Synthetic data – the prerequisite for AI agents in banking and financial services industry

The growing importance of synthetic data generators in BFSI in 2025

Table of Contents

Introduction

According to technology analysts and media professionals, there's a growing consensus that the year 2025 is emerging as the "Year of AI agents". This year, AI agents have successfully transitioned from an "experimental" phase to a "core component of enterprise-level strategy." Agentic AI systems are increasingly automating complex tasks, transforming business workflows, and accelerating business transformation.

This transformational shift is also evident in the banking and financial services industry (BFSI). Business leaders in BFSI are adopting Agentic AI for hyper-personalization, automation, and better productivity.

This eBook explores the growing role of AI-powered agents in the BFSI sector – and why synthetic data is a prerequisite for agents.

Role of AI in BFSI

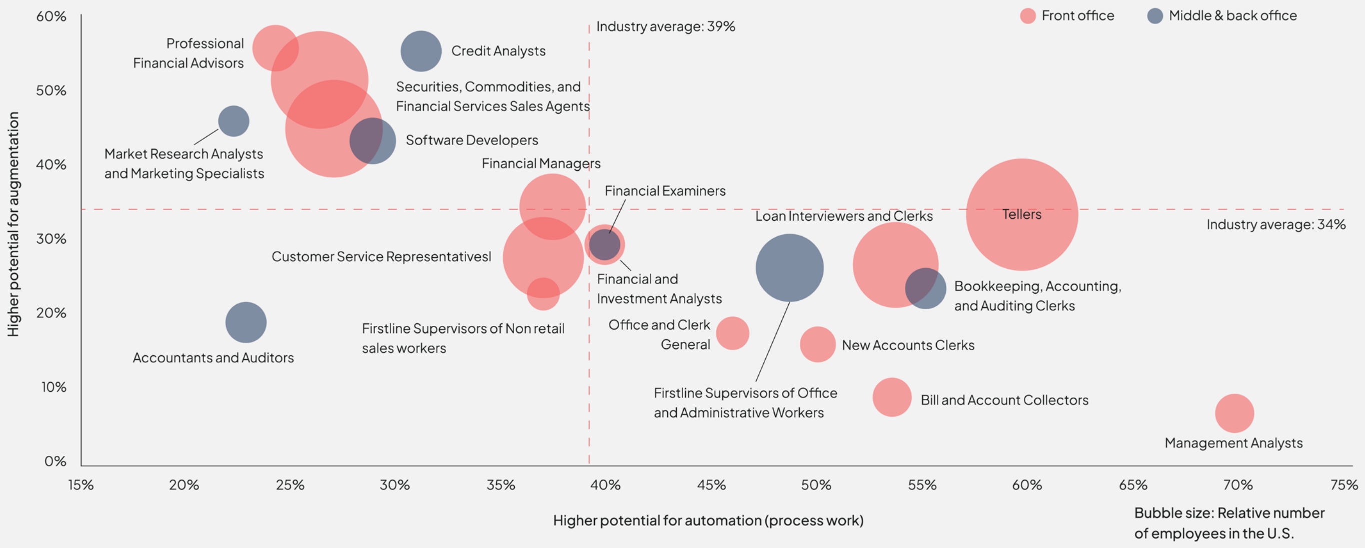
AI technology has a wide variety of use cases in the BFSI industry. AI-powered technologies like computer vision, natural language processing (NLP), and machine learning are driving innovations in areas like customer service, back-office operations, and risk management.

Banks and insurance companies are leveraging AI-enabled solutions like fraud detection, chatbots, and data analytics to deliver benefits like:

- Improved operational efficiency
- Better customer experience
- Data security and privacy

To put things in perspective, Generative AI is not only transforming the BFSI industry, it is reshaping multiple roles across its value chain. This layered impact makes it clear that the inception of AI in BFSI isn't a distant future; it's happening now. Adding to that, Generative AI is transforming BFSI roles as depicted in the following image.

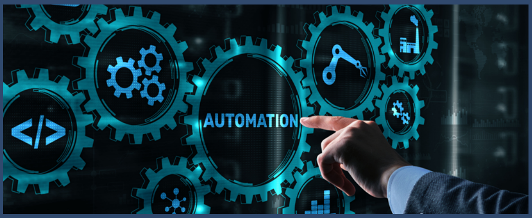
How the top 20 banking industry roles are likely to benefit from generative AI.



Source: Accenture Research analysis and estimates on BLS and O'Net data as of december 2022

The surge of AI agents in BFSI – Core drivers

What's driving the surge of **AI agents in the financial** sector? Here are some of the driving factors:



1. "Round-the-clock" automation

In the BFSI domain, AI agents are handling complex multi-step processes around the clock. This level of automation is eliminating human fatigue, reducing manual errors, and significantly cutting down business costs and time.

2. Productivity boost

Early adopters of Agentic AI solutions have seen a dramatic jump in their productivity levels. For instance, AI agents in the sales function have led to a 40% improvement in team productivity by automating lead generation and follow-ups.

3. Personalization

Personalization, hyper-personalization, and beyond these are no longer just buzzwords. AI agents in the BFSI sector can analyze massive volumes of datasets. This can help BFSI companies customize their marketing strategies, personalize product recommendations, and provide optimal customer support.

4. Boost to human potential

Agentic AI in the banking domain can free up human resources from handling customer queries and complex tasks. Employees can focus their energies on building business strategy and critical thinking, thus boosting their human potential.

