The True Cost of Synthetic Data: A Build vs. Buy Analysis for AI and ML Leaders

The hidden costs, fast wins, and big tradeoffs—revealed.

Table of Contents

Introduction: The Growing Importance of Synthetic Data

As AI adoption accelerates, the demand for high-quality training data has never been greater. Yet real-world data often presents serious challenges—privacy concerns, limited access to rare scenarios, inherent bias, and significant acquisition and labeling costs.

Synthetic data has emerged as a practical, scalable alternative. By simulating the statistical properties of real-world data without exposing sensitive information, it enables organizations to train and test AI models with greater flexibility, speed, and control. From photorealistic images to structured records, synthetic data offers a reliable way to fill data gaps and boost model performance.

With the technology maturing rapidly, organizations now face a critical choice: should they build an in-house synthetic data generation platform tailored to their needs, or leverage a third-party solution that offers faster deployment and built-in expertise?

This white paper explores the trade-offs, costs, and strategic considerations behind the build vs. buy decision—equipping AI and engineering leaders with a framework to make the right call for their teams and goals.

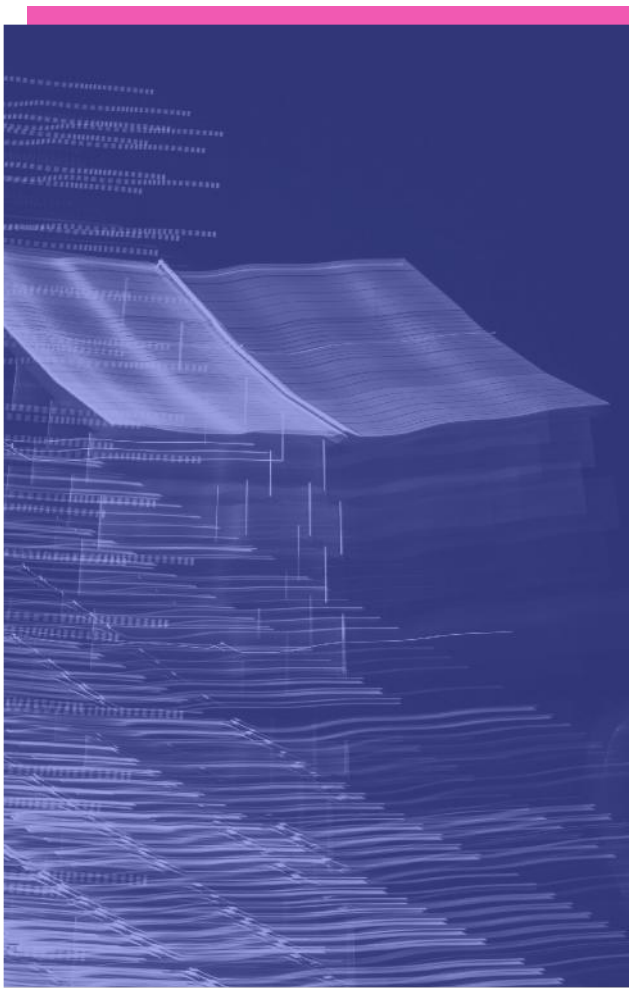
Building a Synthetic Data Platform In-House: A Deep Dive

For organizations with specific and potentially unique requirements, the prospect of developing a synthetic data generation platform internally might seem appealing. This approach offers several potential advantages, primarily centered around the ability to tailor the platform precisely to their needs.

Advantages of In-House Development

Developing a synthetic data generation platform in-house provides a high degree of customization and control over the entire data generation process[1]. Organizations can meticulously design the algorithms and parameters to produce synthetic data that perfectly aligns with their specific business needs and use cases[2]. This level of control extends to the ability to incorporate unique data structures, relationships, and edge cases that might not be adequately addressed by off-the-shelf solutions[3].

For instance, in highly regulated industries or when dealing with proprietary data, the ability to fine-tune the generation process to meet specific compliance standards or replicate very particular data characteristics can be a significant advantage[4]. Furthermore, an internally built platform can be designed for seamless integration with an organization's existing data infrastructure and workflows, potentially streamlining data processing and analysis pipelines.



