SafeData: A Privacy–Utility Preserving Data Publishing Publishing Framework

Adhithya S – 22IT005 B.Tech IT – Chennai Institute of Technology

The Privacy Paradox: Why Anonymization Fails

Microdata is indispensable for cutting-edge research and informed policymaking. However, naive anonymization, merely removing names or IDs, leaves critical vulnerabilities.

The Re-identification Threat

Attackers exploit **quasi-identifiers** (e.g., Age, Gender, ZIP code) to re-identify individuals, compromising privacy even in seemingly anonymized datasets.

Regulatory Mandates

The Digital Personal Data Protection (DPDP) Act, 2023, along with GDPR, mandates robust safeguards, pushing for advanced anonymization beyond simple redaction.

The Anonymization Challenge: Fragmentation & Trade-offs

Existing anonymization techniques are often fragmented and insufficient, failing to offer a comprehensive solution for balancing privacy and data utility.

K-Anonymity

Vulnerable to attribute disclosure attacks.

L-Diversity

Weak performance with skewed data distributions.

Differential Privacy

Can drastically reduce accuracy if improperly tuned.

Federated Analysis

Secure but often limited by query types and complexity.

> The core challenge remains: how to effectively balance rigorous privacy protection with the indispensable utility of data for analysis.

SafeData Objectives: A Holistic Approach

Our SafeData framework aims to overcome current limitations by integrating multiple privacy-enhancing technologies and providing actionable insights.

1 Layered Privacy Framework

Combining K-Anonymity, L-Diversity, T-Closeness, Differential Privacy, and Federated Analysis for robust protection.

2 Quantifiable Trade-offs

Establishing clear metrics to quantify the privacy–utility trade-off, enabling informed decision-making.

3 Real-time Parameter Tuning

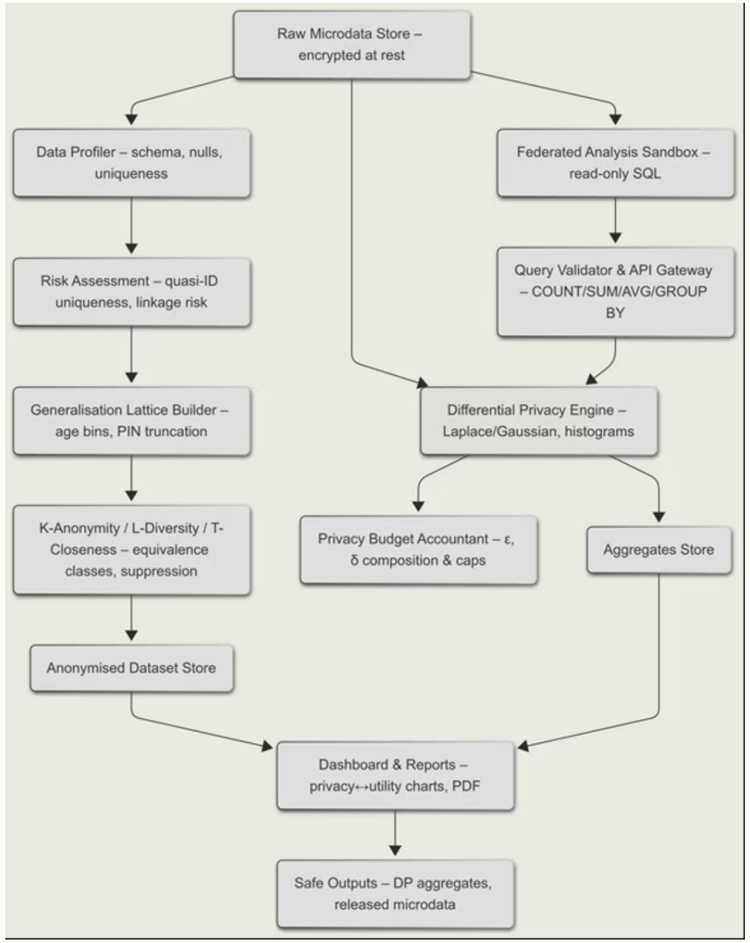
Developing an intuitive dashboard for dynamic adjustment of anonymization parameters (k, L, T, ε).

4 Regulatory Compliance

Ensuring full adherence to the DPDP Act and GDPR guidelines.

System Architecture

The SafeData framework is built on a modular architecture designed for end-to-end data anonymization and secure publication.

