

Democratizing Data Access With Synthetic Data

How synthetic data can power AI models and algorithms



Table of Contents

1	Executive summary	03
2	Introduction	04
3	Synthetic data – the enabler	05
4	Four applications of synthetic data	07
5	Synthetic data generation – methods and approaches	09
6	Overcoming challenges in synthetic data generation	11
7	Addressing enterprise data challenges with Kingfisher	12
8	Conclusion	14

Executive summary

In today's AI-powered world, data accessibility is an essential ingredient for successful AI implementation. An effective AI system is largely dependent on the following three pillars, namely:

- AI models or algorithms
- Computing or processing power
- Data access



Source

Among these three pillars, high-quality data is the only pillar that continues to pose issues for enterprises building AI systems for specific use cases. The current data landscape is limited by the following four constraints:

Data quality

The effectiveness of any AI model is only as good as the data being fed into the system. **85% of AI projects** fail primarily because of poor data quality. Underperforming AI models – built on low-quality data – can cost companies **6%** of their annual revenues. To ensure optimum data quality, enterprises must check the data for completeness, duplication, missing values, and corruption.

Data labelling

Through data labeling, enterprises can train AI models to provide accurate results or outcomes over time. Nonetheless, AI systems don't perform optimally in real-world scenarios without being trained on real-time data, and labeled data can thus produce unanticipated outcomes.

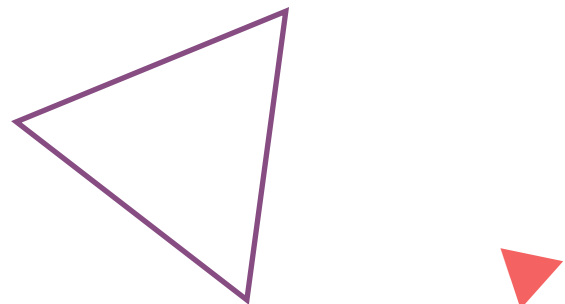
Data bias

AI systems can produce biased outcomes when trained on limited data from selected sources which don't represent the entire data ecosystem, thus producing an incomplete "picture."

Data quantity

Besides quality, AI models also need massive data volumes to produce accurate results. This is achievable only through a centralized repository, which can capture and store the right data for AI models.

Data privacy is another challenge for data democratization in AI-dependent enterprises. Can synthetic data overcome these limitations and enable data democratization? Let's discuss how this can be done in this whitepaper.





Introduction

Data privacy is a growing demand in modern enterprises. As AI-powered systems access more complex datasets, there's a corresponding increase in their ability to process sensitive data. **80% of consumers** are concerned about data privacy, but 55% believe that data privacy is no longer possible.

To ensure data privacy for AI models, **research studies** are also exploring the feasibility of privacy-preserving data sharing for AI projects. Besides privacy, data availability is another pressing challenge for AI-implementing enterprises.

Traditional data handling methods require enterprises to collect, label, train, and maintain massive datasets to ensure data availability. This is both an expensive and time-consuming process. Additionally, enterprises cannot access datasets used for rare real-life scenarios (for example, a severe stock market crash).

Compared to large corporations like Google or Apple, small-to-medium enterprises find it even more challenging and expensive to access proprietary data (due to contractual agreements). Besides, the lack of a common data standard makes it unfeasible to share real-world data.

Four challenges in data accessibility

Data accessibility is the engine that powers the democratization process. However, with the growing number of data sources, enterprises face a host of accessibility challenges in the form of:

Data privacy regulations and compliance

Compliance laws in highly regulated industries can hinder data accessibility. To comply with regulations like GDPR and HIPAA, companies need to invest more time and effort in processing sensitive data securely – with the right security policy. These regulations can also necessitate the use of additional security settings, thus consuming more time in data engineering.

Cost and complexity of data collection

As datasets grow more complex, enterprises are allocating more budget for data collection and integration. Cloud storage can reduce the technical barriers to data exploration. However, it's no longer feasible to "throw" complex data to employees (without any organizational support). Without the democratization process, data-related investments can only produce inaccurate outcomes or insights.

Data bias and imbalances

Besides the complexity factor, data accessibility is impacted by "biased" data collection. For instance, biased datasets can produce discriminatory outcomes from AI systems and algorithms. Under-represented classes like racial minorities or "rare case" scenarios (for example, diagnosis of rare diseases) can contribute to data imbalance.

Data scarcity

In regulated sectors like healthcare and finance, data scarcity limits AI capabilities such as predictive analytics. For example, financial AI models need access to sensitive data for fraud detection, but GDPR restricts data sharing, creating scarcity.

How does synthetic data improve data accessibility? Let's discuss this in the next section.

Synthetic data – the enabler

Is synthetic data the answer to seamless data accessibility? Simply explained, synthetic data is essentially artificial data that can mimic real-world data. Using a variety of techniques, synthetic data can have similar properties, characteristics, and structure as real-world data.

Synthetic data is measurable against the following 3 attributes, namely:

- Fidelity – or how similar the synthetic data is to the original dataset.
- Utility – or how useful synthetic data is for a specific use case.
- Privacy – or has any real-world sensitive data been used for the synthetic data?

Synthetic data can effectively mimic the real world without exposing any sensitive data, thus ensuring data privacy and availability. This characteristic enables highly regulated industries like healthcare and pharmaceuticals to use synthetic data to train their AI models (without violating any regulations).

For example, a pharma company can train its AI models using "artificial" patient data, without compromising the actual patient records.

Among other benefits, synthetic data is cost-effective and faster to generate than real data, which goes through an extensive collection and labeling process. Further, synthetic data is complete – without any missing values or outliers, that may be the case with real-world data. Synthetic data generators can fill in the missing gaps in datasets, thus providing complete data for AI models to produce accurate outcomes.

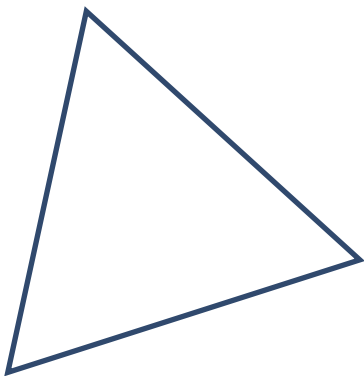
How does synthetic data eliminate bias? Due to its artificial mode of generation, synthetic data generators use diverse and representative datasets, which can handle issues like underrepresentation or limited data. With data generation techniques like Generative Adversarial Networks (GANs), synthetic data provides diverse and balanced datasets with varied data points.



Here's a comparison of real vs synthetic data:

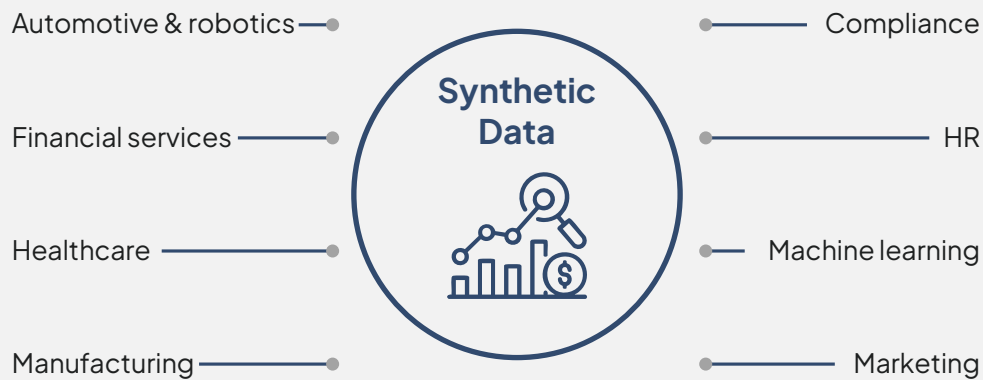
Four challenges in data accessibility

	Synthetic data	Real-world data
Definition	Artificial data used to simulate real-world data patterns	Data collected from real-world applications and systems
Data source	AI algorithms, simulations, or real-world data patterns	Customer interactions, application data, or real-world transactions
Data privacy	High – without any exposure to sensitive data	Low – because of sensitive data
Reidentification risk	Low – as it does not contain real user's information	High – as it may contain personally identifiable information (PII)
Cost	Low – as it can be programmatically generated from real data or code	High – as it requires collecting and labeling data
Scalability	Easier to scale due to massive volumes of data	Limited – depending on data accessibility and availability



Four applications of synthetic data

Synthetic Data Applications



Source

The applications of synthetic data are vast. Depending on the industry and specific use case, its role changes. One thing remains constant: synthetic data often fits as the missing piece in the complex puzzle these industries are trying to solve. Here are four industry-wise examples where synthetic data steps in to bridge critical gaps.

Healthcare

Healthcare providers can generate and feed synthetic patient data into AI-generated models for medical research and training. This allows them to analyze the patient records without violating any data privacy regulations.

In the R&D domain, healthcare companies can access large-scale datasets to:

- Personalize patient treatment.
- Improve drug discovery and development.
- Make accurate predictions about emerging

healthcare trends.

Similarly, AI-powered systems can utilize artificial image data (for example, MRI scans) for accurate diagnosis and surgical interventions.

Financial services

In the financial services domain, AI-based financial models (trained on synthetic data) can improve fraud detection with access to diverse datasets. With synthetic data, fraud detection algorithms can access sensitive data with its privacy-enabled solution.

Synthetic data-powered AI models can also produce accurate outcomes in the following areas:

- Credit scoring
- Anti-money laundering activities
- Regulatory compliance



Software development

Synthetic datasets are a source of diverse augmented data, which can fuel the future development of AI technologies. Effectively, synthetic data is best suited for training and testing powerful AI models and validating advanced algorithms.

For instance, data scientists can utilize synthetic data to generate high-quality datasets, which can further improve the performance of their data models. Software testers and QA teams can use synthetic data to test complex applications for performance and scalability.

Defense and public sector industries

Synthetic data can also transform the capabilities

of AI systems in the military and defense industry. Here are some of its applications:

- Identifying and classifying military terrains, weapons, people, and military equipment.
- Simulating sensor models with accuracy and precision.
- Detecting variability elements in the form of light and weather.

In the public sector, synthetic data can also boost innovation, while protecting public information. It can encourage government agencies to adopt AI technology for making important decisions and strategies.



Synthetic data generation – methods and approaches

Synthetic data generation is not a modern-day invention, but has existed for long, through various techniques like randomization and sampling. Modern techniques now use advanced machine learning algorithms to generate synthetic data.

Here are the two approaches used for synthetic data generation:

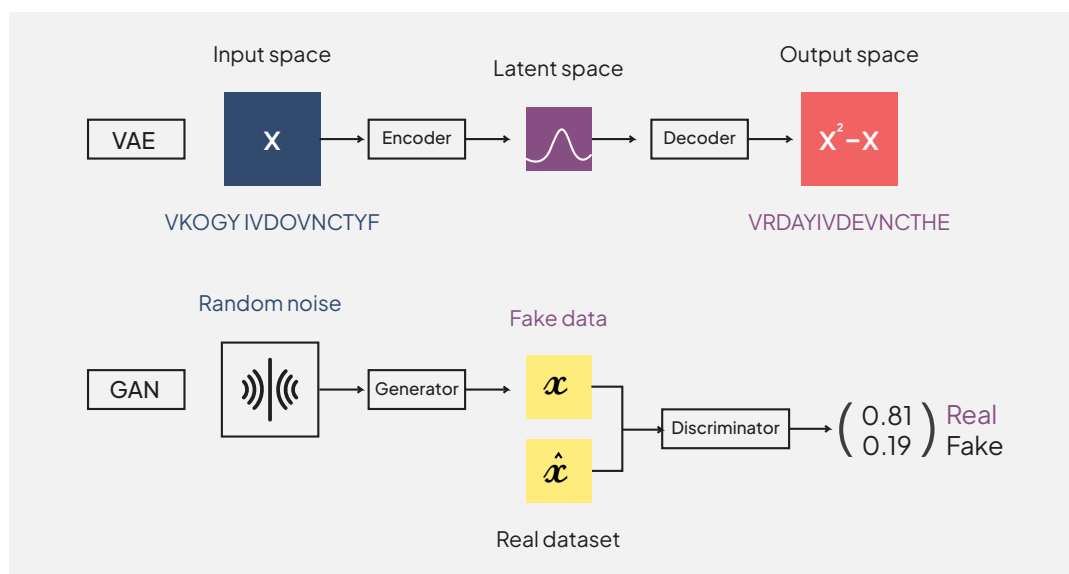
Rule-based approach

This method generates synthetic data based on

predefined rules. This approach is effective in creating synthetic data that follows a specific pattern or logic. Common use cases include software testing and real-life scenario planning.

Model-based approach

This approach generates synthetic data by utilizing statistical models trained on real data. It is ideal for creating large and diverse datasets used in AI model training. The model-based approach can generate realistic data while maintaining privacy.



Image

Here are some of the advanced techniques used in synthetic data generation:

Generative adversarial networks (GANs)

GANs use a machine learning algorithm with separate neural networks as follows:

- Generator is used to synthesize data from a selected dataset – and tries to replicate the

statistical distribution and patterns from the original data.

- Discriminator compares the real and synthetic datasets for differences – and if necessary, notifies the generator to make changes and distinguish the synthetic data from its real counterpart.



Variational Autoencoders (VAEs)

VAEs also deploy neural networks to encode real-world data into a compact format, followed by decoding it to produce similar data instances. This method is suited for encoding complex data representations.

Additionally, synthetic data generators can use the following techniques to ensure data privacy:

- **Differential privacy**

The differential privacy technique produces a synthetic dataset similar to a real dataset – with the same database schema and properties. However, it delivers data privacy by protecting individual data points for AI-powered models learning data patterns and distributions.