While the initial investment is substantial, some organizations might anticipate long-term cost savings by avoiding recurring subscription fees associated with third-party platforms.

The primary motivation behind choosing to build a synthetic data generation platform internally often stems from the necessity for highly specialized synthetic data that is perfectly aligned with distinct business requirements and existing infrastructure.

When standard solutions lack the granularity or specific features required, particularly in complex or heavily regulated sectors, the capacity to customize every facet of the data generation becomes paramount[4].

Disadvantages and Challenges of In-House Development
Despite the allure of complete control, building a synthetic data generation platform in-house presents a multitude of significant disadvantages and challenges.

The initial development costs can be substantial, often ranging from tens of thousands to millions of dollars, depending on the development of a user interface, integration with external systems, rigorous testing and validation, and the necessary infrastructure. Moreover, the time to market for an in-house platform can be considerable. While certain tools might facilitate rapid dataset creation, the development of a comprehensive and robust platform from the ground up typically requires several months, if not longer[5].

A critical factor is the need for specialized expertise in areas such as artificial intelligence, machine learning, and data engineering. Finding and retaining professionals with the specific skills required to design, develop, and maintain a sophisticated synthetic data generation platform can be challenging and expensive. Furthermore, the platform will necessitate ongoing maintenance, updates, and support to ensure its continued functionality and effectiveness[6].

Skills required for Synthetic Data Generation		
Computer Vision Experts	3D Modelers	Texture Artists
Machine Learning Innovators	Animators	Rendering Specialists
Algorithm Development Engineers	Procedural Scripting Specialists	Data Wranglers
GPU Performance Engineers	Lighting Experts	

Ensuring the quality, realism, and privacy of the generated data is another significant hurdle. Synthetic data models must be carefully designed and validated to accurately reflect the statistical properties of real-world data without inadvertently revealing sensitive information or replicating existing biases. The potential for replicating biases present in the original data used to train the generative models is a serious concern that requires careful attention and mitigation strategies[7]. The substantial financial investment, protracted development timelines, and the intricate expertise demanded render the prospect of constructing an in-house synthetic data platform a significant undertaking fraught with considerable risks concerning data quality and the potential for bias.

Leveraging Third-Party Synthetic Data Platforms

An alternative approach to building a synthetic data generation platform internally is to leverage the capabilities of third-party platforms. These solutions offer a range of benefits that can make them an attractive option for many organizations.

Benefits of Utilizing Third-Party Platforms

Third-party synthetic data generation platforms are often characterized by their ease of use and quick deployment. Many platforms offer user-friendly interfaces that allow individuals without extensive coding knowledge to generate synthetic datasets. These platforms are also designed for scalability, enabling organizations to handle large volumes of data as their needs grow. Users gain access to advanced features that have been developed and refined by experts in the field[8].

This can significantly reduce development overhead and lead to a faster time to value compared to building a platform from scratch. A key advantage of many third-party platforms is their strong focus on data privacy and security features, often incorporating techniques like differential privacy to ensure the anonymity of the generated data[9]. Organizations also benefit from access to expertise and support provided by the vendor, which can be invaluable for troubleshooting and maximizing the platform's potential.

Utilizing a third-party platform can be more cost-effective than the significant upfront and ongoing costs associated with in-house development. Third-party platforms democratize access to synthetic data by offering intuitive interfaces, sophisticated functionalities, and integrated privacy safeguards, thereby lowering the entry barriers for organizations lacking extensive internal AI/ML proficiency.

Potential Drawbacks of Third-Party Platforms

While third-party synthetic data platforms offer numerous advantages, it is important to acknowledge their potential drawbacks. Depending on the specific needs of an organization, there might be limitations in customization for highly specific or niche requirements. Organizations must also carefully consider data security concerns related to sharing potentially sensitive data with an external vendor, although many platforms offer on-premises deployment options to mitigate this risk[10].

