

Comparative Evaluation of Privacy-Preserving Techniques for Real-Time Big Data Analytics: An Empirical Study and Hybrid Differential Privacy–Synthetic Data Framework

Rajesh Bishnoi¹, Ashish Sharma²

^{1,2}Department of Computer Science and Engineering, Maulana Azad University, Jodhpur

e-mail: rb0107@gmail.com

Abstract

Big Data analytics creates substantial value but also increases the possibility of identity disclosure, linkage attacks, and misuse of sensitive records. This study analytically compares five privacy-preserving techniques—Random Noise Addition (RNA), Differential Privacy (DP), Data Swapping (DS), K-Anonymity (KA), and Synthetic Data Generation (SDG)—under a common experimental framework designed for real-time Big Data analytics. The experiments use a standardized synthetic dataset containing 500 individual profiles and 15 demographic, financial, and behavioural attributes. Six numerical variables are normalized, categorical variables are encoded, and privacy resistance is tested with a Re-identification Neural Network (RNNR). Privacy, utility, latency, throughput, scalability, distributed-platform execution time, robustness, and regulatory suitability are jointly evaluated. The results show that DP with the Laplace mechanism at $\epsilon=1.0$ preserves numerical means with 99.97–100.00% accuracy and produces 52% RNNR re-identification accuracy, close to the 50% chance baseline. Spark execution is 0.6 s, compared with 3.4 s on Hadoop. SDG using a Gaussian Copula provides the strongest structural privacy and high statistical fidelity but requires 8.4 s on Spark. KA performs poorly for streaming use: $k=5$ requires 48.6 ms per record and remains vulnerable to background-knowledge attacks. A multi-criteria evaluation ranks DP highest among individual techniques (62/70), followed by SDG (57/70), while the proposed Hybrid DP+SDG architecture reaches 67/70. The findings indicate that privacy-preserving analytics should be selected by deployment context rather than by a single universal criterion.

Keywords: Big Data privacy; differential privacy; k-anonymity; synthetic data; privacy–utility trade-off; re-identification; Apache Spark.

INTRODUCTION

Big Data systems combine large volumes of heterogeneous information at high speed. Their analytical value depends on the ability to link records, detect patterns, and infer relationships that may not be visible in isolated datasets. The same characteristics, however, create privacy risks because combinations of apparently harmless attributes can make individuals distinguishable. Re-identification research has shown that de-identified data may remain vulnerable when attackers have access to auxiliary information, especially when the released dataset contains detailed demographic or behavioural variables (Rocher et al., 2019). This problem has become more important as organizations increasingly use cloud platforms, streaming infrastructures, machine learning, and cross-domain data integration.

Traditional anonymization mechanisms remain useful, but the privacy literature increasingly distinguishes between syntactic anonymity and mathematically defined privacy guarantees. K-anonymity protects each released record by ensuring that it shares its quasi-identifier pattern with at least $k-1$ other records, whereas differential privacy limits the influence that any single record can have on released analytical results. Recent reviews show that differential privacy has become one of the most widely studied privacy-enhancing technologies because it supports quantifiable privacy guarantees and can be adapted to data publishing, learning, and distributed settings (Huang et al., 2023; Pan et al., 2024; Yang et al., 2024). At the same time, synthetic data has gained attention as a practical way to release statistically useful information without directly releasing original records (Fonseca & Bacao, 2023; Hu & Bowen, 2024).

A major weakness in many comparative studies is that privacy techniques are evaluated on different datasets, with different threat models, metrics, or hardware environments. This makes direct conclusions difficult. Recent synthetic-data reviews have also noted that privacy is often assumed rather than empirically tested, while utility receives much more systematic evaluation (Kaabachi et al., 2025; Osorio Marulanda et al., 2024). Similarly, industrial experience with differential privacy shows that formal guarantees alone do not ensure deployability; implementation complexity, computational overhead, privacy-budget selection, governance, and analyst usability remain important (Munilla Garrido et al., 2023).

The present study addresses this problem by evaluating five techniques under one analytical framework: Random Noise Addition (RNA), Differential Privacy (DP), Data Swapping (DS), K-Anonymity (KA), and Synthetic Data Generation (SDG). The analysis is grounded in the experimental work of the underlying doctoral thesis and extends it into a compact research-paper format. The study combines privacy–utility analysis, adversarial re-identification testing, real-time latency and throughput measurements, Spark-versus-Hadoop execution benchmarking, scalability evaluation, and multi-criteria deployment scoring. It also evaluates a Hybrid DP+SDG architecture intended to separate real-time internal analytics from high-assurance external publication.