5. Rapid market growth

The global market for AI agents was valued at $5.4 billion in 2024, and is growing at an annual rate of 46%. This massive growth is being fueled by market demands for scalable productivity and real-time personalization.

6. Shift in AI mindset

BFSI companies no longer view AI-powered agents as business tools but more as digital co-workers. They're adopting AI agents to augment human roles, boost decision-making, and transform into future-ready enterprises.

What's hampering the adoption of AI agents in the BFSI domain? Let's discuss some of the key challenges in the next section.

Challenges facing AI agent deployment in BFSI

Even as more BFSI companies realize the value of Agentic AI, it faces a host of regulatory challenges and data privacy issues. Here's a deep dive into the common AI-related barriers in the BFSI domain:

Data privacy and regulatory compliance

The BFSI industry collects and manages massive volumes of sensitive customer data, which must be handled efficiently by AI agents. 39% of banks cite data privacy as the biggest hurdle to AI adoption. Data mismanagement can seriously compromise regulations such as GDPR, which can attract heavy fines and penalties.

Financial companies are justified in being concerned about AI agents having unchecked access to confidential customer data, which can violate compliance-related rules. Besides GDPR in Europe, there are region-specific regulations like CCPA in California and PDPA in Asia that must also be strictly followed by BFSI companies.

Here's a bird's-eye view of the major compliance frameworks across the globe and their purpose.

Data Govern. Act
Data Act
GDPR
AI Act
Cyber Security Act
Digital Markets Act
Database Directive
E-privacy Regulation

Compliance framework	Jurisdiction	Purpose
DORA (Digital Operational Resilience Act)	European Union (EU)	To ensure ICT risk management, resilience, and continuity in financial entities.
GLBA (Gramm-Leach-Bliley Act)	U.S.	To protect consumers' financial privacy and to secure sensitive financial data.
FINRA Rules (Rule 3110, 4511)	U.S. (SCRO)	To maintain supervision and to retain records for brokers-dealers.
Basel III / BCBS 239	Global (Basel Committee)	To strengthen capital requirements and risk data aggregation.
MAS TRM Guidelines	Singapore	To provide technology risk management for financial institutions.
RBI Guidelines on Digital Lending & Cybersecurity	India	To safeguard customer data and ICT security in digital lending.
CCAR (Comprehensive Capital Analysis and Review)	U.S. (Federal Reserve)	To ensure capital adequacy under stressful conditions.
GDPR for financial institutions	E.U.	To protect the personal data and privacy of EU citizens.

Data and infrastructural challenges

BFSI companies also face a host of data and infrastructure-related challenges that hamper the deployment of AI agents in their processes. This includes the following:

1. Data fragmentation across legacy systems

BFSI companies continue to operate on legacy infrastructure and store their financial data on multiple outdated systems. This leads to data fragmentation, which makes it challenging to build a single, structured, and usable dataset for AI agents.

2. Limited data sharing

With the presence of data silos and strict access control, BFSI teams are unable to share data and collaborate with other teams. This can impact the development of cross-functional AI models and their training process.

3. Stringent regulations

As highlighted in the previous section, the BFSI sector faces a host of industry regulations like GDPR and DORA. Hence, financial companies are unable to modernize their legacy infrastructure on a cloud-powered AI infrastructure. This can limit scalability and experimentation with advanced AI models.

4. Data bias and gaps

Even when AI agents have access to real-world data, many datasets are inaccurate, incomplete, or biased. AI agents that are trained on these datasets can produce inaccurate outputs in areas like fraud detection and credit scoring.

