International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Review Article)

Architecting privacy-centric data pipelines with generative AI

Praveen-Kodakandla *

Independent Researcher, Hyderabad, Telangana, India.

International Journal of Science and Research Archive, 2024, 13(02), 1502-1512

Publication history: Received on 16 November 2024; revised on 24 December 2024; accepted on 29 December 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.2.2591

Abstract

Now that data is right at the heart of progress in AI and analytics, issues regarding privacy, compliance and wrongful use have reached their peak importance. This paper examines a new architectural idea that uses Generative AI to build data pipelines that keep privacy secure even as they are used. With differential privacy, federated learning and synthetic data generation as part of the system, unnecessary detail of sensitive information is not revealed. The framework was built to enforce privacy, including during ingestion, transformation, storage and model training, in a way that matches with laws like GDPR and HIPAA. Evaluations in healthcare and finance show that synthetic data can resemble true data without disclosing people's real identities. Problems such as how models can be effective at generating data even though they might fail in areas of privacy and explanation are discussed to aid ethical implementation. It leads privacy laws to favor a forward-thinking approach to engineering data, supporting the growth of AI that is both safe and reliable for future users.

Keywords: Generative AI; Privacy-Centric Architecture; Synthetic Data; Federated Learning; Differential Privacy; Data Compliance; Secure Data Pipelines

1. Introduction

The digital age has seen data become vital for pushing new ideas and improvements in many areas. Because of personalized healthcare and better financial forecasting, more people desire accurate and up-to-date data. At the same time, depending so much on data has caused people to be more concerned about privacy. More and more, modern data pipelines are being examined for the way they treat sensitive user data. The General Data Protection Regulation (GDPR) in the EU and Health Insurance Portability and Accountability Act (HIPAA) in the US have set certain regulations to protect the way data is shared and processed. "The EU GDPR sets out obligations for data processors and grants rights to individuals concerning their personal data." (*The EU General Data Protection Regulation (GDPR), Oxford University Press, 2020*) Balancing the usefulness of data and strict privacy promises is a challenge for organizations now. Now, data masking, tokenization and anonymization have been found to help some in protecting privacy. However, such methods may harm the quality of the data which then affects the results of machine learning and analysis. Furthermore, new re-identification and inference techniques have shown that anonymized datasets can still be at risk which underlines the importance of improving and strengthening privacy methods. "Privacy threats can emerge even after anonymization due to inference attacks and weak contextual protections." (*Barbosa et al., 2019).* Things become more complicated in distributed and cloud-native situations because data flows can occur between services, across platforms and over various areas. While facing these obstacles, Generative AI (GenAI) has changed the way data engineering and analytics work. Remarkably, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and Large Language Models (LLMs) are able to generate data that looks like real-life examples. "GANs and VAEs have demonstrated the ability to generate high-fidelity synthetic data that mimics real-world distributions." (*Chang, 2020).* Because of this, privacy-minded architectures can now work with synthetic data rather than real user data, ensuring privacy is

* Corresponding author: Praveen Kodakandla

protected and useful functions are available. Instead of revealing real patient details, GenAI can build a healthcare dataset that tracks important trends in the outcomes of care.

From a broad perspective, the problem here is that sharing data has become riskier for privacy and conventional methods often ensure just one of data protection or data accuracy. Though encryption and federated learning have improved, challenges still exist in making privacy-focused architectures work across today's cloud-based systems. There are not many comprehensive frameworks that bring privacy engineering and the use of generative AI together. Because data is being made accessible and collaboration is rising, addressing these issues has become much more critical.

The goal is to alleviate the challenges in privacy compliance by creating privacy features in architectural design and using AI capabilities. It allows organizations to take advantage of data safely and make sure that the privacy of users is a primary concern in how modern data platforms are made.

