Synth Studio

Unlocking Innovation through Privacy-preserving Data

Whitepaper by

Mayank Umrao, Lead Data Scientist – Mphasis NEXT Labs | Solution Development
Ayaan Khan, Lead Data Scientist – Mphasis NEXT Labs | Solution Development
Rahul Gupta, Manager – Data Science, Mphasis NEXT Labs | Solution Lead
Dr. Revendranath T, Project Manager – Mphasis NEXT Labs | Advisor
Hareesh PS, AVP – Mphasis NEXT Labs | Advisor
Dr. Udayaadithya Avadhanam, Principal-Cognitive Sciences – Mphasis NEXT Labs | Advisor
Biju Mathews, Partner – Mphasis | Advisor

Mphasis
The Next Applied

Contents

1. Introduction

The increasing adoption of Artificial Intelligence (AI) has spurred a demand for synthetic generation of data to experiment, build and deploy Machine Learning, Deep Learning and large language models to accelerate industrial implementation of data products and solutions. Furthermore, the widespread utilization of connected devices and IoT, further expedites the need for synthetic data to bridge the gap in data provision and generate on-demand data. This gap in data provision arises from factors such as privacy concerns, security issues, data ownership complexities, regulatory compliance and the lack of standardized formats, hindering organizations from freely sharing their data. Synthetic data can help bridge the gap in data provision by providing a privacy-preserving and secure alternative, allowing organizations to share representative datasets without disclosing sensitive information.

Synthetic data is not derived from any actual observations but is artificially generated data to mimic the characteristics of real-world data. Synthetic data is created using mathematical models, computer algorithms or other artificial means, and can be used to train Machine Learning models or test software systems in situations where real data is difficult or costly to obtain, or when privacy concerns prevent the use of real data. For example, generative AI algorithms learn about the characteristics of actual observations and then possess the ability to generate synthetic data mimicking the actual observations while effectively hiding sensitive information.

Synthetic data can take many forms, including images, text, audio or numerical data. It is often used in industries such as healthcare, finance and cybersecurity to test and develop algorithms and models without the risks associated with using sensitive or private data.

2. Background

There is a growing demand for top-notch data in different industries because of the increased use of AI, big data and advanced analytics. Understanding and using this data is crucial for creating new applications. However, this need clashes with the rising challenges of more data breaches, cyber threats and worries about keeping data private. Strict rules like HIPAA, GDPR, CCPA and CPA make it even more important to manage data carefully.

Organizations are confronted with a dilemma: the need to share data for robust system development and leveraging advanced data solutions conflicts with concerns about disclosing sensitive, confidential data. This caution stems from awareness of the risks in a world full of cyber threats and complicated rules. As a result, organizations are balancing the act of protecting the data, complying with privacy regulations and training AI/ML models to leverage the data.

To address challenges in data quality and diversity, practices like augmentation, anonymization and differential privacy are common. Augmentation enhances datasets by creating new points, benefiting Machine Learning models. Anonymization removes identifiable information enhancing privacy. Differential privacy introduces random noise to protect individual data. However, such methods often compromise data utility. Several state-of-the-art techniques are employed in generating synthetic data are as follows:

Generative Adversarial Networks (GANs)

GANs, introduced by Ian Goodfellow in 2014, are a class of Deep Learning models that consist of two neural networks: the generator and the discriminator. These networks are trained simultaneously through a process known as adversarial training. The GAN model structure and training process are presented as shown in Figure 1:

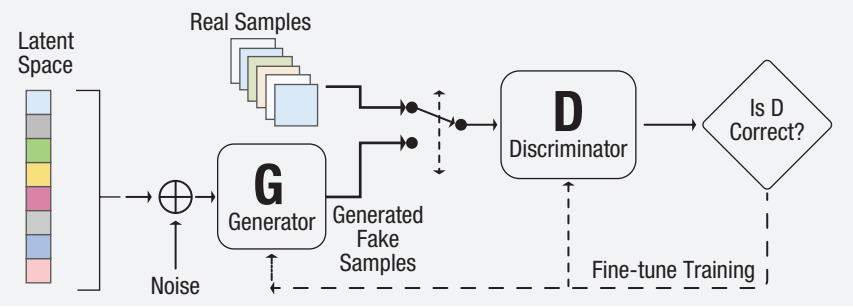

Figure 1: GAN training pipeline [1]

- **Generator:** The generator creates synthetic data samples from random noise with an objective to generate data indistinguishable from real data.
- **Discriminator:** The discriminator evaluates both real and synthetic data, attempting to distinguish between them. Its goal is to correctly identify whether a sample is real or generated.

