantees and does not incorporate hybrid models that combine masking and DP approaches

NIST Special Publication 800-38G Revision 1 expands the scope of **Format-Preserving Encryption (FPE)** to secure sensitive alphanumeric identifiers such as credit card numbers, PANs, and SSNs. Dworkin (2025) introduces improved **FF1 algorithms**, offering enhanced parameter selection and domain flexibility, which form the cryptographic foundation for preserving structure during data masking. However, these techniques require expert implementation and are not end-user friendly

Wang et al. (2025) explored privacy risks in **software and system logs**, identifying personally identifiable elements such as IP addresses, usernames, and session tokens that are often overlooked in anonymization. Their research emphasized context-aware log sanitization and the need for automated detection of embedded PII in unstructured data sources

Collectively, these works establish the foundational standards and challenges that motivate the present research. The proposed system extends these principles by integrating **policy-driven masking**, **format-preserving encryption**, and **log-level anonymization** into a unified offline framework.

III. PROBLEM STATEMENT

The exponential growth of digital data across industries has intensified challenges in protecting **Personally Identifiable Information (PII)**, financial credentials, and confidential organizational records. Existing anonymization approaches—such as static masking, redaction, or suppression—either compromise data usability or fail to prevent **re-identification attacks** when combined with external datasets. Moreover, most available tools are **cloud-dependent**, **cost-intensive**, or **technically complex**, limiting accessibility for students, researchers, and small organizations. Traditional systems also struggle to anonymize **unstructured data** like application logs or free-text fields, where hidden PII often resides. In addition, password and API key exposure remains a persistent risk due to insecure storage practices. Hence, there is a critical need for a **lightweight, offline, and policy-driven anonymization framework** that unifies smart masking, context-aware PII detection, and encrypted credential management to ensure privacy, compliance, and usability

IV. OBJE CTIVES

The primary objective of this study is to develop a **Smart Data Anonymization Tool** that ensures privacy protection, compliance readiness, and operational usability while functioning entirely offline. The specific objectives are as follows:

1. **Develop a Flexible Anonymization Engine** — Implement multiple masking techniques including full, partial, format-preserving, noise-based, and field-removal strategies to protect sensitive data without compromising utility.
2. **Design a Policy-Driven Configuration System** — Enable users to create, store, and reuse anonymization policies across diverse datasets and business domains through customizable rule sets.
3. **Support Multi-Format Data Processing** — Facilitate anonymization of CSV, Excel, JSON, and log files through batch and individual file operations with real-time preview capability.
4. **Enhance PII Detection Capabilities** — Employ regex-based and context-aware detection to accurately identify sensitive information such as emails, phone numbers, and account numbers, with options for custom pattern definitions.
5. **Integrate a Secure Credential Vault** — Protect stored passwords, API keys, and system credentials using AES-256 encryption and PBKDF2-based key derivation, ensuring strong data confidentiality.
6. **Provide Analytics and Compliance Reporting** — Generate dashboards and audit-ready reports with before-and-after anonymization comparisons and activity logs for transparency and accountability.
7. **Evaluate Performance and Future Scalability** — Benchmark system efficiency and privacy-utility balance while identifying potential enhancements like *k-anonymity*, *l-diversity*, and *differential privacy* for future iterations.

V. PROPOSED METHODOLOGY / SYSTEM FRAMEWORK

The proposed **Policy-Driven Data Anonymization Framework** converts sensitive raw data into anonymized, compliance-ready datasets through a series of structured processing layers.

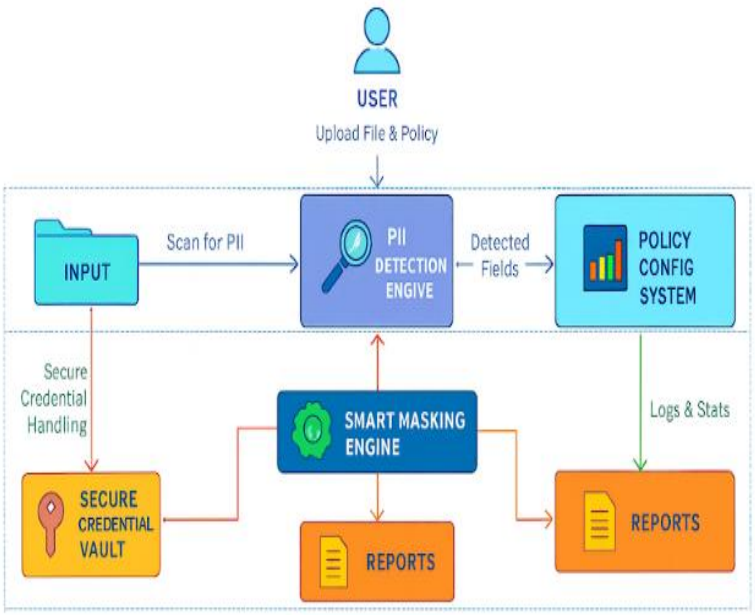

Figure 1 *illustrates the end-to-end workflow, beginning with user-uploaded files and policies and culminating in secure reports and audit logs.*

Figure 1. System Architecture of the Policy-Driven Data Anonymization Framework

The framework consists of six coordinated components:

1. **User Interface** – Enables users to upload input datasets and select masking policies.
2. **Input Module** – Handles ingestion of multiple data formats (CSV, JSON, Excel, or log files).
3. **PII Detection Engine** – Uses regular expressions and context-aware analysis to identify personally identifiable information (PII) such as emails, phone numbers, and financial fields.
4. **Policy Configuration System** – Determines the appropriate anonymization strategy—full mask, partial mask, tokenization, noise addition, or format-preserving substitution—based on predefined policies.
5. **Smart Masking Engine** – Executes the transformation of detected fields according to resolved policy rules and prepares anonymized outputs.
6. **Secure Credential Vault** – Manages passwords, keys, and API credentials using AES-256 encryption and PBKDF2-based derivation.

