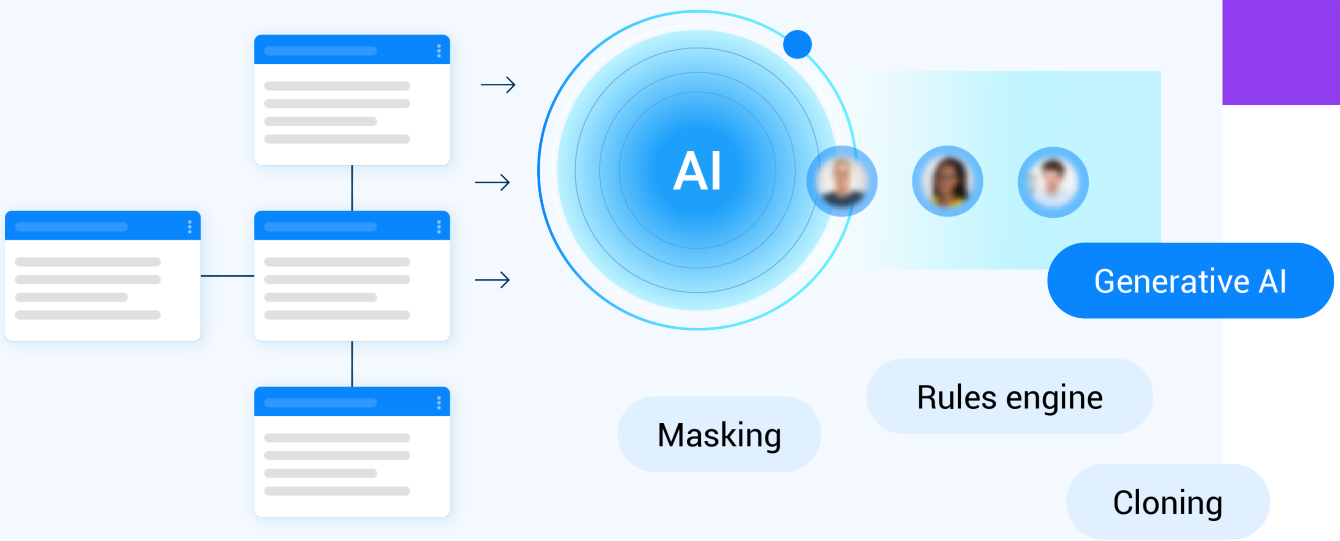
k2view

WHITEPAPER

Tabular Synthetic Data Generation

The Complete Handbook



Synthetic data generation refers to the process of creating new data that mimics the characteristics, patterns, and statistical properties of real data.

INTRO

Synthetic Data Accelerates Innovation

Synthetic data can be described as fake data, generated by computer systems but based on real data.

Enterprises create synthetic data to test software under development and at scale, and to train Machine Learning (ML) models.

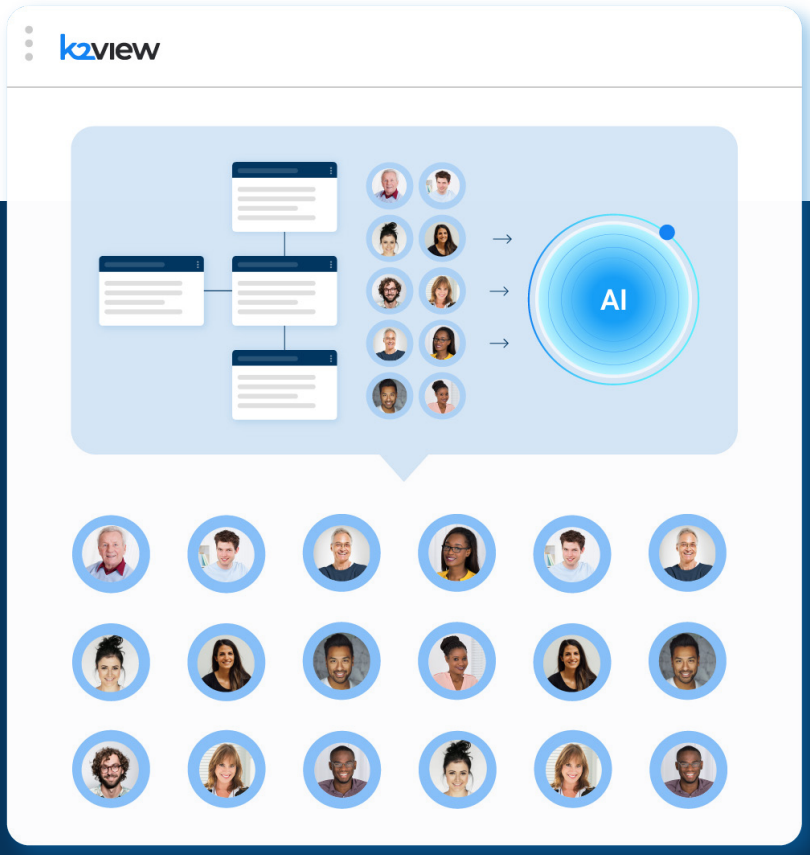
There are 2 types of synthetic data:

