Mphasis SynthStudio

Product Description

Mphasis' SynthStudio generates high-quality synthetic data to monetize trustworthy business insights while preserving privacy and protecting data subjects. Its AI-generated synthetic data enables organizations to leverage data's maximum potential to cross-collaborate and build reliable and highly accurate assets with no loss of data privacy and utility.

Overview

Mphasis' SynthStudio creates synthetic data, based on metadata, constraints, and conditions, for a wide range of requirements - including application testing, creation of POCs, fueling hackathons and innovation initiatives. Synthetic data is created using mathematical models, computer algorithms or other artificial means, and can be used to train Machine Learning models or test software systems in situations where real data is difficult or costly to obtain, or when privacy concerns prevent the use of real data. This synthetic data, artificially constructed, allows businesses to test, learn, and innovate without breaching any real-world data privacy.

The proposed solution incorporates two components: Data Synthetization and Data Enrichment. These components work in tandem to maximize the value and utility of existing data resources. Data Enrichment is a sophisticated process that augments existing datasets with additional, relevant information. This enhancement significantly increases the contextual utility of the data, opening new avenues for innovation and deeper insights. For instance, it can involve generating descriptive tags for images, assigning sentiment labels to text reviews, or adding demographic information to user profiles. By layering these additional dimensions onto the original data, organizations can unlock hidden patterns and correlations, leading to more nuanced understanding and decision-making capabilities. Data Synthetization, on the other hand, is a powerful technique that leverages both representative original data and associated metadata to generate high-quality, synthetic datasets.

Synthetic data that is generated by this solution as a privacy-safe version of the original data unleashing maximum utility for insights. This feature is especially useful when synthetic data is generated from original data while preserving its statistical properties and distributions.

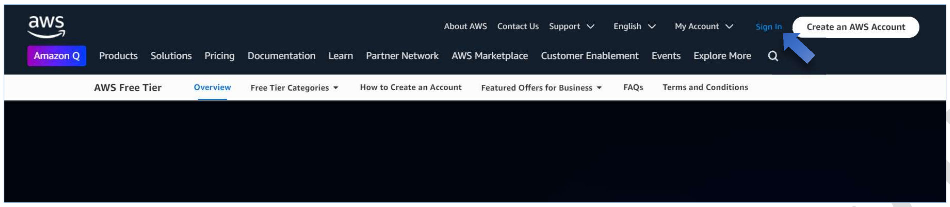
Mphasis' SynthStudio stands out as a synthetic data generation and enrichment solution, utilizing innovative algorithms and proprietary methodologies. This approach establishes a scalable and secure pipeline, ensuring the creation of technically sound and privacy-protected synthetic datasets.

Highlights

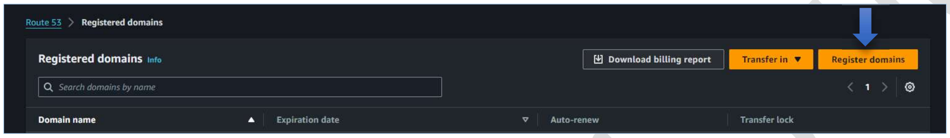
- Privacy-preserving
- Expert-assisted Synthetic Data Generation
- Deep Learning

Prerequisites

1. Domain Setup
   If you already have a domain registered or hosted on Route 53, you can use the existing domain and skip the Domain Setup.
   You can get a dedicated URL with AWS Route 53. Follow “Domain Setup” to know how to setup a domain
2. Domain Setup
   a. Create or sign into your AWS account



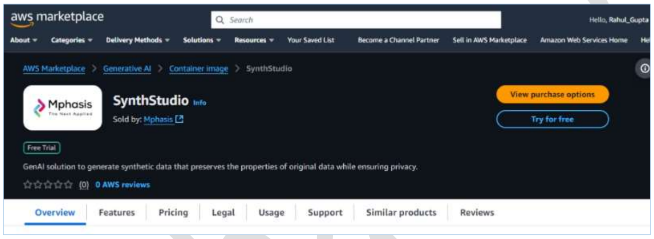
   b. Search Route 53 on console and register a domain.



   For more details steps on how to register domain click here

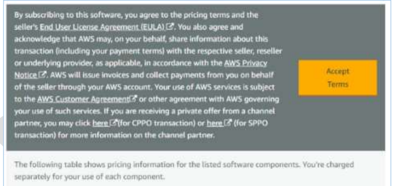
How to Subscribe

1. Navigate to AWS marketplace AWS Marketplace: Homepage
2. Search for SynthStudio Mphasis
3. Subscribe to the product



GenAI solution to generate synthetic data that preserves the properties of original data while ensuring privacy.

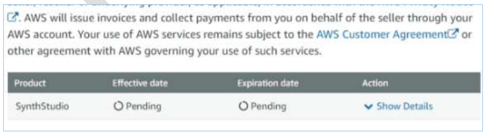
4. Click on Accept Terms



By subscribing to this software, you agree to the pricing terms and the seller's End User License Agreement (EULA). You also agree and acknowledge that AWS may, on your behalf, share information about this transaction (including your payment terms) with the respective seller, reseller or underlying provider, as applicable, in accordance with the AWS Privacy Notice. AWS will issue invoices and collect payments from you on behalf of the seller through your AWS account. Your use of AWS services is subject to the AWS Customer Agreement or other agreement with AWS governing your use of such services. If you are receiving a private offer from a channel partner, you may click here (for CPPO transaction) or here (for SPPO transaction) for more information on the channel partner.

The following table shows pricing information for the listed software components. You're charged separately for your use of each component.

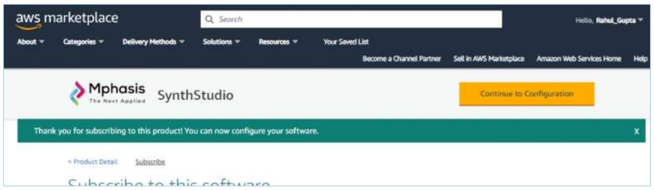
5. Wait for processing to complete



. AWS will issue invoices and collect payments from you on behalf of the seller through your AWS account. Your use of AWS services remains subject to the AWS Customer Agreement or other agreement with AWS governing your use of such services.

Product	Effective date	Expiration date	Action
SynthStudio	○ Pending	○ Pending	Show Details

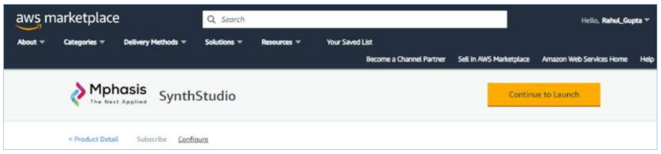
6. Click on Continue to Configuration to start configuring your product.



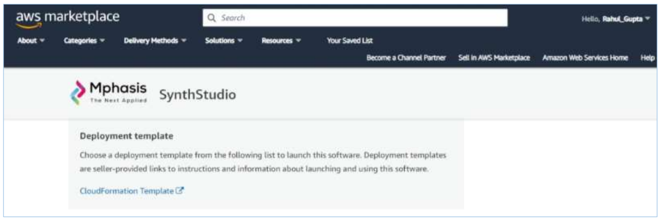
Thank you for subscribing to this product! You can now configure your software.

Subscribe to this software

7. Click on Continue to Launch to initiate the launch process for your product.



8. Click on CloudFormation Template link under Deployment template to access the installation.



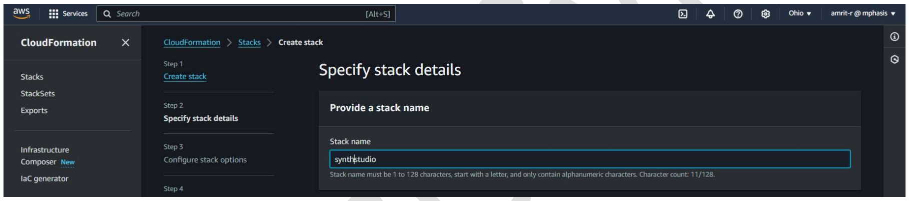
Deployment template

Choose a deployment template from the following list to launch this software. Deployment templates are seller-provided links to instructions and information about launching and using this software.

CloudFormation Template

How to Setup

1. Launch the cloud formation template. You will be redirected to AWS CloudFormation service from the previous step



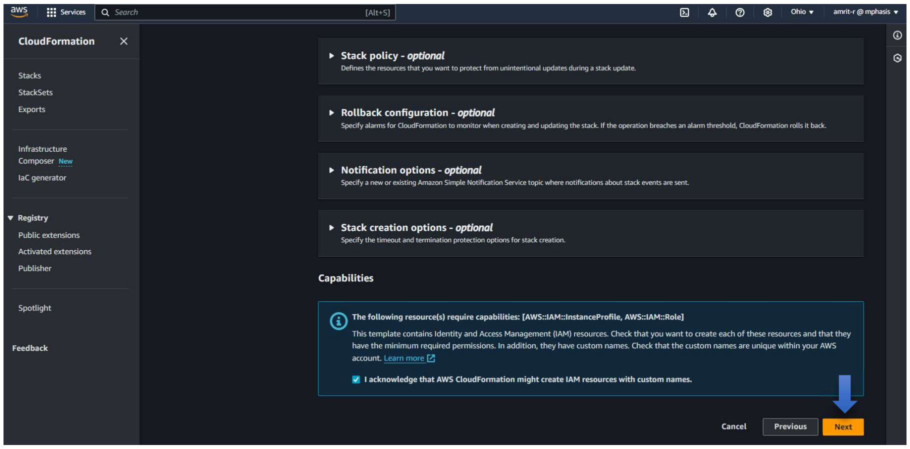
Specify stack details

Provide a stack name

Stack name: synthstudio

Stack name must be 1 to 128 characters, start with a letter, and only contain alphanumeric characters. Character count: 11/128.

2. Fill in the necessary **Parameters** fields required in the cloud formation template.
    a. **Clientname**: The name you give to your application. This name is a unique identifier, and all resources created will be using this name.
    b. **DNSRecord**: Your domain name. This domain name is part of AWS Route 52 Hosted zone.
    c. **InstanceTypeInference**: Options to select your EC2 Instance type for inference. Picking a GPU instance is recommended. Default g4dn.xlarge.
    d. **InstanceTypeTrain**: Options to select your EC2 Instance type while training the model. Picking a GPU instance is recommended. Default g4dn.xlarge.
    e. **Route53HostedZoneID**: The hosted zone ID from route53 where your domain is hosted.
    f. **SSLCertificateARN**: A valid certificate ID from AWS certificate manager which is issued against your domain name.
3. For advanced users only (Optional Step): Fill in the optional parameters if need and click next.



Stack policy - *optional*
Defines the resources that you want to protect from unintentional updates during a stack update.

Rollback configuration - *optional*
Specify alarms for CloudFormation to monitor when creating and updating the stack. If the operation breaches an alarm threshold, CloudFormation rolls it back.

Notification options - *optional*
Specify a new or existing Amazon Simple Notification Service topic where notifications about stack events are sent.

Stack creation options - *optional*
Specify the timeout and termination protection options for stack creation.

Capabilities

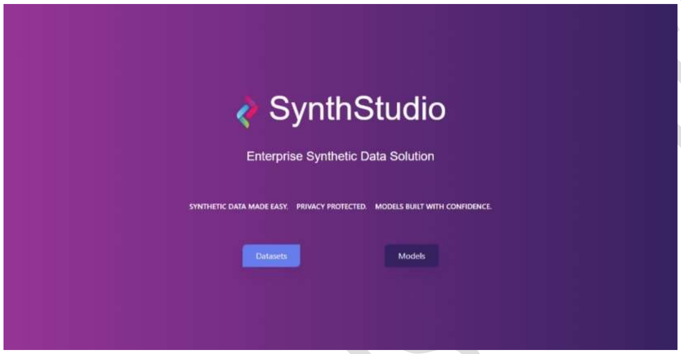
The following resource(s) require capabilities: [AWS::IAM::InstanceProfile, AWS::IAM::Role]
This template contains Identity and Access Management (IAM) resources. Check that you want to create each of these resources and that they have the minimum required permissions. In addition, they have custom names. Check that the custom names are unique within your AWS account. Learn more

☑ **I acknowledge that AWS CloudFormation might create IAM resources with custom names.**

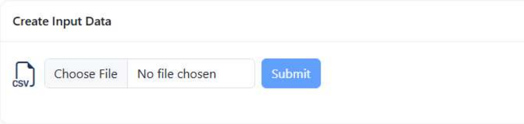
4. Wait till the resources are created and CloudFormation status goes to "CREATE_COMPLETE".
5. (Optional): Modify the load balancer. The Security group attached to the load balancer is open to the public by default and its inbound rules can be modified according to organization policies.
6. Go to your browser and open https://clientname.DNSRecordName.

How to Run

1. Open Homepage


SynthStudio
Enterprise Synthetic Data Solution
SYNTHETIC DATA MADE EASY. PRIVACY PROTECTED. MODELS BUILT WITH CONFIDENCE.
Datasets
Models

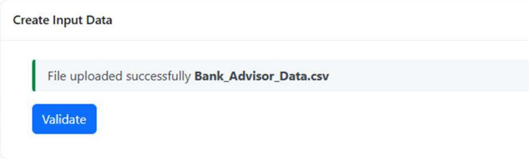
2. Upload a CSV file



Create Input Data

No file chosen

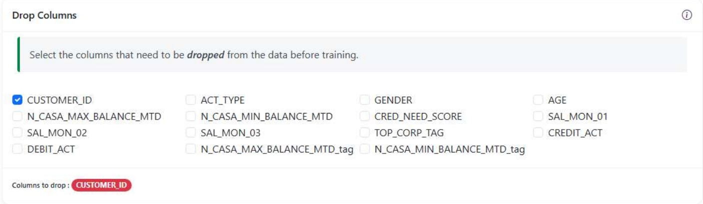
3. Click on Validate



Create Input Data

File uploaded successfully **Bank_Advisor_Data.csv**

4. Select columns to be dropped



Drop Columns

Select the columns that need to be ***dropped*** from the data before training.

☑ CUSTOMER_ID	□ ACT_TYPE	□ GENDER	□ AGE
□ N_CASA_MAX_BALANCE_MTD	□ N_CASA_MIN_BALANCE_MTD	□ CRED_NEED_SCORE	□ SAL_MON_01
□ SAL_MON_02	□ SAL_MON_03	□ TOP_CORP_TAG	□ CREDIT_ACT
□ DEBIT_ACT	□ N_CASA_MAX_BALANCE_MTD_tag	□ N_CASA_MIN_BALANCE_MTD_tag	

Columns to drop : CUSTOMER_ID

5. Select variable categories if not identified correctly. There should be at least one categorical and at least one numerical variable.

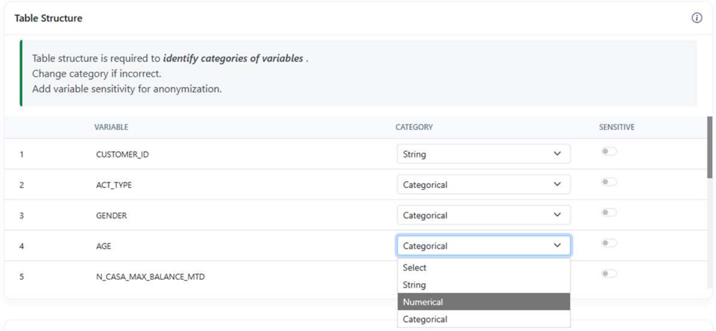
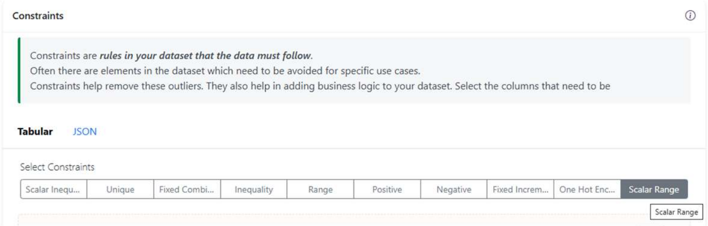


Table Structure

Table structure is required to ***identify categories of variables*** .
Change category if incorrect.
Add variable sensitivity for anonymization.

	VARIABLE	CATEGORY	SENSITIVE
1	CUSTOMER_ID	String	○
2	ACT_TYPE	Categorical	○
3	GENDER	Categorical	○
4	AGE	Categorical	○
5	N_CASA_MAX_BALANCE_MTD		○

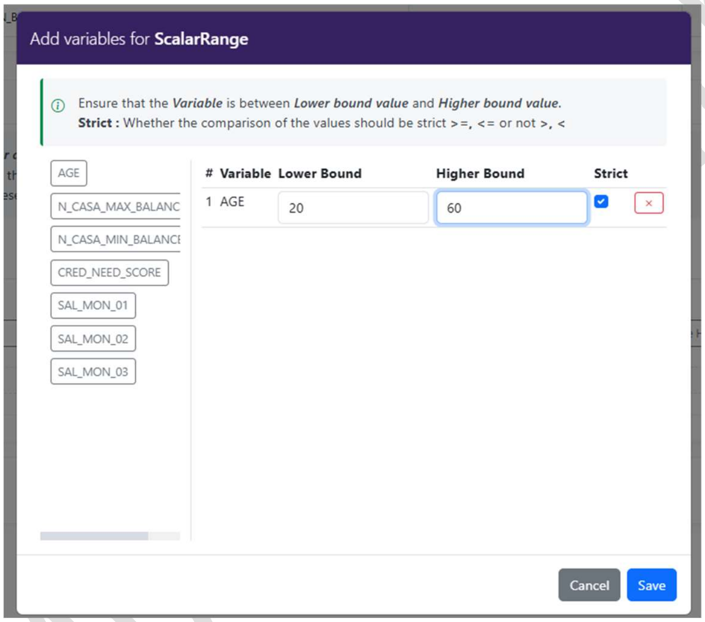
6. Add constraints based on requirement



Constraints

Constraints are ***rules in your dataset that the data must follow.***
Often there are elements in the dataset which need to be avoided for specific use cases.
Constraints help remove these outliers. They also help in adding business logic to your dataset. Select the columns that need to be

7. Add Constraint details

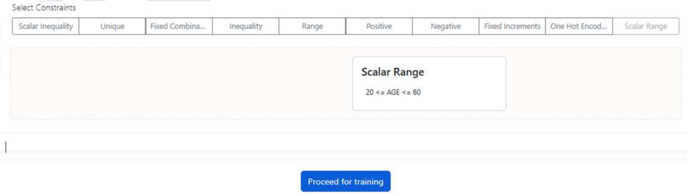


Add variables for **ScalarRange**

Ensure that the *Variable* is between ***Lower bound value*** and ***Higher bound value***.
Strict : Whether the comparison of the values should be strict >=, <= or not >, <

#	Variable	Lower Bound	Higher Bound	Strict
1	AGE	20	60	☑

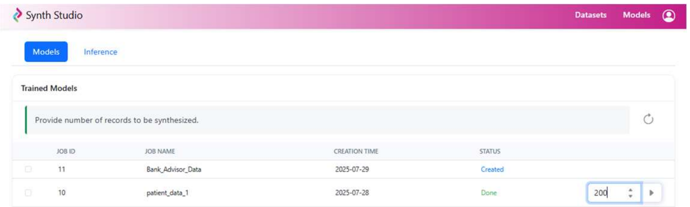
8. Proceed for training



Scalar Range

20 <= AGE <= 60

9. Go to Models page to check the training progress. After training is Done, add number of rows to be generated and proceed

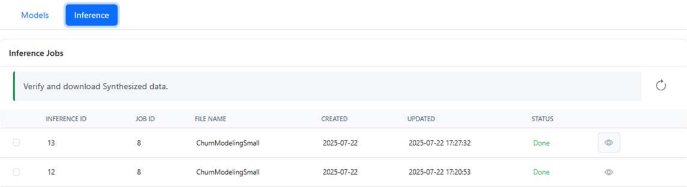


Trained Models

Provide number of records to be synthesized.

	JOB ID	JOB NAME	CREATION TIME	STATUS	
□	11	Bank_Advisor_Data	2025-07-29	Created	
□	10	patient_data_1	2025-07-28	Done	200

10. Under Inference tab, check the inference progress. After inference is Done, proceed to look at the data

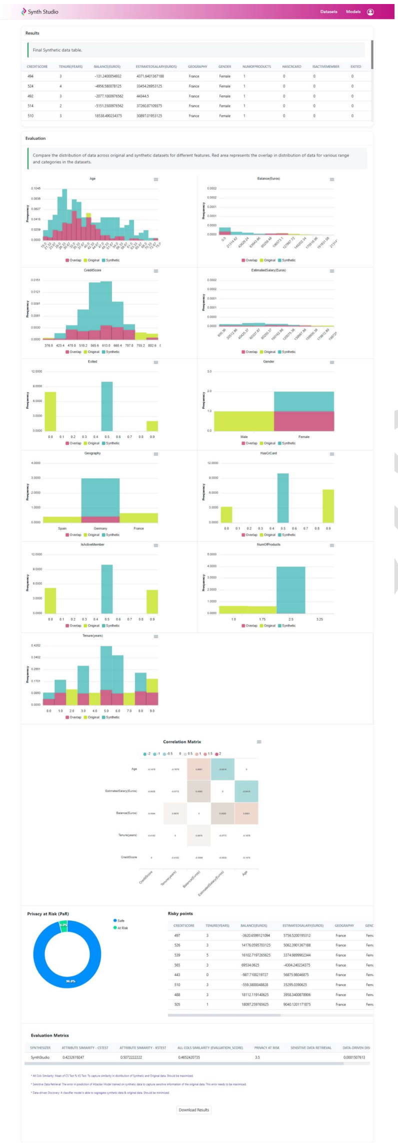


Inference Jobs

Verify and download Synthesized data.

	INFERENCE ID	JOB ID	FILE NAME	CREATED	UPDATED	STATUS
□	13	8	ChurnModelingSmall	2025-07-22	2025-07-22 17:27:32	Done
□	12	8	ChurnModelingSmall	2025-07-22	2025-07-22 17:20:53	Done

11. On the results page you can have a quick look at the generated data, evaluation metrics and download the results



Results

Final Synthetic data table.

CREDITSCORE	TENURE(YEARS)	BALANCE(EUROS)	ESTIMATEDSALARY(EUROS)	GEOGRAPHY	GENDER	NUMOFPRODUCTS	HASCRCARD	ISACTIVEMEMBER	EXITED
494	3	-131.2400054932	4371.6401367188	France	Female	1	0	0	0
524	4	-4956.580078125	33454.26953125	France	Female	1	0	0	0
492	3	-2077.1000976562	44344.5	France	Female	1	0	0	0
514	2	-5151.3500976562	37260.87109375	France	Female	1	0	0	0
510	3	18538.490234375	30897.01953125	France	Female	1	0	0	0

Evaluation

Compare the distribution of data across original and synthetic datasets for different features. Red area represents the overlap in distribution of data for various range and categories in the datasets.

Privacy at Risk (PaR)

Risky points

CREDITSCORE	TENURE(YEARS)	BALANCE(EUROS)	ESTIMATEDSALARY(EUROS)	GEOGRAPHY	GENC
497	3	-3620.6599121094	5756.5200195312	France	Fem
526	3	14176.0595703125	5062.3901367188	France	Fem
539	5	16102.7197265625	3374.9899902344	France	Fem
565	3	69534.0625	-4304.240234375	France	Fem
443	0	-987.7100219727	56875.98046875	France	Fem
510	3	-559.3800048828	35295.0390625	France	Fem
488	3	18112.119140625	3958.3400878906	France	Fem
505	1	18097.259765625	9040.1201171875	France	Fem

Evaluation Metrics

SYNTHESIZER	ATTRIBUTE SIMIARITY - CSTEST	ATTRIBUTE SIMIARITY - KSTEST	ALL COLS SIMILARITY (EVALUATION_SCORE)	PRIVACY AT RISK	SENSITIVE DATA RETRIEVAL	DATA-DRIVEN DIS
SynthStudio	0.4232619247	0.5072222222	0.4652420735	3.5		0.0001507613

* All Cols Similarity: Mean of CS Test % KS Test. To capture similarity in distribution of Synthetic and Original data. Should be maximized.

* Sensitive Data Retrieval: The error in prediction of Attacker Model trained on synthetic data to capture sensitive information of the original data. This error needs to be maximized.

* Data-driven Discovery: A classifier model is able to segregate synthetic data & original data. Should be minimized.