

SOLUTION OVERVIEW

Aindo's Synthetic Data Platform empowers organisations to unlock the full potential of their data. By leveraging generative models developed specifically for structured data it enables:

- unlocking access and reuse of privacy-constrained datasets through synthesis and anonymization/pseudonymization;
- Enhancing data quality through data rebalancing, augmentation and missing values imputation.
- The creation of predictive models on relational data directly, without the need for feature engineering;
- The generation of what if scenarios;
- The identification and redaction of personally identifiable information (PII) within textual data and documents.

Our solutions address key technical challenges – such as limited data accessibility, data scarcity, variable data quality and the creation of machine learning models on structured data – while ensuring compliance with stringent regulatory requirements like GDPR and the upcoming AI Act. By leveraging advanced synthetic data solutions, organisations can confidently harness their data for innovation and decision-making, transforming obstacles into opportunities.

Features

Synthetic Data Generation

Our solution offers a diverse set of generator options, covering most of the market needs.

Relational Synthetic Data Generation – A robust library of data generators for creating high-quality synthetic datasets composed of one or multiple tables connected between them by foreign keys. The generators train directly on the customer's data, without leaving the installation environment. Our generators preserve the statistics of the original tables even when pertaining to correlations between different tables or rows within the same table, supporting time series by default. Importantly, the generated data cannot be linked to any real person, being therefore anonymous according to GDPR. They also preserve the integrity of the keys, while handling different data types such as:

- Categorical
- Boolean
- Date
- Time
- DateTime
- Integer
- Floating Point
- Coordinates
- Free text
- Italian fiscal code

In particular, the free text column is also generated while maintaining the correlations with other columns in the same table.

Our generators offer excellent privacy protections: they support differentially private training for mathematical guarantees on the maximum amount of information being released. Also, to protect

from secondary sources of privacy leaks, the user has the option to generate synthetic data lacking rare values that might be present in the original data.

Finally, our solution includes a module allowing the user to define hard or business rules that might be present in the data and that must be respected by the synthetic data. We currently handle generic mathematical relations between columns that can be expressed as equality or disequality equations (expression such as: “column A / Column C + 3 > Column D”) and mappings between columns.

Partially Synthetic Data Generation – Our solution offers the possibility to synthesize only a set of columns in a dataset while preserving the other columns from the original dataset. This option allows for a fine grained control of the original information being revealed and of the privacy protection offered.

Pre-trained Synthetic Data Generation – Our solution offers the possibility to generate tabular data from natural language description of their content, using a generator pre-trained on a large set of structured data. This allows for generation of data even in the absence of any examples, to support use cases such as software development where one would need data that would not be otherwise available. Differently from other LLM-based solutions available on the market, our solution guarantees the format of rows and columns, ensuring that subsequent preprocessing can occur without incurring in formatting errors.

If deemed useful, the pre-trained solution can be further fine tuned, an option being particularly useful when the user has a limited number of examples available. This fine tuning can also be performed on multiple datasets of different formats, allowing for the generation to learn from multiple sources simultaneously.

The pretrained generator is also capable of augmenting a dataset with columns that were not originally present.

Synthetic data evaluation – Our solution offers a comprehensive set of metrics ensuring they meet essential standards for privacy and utility. This evaluation framework is critical for maintaining data quality and regulatory compliance across use cases. Our evaluations also come in a pdf-report format that is automatically produced at each generation.

Data Curation – Our generators allow the user to modify, balance, and enhance both real and synthetic datasets. Key features include:

- Detecting and correcting imbalances to improve dataset representativeness,
- Smart imputation to handle missing or incomplete data points,
- Augmentation techniques to expand dataset diversity and quality.

Anonymization tools

Beside synthesis options, our platform offers additional tools for enhanced privacy protection.

Classic anonymization or pseudonymization – Our solution includes a set of tools developed for anonymizing or pseudonymizing tabular data, by combining a diverse set of techniques: generalization, randomization, permutation, hashing, mocking. It can be combined with partial synthesis offering great configuration flexibility.

PII redaction in text and documents* – We offer a state of the art text processing tool to identify and remove potentially personally identifiable (PII) information in text and documents. The tool also allows you to substitute the personal information with mock information, preserving the overall document structure.

** under development.*

Predictive modeling

Beside data generation, our generators can also be used to perform predictive tasks on structured data.