Reliance on an external vendor for ongoing support and updates is another factor to consider, as is the potential for costs associated with licensing and usage fees[5]. The risk of vendor lock-in is also a possibility, where transitioning to a different platform in the future could present complexities and additional expenses. Despite the many benefits, organizations must carefully assess the customization options offered by third-party platforms to guarantee they can meet their unique demands.

Real-World Insights: Case Studies of Build vs. Buy

Examining how organizations have approached the decision of building versus buying a synthetic data generation platform provides valuable real-world context and highlights the practical implications of each strategy.

Organizations That Built In-House: Successes and Setbacks
Shopify, a prominent e-commerce platform, undertook an initiative to generate synthetic transaction data internally to test new functionality within their data platform. This case study demonstrated the ability of an in-house solution to effectively replicate the overall distribution of transaction data, providing a useful "digital twin" for testing purposes. However, the study also revealed limitations in accurately representing smaller data segments and preserving the logical relationships between different data dimensions, such as the association between countries and states[11].

Relari AI developed its own platform for generating synthetic data specifically to benchmark Retrieval-Augmented Generation (RAG) systems[12]. This example illustrates a targeted in-house development focused on a very specific application.

These instances suggest that organizations opt for building in-house when they have very specific testing needs, when synthetic data generation is central to their business model, or when they possess significant internal expertise in the field. While these efforts were successful within specific use cases, they also highlight the difficulty of fully replicating the complexity of real-world data.

Organizations That Chose Third-Party Platforms: Outcomes and Lessons Learned
According to a case study, Nationwide Building Society collaborated with a synthetic data platform provider to test a system for identifying vulnerable customers. By using the platform's synthetic data, Nationwide significantly reduced the testing duration from six months to just three days while minimizing the processing of personal data.

Telefónica, Erste Bank, and Anthem have all utilized the third party synthetic data platforms for diverse applications, including customer data analytics, mobile banking app development using synthetic test data, and fraud detection, showcasing the versatility of third-party solutions.

A healthcare institute used a third-party platform to generate synthetic patient data for research, achieving model accuracy comparable to real data while meeting regulatory standards[13]. In financial services, synthetic data has improved model performance by enabling privacy-safe client data anonymization and enhanced fraud detection[14].

These examples of organizations leveraging third-party platforms highlight the advantages of faster implementation, reduced risks associated with handling real sensitive data, and the ability to achieve tangible business outcomes across various industries, thereby validating the value proposition of readily available solutions.

Key Considerations for Choosing Your Synthetic Data Strategy

One of the most compelling advantages of using a third-party solution is speed. Developing an internal platform can take months of engineering effort, infrastructure setup, and testing before usable data is available. Third-party providers, on the other hand, deliver custom, high-fidelity datasets that are production-ready—dramatically accelerating the AI development cycle.

This faster deployment means teams can spend less time building tooling and more time training, testing, and iterating on models. For projects where time-to-market or real-world validation is critical, that acceleration can be a decisive advantage.

A full cost analysis is still essential. In-house solutions require long-term investment in personnel, infrastructure, and maintenance. Third-party platforms primarily involve licensing and customization fees —but eliminate the significant delays and overhead associated with building from scratch.

The table below provides a general comparison of cost and deployment tradeoffs between the two approaches:

Cost Comparison: Build vs. Buy

Cost Component	Building In-House	Using Third-Party Platform
Initial Development	High (Salaries, Infrastructure, Software)	Low (Subscription/Licensing Fees)
Ongoing Maintenance	High (Updates, Bug Fixes, Infrastructure)	Included in Subscription/Usage Fees
Infrastructure	High (Servers, Storage, GPUs)	Typically Managed by Vendor
Personnel	High (Specialized AI/ML Team)	Lower (Existing Data Science Teams)
Licensing (Internal)	Potential Costs for Specific Libraries	Included in Vendor Fees
Licensing (External)	N/A	Subscription/Usage Fees
Customization	Fully Customizable	May Incur Additional Costs/Limitations
Time to Market	Long	Short

The technical expertise and resources available within the organization are another critical consideration. Building a platform in-house necessitates a strong team with specialized skills in AI, machine learning, and data engineering, whereas using a third-party platform often requires less specialized technical knowledge.