- **Structured,** tabular data
- **Unstructured,** image and video data

This paper focuses on structured, tabular data and the 4 methods used to synthesize it:

1. **Generative AI** creates realistic, synthetic data via ML algorithms.
2. **A rules engine** provisions data using user-defined business policies.
3. **Entity cloning** extracts business entity data, masks it, and then replicates it.
4. **Data masking** anonymizes PII and sensitive data to generate new, compliant data.

This guide also covers the end-to-end synthetic data management processes and tools needed to support operational and analytical workloads – with capabilities like multi-source data extraction, data subsetting, data masking, data versioning, rollback, and more.



Generative AI for data augmentation

CHAPTER 01

What is Synthetic Data and How is it Generated?

Synthetic data generation is the process of creating artificial data that mimics the statistical patterns and properties of real-life data.

Synthetic data is generated using algorithms, models, or other techniques. Even though it's usually based on real data, synthetic data often contains no actual data from the original dataset.

Unlike real data, which may contain sensitive or Personally Identifiable Information (PII), synthetic data ensures data privacy, while at the same time enabling data analysis, research, and software testing.

The 4 key synthetic data generation techniques are listed below:

1. **Generative AI** – modeled using Generative Pre-trained Transformers (GPT), Generative Adversarial Networks (GANs), and Variational Auto-Encoders (VAEs) – learns the underlying distribution of real data to generate similarly distributed synthetic data.
2. **A rules engine** creates synthetic data via user-defined business rules. Intelligence can be added to the generated data by referencing the relationships between the data elements, to ensure the relational integrity of the generated data.
3. **Entity cloning** extracts data from the source systems of a single business entity (e.g., customer) and masks it for compliance. It then clones the entity, generating different identifiers for each clone to ensure uniqueness.
4. **Data masking** replaces PII with fictitious, yet structurally consistent, values. The objective of data masking is to ensure that sensitive data can't be linked to individuals, while retaining the overall relationships and statistical characteristics of the data.

These 4 techniques will be discussed in greater detail in chapter 5.

Here are some other, less common methods:

- **Copula models** discover correlations and dependencies within the production data sets, and then use them to generate new, realistic data.
- **Data augmentation** applies supplementary techniques – such as flipping, rotation, scaling, and translation – to existing values to create new data.
- **Random sampling** and noise injection add random sampling data points from known distributions, as well as noise to existing data, to create new data points that closely resemble real-world data.

CHAPTER 02

Synthetic Data Generation Use Cases

There are 2 primary use cases that rely on synthetic data:

1. **Software testing** needs compliant synthetic test data provisioned to test environments, to ensure that the applications being developed perform as expected.
2. **Machine Learning (ML) model training** relies on synthetic data generation to supplement existing datasets when production data is scarce or non-existent.

Each of these use cases will be discussed in greater detail in chapters 3 and 4.

Additional use cases include:

- **Privacy-compliant data sharing** uses synthetic data to distribute datasets internally (to other domains) or externally (to partners or other authorized third parties) without revealing PII. Good examples of this are synthetic financial data and synthetic patient data.
- **Product design** deploys synthetic data to provide standardized benchmarks for evaluating product performance in a controlled environment.
- **Behavioral simulations** employ synthetic data to explore different scenarios, validate models, and test hypotheses without using real-life data.