The training process involves the generator trying to fool the discriminator, while the discriminator aims to accurately differentiate between real and synthetic data. This adversarial process continues until the generator produces highly realistic synthetic data.

Conditional Tabular GAN (CTGAN)

CTGAN is a specialized type of GAN designed to handle tabular data, which often includes both continuous and categorical variables. Developed by researchers at MIT, CTGAN addresses the limitations of traditional GANs in generating realistic tabular data by incorporating several key techniques:

- **Conditional Generation:** CTGAN generates data conditionally, meaning it can produce data samples given specific conditions (e.g., generating synthetic data for a particular category or class). This is particularly useful for imbalanced datasets.
- **Mode-specific Normalization:** To effectively model the distribution of continuous variables, CTGAN uses mode-specific normalization. This technique improves the learning process by normalizing data based on its mode.
- **Training-by-sampling:** CTGAN employs a sampling strategy during training to ensure that both the generator and discriminator are exposed to a balanced representation of all variable types. These enhancements make CTGAN particularly effective for generating high-quality synthetic tabular data, preserving the statistical properties and relationships present in the original dataset.

Variational Autoencoders (VAEs)

VAEs are another popular technique for synthetic data generation. They consist of two main components:

- **Encoder:** The encoder maps input data to a latent space, representing the data in a compressed form.
- **Decoder:** The decoder reconstructs the data from the latent space representation.

VAEs are trained to minimize the difference between the original and reconstructed data while regularizing the latent space to ensure smooth transitions between data points. This makes VAEs well-suited for generating synthetic data that captures the underlying distribution of the original dataset.

Differentially Private Synthetic Data

Differential privacy introduces controlled noise into the data generation process to ensure individual data points cannot be re-identified. Techniques for generating differentially private synthetic data include:

- **Laplace Mechanism:** Adds Laplace noise to the data to achieve privacy guarantees.
- **Exponential Mechanism:** Selects data points based on a probability distribution that balances utility and privacy. These methods ensure that the synthetic data maintains a high level of privacy while retaining the utility needed for analysis and model training.

Synthetic data generation is rapidly emerging as a key solution for addressing data privacy challenges while enabling robust AI and Machine Learning applications. Techniques such as GANs, CTGAN, VAEs and differentially private data generation provide diverse approaches to creating high-quality synthetic data that preserves the statistical properties of real-world data. As organizations continue to navigate the complexities of data privacy regulations, synthetic data offers a promising avenue for safely leveraging data to drive innovation and advanced analytics. This paper emphasizes the importance of synthetic data, compares it with alternatives and explores potential use cases, highlighting its growing role in overcoming data privacy concerns in various business applications.

3. Problem Statement

Developing robust and adaptable AI models depends on access to diverse, high-quality labeled datasets. Gartner's insight, indicating that 85% of AI projects face the risk of generating inaccurate results because of bias in data or algorithms, underscores a significant challenge in AI development. The prevalence of this issue can result in unintended and potentially harmful outcomes, underscoring the need for developers to prioritize the creation of training data free from bias.

The growing need for impartial data has resulted in the emergence of synthetic data generation. Through synthetic data creation, developers can address bias issues and guarantee a dataset that is more comprehensive, varied and of superior quality. Prioritizing unbiased data is essential to prevent the continuation of disparities and foster fair and just AI results.

In conclusion, the journey toward robust AI models requires a commitment to crafting training datasets that are diverse, high in quality and representative but also addresses biases through synthetic data generation. Incorporating these considerations is pivotal for the responsible and effective deployment of artificial intelligence in various applications.

