Frequent Questions and Answers

Infrastructure, Deployment, and Integrations

Do you support on-premise and cloud data sources? What data sources are you able to connect to?

Synthesized is able to connect to both, on-prem and cloud data sources.

- On-premise available data sources:
    - **Immediately:** Oracle, Postgres, DB2, Sybase, SQL Server, MySQL
    - **Sources enable upon request:** Sharepoint / File Shares / SMB, Drill, Druid, Hive, Solr, CockroachDB, CrateDB, Exasol, Elasticsearch, Firebird, BigQuery, Google Sheets, Informix, Netezza, MonetDB, ASE, Hana, Snowflake, Teradata
- Cloud-based available data sources:
    - **Immediately:** Azure Table Storage, Azure Database for PostgreSQL, Azure CosmosDB, Azure SQL Database
    - **Sources enable upon request:** Azure Blob Storage, Azure Files, Data Warehouse, O365 Sharepoint / OneDrive

Can I install the software on-premise? And on our private cloud?

Yes, Synthesized supports on-prem installation, and can be easily deployed into MS Azure, AWS and GCP private cloud.

How do your users interact with the software? Can I plug it into our data pipeline and integrate it with our current CI/CD process?

One can interact with the Synthesized platform from the Web-UI, API, or SDK. When used from the Web-UI, the user will benefit from an easy to use and intuitive interface and enable team collaboration. The platform can be integrated with any external service using the API service, and it can be easily integrated with CI/CD processes with the SDK.

Does the product have a UI with access control and audit capabilities?

The solution supports user accounts and roles, and it can be integrated with an external single sign-on (SSO) service. It also provides full audit capabilities: all actions are written into a service table and can be queried using a REST API.

Synthesized Data Engine

What is synthetic data?

Synthetic data is data generated from a Machine Learning model that looks at the original data, learns and understands it, and is able to generate more data that looks, feels, tastes and even smells like original data.

The new data has the same high level properties as the original, but at the row level it's completely new and artificial.

What can I use synthetic data for? What are the benefits?

Original data can be substituted for Synthetic data, and the same results will be achieved. But it has many other benefits, to name a few:

- As it's synthetic, it is privacy-compliant and can be used and shared freely without going through complex and long compliance processes.
- As it can be manipulated and reshaped for different purposes, for example data re-balancing, bias mitigation, data imputation, or model validation with simulated data scenarios. For Machine Learning and to optimize data coverage for testing.
- Data generation is generally very fast, so one can generate large amounts of data for data augmentation and performance testing.
- As it's based on a model, storing the model parameters is much more lightweight than storing the actual data. Then, the data can be generated on-demand when needed.

What data types are handled by the Synthesized platform?

Synthesized Data Platform can work with all structured data, usually — but not exclusively — this refers to tabular data, including flat files)such as Excel spreadsheets, CSVs, etc) and Relational Databases. All usual data types (integers, floats, characters, uuids, jsons) are handled by the platform

How can I compare my original data to my synthetic data?

For each Synthesized data product generated, Synthesized can automatically generate both a data utility and data privacy report that can also be stored, versioned and documented.

Furthermore, different privacy reports and monitoring features are available on-demand, with an alert system that would notify the users if a specific scenario happens. This can be applied to privacy

How complex is the platform to use? Do I need extensive knowledge and training to be able to use it?

Running the engine against a given data source with the default configuration is straightforward, the user just needs to provide connection details for input and output sources, and that's all!
If needed, the user can still provide some extra configuration parameters to enforce certain behaviour (e.g. strict rules or implicit referential integrity).

Privacy and De-Identification

Why is Synthetic data more secure and privacy-preserving than traditional anonymization techniques?

With traditional anonymization techniques each sample in the output data will have a one-to-one mapping with the original set, which means that they are not robust against complex attacks such as linkage attack or attribute inference.

Synthesized approach, on the other hand, is to learn the data distribution of the underlying phenomena and sample new data points from it. This means that there's no one-to-one mapping with the original data which makes it robust against complex disclosure attacks.
Read more about the topic here.

Does the solution have an option to guarantee that no synthetic data points coincide with any real data point?

Generally speaking, the chances of generating a data point that is present in the original set are insignificant. But if that happens, the Synthesizer can be set to remove those points from the output set.

Does your solution have a means to generate synthetic data with or without differential privacy depending on preference?

Yes! Synthesized solution allows you to define and configure:

- Differential Privacy parameters (such as "epsilon" and "nose_multiplier") can be configured depending on the user's needs. More information here.
- Linkage attack "t-closeness" and "k-distance" parameters can be configured, and linkage attack mitigation can be applied to the data (on-demand feature).

How does Synthesized handle columns that contain PPI such as names and addresses? Is it possible to generate “fake” but realistic values in such cases?

Synthesized supports two ways of attribute-level anonymisation:

- **Data obfuscation:** If required, the user can obfuscate any data field with traditional anonymization techniques such as masking, nulling, and generalization among other available techniques.
- **Fake Data:** Synthesized also supports the substitution of production values with realistic “fake” data, while maintaining continuity across attributes in a row. Fake data keeps the anonymity of underlying PII data hidden while still generating realistic values for certain entities such as names, addresses, bank accounts, etc.