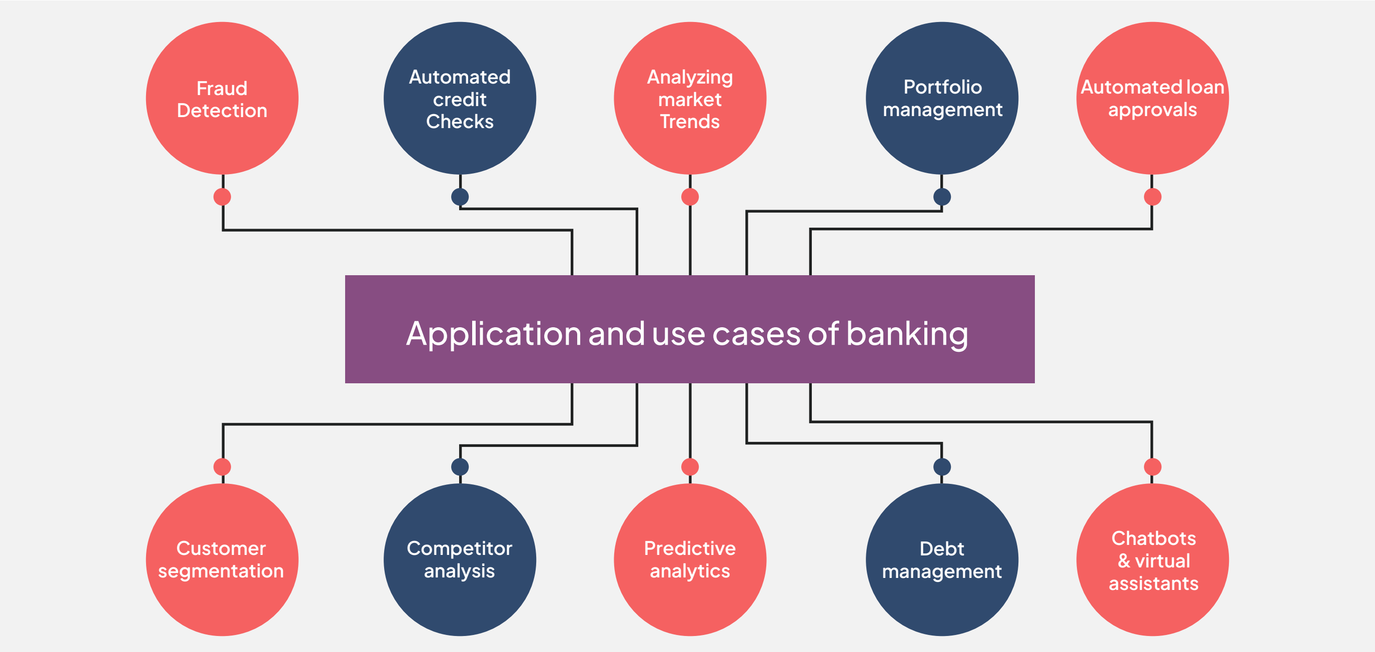
5. Internal constraints

In addition to external industry regulations, BFSI companies can also face internal constraints due to strict policies on the use of sensitive financial data. This can restrict the level of AI development, thus resulting in slower deployment and innovation.

To address these BFSI-specific challenges, **synthetic data solutions** are increasingly being adopted by both banks and financial institutions. In the following sections, let's explore why synthetic data is being regarded as a "prerequisite" for AI agents.

How synthetic data addresses data privacy concerns in BFSI

Synthetic data essentially mimics the statistical properties of real-world datasets, without including any real personal information. Hence, it can be freely used by BFSI firms without violating any privacy regulations. This is why more companies are turning to enterprise-grade synthetic data to train their AI agents.



By deploying high-quality synthetic data generation tools, BFSI firms can generate synthetic data that preserves the analytical integrity of real-world data, while anonymizing its sensitive information. Companies can also share synthetic data with other vendors and cloud providers, without worrying about any compliance breach. As an example, synthetic data on financial transactions can include real-world patterns and outliers, without mapping to the actual customer making the transaction.

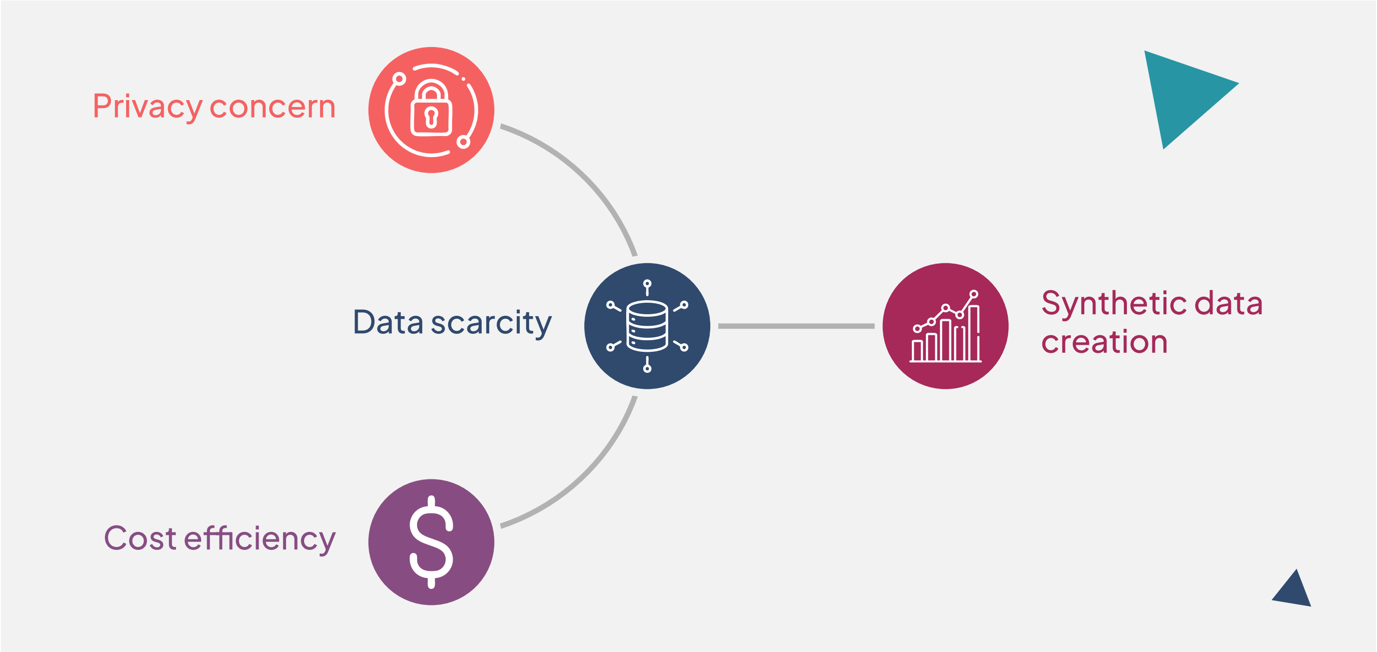
To be effective, synthetic data must fulfill the following requirements:

- Privacy-compliant
- Domain-specific
- Statistically representative

This is only possible by using advanced generative models that can reproduce real-world behaviors without exposing any personal identity. Niraj Kumar, CTO of Onix, explains, “As AI adoption accelerates, relying solely on real-world data is no longer viable. Synthetic data is now essential for secure, scalable, and unbiased model training, especially in regulated industries like finance.”