Synthetic data is increasingly recognized as a transformative solution for addressing data challenges in diverse sectors such as automobile, manufacturing, banking, healthcare, security, surveillance, autonomous vehicles and retail. The adaptability of synthetic tabular data proves highly promising across industries, serving as a robust foundation for innovation in analytics and decision-making while safeguarding sensitive information.

The synthetic data market exhibits steady growth, driven by the availability of numerous open-source tools and techniques. Mphasis' Synth Studio stands out as a synthetic data generation and enrichment solution, utilizing innovative algorithms and proprietary methodologies. This approach establishes a scalable and secure pipeline, ensuring the creation of technically sound and privacy-protected synthetic datasets. The overarching goal of Mphasis' offering is to apply state-of-the-art practices and proprietary methods, effectively assisting clients in meeting their diverse data requirements.

Mphasis' Synth Studio addresses the evolving landscape of synthetic data, emphasizing not only technical soundness but also privacy protection. Through a commitment to innovation and client support, Mphasis contributes to the advancement of synthetic data practices in the realm of data science and analytics.

4. Mphasis Solution

Mphasis' Synth Studio generates high-quality synthetic data to monetize trustworthy business insights while preserving privacy and protecting data subjects. Its AI-generated synthetic data enables organizations to leverage data's maximum potential to cross-collaborate and build reliable and highly accurate assets with no loss of data privacy and utility.

It comprises two effective components - Data Synthetization and Data Enrichment.

Data Enrichment is used to augment available data with more information to enhance its contextual utility and unlock innovation - such as generating tags for images, sentiment tags for text reviews, etc.

Data Synthesis enables the building of smarter Machine Learning models and the generation of realistic test data when there is access to both representative original data as well as metadata and schema information about the data.

The key features include are not limited to the following:

Privacy-preserving Synthetic Data Generation

Synthetic data is generated as a privacy-safe version of the original data unleashing maximum utility for insights. This feature is especially useful when synthetic data is generated from original data while preserving its statistical properties and distributions.

Expert-assisted Synthetic Data Generation

Mphasis' Synth Studio creates synthetic data, based on metadata, constraints and conditions, for a wide range of requirements - including application testing, creation of POCs, fueling hackathons and innovation initiatives.

This synthetic data, artificially constructed, allows businesses to test, learn and innovate without breaching any real-world data privacy. For application testing, it offers a secure and effective way to evaluate the functionality and robustness of new software, ensuring that it performs well under different circumstances and data loads.

The creation of POCs is another significant usage of this synthetic data. It allows businesses to highlight the functionality of a proposed system without having to expose actual data. This helps in demonstrating the feasibility of a project, encouraging stakeholder buy-in, and preventing potential data breaches.

In the context of hackathons and other innovation initiatives, Synth Studio's synthetic data becomes a resourceful tool. Teams participating in these events can use this data to build, evaluate and refine their solutions in a secure environment. Instead of having to worry about data privacy regulations or data sourcing, they can focus entirely on their innovation, thus fostering creativity.

In essence, the synthetic data from Mphasis' Synth Studio is a powerful tool that can support various business needs, from application development and testing to innovation-driven events, all while maintaining data privacy and security.

5. Impact of the Solution

The impact of our synthetic data solution is profound, focusing on enhanced privacy protection and measurement to address the growing concerns of adversarial attacks and model inversion in the realm of data science and analytics.

Privacy is a paramount concern, and our solution prioritizes safeguarding the privacy of data subjects in the original dataset. We employ a Privacy at Risk measurement framework, which serves as a proactive approach to identify potential risks that synthetic data might expose. This framework is instrumental in quantifying and understanding the vulnerabilities of synthetic data, allowing organizations to fortify their privacy strategies.

Auto-synthetization is an intelligence feature in our solution which integrates synthetic data generation with other privacy strategies such as differential privacy and data anonymization. In addition, the auto-synthetization feature enables the solution to automatically select and synthesize data based on data types and metadata, ensuring that the generated synthetic data aligns with the original dataset's characteristics, thus preserving its authenticity and integrity.