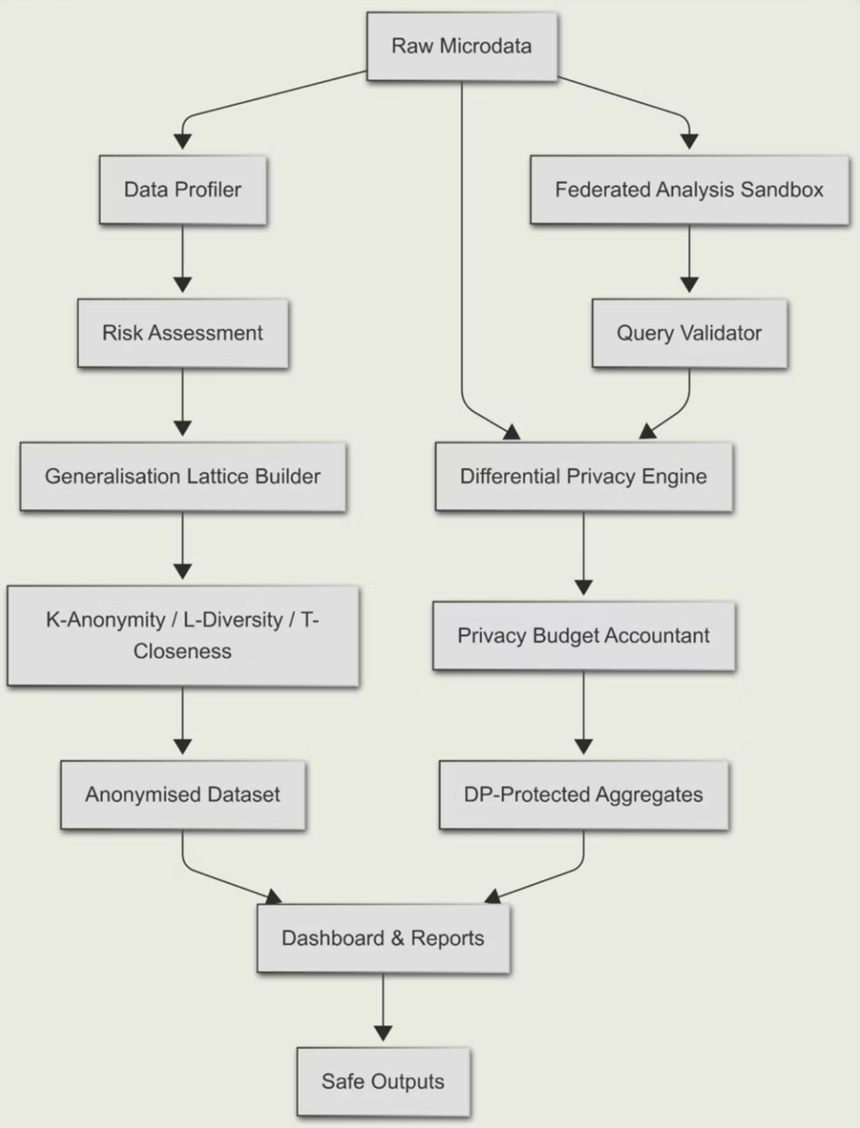


- **Raw Data Store**: Encrypted storage for sensitive microdata.
- **Data Profiler & Risk Assessment**: Identifies quasi-identifiers and estimates re-identification risk.
- **Generalisation Lattice Builder**: Applies K-Anonymity, L-Diversity, and T-Closeness for structured anonymisation.
- **Differential Privacy Engine**: Adds calibrated noise (Laplace/Gaussian) to query results.
- **Privacy Budget Accountant**: Tracks ε values to prevent overuse of queries.
- **Federated Analysis Sandbox**: Enables secure, controlled query execution without releasing raw data.
- **Dashboard & Reports**: Visualises privacy–utility trade-offs, provides safe outputs for stakeholders.

Data Flow Diagram

SafeData ensures privacy and utility through a streamlined data pipeline, from raw input to secure publication and query results.

- **Raw Microdata Entry:** Data undergoes initial profiling and risk assessment.
- **Anonymization Branch:** Data is processed via K/L/T anonymization techniques to create a safe, published dataset.
- **Query Branch:** Queries are securely executed in the Federated Sandbox, with differential privacy applied to aggregates.
- **Converged Insights:** Both branches feed into the dashboard for real-time privacy–utility visualization and parameter tuning.



Implementation Details

SafeData leverages a robust stack of open-source technologies and libraries, ensuring flexibility, scalability, and adherence to best practices in data science and privacy engineering.

Language	Python
Libraries	`pandas`, `NumPy`, `scikit-learn`, `PyDP/diffprivlib`
Backend	FastAPI (for secure query validation and API gateway services)
Database	PostgreSQL
Dashboard	Streamlit (for intuitive parameter tuning and interactive reports)
Dataset	UCI Adult Census Dataset (for comprehensive testing and validation)

Example Data: Unsecured Vs Secured

Name	Age	Gender	ZIP Code	Disease
Ramesh K	32	M	600045	Cancer
Priya S	29	F	600046	Diabetes
Arjun M	41	M	600048	Flu
Kavya L	36	F	600049	Cancer
Suresh P	28	M	600047	Flu
Divya R	31	F	600045	Diabetes
Manoj T	34	M	600046	Flu
Anitha V	27	F	600048	Cancer

Age	Gender	ZIP Code	Disease
30-35	*	60004*	C/F/D
20-30	*	60004*	D/F/C
40-45	*	60004*	F/D/C
35-40	*	60004*	C/D/F

Navigating the Privacy–Utility Trade-off

The inherent challenge in data anonymization lies in the inverse relationship between privacy strength and data utility. SafeData provides a crucial tool to navigate this complex balance.

Stronger Privacy: Implies more generalization or added noise, leading to a corresponding reduction in data utility. This is critical for sensitive datasets but can limit analytical depth.

Relaxed Privacy: While yielding higher data utility and rich insights, it inevitably results in weaker privacy protection, increasing re-identification risks.

The **SafeData Dashboard** is designed to empower users to dynamically select the select the optimal trade-off for their specific use-case and risk appetite, ensuring both ensuring both compliance and analytical value.

Future Scope: Extending SafeData's Reach

SafeData is a foundational framework, poised for significant expansion to address emerging challenges in privacy-preserving data publishing.

Scalability & Streaming

Extend capabilities to large-scale data and real-time streaming platforms (e.g., Spark, Kafka, Flink).

Advanced Cryptography

Integrate Homomorphic Encryption (HE) and Multi-Party Computation (MPC) for enhanced secure computation.

Unstructured Data Support

Develop methods for anonymizing and publishing unstructured data (text, images, IoT sensor data).

Adaptive Privacy Budgets

Implement dynamic privacy budget allocation based on query patterns and data sensitivity.

Privacy-Preserving ML

Seamless integration with privacy-preserving machine learning models for secure analytics and model training.