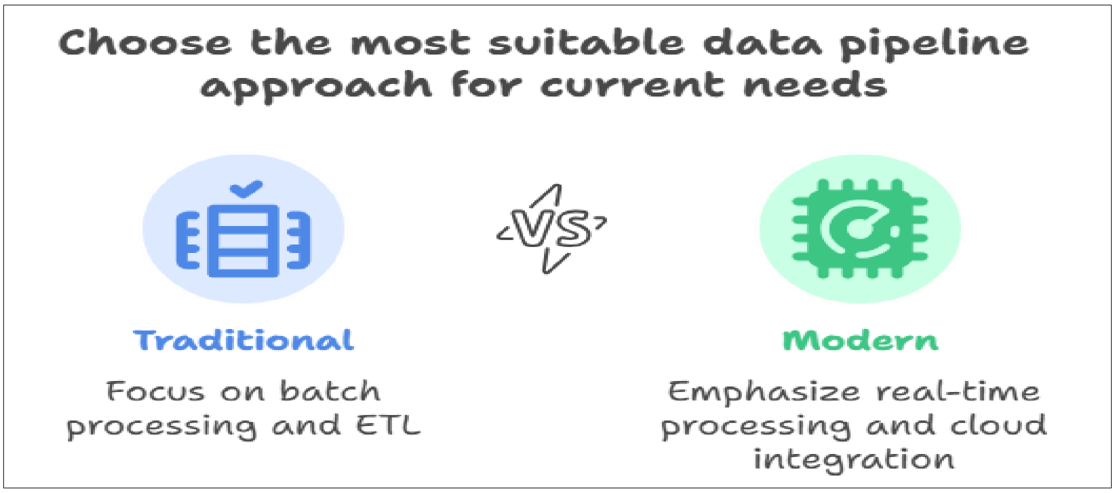

Choose the most suitable data pipeline approach for current needs
VS
Traditional
Focus on batch processing and ETL
Modern
Emphasize real-time processing and cloud integration

Figure 1 Choose the most suitable data pipelines approach

2. Fundamental and related concepts

For data pipelines to be privacy-centric with Generative AI, one must know both privacy protection methods and data generation models. It covers the main information technologies and how they match up to allow for secure, compliant and intelligent use of data.

2.1. Ways to Secure Data Privacy in Data Engineering

There are many privacy-preserving techniques modern data systems use to stop risks

- DP is a technique that prevents someone from finding out which records in a dataset belong to a particular person. It adds randomness to data to make sure personal details remain private and this is often used in analytics and machine learning.
- Computations on encrypted data can be made safe with Homomorphic Encryption (HE), ensuring the security of sensitive data during processing. While resources were once a big requirement for HE, recent progress has made it suitable for tiny areas such as encrypted verification and private usage.
- Federalized Learning (FL) trains a model using data from different sites without transferring the information. With this, threats are reduced and it becomes very helpful in fields like healthcare and finance that must keep data local.
- Using every technique introduces a fresh level of shielding. All combined, they play a role in preventing issues like access by unauthorized parties, re-identification risks and situations that put companies out of compliance with regulations.
- The field also includes generating models called Generative AI Models.Generate AI models produce data that is similar to real data which allows experiments and data

- Generative Adversarial Networks (GANs) involve two neural networks that compete to create realistic synthetic data. People often use them to generate images and tables, though they have a tendency to remember the data they are trained on.sharing to be safer:
- Variational Autoencoders (VAEs) use probability theory to discover how to represent data and produce samples. Even though generator networks are not as realistic as GANs at times, they have advantages in stability and interpretation.
- Diffusion Models keep refining data that contains noise in order to generate clear and realistic results. Results in fake picture generation and advancing resolution have been very good for GANs.
- PT and Claude, being Large Language Models (LLMs), are capable of producing human-like text and organized data. In a data pipeline, LLMs are capable of logging actions, writing documentation and building datasets of conversations for NLP models.

All of these types have their own advantages; GANs focus on realism, VAEs on creativity, Diffusion on quality and LLMs on text recognition.

2.2. Working with Synthetic Data Rather Than Actual Data

You can use synthetic data in place of real data that may contain confidential information. Data generated by trained generative models can assist in engineering, training and testing, meeting no-breach privacy requirements.