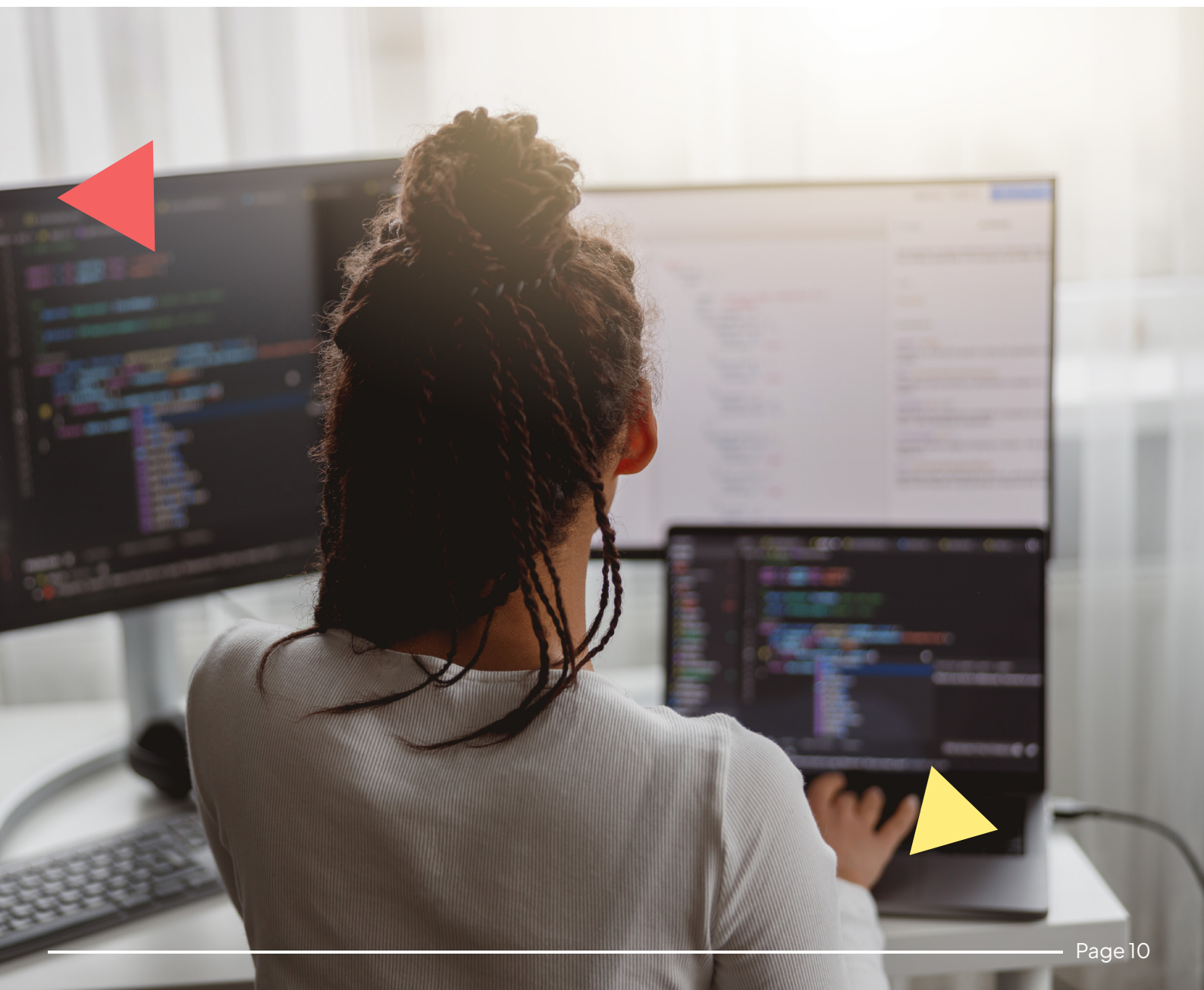
- **Data anonymization**

Data anonymization is the process of creating

fake data under certain circumstances. The generated data looks like the original data – but without any identifiable information.

How can enterprises authenticate the quality of the generated synthetic data? By using evaluation frameworks, which track the following metrics to authenticate the data:

- Fidelity metrics are used to ensure that the synthetic data retains the essential properties or characteristics of the original data.
- Utility metrics focus on the performance of synthetic data – through generalization and ranking of the trained AI models.
- Privacy compliance metrics are used to ensure that synthetic data models do not replicate the real data points.





Overcoming challenges in synthetic data generation

Despite its utility and benefits, enterprises can face challenges when generating synthetic data for their AI projects. Here are 3 challenges to address:

Data utility

Synthetic data can suffer from various shortcomings that can restrict its usability for real-world applications. Here are some of its limitations:

- The incongruity between the real and synthetic datasets with disparities in data distribution, feature distribution, and other statistical properties.
- Incomplete data in synthetic datasets due to failure of encapsulating changes in the authentic dataset during data generation.
- Data inaccuracy due to algorithmic errors and noise injection.

The solution is to update synthetic datasets for new scenarios and identify changes in data distribution to maintain model performance.

Ethical and security concerns

The widespread use of synthetic data can raise ethical concerns about the role of AI systems in creating fictional content. This can lead to social concerns about the spread of misinformation and disinformation.

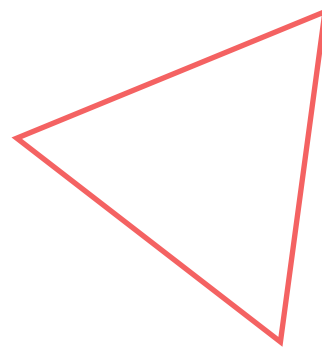
An ethical AI framework using synthetic data is required based on the principles of responsibility, privacy, fairness, and non-maleficence.

Standardization and trust building

Currently, synthetic data generation lacks standardization and guidelines. To address this problem, a standardized framework is required to develop the best practices for data generation. This must cover aspects like:

- Selecting the data generation models.
- Configuring parameter settings.
- Correlating between real and synthetic data.

Synthetic data, like any transformative technology, comes with its own set of challenges, but its long-term value depends on how effectively organizations address them. With this understanding, Onix has applied years of hands-on experience and research to design **Kingfisher**: A solution purpose-built to meet enterprise needs with security, scalability, and reliability at its core.



Addressing enterprise data challenges with Kingfisher

Synthetic data has shown promise in overcoming many of the structural limitations around data accessibility and privacy. However, enterprise adoption depends not just on theoretical utility, but on how well a solution fits within the existing constraints, systems, and compliance frameworks. **Kingfisher**, developed by Onix, is a synthetic data generator built with this context in mind.

It supports both data-driven and logic-driven generation techniques, helping organizations adapt to different types of data gaps, whether due to privacy restrictions, unavailable real-world samples, or early-stage development.

Dual Approach to Data Generation

Kingfisher offers two complementary modes of synthetic data generation, enabling it to fit into a wide variety of enterprise use cases:

Mode	Description	When It's Useful
Data from Data	Learns statistical properties, schema, and patterns from real datasets.	When anonymized access to representative data is available.
Data from Code	Synthesizes data using metadata, schema definitions, and logic, no data needed.	When no sample data can be accessed, or in early dev/test cycles.

By offering both approaches, Kingfisher can support projects across different maturity levels and data sensitivity thresholds, from legacy system modernization to AI model training.

Capabilities Aligned to Enterprise Needs

Rather than focusing on synthetic data as a generic concept, Kingfisher addresses the practical aspects enterprises often face when trying to implement data-driven workflows in secure or regulated environments.

Kingfisher can be seamlessly deployed within customer environments, ensuring data never leaves their secure perimeter. Its zero-code interface empowers users across functions to generate synthetic data without relying on engineering teams.





Capability	Enterprise Need Addressed
Schema-aware synthesis	Maintains structure, types, keys, and constraints for downstream compatibility.
Constraint-preserving logic	Ensures referential integrity and valid inter-table relationships.
Zero-code Interface	Generate synthetic data easily through a user-friendly interface, no coding required.
Scalability across volumes and systems	Suitable for large datasets and complex cloud-native architectures.
Privacy-first generation	Avoids PII/PHI use altogether, reducing risk and easing compliance efforts.

These features are particularly relevant in sectors like financial services, healthcare, and telecom, where data regulations, legacy architectures, and growing AI adoption intersect.

Common Use Cases

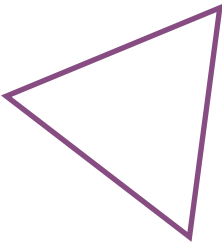
Kingfisher’s implementation across various environments shows how different types of teams benefit from targeted synthetic data generation.

- **Data Science & AI Teams**– Train models using diverse, realistic inputs, even when actual data is limited, biased, or restricted.
- **DevOps & QA**– Enable test environments with production-like datasets that reflect real scenarios without exposing sensitive records.
- **Cloud Migration & Platform Engineering**– Simulate workloads, validate pipelines, or stress-test new architectures before ingesting actual data, especially useful in greenfield or staging environments.
- **Compliance and Risk Management**– Replace sensitive datasets in regulatory sandboxes or cross-border environments with synthetic equivalents to reduce risk exposure.

Supporting a Broader Data Strategy

According to a [Gartner](#) projection, synthetic data will replace real data in AI-powered models by 2030. [Forbes](#) has also ranked synthetic data among the “top 5 trends in data science.” As synthetic data continues to evolve, organizations require tools that go beyond generation and contribute to a responsible, scalable, and compliant data strategy. Kingfisher is designed to be part of that broader transformation, where data usability and data responsibility can co-exist.

By aligning with enterprise architecture, DevSecOps practices, and data governance frameworks, it helps teams adopt synthetic data in a way that’s sustainable, not just technically, but operationally and ethically as well.






Conclusion

Synthetic data is not simply another tool – but a game changer in providing secure and scalable datasets to data-dependent enterprises. This technology can power the next phase of data-driven decision-making by delivering faster insights that would otherwise require extensive preparation.

At Onix, our proprietary synthetic data generation tool, **Kingfisher** is designed to generate synthetic data from code. Here's a **case study** of how a global bank reduced its data migration time by 85% by using Kingfisher.

Learn more about our Kingfisher tool. **Get in touch** with us.



 onixnet.com
 connect@onixnet.com
 800.664.9638

Get in touch

Follow us:



Copyright © 2025 Onix . All Rights Reserved.