How synthetic data addresses data scarcity and quality

In addition to data privacy, synthetic data addresses the operational barrier of data scarcity in training AI models. To be effective, AI agents need a continuous flow of diverse high-quality data. In reality, data in the BFSI domain is often limited, fragmented, or biased.



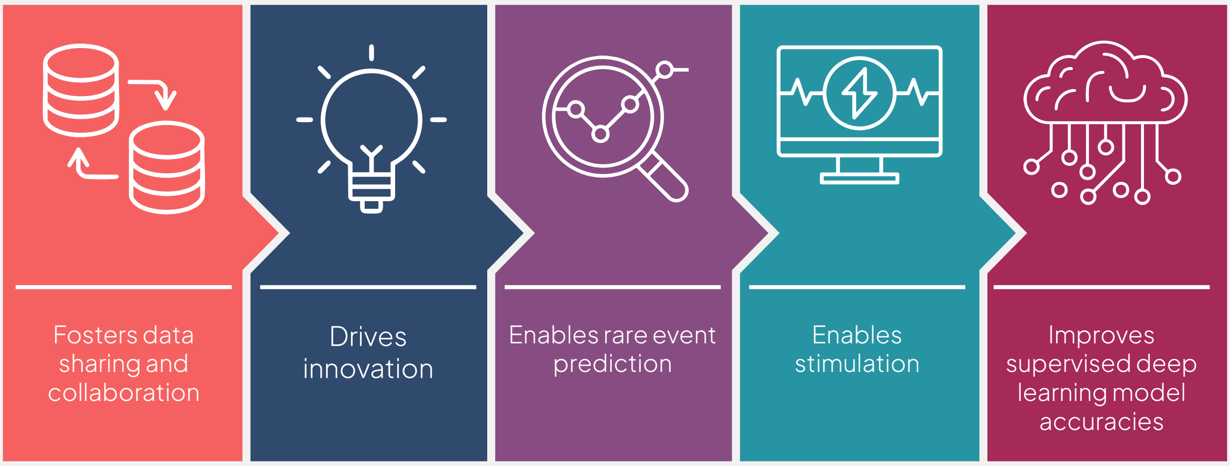
Synthetic data generation tools can generate massive volumes of synthetic data, which can also be applied to rare-world scenarios. For instance, an AI-powered fraud detection model can be trained on specific fraud patterns (from synthetic data) without waiting for any real-world event. Similarly, synthetic data can mimic historical data that lacks balance across demographics or geographical conditions to address any data gap and improve model accuracy and fairness.

BFSI companies are also using synthetic data to simulate "what-if" scenarios – like stock market crashes or new fraud tactics – to test their AI systems. Essentially, synthetic data eliminates the constraints of real-world data and privacy laws, thus enabling a safe and scalable AI system.

Next, let's look at some of the use cases of synthetic data across financial applications.

Four use cases of synthetic data in core financial applications

Common applications of synthetic data in finance



Fosters data sharing and collaboration

Drives innovation

Enables rare event prediction

Enables stimulation

Improves supervised deep learning model accuracies

Synthetic data has effectively proved to be a “game-changer” across BFSI applications. Financial institutions can also transform AI models in ways that were impractical before the emergence of synthetic data. Let’s look at 4 core BFSI applications that are most impacted by synthetic data.

Fraud detection and financial crime prevention

Financial fraud and crimes are constantly evolving in the BFSI domain. The primary challenge for AI agents (or models) in fraud detection is the rare occurrences of fraudulent transactions and sensitive information (including personal identifiers).

Synthetic data can enable AI agents to simulate a wide range of fraudulent transactions without depending on real-world transactions. For example, banks and credit card companies can simulate millions of credit card transactions (using synthetic data) that can emulate fraudulent patterns such as:

- Micro charges
- Unusual locations
- Rapid purchases

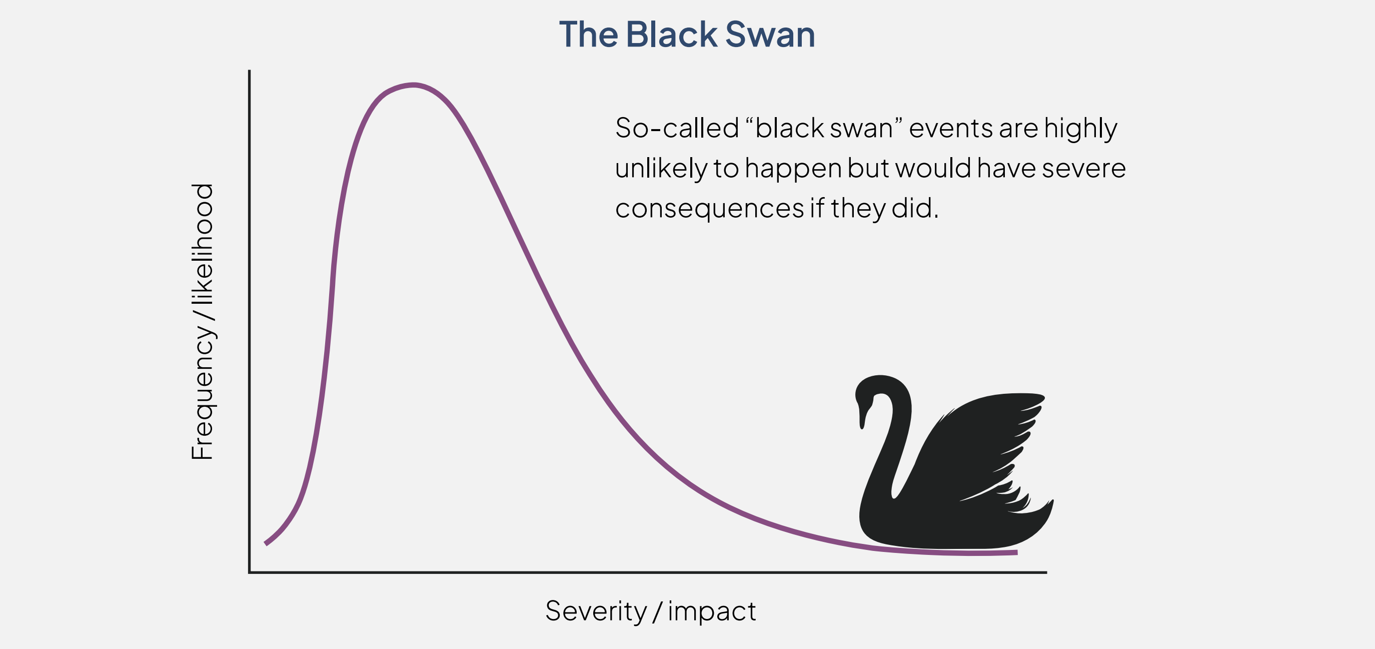
When these patterns are fed into any AI model, it can easily detect and report anomalies more effectively. An AI research team at J.P. Morgan also noted that synthetic datasets can accelerate model development in financial services by providing abundant volumes of data for algorithms.

Synthetic data can also reduce the class imbalance problem. Instead of training models using datasets (with 0.1% fraudulent transactions), synthetic datasets can provide models with abundant examples of fraudulent activities. It can also reduce false positives by enriching model training data using a variety of legitimate behavior patterns.

Risk modeling and stress testing

Risk management is the “core” of any BFSI operation, be it in the form of credit risk, market risk, or operational risk. Traditional risk models are constrained by the volume of historical data. With synthetic data, risk management teams can go beyond historical scenarios and test AI models with hypothetical risk events.

For instance, banks can use synthetic data to simulate a “Black swan” event and test their AI-powered risk models.



Typically, the Black Swan event involves creating time-series synthetic data for stock prices, interest rates, or FX rates, with an imagined 40% market drop in a day or any other stress condition. This type of risk modeling allows them to monitor the performance of their trading algorithms (or portfolio) in stressful conditions.

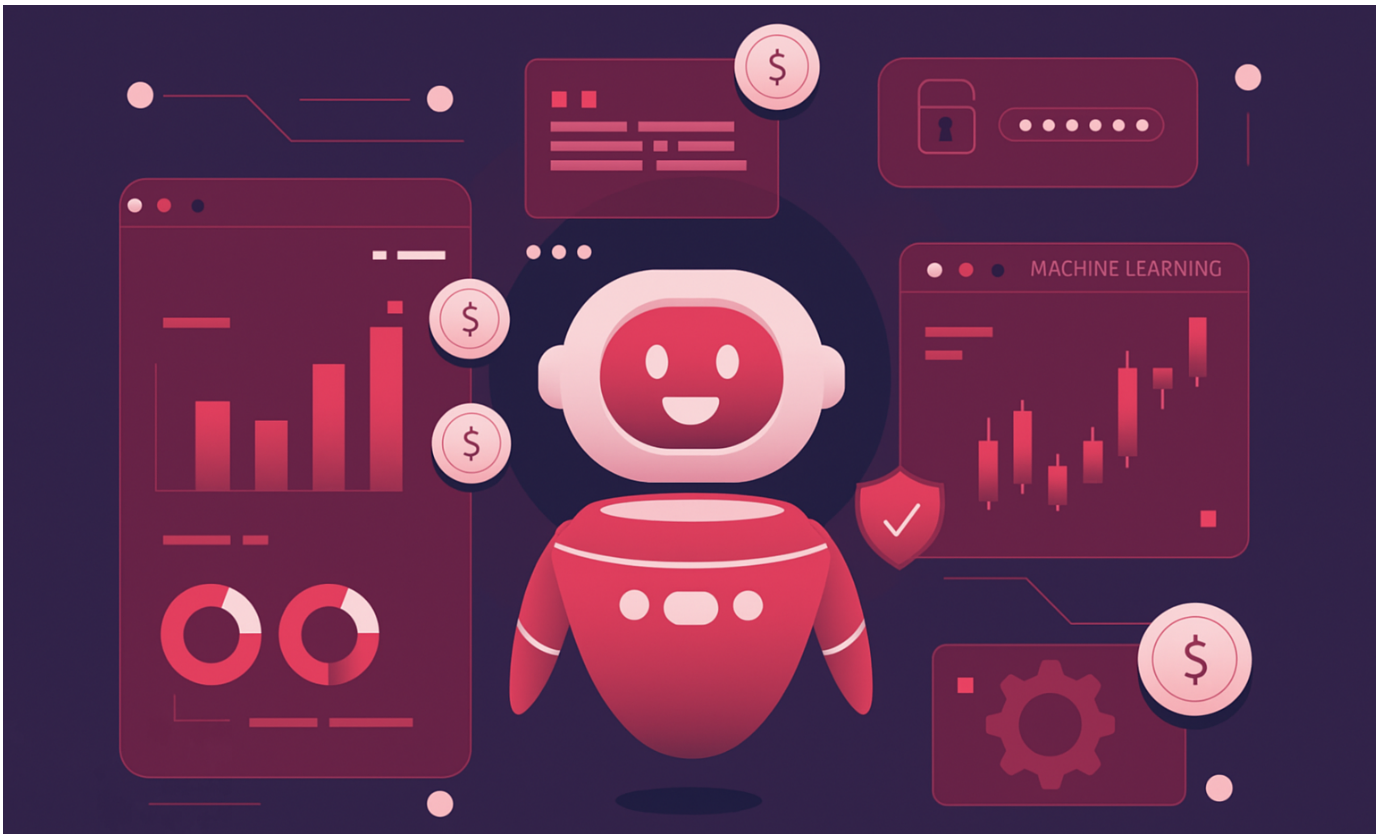
Besides, regulators and risk management teams can use this application to check if the company can handle worst-case scenarios. Synthetic data can also generate rare event samples for modeling credit risk - for instance, simulating the impact on the loan portfolio whenever there’s a 5% spike in unemployment in selected industries.