Conditional synthetization adds another layer of sophistication to our solution. Tables can be synthesized based on user-declared conditions or constraints, whether at the single attribute level or across multiple attributes. For instance, our model can protect information on minority data patterns through intelligent data sampling strategies, such as synthesizing data for specific conditions like age greater than 18 or date of joining less than the date of leaving.

Our solution seamlessly supports relational tables, allowing for the synthesis of multiple tables while preserving relational coherence. This includes maintaining primary and foreign key relationships, ensuring that the synthetic data accurately reflects the structure of the original dataset.

Furthermore, our solution supports multiple data structures and formats, extending its versatility. Whether dealing with network data, images or text, our solution adapts to various data types, providing a comprehensive approach to synthetic data generation.

Tightly integrated with downstream requirements, our solution aligns synthetization methods and privacy metrics with the specific purpose of downstream activities. For instance, customer segmentation needs may require preserving information about groups, while Anti-Money Laundering (AML) activities may demand the preservation of network topology.

Our Expert Assisted Data Generator enhances the user experience by allowing for the generation of synthetic data based on metadata structures, constraints and conditions. AI/ML-based synthesizers preserve data patterns at cross-attribute levels, ensuring consistency and accuracy in the generated synthetic data.

Data enrichment is a significant aspect of our solution, allowing for the generation of synthetic data along with data-driven tags. This enriches data quality for downstream activities, such as image data with object annotations or text data with sentiment tags.

Moreover, our solution excels in multi-modal data synthesis, providing consistent and enriched data generation across multiple modes. For instance, synthesizing x-ray data with radiology reports or 3D MRI data adds a layer of depth and richness to the synthetic data, enhancing its utility for a wide range of applications.

6. Conclusion

Our synthetic data solution - Synth Studio goes beyond mere generation, encompassing privacy protection, intelligent synthesis and seamless integration with downstream requirements. Its impact is far-reaching, addressing the intricate challenges of modern data science while providing a robust foundation for innovation and decision-making.

Explore the capabilities of Mphasis' Synthetic Data Solution by subscribing to our specialized components on the AWS Marketplace for Machine Learning. Engage with the Tabular Synthetic Data Generator, Relational Synthetic Data Generator or the Structured Synthetic Data Evaluator to meet specific data generation and evaluation needs. These components cater to single tabular data, relational tabular data and provide an evaluation platform for your generated data.

For a comprehensive understanding of our solution, visit the Synth Studio solution page, where one can delve deeper into the features and benefits of our state-of-the-art synthetic data generation and enrichment solution.

7. References

[1] Jonathan Hui. GAN – What is Generative Adversarial Networks GAN? (2018), medium article.

About Mphasis

Mphasis' purpose is to be the *"Driver in the Driverless Car"* for Global Enterprises by applying next-generation design, architecture and engineering services, to deliver scalable and sustainable software and technology solutions. Customer centricity is foundational to Mphasis, and is reflected in the Mphasis' Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized ($C = X2C^{2}_{TM} = 1$) digital experience to clients and their end customers. Mphasis' Service Transformation approach helps 'shrink the core' through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis' core reference architectures and tools, speed and innovation with domain expertise and specialization, combined with an integrated sustainability and purpose-led approach across its operations and solutions are key to building strong relationships with marquee clients. Click here to know more. (BSE: 526299; NSE: MPHASIS)

For more information, contact: marketinginfo.m@mphasis.com

USA
Mphasis Corporation
41 Madison Avenue
35th Floor, New York
New York 10010, USA
Tel: +1 (212) 686 6655

UK
Mphasis UK Limited
1 Ropemaker Street, London
EC2Y 9HT, United Kingdom
T : +44 020 7153 1327

INDIA
Mphasis Limited
Bagmane World Technology Center
Marathahalli Ring Road
Doddanakundhi Village, Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000

WAS 09/09/24 US LETTER BASIL 9224

www.mphasis.com