Data sensitivity and privacy requirements will also influence the decision. Organizations handling highly sensitive data might prefer the greater control offered by an in-house solution or need to carefully evaluate the security protocols and deployment options of third-party vendors.

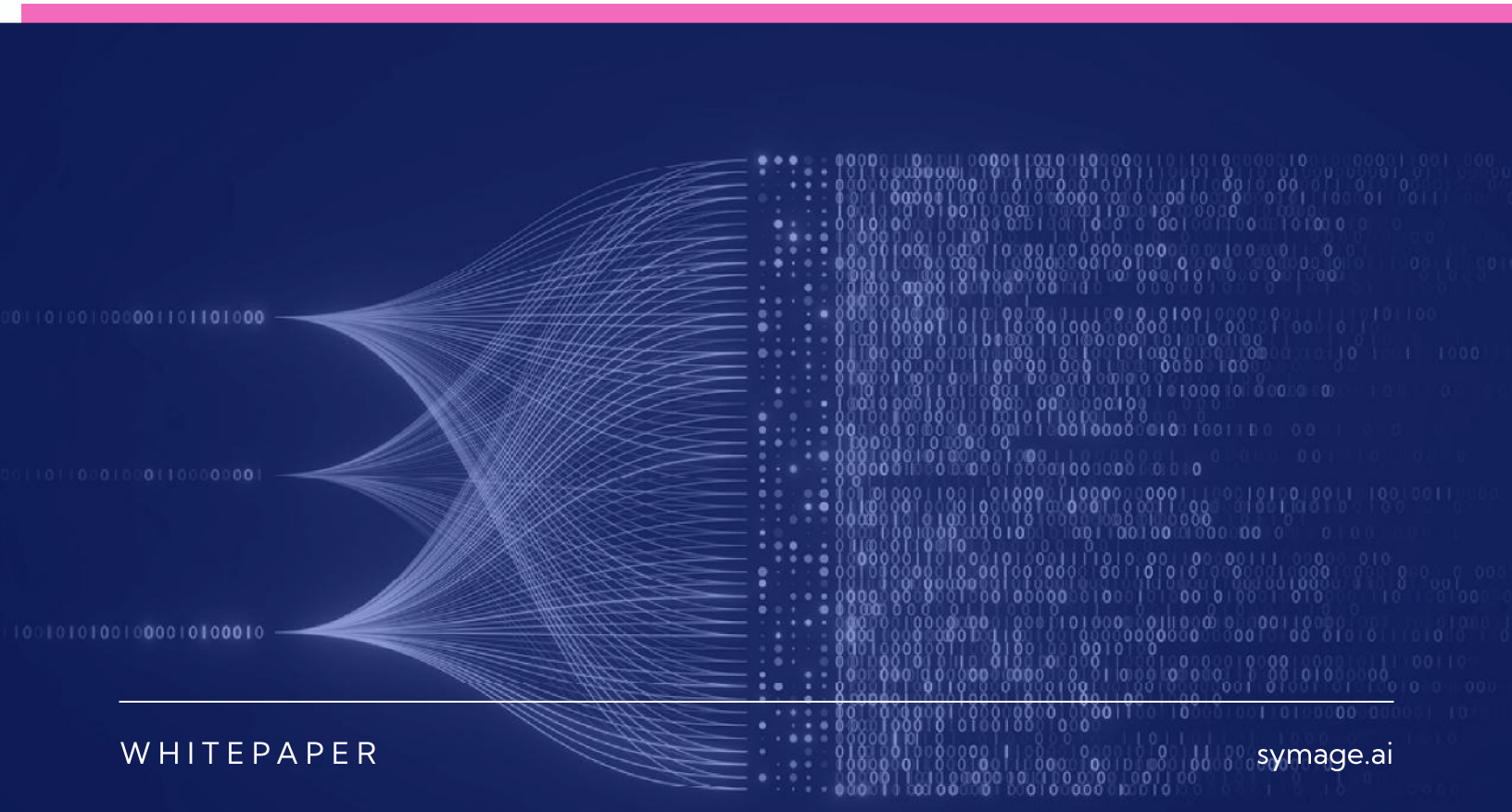
The specific use cases and the required level of customization are paramount. If off-the-shelf solutions can adequately address the organization's needs, a third-party platform might be the more efficient choice. However, highly unique or complex requirements might necessitate the flexibility of an in-house build.

Time to market is a crucial factor, especially for organizations needing to implement a synthetic data solution quickly. Third-party platforms generally offer a much faster deployment timeline.

Scalability requirements for future data needs should also be considered, ensuring that the chosen solution can adapt to growing demands.

Finally, the ease of integration with existing systems and workflows can significantly impact the efficiency of the overall data pipeline.

The decision ultimately hinges on a comprehensive assessment of an organization's financial resources, technical capabilities, data governance policies, specific application needs, and strategic priorities.



Symage: Empowering AI with High-Quality Synthetic Image Data

Symage and SymageDocs, developed by Geisel Software, are two specialized platforms designed to solve the data challenges AI teams face in both vision and document-based applications. Backed by decades of expertise in AI, machine learning, computer vision, and robotics, these platforms deliver limitless, high-fidelity custom synthetic data to accelerate model training—without the limitations of real-world data.

Symage: 3D photorealistic synthetic image data for computer vision
Symage generates 3D, physics-based environments with pixel-perfect automated intelligent labeling, enabling training on rare, complex, or hard-to-capture scenarios. It's ideal for industries like robotics, healthcare, and logistics where real-world data is costly, sensitive, or difficult to obtain.

Key benefits:
- 3D photorealistic, bias-free image data
- Pixel-perfect intelligent labeling with no manual effort
- Simulates edge cases and extreme conditions
- Scalable, cost-effective alternative to field data collection

SymageDocs: Structured synthetic data for documents, forms, records, and IDs
SymageDocs creates highly realistic, privacy-safe synthetic documents to support document AI, OCR, and NLP models. From invoices and ID cards to complex medical or legal forms, it replicates the variability and structure of real-world data—without exposing sensitive information.

Key benefits:
Realistic formats across a wide range of document types
Built-in privacy—no real personal identifying information
Supports rare layouts and edge cases
Streamlines training for OCR, parsing, and classification

Why Choose Symage or SymageDocs as Your Third-Party Solution?
Choosing Symage or SymageDocs as your synthetic data partner offers several compelling benefits. It enables faster AI development by significantly reducing the time spent on collecting, cleaning, and labeling real-world data. The platform generates customizable datasets to address gaps in real-world data availability and to simulate a wide range of real-world and edge-case scenarios, leading to higher accuracy in trained AI models

Together, Symage and SymageDocs give AI teams the data they need—when they need it—without the constraints of real-world collection.

Conclusion: Navigating the Path to Effective Synthetic Data

The decision of whether to build a synthetic data generation platform in-house or to utilize a third-party solution is a strategic one that carries significant implications for an organization's AI initiatives. Building in-house offers the advantage of complete customization and control but comes with substantial costs, long development timelines, and the need for specialized expertise.

Conversely, third-party platforms provide ease of use, scalability, access to advanced features, and often a faster time to value, though they might present limitations in customization and raise data security considerations.

Ultimately, the optimal strategy hinges on a careful assessment of an organization's unique needs, technical capabilities, budget constraints, and strategic goals. For organizations focused on image data-intensive applications, third-party solutions like Symage offer a compelling value proposition. Symage's specialized focus on generating high-quality, photorealistic synthetic image data with automated labeling addresses the core challenges of data scarcity, bias, and the time-consuming nature of real-world data collection in this domain.

The future of artificial intelligence is increasingly intertwined with the power of synthetic data. Predictions indicate that a significant portion of data used in AI and analytics projects will be synthetically generated in the coming years.

As organizations navigate this evolving landscape, understanding the strengths and weaknesses of different synthetic data generation strategies will be crucial for unlocking the full potential of AI. Organizations are encouraged to explore how specialized third-party platforms like Symage can accelerate their AI initiatives and drive innovation in their respective industries.

References

1. Synthetic Data Generation: Definition, Types, Techniques, & Tools - Turing, accessed March 25, 2025, https://www.turing.com/kb/synthetic-data-generation-techniques
2. Should You Build vs. Buy Generative AI? The Pros and Cons - MOHARA Insights, accessed March 25, 2025, https://mohara.co/should-you-build-vs-buy-generative-ai-the-pros-and-cons/
3. Build or buy AI: Rethinking the conventional wisdom - Algorithma, accessed March 25, 2025, https://www.algorithma.se/our-latest-thinking/build-or-buy-ai-rethinking-the-conventional-wisdom
4. Buy or build: Key considerations for investing in AI/ML - Slalom Consulting, accessed March 25, 2025, https://www.slalom.com/us/en/insights/buy-or-build
5. AI Development Cost: A Comprehensive Overview for 2025 - UpsilonIT, accessed March 25, 2025, https://www.upsilonit.com/blog/how-much-does-it-cost-to-build-an-ai-solution
6. How Much does it Cost to Build a Generative AI in 2025 : Aalpha, accessed March 25, 2025, https://www.aalpha.net/blog/cost-to-build-a-generative-ai/
7. What is synthetic data? Types, challenges, and benefits - Sigma AI, accessed March 25, 2025, https://sigma.ai/synthetic-data/
8. How to Choose the Right Synthetic Data Company - K2view, accessed March 25, 2025, https://www.k2view.com/blog/synthetic-data-companies/
9. Synthetic Data: Everything You Need to Know - Splunk, accessed March 25, 2025, https://www.splunk.com/en_us/blog/learn/synthetic-data.html
10. Synthetic data generation: Building trust by ensuring privacy and quality | IBM, accessed March 25, 2025, https://www.ibm.com/products/blog/synthetic-data-generation-building-trust-by-ensuring-privacy-and-quality
11. Synthetic Data in Practice: A Shopify Case Study - Towards Data Science, accessed March 25, 2025, https://towardsdatascience.com/synthetic-data-in-practice-a-shopify-case-study-79b0af024880/
12. Case Study: Using Synthetic Data to Benchmark RAG Systems | by Yi Zhang | Relari Blog, accessed March 25, 2025, https://blog.relari.ai/case-study-using-synthetic-data-to-benchmark-rag-systems-be324904ace1
13. Real-World Case Studies of Synthetic Data Generation with 'Gretel'(PART-4) - Medium, accessed March 25, 2025, https://medium.com/@mauryaanoop3/real-world-case-studies-of-synthetic-data-generation-with-gretel-part-4-3ac22d534629
14. Use Cases of Synthetic Data and Generative AI in Data Security - ScikIQ, accessed March 25, 2025, https://scikiq.com/blog/use-cases-of-synthetic-data-and-generative-ai-in-data-security/

Let's Talk About Your Synthetic Data Strategy

Choosing whether to build or buy your synthetic data generation platform is a pivotal decision—one that can impact your AI initiatives for years to come. While both paths offer distinct advantages, the right choice ultimately depends on your organization's technical requirements, resource availability, compliance needs, and long-term goals.

If your team is looking for a faster, more scalable way to generate high-quality synthetic image data, Symage is engineered to deliver. With pixel-perfect intelligent labeling, photorealistic simulation, and limitless custom data generation, Symage helps organizations accelerate model training and reduce reliance on real-world data—all while ensuring precision and privacy.

We'd love to learn more about your challenges and show you how Symage can support your AI development efforts.

Contact the Symage Team

Email: realresults@symage.ai

Website: www.symage.ai

JR Rodrigues: (508) 936-5097