Instead of relying on historical data or limited scenarios, banks can use synthetic data to algorithmically generate fresh datasets, which cover the entire spectrum of risk distribution. Effectively, robust risk models can be tested against the complete range of market conditions.

Further, banks can now share extreme stress scenarios and results with regulators or auditors without using the actual customer data. This helps them comply with capital requirements and proper risk governance.

Customer support and chatbot training

As more customers demand instant personalized support, BFSI companies are deploying AI-powered chatbots and virtual assistants in customer service. AI agents need a high volume of conversational data for training purposes. However, data like customer chat logs can contain sensitive data.



Synthetic data can fill this data gap by simulating customer interactions for training and testing AI-enabled customer support systems. Right from queries about account balance to lost cards, banks can now generate synthetic data with the appropriate responses.

Synthetic data can also simulate various customer personas and interactions, thus allowing AI-enabled chatbots to train using realistic scenarios without relying on actual customer data. This effectively means that any virtual assistant can handle any situation – from an upset customer to an unusual request.

This is useful for protecting the customer's privacy, as no "real" chat transcripts are used in the simulation – hence there is no risk of leaking the customer's personal or account details. Banks can also improve the performance of their customer support by generating large volumes of sample interactions – including rare scenarios.

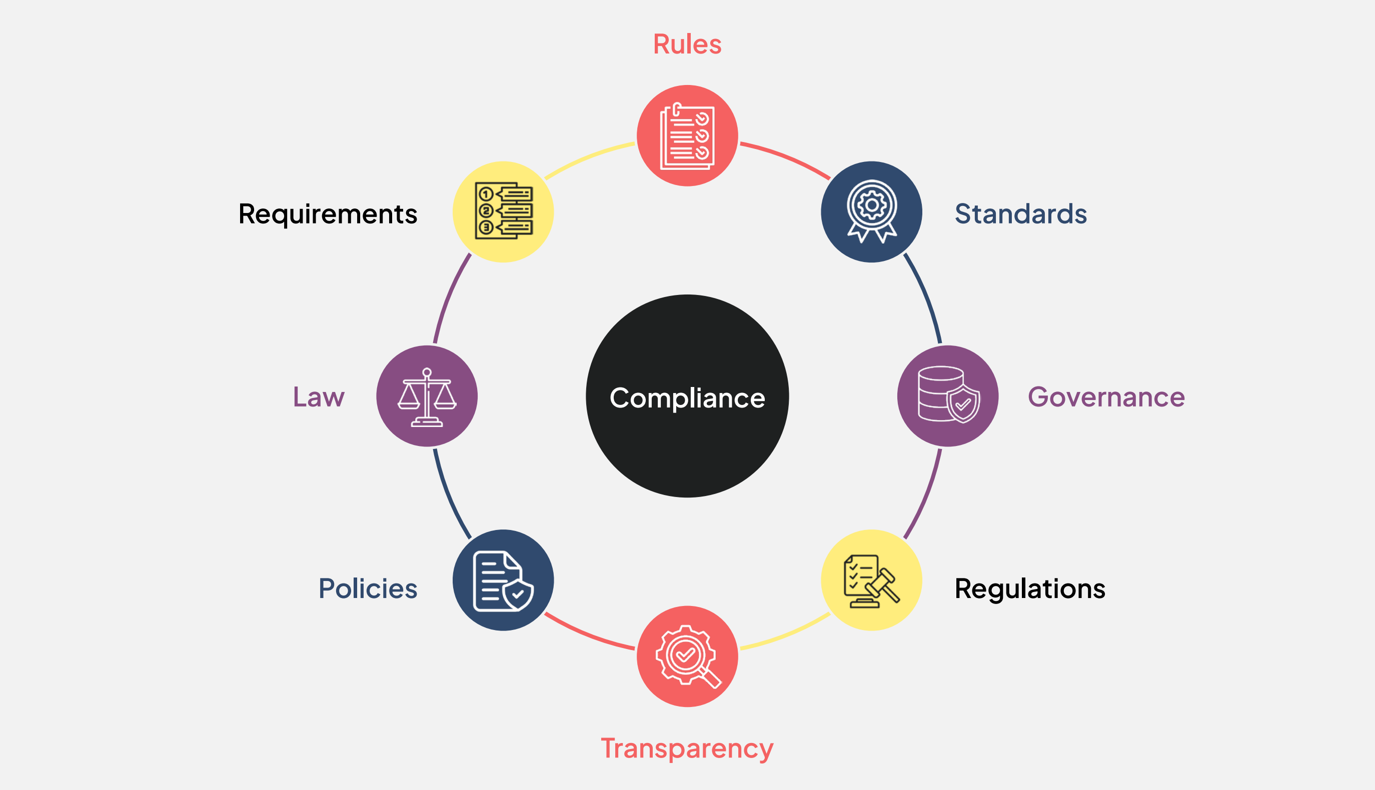
Synthetic data can help QA teams test their chatbot before deployment. They can run the chatbot through synthetic test queries and check for appropriate responses and regulatory compliance.