While synthetic data is useful, it doesn't replace real data. **Both are needed.** The quality of the synthetic data depends on the selected generation method and the accuracy with which it captures the underlying patterns of the real data.

CHAPTER 03

Synthetic Data Generation for Software Testing

Test data generation plays a critical role in software testing by creating representative datasets specifically designed to evaluate the functionality, performance, and reliability of an application under development. It's typically used when production data isn't accessible, or when testing new functionality for which production data isn't available.

> **Test data generation must be configurable, enabling data teams to request the amount, and type, of data they want to generate, as well as the characteristics it should have.**

Synthetic test data, which is commonly used to test applications in a closed test environment before deployment to production, provides several benefits to testing teams. Not only does it give them total control over the datasets in use, but it also protects data privacy by delivering fake, but lifelike, information. Synthetic data can also be more efficient to use than production data because it allows testers to generate large amounts of test data quickly and easily.

Synthetic test data is used for:

- **Progression testing**, which tests new functionality that has been developed
- **Negative testing**, which ensures that the application handles invalid inputs correctly
- **Boundary testing**, which tests the limits an app can handle, like the largest input, or the maximum number of users
- **Load testing**, which stress-tests a system to ensure that it can accommodate massive amounts of data or simultaneous users

Synthetic test data is critical to data testing and DevOps teams, in cases where there's not enough complete or relevant real-life data at hand. Not only does synthetic test data reduce the non-compliance and security risks (associated with using actual data in a test environment), but it's also ideal for validating new applications, for which no data exists. In such cases, testing teams can match their requirements to the closest profiles available.

For test data management teams, whether the data is real or fake may not be an important criterion. More critical are the balance, bias, and quality of the datasets - and that the data maximizes test coverage. The benefits of synthetic test data include:

Extended data coverage

By controlling the specifications of the generated data, testing teams can synthesize the precise data they need for their specific test cases, to ensure full test coverage.

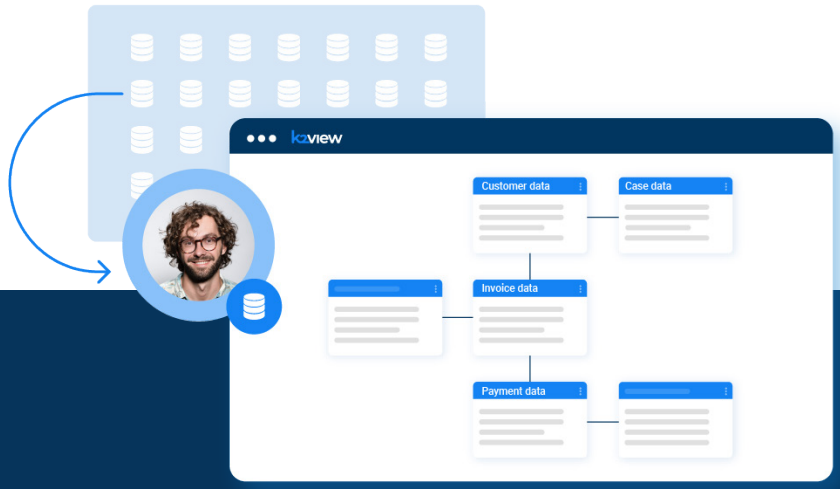
Increased speed

It's a lot quicker and easier to define the parameters for synthetic data generation than to provision data subsets from disparate systems in the higher environment.

Enhanced data protection

As opposed to data masking tools that are used to protect production data, synthetic data generation tools create fake data that closely resembles real data, but without actual values that could lead to the identification of individuals.

In the synthetic test data vs test data masking comparison, IT and testing teams must decide which model suits their specific needs best.



CHAPTER 04

Synthetic Data Generation for Training ML Models

Synthetic data is used to train ML models because it's typically provisioned more quickly and easily than real data. The synthetic datasets let the model practice its skills in a closed environment before going live in production. The model learns patterns from the synthetic data and gets better at its task.

Data scientists often prefer synthetic over production data for training ML models, due to data:

Augmentation

Synthetic data is often used to supplement the original dataset by adding anomalies, noise, or other variations. Such additions improve the model's ability to manage a wider range of input conditions, making it more versatile and robust.