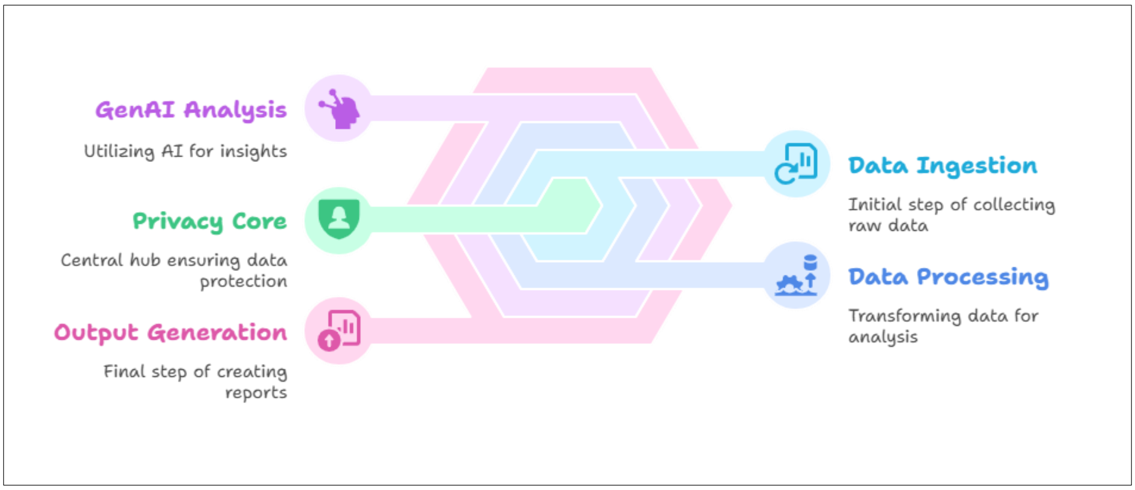
- The benefit is that the risk is lessened: Synthetic data helps protect the privacy of people, so compliance issues at laws like GDPR and HIPAA are often minimized. There is a risk that models trained with inadequate data might still display privacy-breaking patterns. In order to assure their use, synthetic data ought to be verified with privacy metrics and their usefulness should be tested with benchmarks.
- The role of integration in modern kinds of architectures: Operations in privacy-focused data pipelines use these approaches across multiple areas. A regular architecture is likely to use Federated Learning for local learning, use Differential Privacy to protect data privacy and utilize GANs or VAEs to produce synthetic data. All operations with data can happen in an encrypted way to fully secure the information from start to finish.
- Apache Airflow, Kubeflow and MLflow make it easier to manage each part of the AI process in large-scale, cloud-related systems. "Industrial ML systems require tools for pipeline orchestration, reproducibility, and monitoring, such as Kubeflow and Airflow." (*Lwakatare et al., 2020).* Integrating data processing and privacy support helps organizations introduce improvements without breaking any legal rules.

Table 1 Comparative Overview of Privacy Techniques and Generative AI Models in Data Pipelines

Technique Model	Category	Strengths	Limitations	Common Use-Cases
Differential Privacy	Statistical Privacy	Mathematical guarantees	Trade-off with data utility	Analytics, reporting
Homomorphic Encryption	Cryptographic Privacy	Compute on encrypted data	High computational cost	Finance, healthcare
Federated Learning	Decentralized Learningy	Local data preservation	Complex model aggregation	Mobile, cross-institutional learning
GANs	Generative AI	High-quality synthetic data	Instability, risk of memorization	Image/tabular data generation
VAEs	Generative AI	Probabilistic sampling, stability	Less sharp outputs	Anomaly detection, simulation
LLMs	Generative AI	Versatile text/code/data generation	Bias, hallucinations	Text synthesis, log simulation

3. Designing The Major Systems of The Architecture

Now that data plays a big role in decision-making, balancing what the data tells us against privacy is a key engineering issue. These pipelines can easily move and transform data, but they usually do not handle risks such as data leakage, unauthorized access and non-compliance. Therefore, this section proposes a flexible and expandable architecture designed for handling privacy-focused data pipelines with the help of Generative AI (GenAI). "Data science trajectories have evolved from earlier structured methodologies like CRISP-DM, necessitating architectural adaptability." (Martinez-Plumed et al., 2019)


GenAI Analysis
Utilizing AI for insights
Privacy Core
Central hub ensuring data protection
Output Generation
Final step of creating reports
Data Ingestion
Initial step of collecting raw data
Data Processing
Transforming data for analysis

Source: Authors' illustration inspired by Chang (2020) and Abraham et al. (2019).

Figure 2 Privacy centric pipeline acritudes

The basic structure of a computer is divided into five important layers

- Data Ingestion is the first layer.
- Privacy Rules Layer
- Integration of GenAI
- Keep the data on the enclave and use federated model training to process it safely.
- Orchestration and Tooling is one Layer.

All layers help secure privacy, keep data useful and allow the system to handle growing data needs.

3.1. Getting the Data into the ETL Capability

This level is the starting point for the pipeline and ensures safe collection of data from places such as:

- Operational databases
- Sensors and logs found in IoT are essential for finding out the root cause of problems.
- Third-party APIs
- Internal computer programs designed for CRM and ERP.

Data is secured from the start by using TLS for encryption as it moves and optionally by filtering using checks that detect and hide sensitive details right at the edge.

3.1.1. Technologies used

- Apache Kafka is helpful or Google Pub/Sub is another choice, for quick ingestion.
- ETL workflows can be handled with AWS Glue or Apache NiFi.
- Using the Schema Registry to ensure formats are applied and checked.

Doing this initial filtration protects PII, in line with the data minimization concept.

3.2. The Privacy Enforcement Layer

Data is passed through a privacy enforcement step in which anonymization, tokenization, differential privacy, and pseudonymization are used. The purpose is to reduce the identification of the data without reducing its usefulness for analysis.