Prediction on relational data – Our solution allows the use of the synthetic data generators in predictive mode. In this mode, they can perform predictions on the value of certain columns of a relational dataset based on the value of all the other columns. This is performed directly on the relational data, without any feature engineering, while maintaining a very high level of performance. This feature can also be used as a data imputation tool, for inferring missing information that might be present in a dataset.

Forecast on relational data – Our solution includes a module capable of forecasting relational data. In this mode, they enable the prediction of future trends based on historical data. This module uses a specific generator trained on past data to provide accurate and contextually relevant predictions. Because the forecast is performed on complex structured data it allows the inclusion of information of variable formats into the forecasts.

The module is also capable of generating what if scenarios, allowing for the evaluation of the difference of outcomes should a particular event occur.

Installation



















The Aindo Synthetic Data platform is powered by Docker containers running in a Kubernetes cluster, ensuring scalability and reliability across a variety of use cases. It can therefore be used as a SaaS, be hosted on private clouds or installed on premises.

<p>① SaaS</p> 	<p>② Self hosted</p> 	<p>③ On-Prem</p>
---	--	------------------

User Interface, REST API, SDK – The platform is intended to support both development and production projects, therefore offers diverse interaction options to accommodate the customers' needs.

The platform has an intuitive user interface (UI) allowing non expert users to leverage most of the available features. The functionalities available in the UI are also available via a REST API for machine-to-machine interactions. Finally, we also offer an SDK suitable for developers. The functionalities available in the various deployment options are listed in the following table.

ITEM	SDK	UI / API
------	-----	----------

Relational Synthetic Data Generation		
Differentially private training		
Rare value protection		
Constraints handling		
Partially Synthetic Data Generation		
Pre-trained Synthetic Data Generation		
Synthetic data evaluation		
Data Curation		
Classic anonymization or pseudonymization		
PII redaction in text and documents	Under development	
Prediction on relational data		
Forecast on relational data		

Scalability – The solution can be executed on CPU, but also supports multi-GPU training for the generators, allowing the synthetization of highly complex and large volume data.

Connectors – The platform offers multiple connectors for data input/output.

We currently have implemented the following database connectors:

- PostgreSQL
- MySQL
- MariaDB
- Google Big Query
- Microsoft SQL Server
- Oracle Database

We also support connections to the following object storage systems:

- Amazon S3
- Google Cloud Storage

Encryption – Our platform encrypts all data at rest and in transit, thus providing a high level of data protection and preventing unauthorized access to data.

Certifications

We are the first synthetic data company to be EuroPrivacy certified. This certifies the GDPR compliance of our synthetic data generation protocols in the healthcare domain. Aindo is also holds the ISO 9001 and ISO 27001 certifications.

Support

We provide support services to customers having an active contract (i.e. maintenance, or annual subscription). Support is available in Italian and English.

Use cases

(1) AI/ML–model development & Analytics – Generate an anonymous version of an input dataset that preserves the patterns in the data, while being void of personal identifiers, ensuring privacy compliance. Enhance the quality of the synthetic dataset to maximise its value for

downstream analysis, using functionalities such as rebalancing to ensure representativeness, or smart-imputation to fill gaps from missing data. This enables developers to build and validate AI / ML models in privacy-safe environments.

Example: **Anonymisation of Real World Data** for medical research.

(2) Data Sharing for Collaborative Research and Development – Allow organisations to share realistic synthetic versions of sensitive datasets with external collaborators or stakeholders without exposing any personal information. This supports open collaboration in fields like medical research or academic studies, where privacy concerns often restrict data sharing.

Example: share secure data with external researchers, suppliers and vendors.

(3) Data augmentation / rebalancing – Rebalance the subject preset in a dataset to reduce biases and increase fairness of a dataset. Augment rare or edge cases in a dataset for machine learning training purposes.

Example: rebalance the number of males and females in a dataset to improve the fairness of a machine learning algorithm.

(3) Predictive modeling – Predict an individual's likely future trends based on historical data. By analysing patterns learned from similar data, our platform provides personalised predictions, Useful in applications such as patient health monitoring, customer behaviour forecasting or financial trend analysis.

Example: estimating the risk of developing diseases to optimise screening campaigns, or estimating the risk of adverse events for optimal planning of healthcare resources (*programmazione sanitaria*)

(4) PII redaction in text and documents* – Detect and protect personally identifiable information (PII) within textual data and documents. Users can choose to either mask the PII for security purposes or replace it with realistic synthetic equivalents, enabling safe sharing and processing of sensitive documents across teams or organisations without compromising privacy.

Example: remove all the PII information from electronic health records

* *under development.*