The paper makes three contributions. First, it provides a direct empirical comparison of five privacy-preserving techniques under a common dataset and evaluation protocol. Second, it treats re-identification resistance and computational feasibility as equally important deployment criteria rather than evaluating privacy in isolation. Third, it proposes and evaluates a two-layer DP+SDG model that combines a fast formal-privacy layer for internal analytics with a synthetic publication layer for external data sharing. This design is especially relevant because recent evidence shows that synthetic data can retain useful statistical and machine-learning properties, but its privacy should still be tested rather than assumed (Pereira et al., 2024; Kiran et al., 2025; Liu et al., 2025).

RELATED LITERATURE AND RESEARCH GAP

Privacy-preserving data analytics has evolved from identity masking toward risk-aware and mathematically formal methods. K-anonymity, l-diversity, and t-closeness remain foundational approaches for microdata release because they control distinguishability and sensitive-value disclosure within equivalence classes (Sweeney, 2002; Machanavajjhala et al., 2007; Li et al., 2007). Their principal limitation is that privacy protection depends on the structure of the released dataset and on assumptions about an adversary’s background knowledge. Modern linkage attacks weaken these assumptions, particularly in high-dimensional data.

Differential privacy was developed to provide a stronger guarantee that is independent of an attacker’s external knowledge. A randomized mechanism is differentially private when the probability distribution of its output changes only within a controlled bound when one individual is added to or removed from the dataset. The practical issue is not whether DP offers a formal guarantee, but how to select privacy parameters while retaining adequate utility. Recent studies continue to identify ϵ selection as a central deployment problem and show that utility can vary substantially across tasks and privacy budgets (Li et al., 2023; Subramanian, 2023). Reviews of DP in deep learning and local settings further emphasize sensitivity calibration, composition, mechanism choice, and attack models as key design decisions (Pan et al., 2024; Yang et al., 2024).

Synthetic data provides a different form of protection. Instead of releasing transformed copies of original records, a generative model learns statistical structure and produces new records. Reviews published in 2023–2025 show rapid growth in tabular, healthcare, and general-purpose synthetic-data research, but they also stress that privacy, fidelity, and downstream utility must be evaluated separately (Fonseca & Bacao, 2023; Hu & Bowen, 2024; van Breugel et al., 2024; Kaabachi et al., 2025). Osorio Marulanda et al. (2024) found wide variation in privacy metrics across synthetic-data studies, while Pereira et al. (2024) demonstrated that some differentially private tabular synthesizers can support machine-learning utility close to real-data baselines.

The literature therefore reveals a clear research gap: few studies compare classical perturbation, formal differential privacy, syntactic anonymization, and synthetic generation under one threat model while also measuring real-time computing performance. The present study fills that gap by using the same 500-profile dataset, preprocessing pipeline, adversarial evaluator, and platform benchmarks for all five techniques.

Table 1. Recent research context for the present comparative study

Study	Focus	Relevance
Rafiq et al. (2022)	Big Data privacy review	Shows continued use of DP, k-anonymity, l-diversity and t-closeness; highlights method-specific limitations.
Munilla Garrido et al. (2023)	Industrial DP adoption	Finds a gap between theoretical DP maturity and practical enterprise adoption.
Fonseca & Bacao	Synthetic tabular data	Reviews generation methods and the disclosure-risk/utility trade-off.

(2023)		
Pan et al. (2024)	DP and deep learning	Systematizes privacy attacks and DP-based defences.
Hu & Bowen (2024)	Synthetic microdata	Emphasizes generation, evaluation, practical use and disclosure risk.
Osorio Marulanda et al. (2024)	Synthetic-data privacy metrics	Shows wide diversity in privacy evaluation and strong use of DP.
Pereira et al. (2024)	DP synthetic tabular data	Finds that selected synthesizers can retain downstream utility and fairness.
Kaabachi et al. (2025)	Synthetic-data privacy/utility metrics	Shows privacy evaluation is less consistent than utility evaluation.

MATERIALS AND METHODS

Dataset and Pre-processing

The experiment used a standardized synthetic dataset containing 500 individual profiles and 15 attributes. The attributes span demographic, financial, health-related, and behavioural domains. The use of synthetic source data avoided direct processing of real personally identifiable information while preserving the heterogeneity needed to test privacy mechanisms. The dataset contains six numerical and nine categorical variables. ID is treated as a direct identifier, while Age, Income, Loan Amount, Credit Score, and Investment Portfolio Value carry high sensitivity because of their strong discriminating or financial characteristics.