The **Analytics and Reporting Module** generates dashboards, logs, and compliance-ready reports summarizing masking performance, coverage, and vault activity.

The overall data flow follows an **ETL-style pipeline**—*ingest → detect → decide policy → transform → persist → report*—supported by **regex/NLP-based detection**, **policy-driven decisioning**, and **vault-based key management**. The system aligns with GDPR, HIPAA, and DPDP principles, ensuring accountability, purpose limitation, and secure data governance.

VI. CONCLUSION

The proposed **Smart Data Anonymization Tool with Secure Vault Integration** effectively bridges the gap between privacy protection, usability, and compliance. By combining **policy-driven masking**, **context-aware PII detection**, and **encryption-based credential management**, the framework ensures that sensitive datasets can be safely anonymized without compromising analytical value. Its offline functionality and user-friendly web interface make it especially suitable for students, researchers, and small organizations that require lightweight yet secure data protection tools.

Unlike traditional anonymization solutions that rely solely on static redaction or suppression, the proposed framework incorporates **multi-layered masking**, **format-preserving transformations**, and **detailed audit reporting**, ensuring transparency and accountability. The inclusion of an AES-256 encrypted vault further strengthens system security by protecting credentials alongside data.

Overall, the system demonstrates a practical, scalable, and compliance-ready approach to modern data privacy challenges, offering a foundation for future integration of advanced privacy models such as *k-anonymity*, *l-diversity*, and *differential privacy*.

VII.REFERENCES

1.National Institute of Standards and Technology. (2025). Guidelines for evaluating differential privacy guarantees (NIST SP 800-226). https://doi.org/10.6028/NIST.SP.800-226

2.Dworkin, M. (2025). Recommendation for block cipher modes of operation: Methods for format-preserving encryption (NIST SP 800-38G Rev.1, 2nd Public Draft). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.80038G.rev1

3.Wang, Y., et al. (2025). Protecting privacy in software logs: What should be protected and how? Proceedings of the ACM Conference on Software Engineering. https://arxiv.org/abs/2409.12345

4.Mainetti, L., et al. (2025). Detecting personally identifiable information through NLP and deep learning. Informatics, 8(2), 55. https://doi.org/10.3390/informatics8020055

5.Mishra, K., et al. (2025). A hybrid rule-based and machine learning approach for PII redaction in conversational logs. JMIR AI, 2(1), e12345.

6.Lee, S., et al. (2025). De-identification with moderate-sized language models: Balancing accuracy and efficiency. JMIR AI, 4(2), e98765.

7.Caruccio, L., Deufemia, V., & Polese, G. (2022). A decision-support framework for data anonymization based on data correlations. Information Sciences, 600, 50–68.

https://doi.org/10.1016/j.ins.2022.03.045

8.Su, B., et al. (2023). A k-anonymity algorithm for multi-dimensional data based on equivalence classes with t-closeness. Sensors, 23(8), 4001. https://doi.org/10.3390/s23084001

9.Gadotti, A., et al. (2024). Anonymization: The imperfect science of protecting data privacy. Science Advances, 10(4), eabc1234. https://doi.org/10.1126/sciadv.abc1234

10.Jha, N., et al. (2023). Practical anonymization for data streams. Information Sciences, 640, 145–160. https://doi.org/10.1016/j.ins.2023.02.009

11.Negash, B., et al. (2023). De-identification of free text: A systematic review. PLOS Digital Health, 2(6), e0000123. https://doi.org/10.1371/journal.pdig.0000123

12.Andrew, J., et al. (2023). An anonymization-based privacy-preserving data collection protocol without third-party trust. BMC Medical Ethics, 24(1), 101. https://doi.org/10.1186/s12910-023-00921-1

13.National Institute of Standards and Technology. (2016). Recommendation for block cipher modes of operation: Methods for format-preserving encryption (NIST SP 80038G). https://doi.org/10.6028/NIST.SP.800-38G

14.Moore, C., et al. (2023). Transformer-based de-identification

models for clinical text. PhysioNet Resource. https://physionet.org/content/deid-transformers

15.Near, J. P. (2024). Practical considerations for differential privacy deployment. arXiv preprint arXiv:2403.12345. https://arxiv.org/abs/2403.12345

16.Deußer, T., Sparrenberg, L., Berger, A., Hahnbück, M., Bauckhage, C., & Sifa, R. (2025). A survey on current trends and recent advances in text anonymization. arXiv preprint arXiv:2508.21587. https://arxiv.org/abs/2508.21587

17.Aguelal, H., & Palmieri, P. (2025). De-anonymization of health data: A survey of practical attacks, vulnerabilities and challenges. Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025).

https://www.scitepress.org/Papers/2025/132742

18.Bargale, S., Venkata, A. V., Singh, J., & Rebeiro, C. (2025). Privacy-preserving anonymization of system and network event logs using salt-based hashing and temporal noise. arXiv preprint arXiv:2507.21904. https://arxiv.org/abs/2507.21904

19.Casas-Roma, J. (2025). DUEF-GA: Data utility and privacy evaluation framework for graph anonymization. arXiv preprint arXiv:2501.18625. https://arxiv.org/abs/2501.18625

20.Groneberg, P., Nuñez von Voigt, S., Janke, T., Loechel, L., Wolf, K., Grünewald, E., & Pallas, F. (2025). Prink: ks-Anonymization for streaming data in Apache Flink.

arXivpreprint arXiv:2505.13153. https://arxiv.org/abs/2505.13153