



Create high quality, compliant synthetic data at scale with Synthesized Scientific Data Kit

Business challenges

Synthesized's Scientific Data Kit (SDK) helps some of the world's largest and most highly-regulated organizations improve data quality and quantity, increase speed to data access, and ensure compliance of sensitive data for use in supporting their AI/ML initiatives.





The SDK solves these challenges by enabling the generation of exceptionally high-quality synthetic data for machine learning and data science tasks. It can be used to bootstrap data where the density of records is low, automatically rebalance data to improve model performance, and anonymize data to allow greater access.

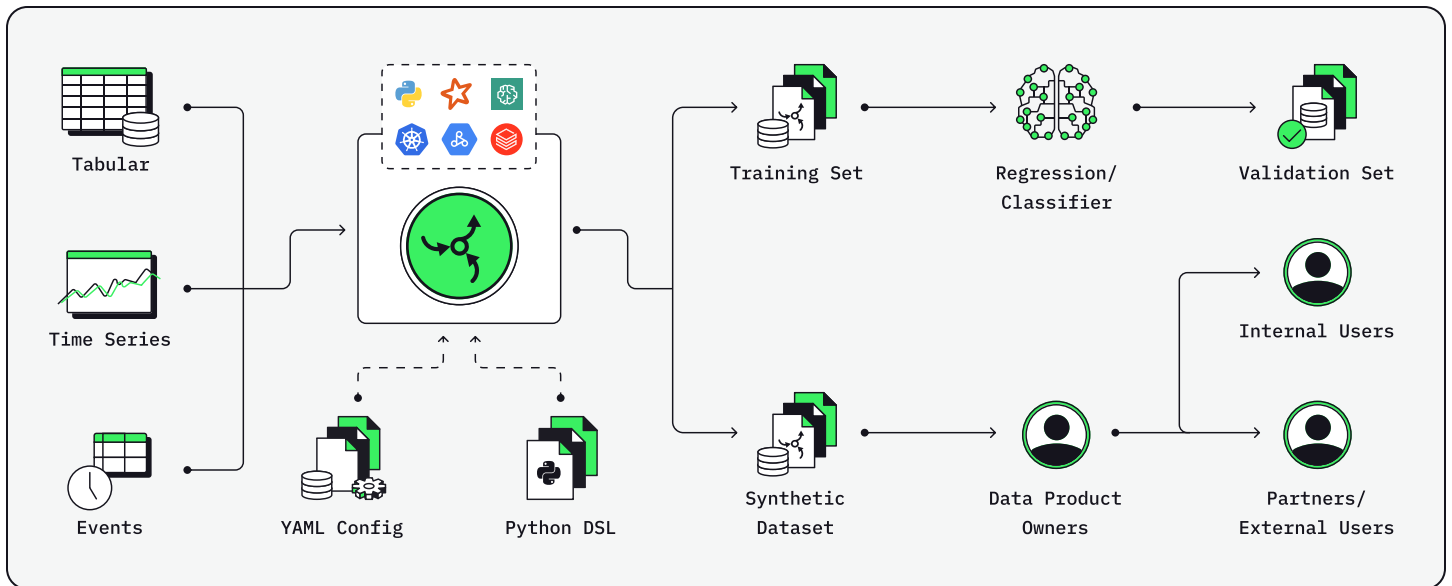
Solution overview

The SDK can be run fully isolated on-premise or in the cloud, on a single CPU/GPU or scaled across existing Spark clusters. The ability to configure the SDK with YAML, extend its functionality in python and integrate it within big data ecosystems, makes it the only platform of its kind which is equipped to handle the data privacy, quality, scale, and governance requirements of the world's most demanding financial services organizations.

Synthesized can be used to quickly generate both tabular and time-series data, through a combination of high dimensional, deep-generative models that are able to preserve the statistical properties, distributions and correlations between properties of the data.

KEY FEATURES

-  **Improved model performance**
Synthesize highly representational synthetic data that can be used to supplement or replace original data. Benefit from up to 15% uplift in model performance with data rebalancing, data imputation, and high-quality synthetic data generation. SDK helps increase revenue across conversion, fraud, revenue recovery, and more.
-  **Scalable processing of big data**
Use existing Spark clusters to train deep generative models and synthesize enhanced data at scale.
-  **Guaranteed compliance**
Configure a combination of data masking alongside privacy preserving synthesis to codify complex compliance requirements into concrete data transformations, eliminating the costly risks that comes with sensitive data.
-  **Flexible, fully isolated deployment options**
Deploy on premise or in the cloud using docker, kubernetes, openshift or simply as a python package. Configure any synthesis workflow in YAML allowing for fast and easy synthetic data.



CASE STUDY

Challenge

A financial institution operating in Latin America rapidly expanding its operations lacked labeled transactional data to develop new product offerings at the speed demanded by its users; which led to low predictive performance of customer-facing models.

Specifically, transactional bank fraud is a notoriously difficult and complex problem to address.

The performance of fraud detection and AML models is only as good as the quality and amount of training and test data. Data acquisition and provisioning is often slow or impossible due to the highly controlled records.

Solution

Synthesized SDK automatically extracted the deep statistical properties of the available fraud data to create highly representative new examples of fraud for training and testing of the fraud detection system and to augment the existing data. The SDK learned to generate 5 million complex fraud data records in less than 10 minutes.

Impact & Benefits

- ✓ 5 times fewer errors on the augmented fraud data;
- ✓ Increased developer productivity & speed-to-market;
- ✓ Lowered data acquisition costs.

Trusted by data teams at the world's leading companies:



Data for what's next.®

Synthesized delivers the fastest way to create and share trusted data.