Diversity

Real-life datasets don't necessarily capture all possible scenarios, leading to biases or limited generalization. Synthetic data can introduce more diverse scenarios to ensure the model learns how to deal with a broader range of situations and inputs.

Imbalance

At times, a dataset has imbalanced classes, for example, when a particular class is either over- or under-represented. Synthetic data addresses the imbalance by distributing majority and minority classes more equally.

Privacy

When handling sensitive information, synthetic data can be used to generate similar data without revealing an individual's personal details. This anonymity lets model developers and testers use the data freely while also maintaining data privacy.

Scarcity

Sometimes, provisioning enough real-world data for training ML models can be extremely difficult. Synthetic data can augment available real data, increasing the size of the dataset to enable the model to learn more effectively.

CHAPTER 05

Synthetic Data Generation Techniques

As mentioned in chapter 1, the 4 key synthetic data generation techniques are:

1. **Generative AI**
   Generative AI uses ML models, including Generative Pre-trained Transformer (GPT), Generative Adversarial Networks (GANs), and Variational Auto-Encoders (VAEs). The models' algorithms learn from existing data to create an entirely new synthetic dataset that closely resembles the original.

   - **GPT** is a language model trained on extensive tabular data, capable of generating lifelike synthetic tabular data. GPT-based synthetic data generation tools rely on understanding and replicating patterns from the training data, useful for augmenting tabular datasets and generating realistic tabular data for ML tasks.
   - **GANs** are based on "generator" and "discriminator" neural networks. While the generator creates realistic synthetic data, the discriminator distinguishes real data from synthetic data. During training, the generator competes with the discriminator to produce data that attempts to fool the model, eventually resulting in a high-quality synthetic dataset that closely resembles real data.
   - **VAEs** are based on an "encoder" and a "decoder". The encoder encapsulates the patterns and characteristics of actual data into a summary of sorts. The decoder attempts to turn that summary back into realistic data. In terms of tabular data, VAEs create fake rows of information that still follow the same patterns and characteristics as the real data.

2. **Rules engine**
   Rule-based techniques create data via user-defined business rules. Intelligence can be added to the generated data by referencing the relationships between the data elements, to ensure the relational integrity of the generated data.

3. **Entity cloning**
   Entity cloning extracts the data for a selected business entity (e.g., specific customer or loan) from all underlying sources, masking and cloning it on the fly. Since unique identifiers are created for each cloned entity, it's ideal for quickly generating the massive amounts of data needed for load and performance testing.

4. **Data masking**
   Data masking retains the statistical properties and characteristics of the original production data, while protecting sensitive or personal information. It replaces private data with pseudonyms or altered values, ensuring privacy while preserving utility.

Here's a synopsis of the pros and cons for each synthetic data generation method:

Method	Pros	Cons	Key reason for use
Generative AI	Speed (time to data)	• Limited by the diversity and size of the real data • May not generate the data needed for maximum testing coverage • Needs access to production data • Requires specialized skills	• Real data is scarce or non-existent • Need for complex data distributions • Requirement for diverse synthetic datasets
Rules engine	Creates large quantities of data, without having to access production data	• Requires detailed knowledge of the data and the logic needed to create it • Labor-intensive and time-consuming	• No access to production data • New functionality testing • Negative testing • Well-defined data generation process
Entity cloning	Instantly generates large datasets for testing and ML training	• Lacks variation and diversity • Can't generate new information or scenarios • Can pose a privacy risk if the cloned data is not properly masked	Performance and load testing
Data masking	• Ensures data privacy • Maintains the statistical properties and distribution of the original data	• Risk of re-identification • Might distort data, affecting its quality and integrity	• Software testing • Loading compliant data into data lakes and data warehouses for analytical workloads

CHAPTER 06

Synthetic Data Generation Tool Capabilities

There are various synthetic data generation tools on the market today. Before selecting one, make sure it can:

1. **Generate synthetic data to support a broad set of use cases:**
   - Software testing
   - Training ML models
   - Privacy-compliant data sharing
   - Building product demos and prototypes
   - Behavioral simulations
2. **Support the 4 key synthetic data generation techniques:**
   - Generative AI
   - Rules engine
   - Entity cloning
   - Data masking
3. **Manage the synthetic data lifecyle end to end:**
   - Extract training data from all relevant source systems
   - Subset the training data to improve accuracy
   - Mask data to ensure compliance
   - Generate new data via any combination of data generation techniques
   - Reserve synthetic data to avoid users from overriding each other's data
   - Version the generated data to allow for rolling back the data to prior versions
   - Load the data to the target systems
4. **Mask data**
   Automatically discover and mask personal identifiable information (PII) and sensitive data from the source data used to generate synthetic data, to ensure data privacy and compliance of production-sourced data.
5. **Maintain relational integrity**
   Ensure relational integrity – by leveraging metadata, schemas, and rules to learn the hierarchy and relationship between data objects across data sources – and preserve relational consistency of the data.
6. **Access all data sources and targets**
   Easily connect to, and integrate with, the source and target data stores relevant for the synthetic data.
7. **Serve yourself for greater agility**
   Provide testing and data science teams with self-service tools to control and manage the data generation process, e.g., configure and customize data generation functions, without being dependent on data & AI teams.
8. **Integrate into CI/CD and ML automation pipelines**
   Easily integrate the synthetic data management process into your testing CI/CD and ML pipelines via APIs

CHAPTER 07

Synthetic Data Generation by Business Entities

Enterprises are turning to entity-based synthetic data generation tools to create realistic but fake data, whose referential integrity is always enforced.

When synthetic data is generated by business entity (customer, device, order, etc.), all the relevant data for each business entity is always generated, is contextually accurate, and consistent across systems.

The entity data model captures and classifies all the fields to be generated, across all systems, databases, and tables. It enforces referential integrity of the generated data by serving as a blueprint for the data generator – regardless of the synthetic data generation technique(s) employed:

Generative AI

The entity data model is used in the generative model training as a means to generate valid, consistent, and accurate data.

Rules engine

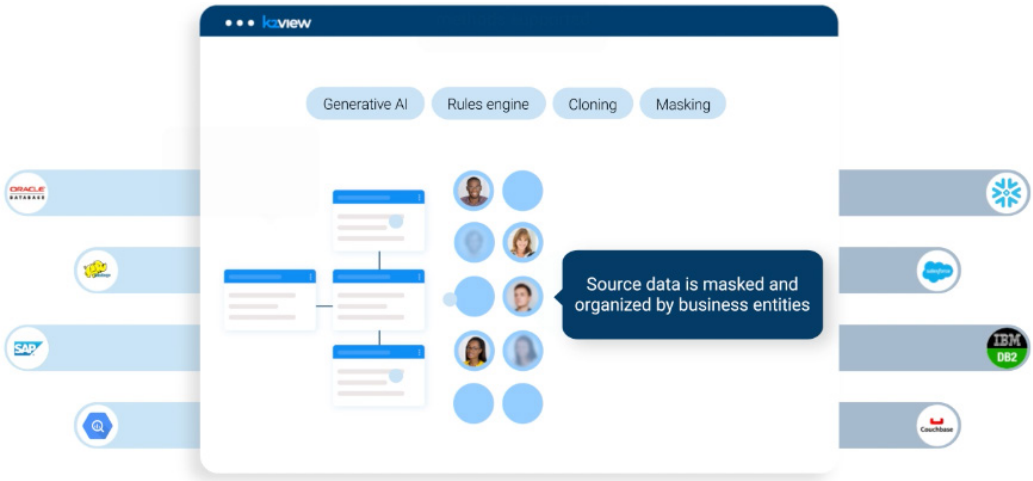
A data generation rule is associated with each field in the entity schema, and the rules are auto-generated based on the field classifications.

Entity cloning

A single instance of a business entity is extracted and masked from the source systems, and is then cloned however many times are required, while generating new identifiers for each cloned entity.

Data masking

When provisioning production data for software testing and analytics, the entity data is masked as a unit, ensuring referential integrity of the masked data.



CHAPTER 08

Tabular Synthetic Data Companies

There are many providers of tabular synthetic data generation tools. The market is continually evolving, with vertical specialization tools, pure-play horizontal platforms, and extensions of existing data management solutions.