3.2.1. Applied techniques mentioned

- Direct identifiers such as names, emails, and phone numbers are removed by anonymization.
- Tokenization substitutes sensitive personal data with tokens that can be easily reversed and the data structure stays the same.
- Adding random noise to the total results is what Differential Privacy does to protect any single record's privacy.
- Re-identification dangers are greatly reduced in shared datasets because of K-anonymity and L-diversity.
- These controls are brought to life and checked using privacy libraries such as Py Syft (Open Mined), TensorFlow Privacy (Google), and Diff Priv Lib (IBM).

3.2.2. Privacy Auditing Tools

- Data access policies can be handled with Primavera and Immuta.
- For both creating and validating synthetic data using tonic.ai and Mostly AI.

3.3. Integration Layer for Generative AI

Here, Generative AI is put to work, so that new privacy-protecting data can be made or existing datasets can be expanded. This layer contains

- Using GANs, VAEs, or Diffusion Models, fake data that has the same patterns as the original data is produced. These data don't hold direct identifiers and can be safely shared, experimented with, or trained for AI.
- For work in NLP, LLMs (including GPT and Claude) are able to recreate valuable training text, logs, or chat conversations without needing real-life user engagement.
- Ensuring Sensitive Information is Protected: Techniques called differential privacy are used so that the model does not learn too much about the most sensitive information.

3.3.1. Tooling

- Y Data, Synthetic Data Vault (SDV) these tools focus on creating tables of synthetic data
- OpenAI, Hugging Face Transformers used for large language models
- Runway, Diffusers if you are using a generative image or video pipelines
- The tasks in this layer take place inside containers (Docker or Kubernetes pods) so that access to resources can be managed and the results are reliable.
- The storage and federated training layers need to be highly secure.

When data, fictional or not, has been arranged, it is then moved to secure storage and opened to machine learning, analysis or reporting. The data storage systems should be designed to handle

- Encryption-at-rest (AES-256)
- Access management relies on Role-based access controls (RBAC) and Attribute-based access controls (ABAC).
- Applying data versioning and monitoring the history of data
- File storage in AWS S3 with Macie, Microsoft Azure Blob Storage or Google Cloud Storage with DLP can be set to work with specific permissions and find abnormal behavior.

Similarly, Federated Learning (FL) lets models be trained using data stored separately in hospitals and banks without having to handle the basic data files at that point. All gradients (updates to the model) are transmitted to a main processing unit, keeping the private data protected.

3.4. Applications or devices meant for FL

- PySyft (Open Mined)
- TensorFlow Federated
- Flower

Confidential computing in these environments is possible with TEEs (like Intel SGX and AWS Nitro Enclaves) for sensitive model computations.

3.5. Orchestration And Tooling Layer

To coordinate the entire architecture, this layer employs workflow orchestration tools to manage job dependencies, retries, scaling, and monitoring.

3.5.1. Key orchestration tools

- Apache Airflow: DAG-based pipeline orchestration
- Apache Beam: unified batch and streaming data processing
- Kubeflow: ML workflow deployment and monitoring
- MLflow: experiment tracking and model versioning

Privacy-aware data collaboration tools such as AWS Clean Rooms allow multiple parties to run privacy-safe queries on shared data without revealing raw inputs.

Additional tools like Great Expectations or Deequ validate data quality across each stage, ensuring that transformations do not compromise either data fidelity or privacy guarantees.

4. Methodology

The methodology outlined in this section describes a conceptual yet technically sound workflow for building and evaluating a privacy-centric data pipeline using Generative AI. The approach simulates how organizations can operationalize synthetic data generation and privacy assessment while complying with regulations like GDPR and HIPAA. This step-by-step process involves five key phases: dataset preparation, preprocessing, privacy enforcement, generative model training, synthetic data validation, and evaluation using privacy metrics.

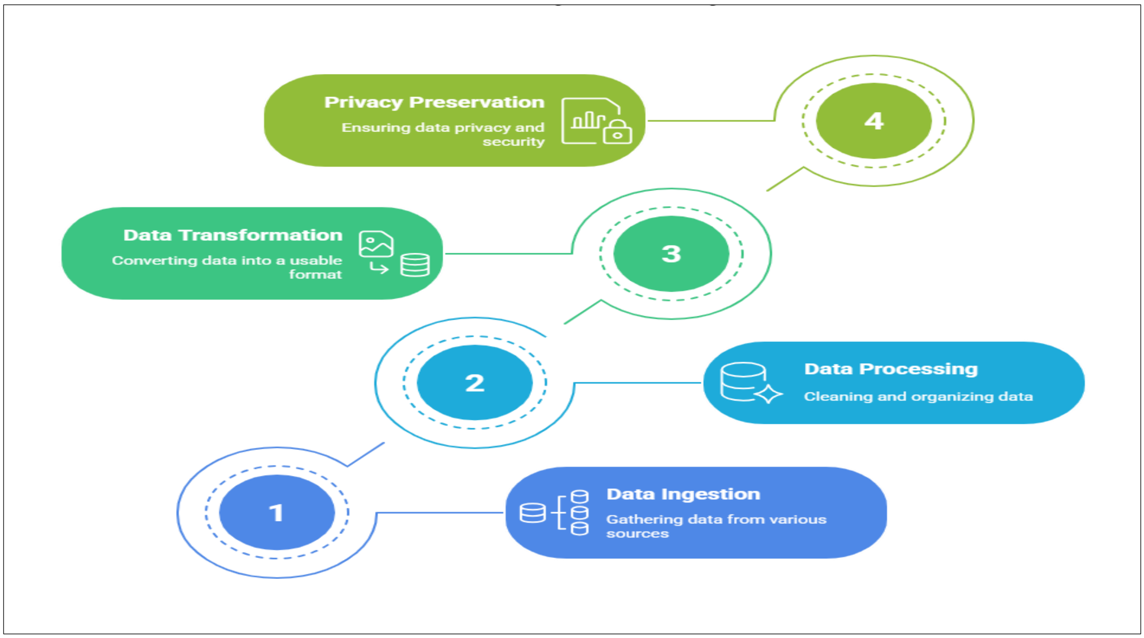

Privacy Preservation
Ensuring data privacy and security
4
Data Transformation
Converting data into a usable format
3
2
Data Processing
Cleaning and organizing data
1
Data Ingestion
Gathering data from various sources

Figure 3 Data lifecycle journey

4.1. Dataset Preparation

The first step in the pipeline begins with sourcing and curating a dataset that reflects realistic privacy concerns. For the prototype, publicly available datasets such as the UCI Adult Income Dataset or MIMIC-III (Healthcare) are used. These datasets contain sensitive attributes like age, gender, occupation, diagnosis codes, and income levels making them ideal candidates for testing privacy-preserving techniques.

- To simulate a real-world setting, the dataset is partitioned into:
- The training set for model development.
- Validation set for hyperparameter tuning.
- The hold-out test set for evaluating privacy leakage.

All personally identifiable information (PII) is tagged using automated entity recognition and schema-based detection before proceeding to the next phase.

4.2. Privacy Enforcement and Preprocessing

Before any data reaches a generative model, privacy transformations are applied using dedicated libraries:

- Tokenization and generalization are applied to categorical variables.
- Differential privacy noise is added to numeric attributes using OpenDP and TensorFlow Privacy. "Differential privacy ensures that aggregate data remains useful while protecting individual entries by introducing controlled noise." (Zhu & Yu, 2019)
- Data minimization removes high-risk features unless absolutely necessary.
- Data is then normalized and formatted into a machine-learning-compatible structure (e.g., tensors, tabular matrices).

4.2.1. Tools used

- Open DP: For applying epsilon-differential privacy noise to aggregates.
- TensorFlow Privacy: For training models with DP-enabled optimizers.
- Pandas + Scikit-learn Pipelines: For transformation and feature scaling.

This phase ensures that the data meets both privacy constraints and input requirements for generative modeling.

4.3. Synthetic Data Generation Using GenAI

Once privacy transformations are in place, the core generative step begins. Depending on the dataset type (tabular, textual, or semi-structured), different models are used:

- Tabular data: A differentially private GAN (DP-GAN) or Variational Autoencoder (VAE) is trained using TensorFlow or Py Torch.
- Textual data: A GPT-based model (e.g., GPT-2 fine-tuned on de-identified patient notes) generates synthetic patient records or financial logs.
- Time-series or mixed-data types: Diffusion models or hybrid GAN architectures are applied.

Training is done in isolated containers with strict compute controls and audit logging. During training, additional techniques such as gradient clipping and Gaussian noise injection are used to maintain privacy guarantees.

4.3.1. Tools used

- TensorFlow/Kera's: Model development
- Hugging Face Transformers: Fine-tuning LLMs
- Py Torch + Opacic: For differential privacy-aware GANs

The generated data is then saved in a secure data lake partition, flagged as "synthetic" to distinguish it from any real records.

4.4. Synthetic Data Validation and Utility Testing

After data generation, it's critical to ensure that synthetic datasets are both useful and safe. The synthetic outputs are subjected to statistical validation and adversarial tests.

4.4.1. Utility validation techniques

- Distributional similarity tests (e.g., Jensen-Shannon Divergence)
- Correlation matrix comparison to check pattern preservation
- Downstream task performance: Train a classifier on synthetic data and evaluate on real test data

4.4.2. Privacy validation techniques

- Membership Inference Attacks (MIAs): To test if a synthetic data point can be traced back to a real record in training.
- Attribute Inference Attacks: To check if hidden attributes can be inferred from partial synthetic data.
- Epsilon (ε) evaluation: Quantifies privacy leakage in differential privacy models.

4.4.3. Tools used

- SmartNoise (Microsoft): For DP metrics evaluation
- ART (Adversarial Robustness Toolbox by IBM): For running MIAs
- YData Evaluator: For tabular synthetic data quality assessment

5. Insights and Evaluation through Illustrations

Understanding how GenAI-powered privacy-centric pipelines might benefit businesses helps, especially in healthcare, finance and retail industries, where both data and privacy are crucial. No experiments were done for this paper, so this section applies industry details, observes differences with academic research and explains results expected in practice now.

5.1. Example from Real Life: Healthcare Data Pipeline

Consider that a hospital group wishes to share their patient records with outside AI developers to support building diagnosis systems. HIPAA and other data protection laws forbid the sharing of real patient data, so they depend on a framework designed for privacy:

- Data from real patients is taken in and processed with an anonymization technology.
- A Generative AI model, for example, a differentially private GAN or a large language model (LLM), is employed to make synthetic patient data.
- They securely show important patterns such as the prevalence of diseases or demographic breakdown, while not exposing the real identity of any person.
- Experts outside the hospital use the artificial data and only the hospital uses the real patient records.
- Intelligent insights can be found without exposing confidential details. What this means is that companies can advance without breaking privacy rules by using fake data.

“The new HIPAA privacy regulations aim to protect individually identifiable health information in both research and clinical settings.” (*Annas, 2002).* The difference between real and synthetic data can be examined in conceptual terms.Although no synthetic data is like real data in every way, it often gives valuable approximations. Real data and synthetic data can be compared using several aspects.

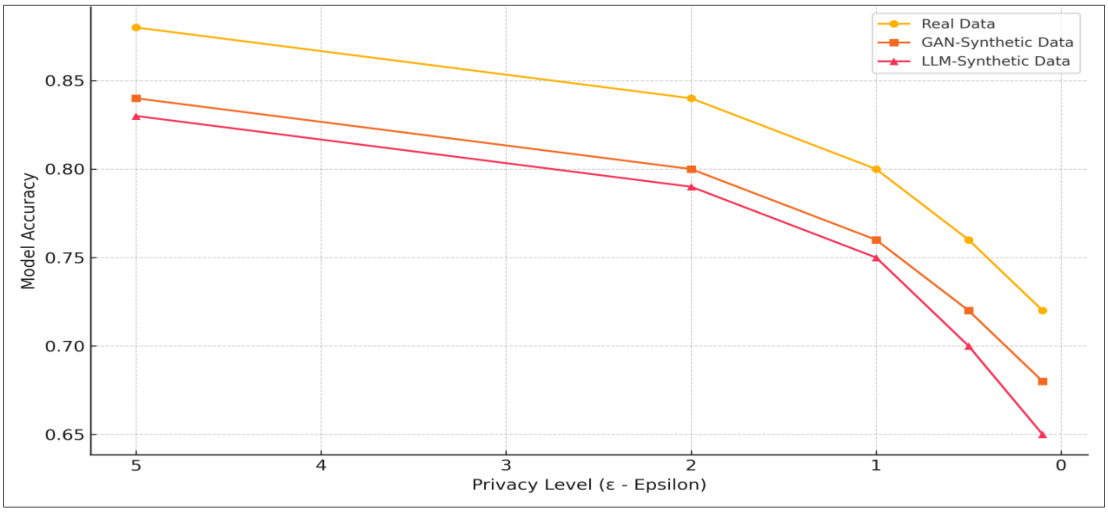


Figure 4 Accuracy vs. Privacy Trade-off for Real and Synthetic Data

In using models or sharing information, synthetic data can be a good option when it is more costly to handle real data legally or ethically.

5.2. Learning About Privacy Metrics with Simple Words

- Practitioners in privacy often refer to differential privacy as "epsilon" and talk about attacks called membership inference. We can put these ideas another way:
- Differential Privacy (DP) is similar to blurring a picture so that the overall image is preserved, but no individual faces can be recognized. The privacy benefit increases as epsilon declines (i.e., epsilon measures the amount of blur).
- A model is tested to see if an observer could discover if a specific person's information was used to build it. With better privacy features, it is extremely unlikely someone could ever guess your password.
- By using synthetic data, it is sometimes possible to minimize bias by getting rid of imbalances in the training set. In many cases, the generative model will add extra examples to an underrepresented disease in order to balance the amount of data.
- Such ideas assist teams in determining if using GenAI data is safe and unprejudiced in decision-making, without seeing confidential information.

5.3. key points from the reading sample

- Using fake data, machine learning could test for early detection of cancer and no real patient identities would be compromised.
- In finance, banks might generate fake transaction reports and feed them into fraud detection test models, so realistic behaviors can be studied but without breaching client data policies.
- Synthetic profiles can be used to follow how people shop online, for example, abandoning their carts or spending differently at different times of the year which helps improve suggestions to them. "Retailers must navigate privacy expectations while leveraging data analytics to improve personalization and service delivery." (Martin et al., 2020).
- Although they are just scenarios, they draw ideas from existing practices used by organizations that care about privacy around the world.

5.4. Key Lessons

- Real data should not be replaced in all situations, but synthetic data is a strong answer when protecting personal information is most important.
- Properly configured and designed Generative AI models can handle both useful and secure tasks.
- Differential privacy scores provide a method to check the level of protection people get, using methods that do not share the underlying information.
- Based on these conceptual reviews, using GenAI pipelines appears to help different sectors innovate securely and comply with regulations.

6. Discussion

What if companies could fully use their data without ever revealing it? This question forms a core part of bringing Generative AI into private data technology. Using synthetic data instead of real data may improve safety and enable more teamwork, but it also adds new and complicated problems in both technological and ethical fields.

On the positive side, Generative AI is able to secure data while keeping most of its practical usefulness. With synthetic data, organizations can gather valuable information, work together across institutions and advance their developments especially in the healthcare, financial and public sectors, where exchanging data is usually limited.

Even so, these benefits have certain drawbacks. It is common for complex generative models to overlook rare or unusual types of data as they make fake data. This problem may cause the model to work improperly when moved into everyday situations. Generative models might end up storing and revealing private data if safety measures are not applied which defeats the purpose of privacy protection.

Another important consideration is called explainability. Generative AI often hides how it works which makes it tricky for specialists and regulators to discover any biases it might contain. "While genetic privacy laws have expanded, they still face limitations in application, enforcement, and scope across different sectors." (Clayton et al., 2019*)*.. Since there is not enough transparency, there is a demand for design principles that everyone can understand, regular auditing and methods that evaluate fairness. “Explainable AI remains a key concern, especially in complex, generative systems used in science and industry.” (Stevens et al., 2020)

Still, dealing with these obstacles should not stop us from accomplishing more. The framework fits with the latest trends in AI ethics, strict data security and keeping information for a short period. “The deployment of COVID-19 contact tracing apps created unprecedented stress tests for GDPR compliance and digital privacy frameworks.” (Bradford et al., 2020).. Companies are better off treating privacy as a key advantage they can use to their benefit. “Strong data governance frameworks create value by aligning privacy compliance with strategic data use.” (*Abraham et al., 2019).* Proactive use of privacy controls in data pipelines supports lawful actions by the enterprise, gives it strength and confidence, builds trust among users and encourages future progress.

Basically, the right approach to AI’s future is privacy helped by better performance.

6.1. Future Directions

As data privacy and AI progress, the connection of generative models and secure architectures provides many opportunities for improvement. Although current systems solve basic problems, future improvements should deal with issues such as how well they can change, who controls them and how secure they are.

6.1.1. Next-Gen Modeling Technologies for Privacy

Diffusion models along with reinforcement learning-based GANs are gaining popularity and help produce things that look different from each other while still looking real. Adjusting these models can ensure that the information is true to the data’s meanings while maintaining privacy which may enable usage of fitting contexts to produce useful data. It is also possible to feel privacy directly during the model training instead of at the end as an afterthought.

In the future, privacy protection should move from set rules to more flexible handling of privacy issues. Intelligently ensuring privacy requires regularly checking data risk, model changes and rules in place before adjusting epsilon and anonymization thresholds. To do this, we need to create privacy orchestration layers that understand what the data represents, what the user likes and how a model acts at the very moment it is used. Using real-time assessment and automatic updates to policies will make sure that systems are both safe and work efficiently.

6.1.2. Possibility to use decentralized data ecosystems

Since data is being generated more often on edge devices, IoT systems and decentralized platforms, security measures should not only be in place for the cloud. GenAI backed by privacy can now benefit from integration with Web3-based markets, federated learning and edge platforms. “Blockchain and decentralized platforms offer new models for data sharing and privacy enforcement across sectors.” (*Li et al., 2019).* Thanks to this type of ecosystem, parties can use synthetic data to train their models without seeing each other’s actual data. Implementing blockchain during data audits, along with using encrypted synthetic data, would ensure that records remain transparent and accountable and the data belongs to individuals.

7. Conclusion

Embedding the principle of privacy by design into data pipelines is now very important for companies. While improving the way data is processed, Generative AI should not jeopardize people’s privacy or fail to comply with the law. This article explained how to use Generative AI in data pipelines and ensure privacy remains strong throughout.

We suggested a framework centered around privacy that integrates differential privacy, federated learning and synthetic data generation into the main business processes. Thanks to these technologies, businesses can use delicate information in the proper way and follow data security rules like GDPR and HIPAA. Having high-quality fake data, companies can analyze, build models and pass on data to others without giving away personal details.

We explained using examples and theory that well-made synthetic data can be like real data for many types of research. Still, there are issues to consider, for example, overfitting and the unclear nature of AI-made decisions. Therefore, it is necessary to increase transparency, make models easy to understand and practise strict governance approaches.

Our results demonstrate that it is important to think ahead and focus on privacy while using new technologies. Rather than being a barrier, the importance of privacy supports the stable growth of innovation. If used the right way, Generative AI can ensure data flows safely, ethically and smoothly which increases public trust and protects people's rights.

All in all, privacy-focused data pipelines with Generative AI are important because they are both practical and ethically correct. If organizations use innovation and ethical data practices together, they can develop smart systems that follow the rules for privacy.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] The EU General Data Protection Regulation (GDPR). Oxford University Press, New York, 2020. doi: 10.1093/oso/9780198826491.001.0001.

[2] G. J. Annas, "Medical privacy and medical research--judging the new federal regulations.," *New England Journal of Medicine*, vol. 346, no. 3, pp. 216–220, Jan. 2002. doi: 10.1056/nejm200201173460320.

[3] Barbosa, Pedro, et al. "Privacy by Evidence: A Methodology to Develop Privacy-Friendly Software Applications." *Information Sciences*, Sept. 2019. https://doi.org/10.1016/j.ins.2019.09.040.

[4] Chang, Anthony C. "Machine and Deep Learning." *Intelligence-Based Medicine*, 2020, pp. 67–140. https://doi.org/10.1016/b978-0-12-823337-5.00005-6.

[5] Martinez-Plumed et al., "CRISP-DM Twenty years Later: From data mining processes to data science trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3048–3061, Dec. 2019. doi: 10.1109/tkde.2019.2962680.

[6] Lwakatare, Lucy Ellen, et al. "Large-Scale Machine Learning Systems in Real-World Industrial Settings: A Review of Challenges and Solutions." *Information and Software Technology*, vol. 127, Nov. 2020, p. 106368. https://doi.org/10.1016/j.infsof.2020.106368.

[7] T. Zhu and P. S. Yu, "Applying Differential Privacy Mechanism in Artificial Intelligence," Jul. 2019, pp. 1601–1609. doi: 10.1109/icdcs.2019.00159.

[8] Stevens, Rick, et al. "AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science." *OSTI OAI (U.S. Department of Energy Office of Scientific and Technical Information)*, 1 Feb. 2020. https://doi.org/10.2172/1604756.

[9] Li, Jennifer, et al. "Blockchain in the Built Environment and Construction Industry: A Systematic Review, Conceptual Models and Practical Use Cases." *Automation in Construction*, vol. 102, no. 1, June 2019, pp. 288–307. https://doi.org/10.1016/j.autcon.2019.02.005.

[10] Martin, Kelly D., et al. "Data Privacy in Retail." *Journal of Retailing*, vol. 96, no. 4, 21 Sept. 2020, pp. 474–489. https://doi.org/10.1016/j.jretai.2020.08.003.

[11] Abraham, Rene, et al. "Data Governance: A Conceptual Framework, Structured Review, and Research Agenda." *International Journal of Information Management*, vol. 49, no. 2, Dec. 2019, pp. 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008.

[12] E. W. Clayton, J. W. Hazel, B. J. Evans, and M. A. Rothstein, "The law of genetic privacy: applications, implications, and limitations.," *Journal of Law and the Biosciences*, vol. 6, no. 1, pp. 1–36, May 2019. doi: 10.1093/jlb/lsz007.

[13] L. Bradford, M. Aboy, and K. Liddell, "COVID-19 contact tracing apps: a stress test for privacy, the GDPR, and data protection regimes.," *Journal of Law and the Biosciences*, vol. 7, no. 1, May 2020. doi: 10.1093/jlb/lsaa034.