With the use of AI agents in customer service, BFSI companies can improve their customer response time and first-call resolution rate, thus boosting customer satisfaction. Banks like HSBC are using synthetic data to "fine-tune" their AI assistants – including training the models using "fake" customers that behave like real-world ones.

Regulatory compliance analytics and audit simulations

As noted, banks and financial institutions face intense scrutiny on regulatory compliance. They need to constantly demonstrate their compliance with laws like anti-money laundering (AML), KYC, and sanctions screening.

Synthetic data has emerged as a viable tool to simulate regulatory compliance. Banks can generate synthetic datasets that reflect compliance scenarios like money-laundering patterns and customers with high-risk profiles. They can follow up by running their compliance monitoring systems on this data to validate if the system detects any suspicious transactions or high-risk customers to meet regulatory requirements.



Synthetic data can also be safely shared with regulators and auditors without exposing any real-world customer data. For example, banks can use a synthetic data generator to create an audit dataset containing policy violations or errors. Regulators are also supporting the use of “sandboxes,” where BFSI companies can upload their synthetic data and AI models to check for AI “compliance and fairness” before deployment. With synthetic data, AI models can also be validated for regulatory compliance - for example, a credit model that is used for approving loans. Banks can validate whether this model discriminates against protected classes by including multiple demographic combinations.

Banks can also fulfil regulators’ demands for rigorous testing and validation of AI systems without any risk to data privacy. This leads to benefits like stronger compliance, lower audit costs, and increased confidence in AI agents. The C-suite is fully backing the use of synthetic data; in the next section, we’ll see how CIOs and CEOs are becoming the driving wave of this change.

How CIOs and CTOs are driving the rise of synthetic data

The global market for synthetic data generation is set to grow from $432 million (in 2024) to $8.9 billion by 2034. This reveals that synthetic data adoption is not a "theoretical future," but is actually "happening" across enterprises.

Recent surveys illustrate the growing trend toward synthetic data solutions because of data privacy and scarcity concerns. Gartner analysts have observed that in 2023, "20% of data used to train AI models is synthetic," which will rise to 80% by 2028. Gartner also forecasts that by 2026, 60% of data used in AI and analytics projects will be synthetic data, instead of real-world data.

CIOs and CTOs are driving the rise of synthetic data as the preferred way to facilitate advanced AI initiatives with strict data regulations. A recent CIO Dive survey found that 4 in 5 decision-makers acknowledge the need to improve business processes using AI technology – with synthetic data as the answer. Similarly, a 2024 EY survey of CIOs reveals that a significant number of enterprises are planning to expand their synthetic data programs as a solution to AI governance. In specific domains like consumer research, 70% of professionals expect synthetic data to dominate data collection within 3 years.

Looking ahead as data privacy laws tighten and AI models demand more diverse datasets, the BFSI industry will turn to synthetic data. CIOs and CDOs are building internal data factories to fuel testing, AI training, and market simulations. The C-suite needs a solution built for this shift, Kingfisher: The synthetic data generator is that missing puzzle piece.

Recently, a leading global bank used Kingfisher to replace sensitive PII with synthetic data during its cloud migration. Tasks that once took weeks, approvals, de-identification, and preparation, were completed in minutes. The result: 85% faster turnaround, 90% efficiency in scaling, and the workload of 100 staff eliminated, all while ensuring secure, compliant testing.