Here are the top 6 companies specializing in tabular synthetic data generation:

1. **K2view**
   K2view manages the entire synthetic data lifecycle, including data extraction, subsetting, pipelining, and operations. It uses a combination of techniques to generate accurate, compliant, and realistic synthetic data for software testing and ML model training. Referential integrity of the generated data is ensured by creating a schema which serves as a blueprint for the data model.
2. **Gretel**
   Gretel provides a synthetic data platform for developers and ML/AI engineers who use the platform's APIs to generate anonymized and safe synthetic data while preserving data privacy.
3. **MOSTLY AI**
   The MOSTLY AI synthetic data platform enables enterprises to unlock, share, fix, and simulate data. Although similar to actual data, its synthetic data retains valuable, granular-level information, while protecting private information.
4. **Syntho**
   The Syntho AI-based engine generates a completely new, artificial dataset that reproduces the statistical characteristics of the original data. It allows users to minimize privacy risk, maximize data utility, and innovate through data sharing.
5. **YData**
   The YData data-centric platform enables the development and ROI of AI solutions by improving the quality of training datasets. Data teams can use automated data quality profiling and improve datasets, leveraging state-of-the-art synthetic data generation.
6. **Hazy**
   Hazy models are capable of generating high quality synthetic data with a differential privacy mechanism. Data can be tabular, sequential (containing time-dependent events, like bank transactions), or dispersed through several tables in a relational database.

CHAPTER 09

The Future of Synthetic Data Generation

Synthetic data generation processes and tools are evolving at a rapid pace.
The following areas promise to introduce innovation that delivers better business outcomes across synthetic data use cases.

Synthetic data operations

The generation of synthetic data is just one step in the synthetic data lifecycle. Data teams are now seeking methods and tools to manage and automate the entire synthetic data lifecycle.

Improved data quality, accuracy, and reliability

Since data professionals rely on accurate and high-quality data for their workloads, synthetic data companies will be driven to continually optimize their synthetic data generation algorithms, and tools will emerge that generate vertical-specific synthetic data.

Ethical and legal perspectives

With the spread of synthetic data, legislators and regulators are paying more attention to its ethical and legal implications. IT and business teams need to be aware of these issues, and take them into account, as they develop.

Integration with production data

By integrating synthetic data with real-life data, data teams hope to generate more realistic and comprehensive datasets. For example, synthetic data could be used to close any gaps in actual datasets, augment real-life information to cover a broader scope of edge cases, and create test data to cover new application functionality being developed.

Emerging use cases

Synthetic data is increasingly being used in new applications, such as autonomous automobiles and virtual reality. Researchers are exploring how synthetic data can be used to improve the performance of AI systems in these, and other, emerging technologies.

CHAPTER 10

Summary

Due to more and more stringent data privacy laws and the ever-increasing complexity of accessing and masking multi-source production data, the need for synthetic data generation and management is growing.

Forward-looking enterprises shouldn't settle for a point solution to generating tabular synthetic data for a specific use case, but should rather be seeking a future-ready solution that can address a multitude of use cases with the needed accuracy and agility, while managing the entire synthetic data lifecycle.

K2view synthetic data management is the only end-to-end synthetic data management solution, with support for the broadest spectrum of use cases required by enterprises. It's also the only solution that combines the 4 leading synthetic data generation techniques with a business entity approach, to deliver the most accurate, compliant, and reliable synthetic data.

About K2view

At K2view, we believe that every enterprise should be able to leverage its data to become as disruptive and agile as the best companies in its industry.

We make this possible through our patented Data Product Platform, which creates and manages a complete and compliant dataset for every business entity – on demand, and in real time. The dataset is always in sync with its underlying sources, adapts to changes in the source structures, and is instantly accessible to any authorized data consumer.

Data Product Platform fuels many operational use cases, including test data management, data masking, data tokenization, Customer 360, data migration, legacy application modernization, data pipelining and more – to deliver business outcomes in less than half the time, and at half the cost, of any other alternative.

The platform inherently supports modern data architectures – data mesh, data fabric, and data hub – and deploys in cloud, on-premise, or hybrid environments.