Table 2. Experimental dataset attributes

Attribute	Type	Sensitivity	Description
ID	Categorical	Critical	Unique profile identifier
Age	Numerical	High	Age in years
Gender	Categorical	Low	Gender
Income	Numerical	High	Annual income
Education Level	Categorical	Medium	Highest education level
Employment Status	Categorical	Medium	Current employment status
Marital Status	Categorical	Medium	Marital status
Number of Children	Numerical	Low	Dependent children
Loan Amount	Numerical	High	Outstanding loan
Credit Score	Numerical	High	Credit score
Residence Type	Categorical	Medium	Residential ownership
Vehicle Ownership	Categorical	Low	Vehicle ownership
Investment Portfolio Value	Numerical	High	Portfolio value
Health Insurance	Categorical	Medium	Insurance coverage
Travel Frequency	Categorical	Low	Travel behaviour

Pre-processing consisted of IQR-based outlier inspection, label encoding for categorical variables, and Min–Max normalization of all six numerical variables to [0,1]. Normalization was essential for differential privacy because the raw numerical ranges differed greatly. After normalization, the global sensitivity of a mean query for each numerical attribute was $\Delta f = (x_{\max} - x_{\min})/n = 1/500 = 0.002$. This produced a common and interpretable basis for Laplace-noise calibration.

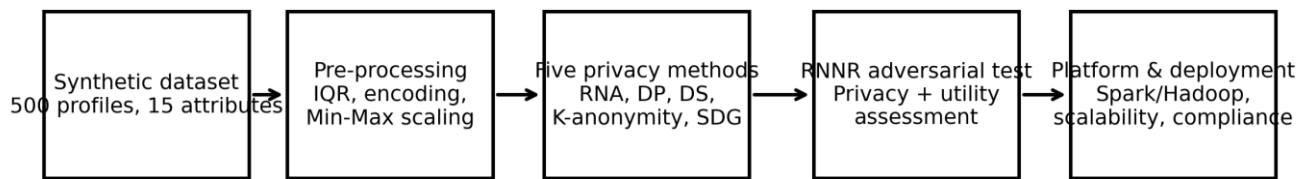


Figure 1. End-to-end experimental workflow used in the comparative analysis

Privacy-Preserving Techniques

RNA added independent Gaussian noise to each normalized numerical variable. The noise standard deviation σ was varied from 0.01 to 0.50. DP used the Laplace mechanism with noise scale $b = \Delta f/\epsilon$ and evaluated $\epsilon \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$. Data Swapping permuted sensitive categorical associations, with practical swap rates concentrated around 70–80%. K-Anonymity used an AnonyOPT-k procedure that generalized or grouped quasi-identifiers for k values of 2, 5, 10 and 20. SDG fitted a Gaussian Copula to the preprocessed data and generated 500 new records while retaining marginal distributions and inter-attribute dependence.

Adversarial and Utility Evaluation

A Re-identification Neural Network (RNNR) was used as the common adversarial evaluator. It contains two ReLU hidden layers with 64 and 32 units and one sigmoid output unit, totaling 3,073 trainable parameters. The model was trained for 50 epochs with batch size 32 and an 80:20 train–validation split. Re-identification accuracy was interpreted relative to a 50% binary chance baseline: values near 50% indicate that the adversary is unable to reliably distinguish target records, whereas values moving toward 100% indicate weaker protection.

Utility was assessed using mean preservation, mean squared error (MSE), distributional similarity and task-oriented scores. Execution time, latency and throughput were also recorded because a privacy mechanism that is secure but too slow for the intended pipeline is not operationally useful. The distributed-computing comparison used simulated Hadoop and Spark execution characteristics within the experimental environment. All experiments used a fixed random seed of 42 for reproducibility.

Table 3. Evaluation dimensions and interpretation

Dimension	Metric	Interpretation
Privacy resistance	RNNR re-identification accuracy; attack success	Lower adversarial success is better; ~50% RNNR accuracy indicates chance-level discrimination.
Utility	MSE, mean preservation, distribution similarity	Higher statistical preservation and lower error are better.
Real-time performance	Latency (ms/record), throughput (records/s)	Low latency and high throughput are required for streams.
Platform efficiency	Hadoop and Spark execution time	Lower execution time indicates better deployment efficiency.
Scalability	500, 5,000 and 50,000 records	Tests whether computational cost remains manageable as n grows.
Robustness	Resistance to background knowledge and inference	Higher robustness reduces dependence on attacker assumptions.
Regulatory suitability	Formal guarantee and release	Formal DP and well-governed synthetic release are

	model	favoured for sensitive use.
--	-------	-----------------------------

RESULTS & DISCUSSION

Random Noise Addition

RNA was computationally efficient, but its privacy behaviour was not monotonic. At $\sigma=0.01$ the average MSE was only 0.00009 and utility was 0.999, but the privacy score was 0.51. Raising σ to 0.10 increased MSE to 0.00866 while utility remained 0.913 and RNNR accuracy was 53%. At $\sigma=0.50$ utility collapsed to zero under the study's normalized scoring rule, yet privacy did not improve consistently. This result is important because it shows that simply adding more random noise does not guarantee stronger privacy against a learning-based adversary.

Table 4. RNA sensitivity analysis

σ	Avg. MSE	Privacy score	Utility score	RNNR re-id
0.01	0.00009	0.51	0.999	49%
0.05	0.00228	0.40	0.977	60%
0.10	0.00866	0.47	0.913	53%
0.20	0.03196	0.50	0.680	50%
0.50	0.11552	0.43	0.000	57%

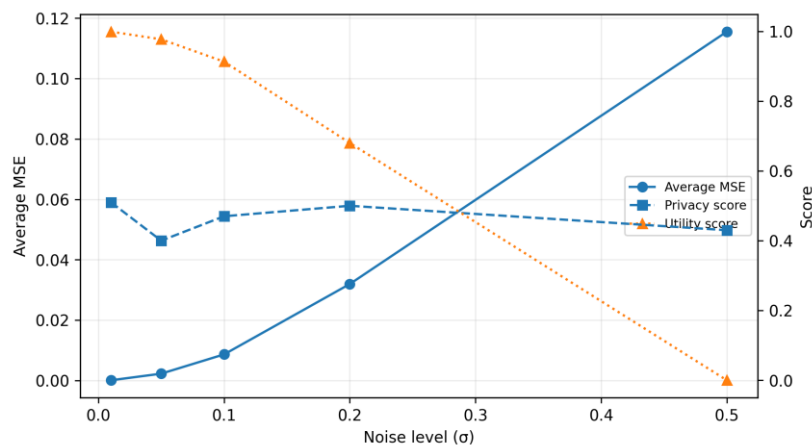


Figure 2. Effect of RNA noise level on error, privacy and utility

Differential Privacy

DP produced the best overall balance among the individual methods. At $\epsilon=1.0$, all six normalized numerical means were preserved with 99.97–100.00% accuracy. Credit Score reached 100.00% mean preservation, while Investment Portfolio Value showed the largest but still very small deviation at 99.97%. The epsilon sensitivity analysis also showed that utility remained effectively unchanged across the tested range, with a utility score of 0.992 at $\epsilon=0.1$ and 1.000 at $\epsilon \geq 0.5$. The RNNR result at $\epsilon=1.0$ was 52%, closest to the 50% chance baseline among the high-utility DP settings.

Table 5. Differential Privacy sensitivity and mean-preservation results

ϵ	Utility	Avg. MSE	Privacy score	RNNR re-id
0.1	0.992	0.00075	0.44	56%
0.5	1.000	0.00003	0.41	59%
1.0	1.000	0.00001	0.48	52%

2.0	1.000	≈ 0.00000	0.47	53%
5.0	1.000	≈ 0.00000	0.41	59%

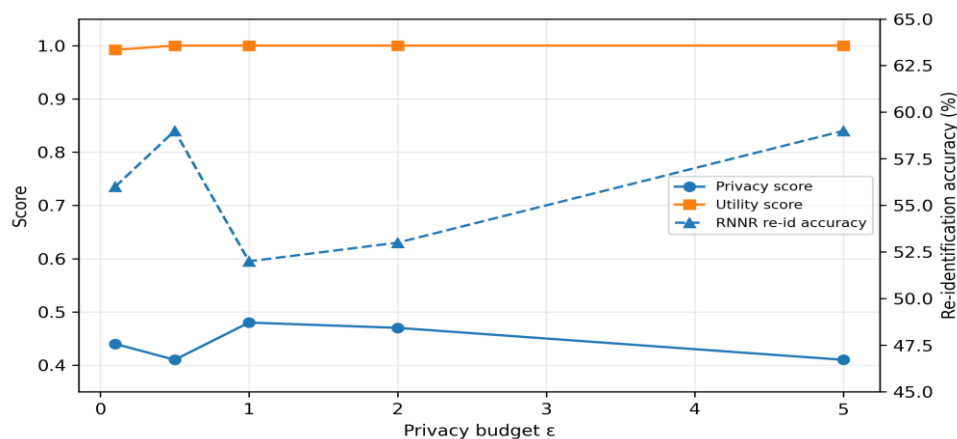


Figure 3. Differential Privacy sensitivity across ϵ values

The experimental results support an operational zone around $\epsilon=0.5$ – 1.0 for this dataset, with $\epsilon=1.0$ selected as the balanced setting. This should not be interpreted as a universal privacy budget. Recent practical DP research emphasizes that ϵ must be selected with reference to the query, sensitivity, threat model, repeated releases, and acceptable analytical error (Li et al., 2023; Munilla Garrido et al., 2023). The contribution here is empirical: under the study conditions, $\epsilon=1.0$ produced a strong combination of low error, high mean preservation, and near-chance RNNR performance.

K-Anonymity and Re-identification Attacks

K-Anonymity produced the clearest real-time limitation. At $k=5$, per-record latency was 48.6 ms and throughput only 21 records/s. At $k=10$, latency rose to 97.2 ms and throughput declined to 10 records/s. These values are much weaker than RNA (0.8 ms; 1,250 records/s), DP (1.2 ms; 833 records/s), Data Swapping (4.5 ms; 222 records/s) and SDG (3.8 ms; 263 records/s) in the thesis benchmark. The reason is structural: maintaining equivalence classes requires global or group-level operations that do not map naturally to continuous low-latency updates.

Privacy protection also remained vulnerable at common k values. Background-knowledge attack success fell from 74.9% at $k=2$ to 61.1% at $k=5$ and 44.2% at $k=10$. Homogeneity-attack success was 55.6%, 47.3% and 31.6%, respectively. Only $k=20$ reduced both attacks below 10%, but that stronger anonymity required more generalization and substantially higher processing cost. This finding is consistent with the long-standing literature on the limitations of k -anonymity and its extensions: group indistinguishability does not, by itself, prevent attribute inference or attacks based on auxiliary knowledge (Machanavajjhala et al., 2007; Li et al., 2007).

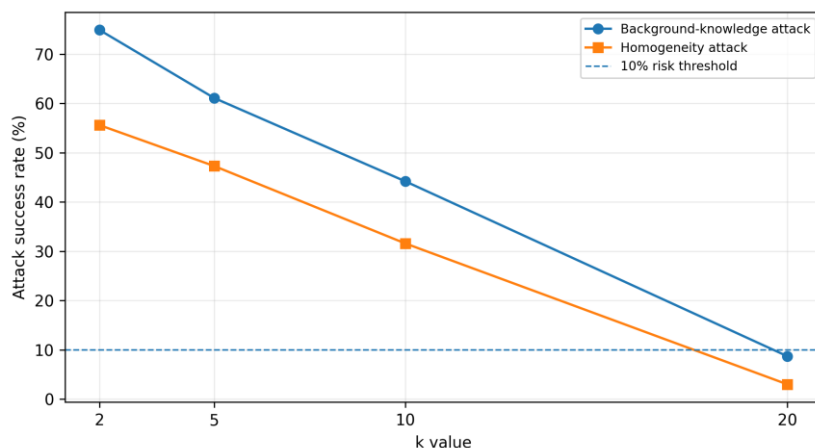


Figure 4. K-Anonymity attack success as k increases

Platform Performance and Scalability

Spark was faster than Hadoop for every technique. RNA improved from 2.1 s on Hadoop to 0.4 s on Spark; DP from 3.4 s to 0.6 s; Data Swapping from 8.7 s to 2.2 s; K-Anonymity from 45.2 s to 24.3 s; and SDG from 22.6 s to 8.4 s. The largest speedups occurred for RNA and DP because their element-wise perturbation operations are highly parallelizable. KA showed the smallest speedup because global equivalence-class construction requires expensive repartitioning and coordination.

Table 6. Hadoop and Spark execution-time comparison

Technique	Hadoop (s)	Spark (s)	Speedup
RNA	2.1	0.4	5.25×
Differential Privacy	3.4	0.6	5.67×
Data Swapping	8.7	2.2	3.95×
K-Anonymity	45.2	24.3	1.86×
Synthetic Data Generation	22.6	8.4	2.69×

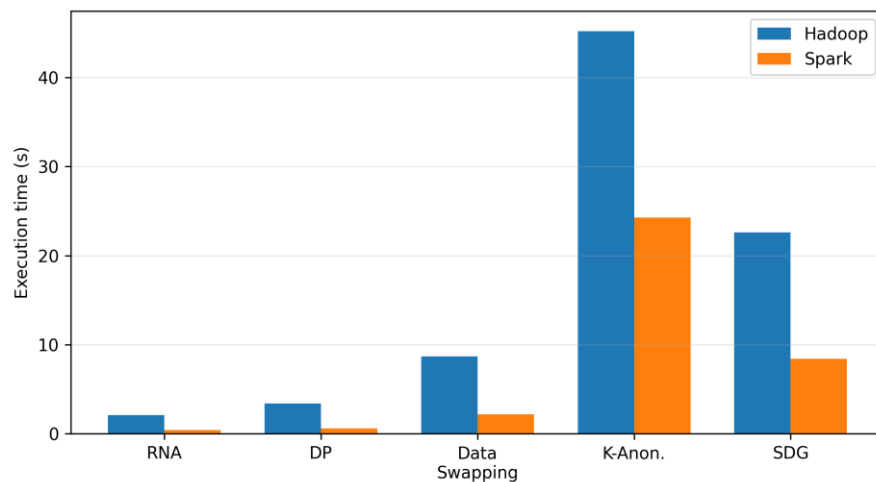


Figure 5. Execution-time comparison on Hadoop and Spark

Scalability analysis over 500, 5,000 and 50,000 records showed approximately linear growth for the five techniques, but the absolute computational cost differed sharply. RNA remained fastest, DP tracked closely, Data Swapping occupied the middle tier, and SDG incurred a larger fixed generative-model cost. KA became operationally impractical at large scale, with a projected execution time of approximately 45,000 s (about 12.5 h) at 50,000 records in the Hadoop-oriented projection. Thus, asymptotic growth alone is not sufficient to judge deployability; the constant factors and synchronization pattern matter greatly.

Integrated Multi-Criteria Evaluation

A seven-dimension scoring matrix was used to integrate privacy, utility, execution speed, scalability, robustness, real-time suitability and regulatory compliance. DP achieved 62/70, the highest score among the five individual techniques. SDG followed at 57/70, RNA at 50/70, Data Swapping at 46/70 and KA at 28/70. The radar representation shows why the ranking is not reducible to one metric: RNA is fast but lacks a formal guarantee; SDG is very strong on privacy and robustness but slower; KA provides syntactic anonymity but has severe speed and utility constraints; DP remains the most balanced.

Table 7. Multi-criteria scoring matrix

Dimension	RNA	DP	DS	KA	SDG
Privacy	6	9	6	8	10
Utility	6	9	7	3	9
Speed	9	8	7	2	6
Scalability	8	8	7	2	7
Robustness	6	9	6	5	10
Real-time	9	9	7	1	6
Compliance	6	10	6	7	9
Composite	50	62	46	28	57

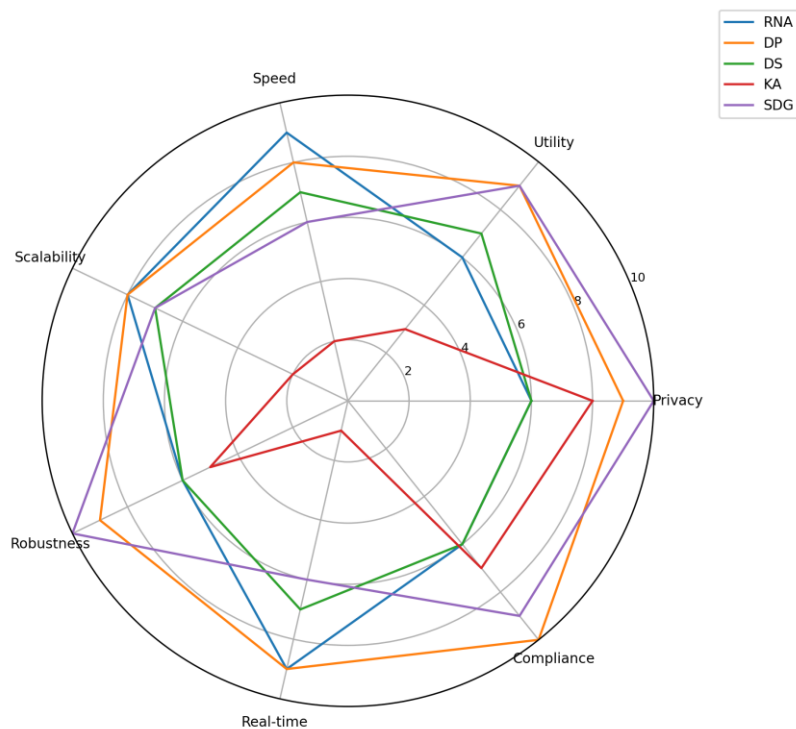


Figure 6. Seven-dimensional performance profiles of the five techniques

Hybrid DP+SDG Model

The Hybrid DP+SDG model separates two organizational needs that are often forced into one privacy mechanism. Layer 1 applies DP to support fast internal or streaming analytics. Layer 2 trains a Gaussian-Copula synthesizer on the DP-protected intermediate data and generates a separate publication dataset. This design avoids requiring a single output to be simultaneously optimized for real-time execution and external release.

Table 8. Hybrid DP+SDG compared with its component techniques

Metric	DP alone	SDG alone	Hybrid DP+SDG
Privacy score	9/10	10/10	10/10
Utility score	9/10	9/10	9/10

RNNR re-id accuracy	52.0%	56.0%	59.0%
Execution time	0.6 s	8.4 s	~9 s batch
Regulatory suitability	Full	Full	Full
Composite score	62/70	57/70	67/70

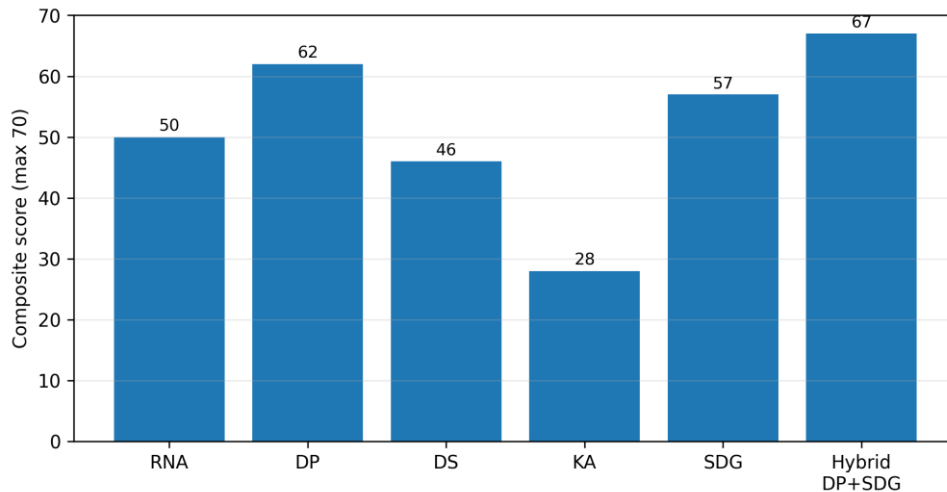


Figure 7. Composite performance ranking including the Hybrid DP+SDG model

The Hybrid architecture achieved the highest composite score, 67/70, while retaining a 9/10 utility rating and 10/10 privacy rating. Its approximately 9 s batch execution reflects the combined DP and SDG stages. The RNNR value of 59% on the final synthetic output requires careful interpretation: the generated records are not source records, so the classifier can detect distributional similarity without proving that an original individual has been recovered. This distinction is consistent with recent literature warning that synthetic-data privacy must be evaluated with explicit attack models and not inferred from realism alone (Kaabachi et al., 2025; Osorio Marulanda et al., 2024).

Discussion

The results show that privacy-preserving analytics is a multi-objective engineering problem. A technique can perform well on one dimension and fail on another. RNA is extremely fast and easy to implement, but the experiment demonstrates that privacy does not improve monotonically with larger noise. This makes uncalibrated randomization unsuitable as the sole protection mechanism for high-sensitivity data. Data Swapping is similarly useful as a supplementary obfuscation technique, particularly for categorical variables, but it lacks the formal privacy accounting available in DP.

DP offers the strongest overall balance in the present experiment because its protection is formally parameterized and its core operations are computationally simple. The observed 99.97–100.00% mean preservation at $\epsilon=1.0$ should not be generalized beyond the tested data and query structure, but it demonstrates that calibrated DP can preserve aggregate information effectively when sensitivity is correctly controlled. This finding aligns with recent research that focuses on practical privacy-budget selection and the possibility of achieving useful analytics under DP (Li et al., 2023; Subramanian, 2023). It also supports the argument that normalization and query sensitivity must be treated as first-order design decisions rather than implementation details.

The negative result for K-Anonymity is equally important. KA is often attractive because it is intuitive and directly operates on released tables, but the experiments reveal two simultaneous problems: substantial record-level latency and persistent attack success at ordinary k values. Increasing k improves resistance to the tested attacks, yet it also increases processing cost and generalization. This is a classical privacy–utility tension, but in a streaming environment it becomes a privacy–utility–latency problem. For static administrative microdata, KA can still be useful, especially when strengthened by l-diversity or t-closeness; however, the present results do not support it as the primary mechanism for high-velocity streams.

SDG offers a different privacy model. Its main advantage is structural separation between source and released records. The Gaussian Copula retained the broad statistical structure of the original data and achieved high privacy and robustness scores, but its fitting and sampling overhead made it less suitable than DP for the low-latency path. Current literature reaches a similar conclusion: synthetic data is promising for sharing and secondary analysis, but privacy and utility require explicit evaluation, and no single fidelity metric is sufficient (Hu & Bowen, 2024; Kaabachi et al., 2025; Liu et al., 2025).

The hybrid result is therefore more than a numerical score improvement. It represents a deployment principle: use a fast formally private transformation for internal analytics, and use a generative release layer when data must leave the protected analytical environment. This is consistent with emerging privacy-engineering practice in which different privacy-enhancing technologies are combined according to data flow and disclosure context rather than treated as substitutes. Recent work on secure synthetic-data analytics and differential privacy similarly emphasizes layered architectures and context-specific controls (Tekchandani et al., 2024; Pereira et al., 2024).

CONCLUSION

This research provides an integrated empirical comparison of five privacy-preserving techniques for real-time Big Data analytics. Differential Privacy produced the strongest overall performance among the individual methods, reaching 62/70 in the multi-criteria framework, preserving means at 99.97–100.00% under $\epsilon=1.0$, and executing in 0.6 s on Spark. Synthetic Data Generation ranked second at 57/70 and offered the strongest structural privacy for publication. Random Noise Addition and Data Swapping were efficient but better suited to supplementary roles. K-Anonymity was not suitable for the tested real-time setting because its latency, throughput and attack-resistance characteristics deteriorated at operationally useful k values. The proposed Hybrid DP+SDG architecture achieved the highest composite score of 67/70 by separating real-time internal analytics from external synthetic publication. The central conclusion is therefore not that one mechanism should replace all others. Privacy-preserving analytics should be designed as a context-specific architecture in which formal guarantee, statistical utility, adversarial resistance, computational cost, and release purpose are evaluated together.

REFERENCES

1. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407.
2. Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: A literature review. *Journal of Big Data*, 10, 115.
3. Hu, J., & Bowen, C. M. K. (2024). Advancing microdata privacy protection: A review of synthetic data methods. *WIREs Computational Statistics*, 16(1), e1636.
4. Huang, W., Zhuo, M., Zhu, T., Zhou, S., & Liao, Y. (2023). Differential privacy: Review of improving utility through cryptography-based technologies. *Concurrency and Computation: Practice and Experience*, 35(5), e7565.
5. Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Kulynych, B., Prasser, F., & Raisaro, J. L. (2025). A scoping review of privacy and utility metrics in medical synthetic data. *npj Digital Medicine*, 8, 60.
6. Kiran, A., Rubini, P., & Saravana Kumar, S. (2025). Comprehensive review of privacy, utility, and fairness offered by synthetic data. *IEEE Access*, 13, 15795–15811.
7. Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy beyond k-anonymity and l-diversity. In 2007 IEEE 23rd International Conference on Data Engineering (pp. 106–115). IEEE.
8. Li, Y., Liu, Y., Li, B., Wang, W., & Liu, N. (2023). Towards practical differential privacy in data analysis: Understanding the effect of epsilon on utility in private ERM. *Computers & Security*, 128, 103147. <https://doi.org/10.1016/j.cose.2023.103147>
9. Liu, Y., Acharya, U. R., & Tan, J. H. (2025). Preserving privacy in healthcare: A systematic review of deep learning approaches for synthetic data generation. *Computer Methods and Programs in Biomedicine*, 260, 108571.
10. Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkatasubramanian, M. (2007). l-Diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), Article 3.
11. Munilla Garrido, G., Liu, X., Matthes, F., & Song, D. (2023). Lessons learned: Surveying the practicality of differential privacy in the industry. *Proceedings on Privacy Enhancing Technologies*, 2023(2), 151–170.

12. Osorio Marulanda, P. A., Epelde, G., Hernández, M., Isasa, I., Moreno, N., & Beristain, A. (2024). Privacy mechanisms and evaluation metrics for synthetic data generation: A systematic review. *IEEE Access*, 12, 88048–88074.
13. Pan, K., Ong, Y.-S., Gong, M., Li, H., Qin, A. K., & Gao, Y. (2024). Differential privacy in deep learning: A literature survey. *Neurocomputing*, 589, 127663.
14. Pereira, M., Kshirsagar, M., Mukherjee, S., Dodhia, R., Lavista Ferres, J., & de Sousa, R. (2024). Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data. *PLOS ONE*, 19(2), e0297271.
15. Rafiq, F., Awan, M. J., Yasin, A., Nobanee, H., Zain, A. M., & Bahaj, S. A. (2022). Privacy prevention of big data applications: A systematic literature review. *SAGE Open*, 12(2).
16. Rocher, L., Hendrickx, J. M., & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10, 3069. Subramanian, R. (2023). Have the cake and eat it too: Differential Privacy enables privacy and precise analytics. *Journal of Big Data*, 10, 117.
17. Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557–570.
18. Tekchandani, P., Das, A. K., & Kumar, N. (2024). Large-scale secure model learning and inference using synthetic data for IoT-based big data analytics. *Computers & Electrical Engineering*, 119, 109565.
19. van Breugel, B., Liu, T., Oglic, D., & van der Schaar, M. (2024). Synthetic data in biomedicine via generative artificial intelligence. *Nature Reviews Bioengineering*, 2, 991–1004.
20. Yang, M., Guo, T., Zhu, T., Tjuawinata, I., Zhao, J., & Lam, K.-Y. (2024). Local differential privacy and its applications: A comprehensive survey. *Computer Standards & Interfaces*, 89, 103827.