Kingfisher – Generating synthetic data from code and data

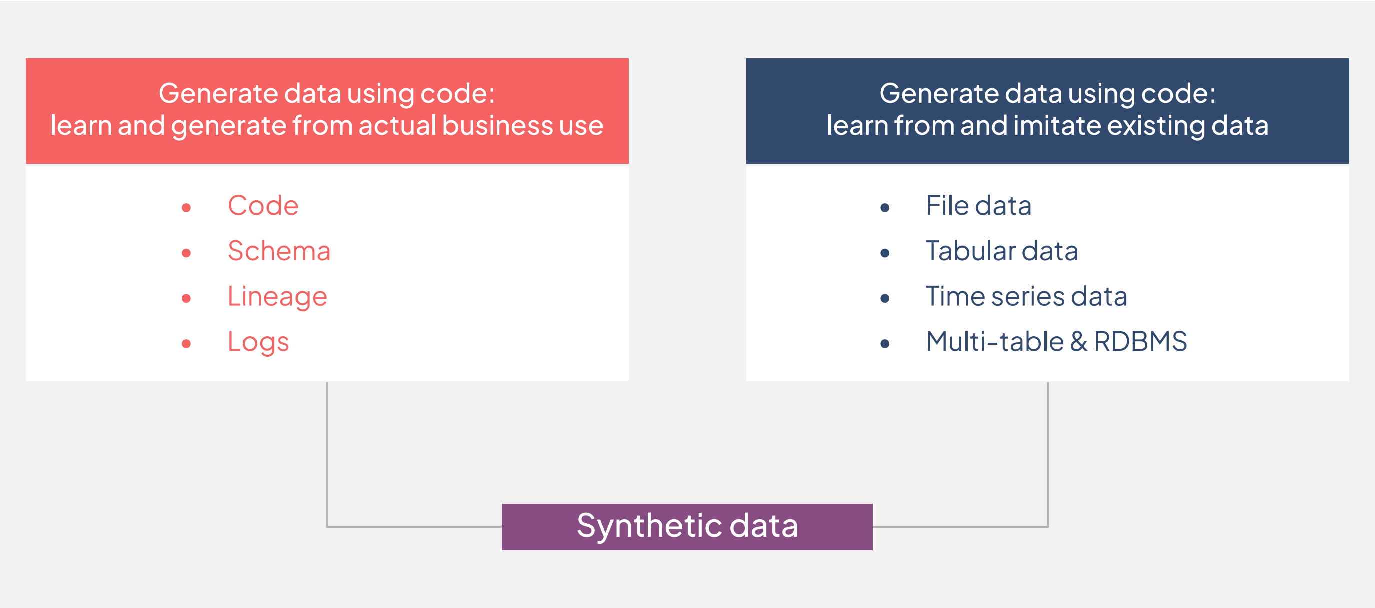
As Onix's proprietary synthetic data generator, Kingfisher can generate synthetic data from enterprise data and code. Besides extracting data from raw tables, Kingfisher can analyze the production code (SQL code) and business logic to generate synthetic data. This means that this tool can detect intricate data patterns and edge cases that other tools may overlook. This is why Onix's CTO says, "Kingfisher addresses the need for secure, privacy-compliant data by generating synthetic data that mirrors the characteristics of real data without compromising sensitive information."

This AI-powered tool can create comprehensive datasets statistically identical to real-world data. For example, in the case of banking applications with complex rules for loan approvals and cash transactions, Kingfisher can examine each rule and include the possible scenarios (including rare cases) in its synthetic dataset.

With its code-driven approach, Kingfisher can be a "game-changer" for ensuring data quality and completeness, where synthetic data is grounded in business logic.

Delivering enterprise-grade synthetic data in BFSI

Besides being a synthetic data generator, Kingfisher is a complete data engineering tool designed to meet the fluctuating demands of the BFSI industry. Effectively, Kingfisher can deliver statistically accurate synthetic data that maintains referential integrity across complex BFSI relational data, thus ensuring both data privacy and auditability.



Generate data using code:
learn and generate from actual business use

- Code
- Schema
- Lineage
- Logs

Generate data using code:
learn from and imitate existing data

- File data
- Tabular data
- Time series data
- Multi-table & RDBMS

Synthetic data

Here's what sets Kingfisher apart from other synthetic data generation tools:

1. Code-based data generation

Kingfisher can generate synthetic data from existing code, along with generating detailed logs and data lineage for every record. This helps enterprises perform accurate debugging, as well as ensure compliance and auditability.

2. Support for complex relational schemas

Kingfisher can handle a variety of financial data structures, along with customer accounts, transactions, and customer records (across multiple tables), thus maintaining referential integrity.

3. Diverse data types

Right from structured tables to time-series data, Kingfisher can preserve the statistical fidelity of the source datasets, thus delivering synthetic data that's similar to real-world data.

4. Drop-in replacement for production data

With Kingfisher, data scientists and engineers can use synthetic datasets completely to test and train AI models, without accessing any real sensitive data.

5. Compliance with privacy regulations

The Kingfisher tool can be deployed in the customer's environment, thus ensuring compliance with regulations like CCPA and GDPR. This means sensitive data remains in the safety of their IT environment.

6. Scalability

Kingfisher is also designed to generate massive volumes of synthetic data. Equipped to manage billions of records, this tool can accelerate AI development projects with real-time data provisioning.

7. Data fidelity

Enterprises using Kingfisher ensure real-world fidelity by replicating data distribution and relationships, without retaining any real identity or sensitive values.

8. Designed for highly regulated industries

Kingfisher is built for highly regulated industries where customer data is both a "goldmine" and a "minefield," thus ensuring the balance between innovation and compliance.

How does Kingfisher impact enterprises from a leadership perspective? Onix's CEO explains that with Kingfisher, CIOs and CTOs can "innovate at scale with absolute confidence in data privacy and efficiency." Enterprises can train their AI agents or projects on a rich trove of data, without waiting for approvals or worrying about a compliance breach. AI project teams don't need to wait for access to production data, nor spend time in "anonymizing" real data for their use cases.

Further, with its detailed logs and lineage, regulators or auditors can validate every source of synthetic data generated by Kingfisher. Finally, this tool effectively bridges the gap between regulatory compliance and utility, thus allowing BFSI companies to fully leverage the power of AI agents.

Conclusion

With the convergence of AI agents and synthetic data, BFSI companies can usher in the next wave of digital transformation. With AI-powered agents, banks can bring in advanced automation, personalization, and productivity levels to gain a competitive edge. Synthetic data generators like Kingfisher deliver a robust data strategy with in-built privacy.

If you want more information about Onix's Kingfisher tool, contact our team today.

onixnet.com

connect@onixnet.com

800.664.9638

Get in touch

Follow us: