

Synthetic Data Generation: A Process Paradigm

Why just 'throwing an algorithm' at data is not enough...

Sundaresk Sankaran
Sherrine Eid
SAS Institute



Synthetic Data Generation

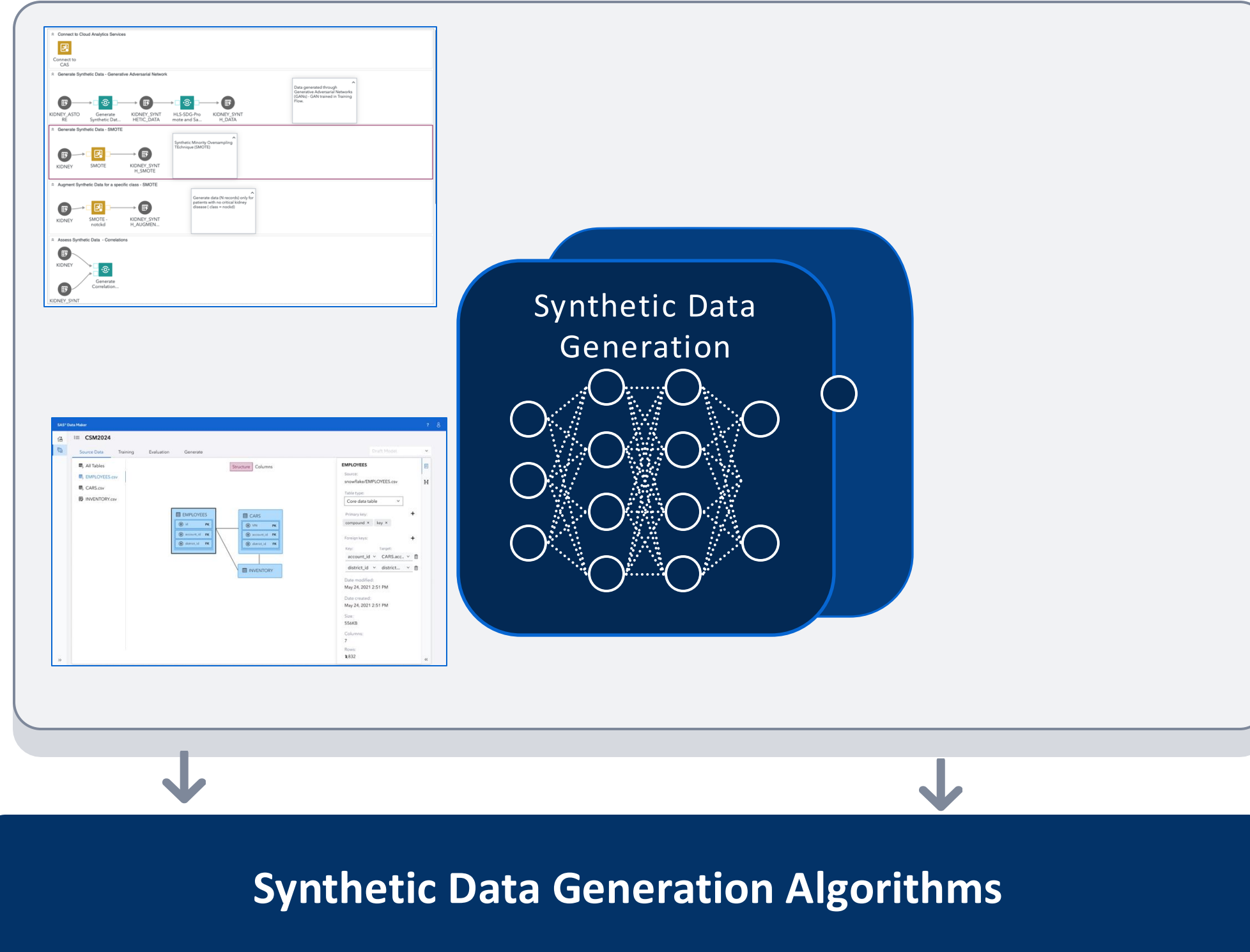
Enables Analytics Innovation

Restricted Data

Imbalanced Data

Rare Events

Insufficient Data



Synthetic Data

ID1

ID2

ID3

ID4

IDN

- Generate high quality synthetic data
- Validate the quality of the generated data
- 'Market' data downstream

Synthetic data can address these challenges



Imbalanced Data

Increasing model bias

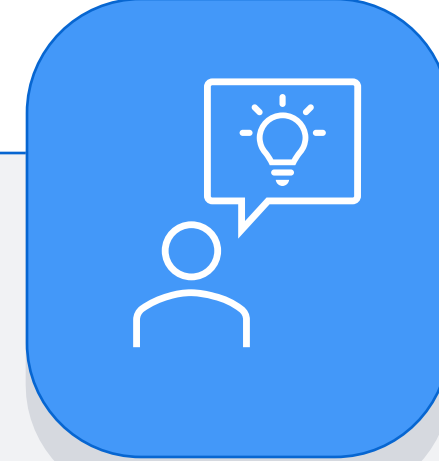
Your dataset is imbalanced, with some classes or categories underrepresented



Rare Events

Testing edge cases

You need to test how your solution handles rare or unusual scenarios that are not well-represented in your existing data.



Restricted Data

Privacy concerns limiting use

You're working with sensitive or personal data but can't use real data due to privacy concerns or regulations.



Insufficient Data

For model training

You have limited or incomplete datasets that don't provide enough examples for training or testing models effectively.

Yet, Risks Abound

‘Throwing an algorithm at data’ is not enough

Data Quality

- Lack of Pre-processing
- Poor knowledge of original data

Privacy Risks

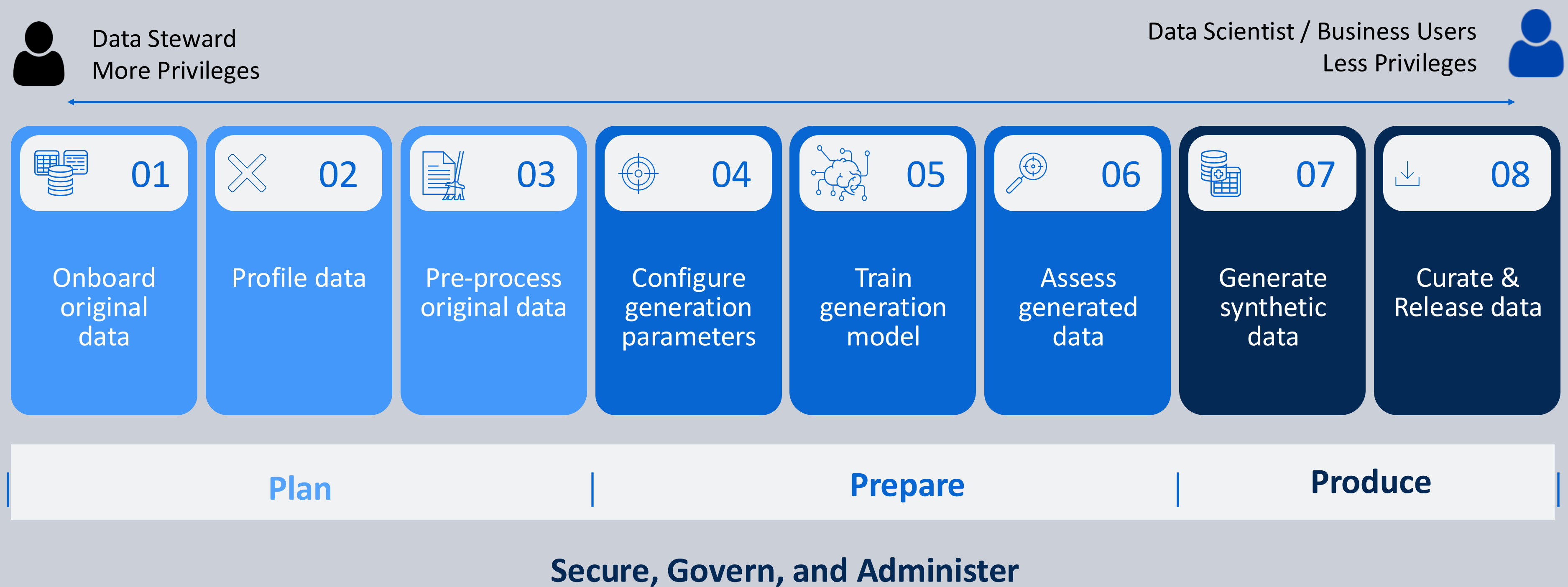
- Data Leakage
- Re-identification Risk

Governance Risks

- Who gets to generate data?
- Who gets to use data?
- How secure is the data?

A Process Paradigm for Synthetic Data

Some common steps recommended for synthetic data generation

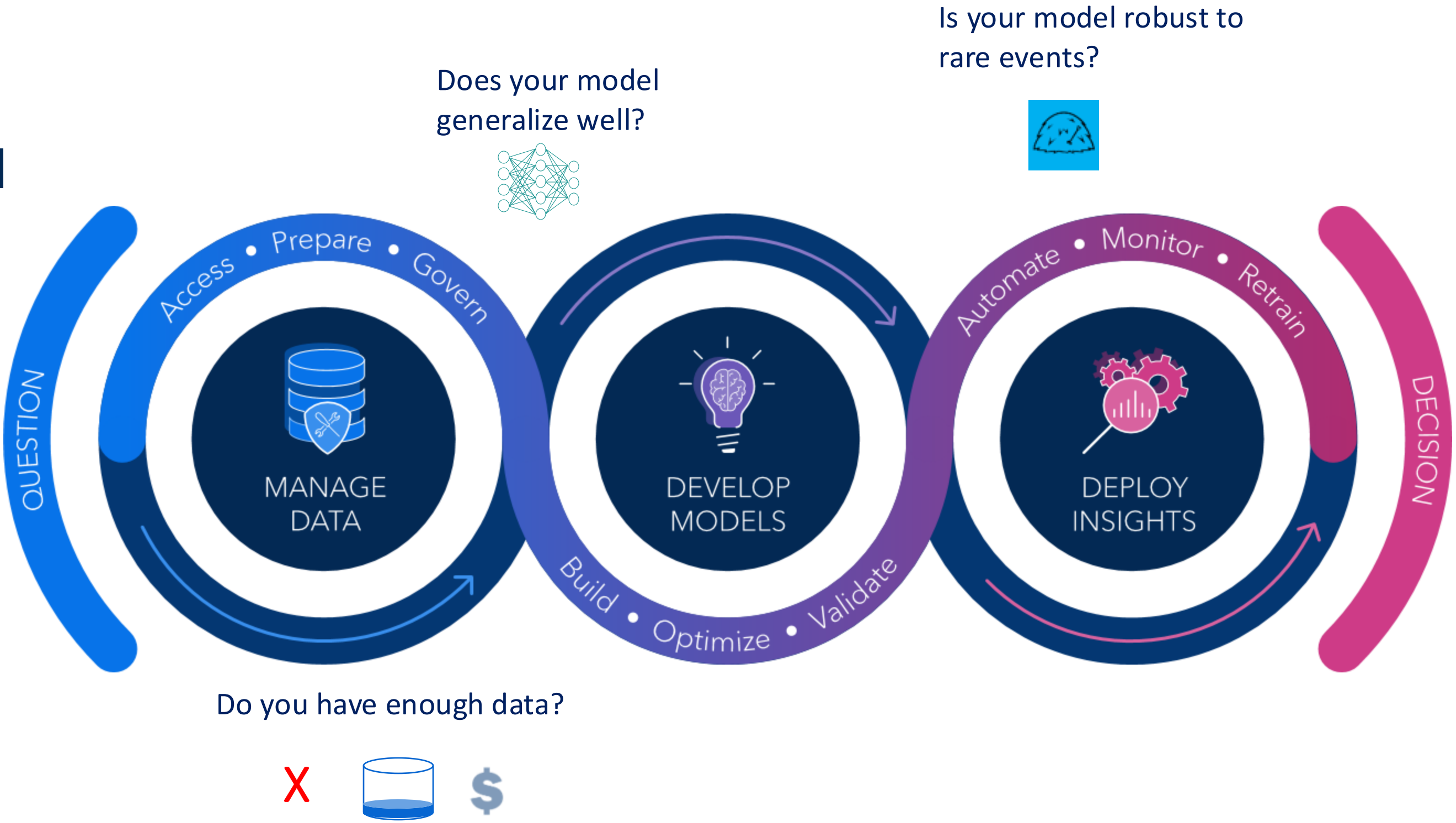


Synthetic Data Raises Questions, Across the Life Cycle!

Is there sufficient representation among sensitive groups?



Do privacy issues prohibit access to or usage of real data?



Pre-generation tasks

Right Data begets Right Data

Data Assessment and Quality Checks

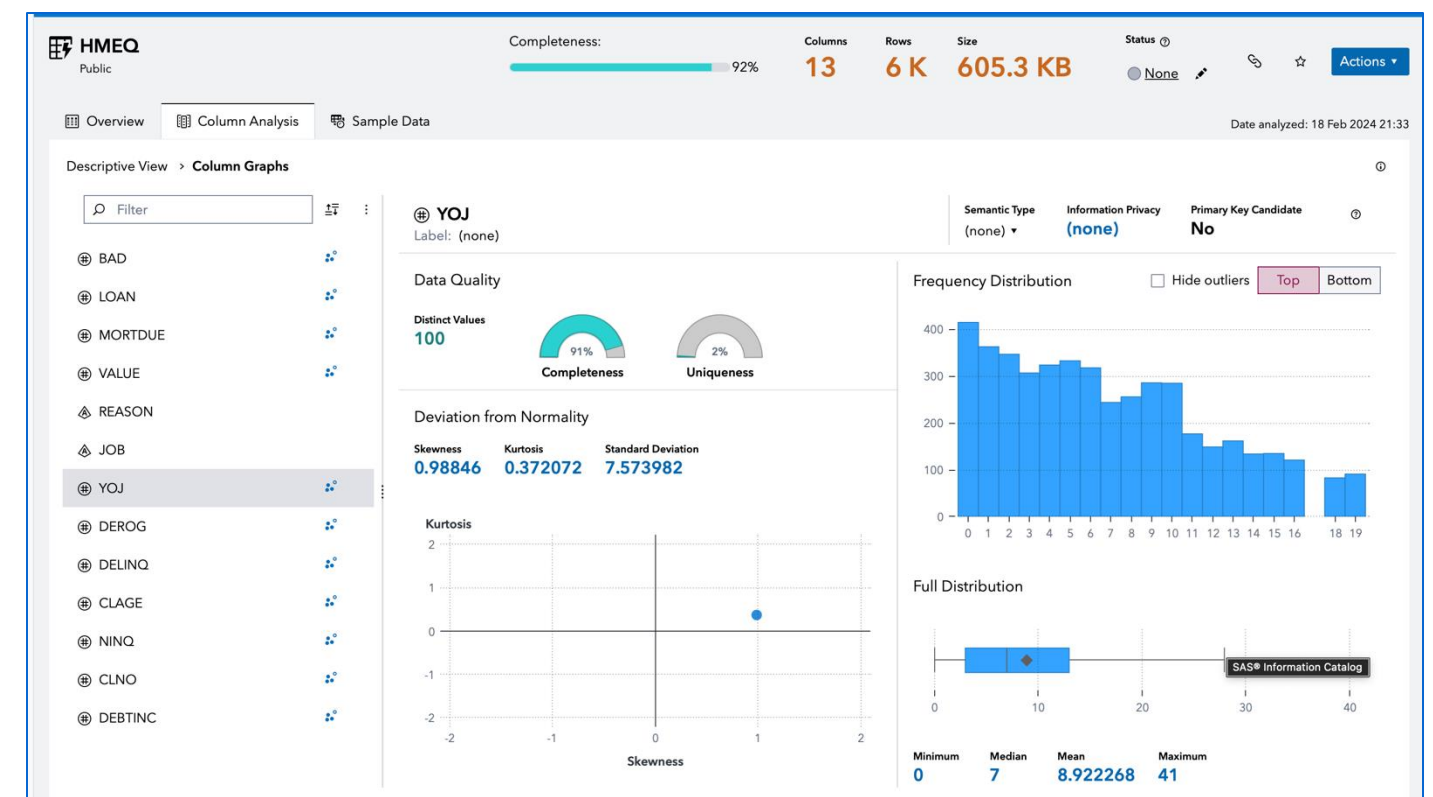
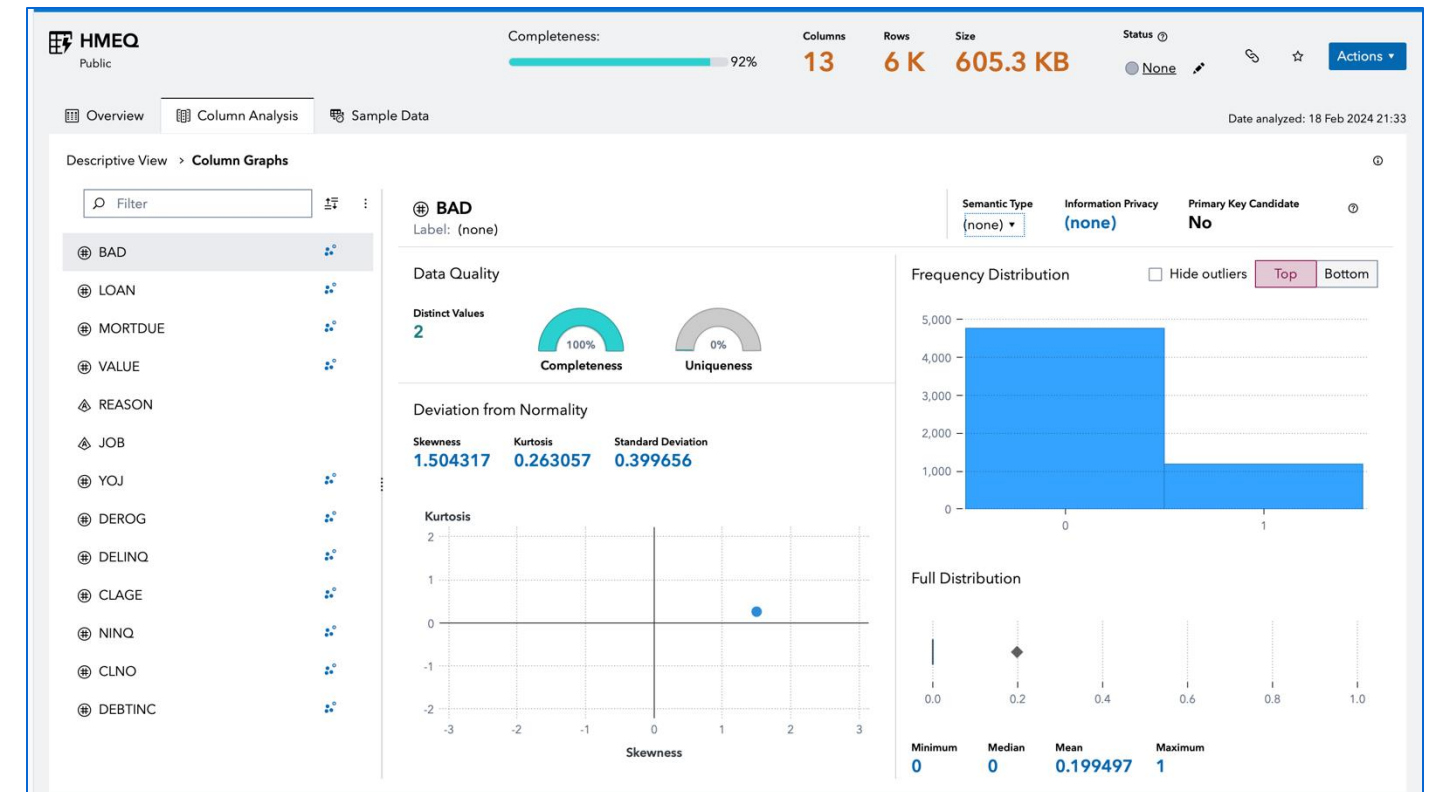
- De-identification
- Imputing missing values
- Identifying anomalies

Data Transformation

- Variable transformation and encoding
- Normalisation

'Learning' the Distribution

- Joint Probability Distributions
- Clustering & 'right' number of clusters





New ▾

Search



Quick Access



Discover Information Assets

Reports

Build Models

Manage Models

Build Decisions

Develop Code and Flows

Risk Explorer

Build Custom Graphs

All

Recent



My Favorites

My Folder

SAS Content

Shared

Recommended

Recycle Bin

My Folder

Sort by:

Name



Customer Propensity A...

Customer Segmentation

git

My Snippets

My Tasks

My Tasks

SAS Videos

Census_Train_Initial

Select an item to see its information.

/*****

This block represents
code provided by Bruno -
helps
us create a table of

A PROJECT-CENTRIC structure is crucial.

SAS® Data Maker

PROJECTS

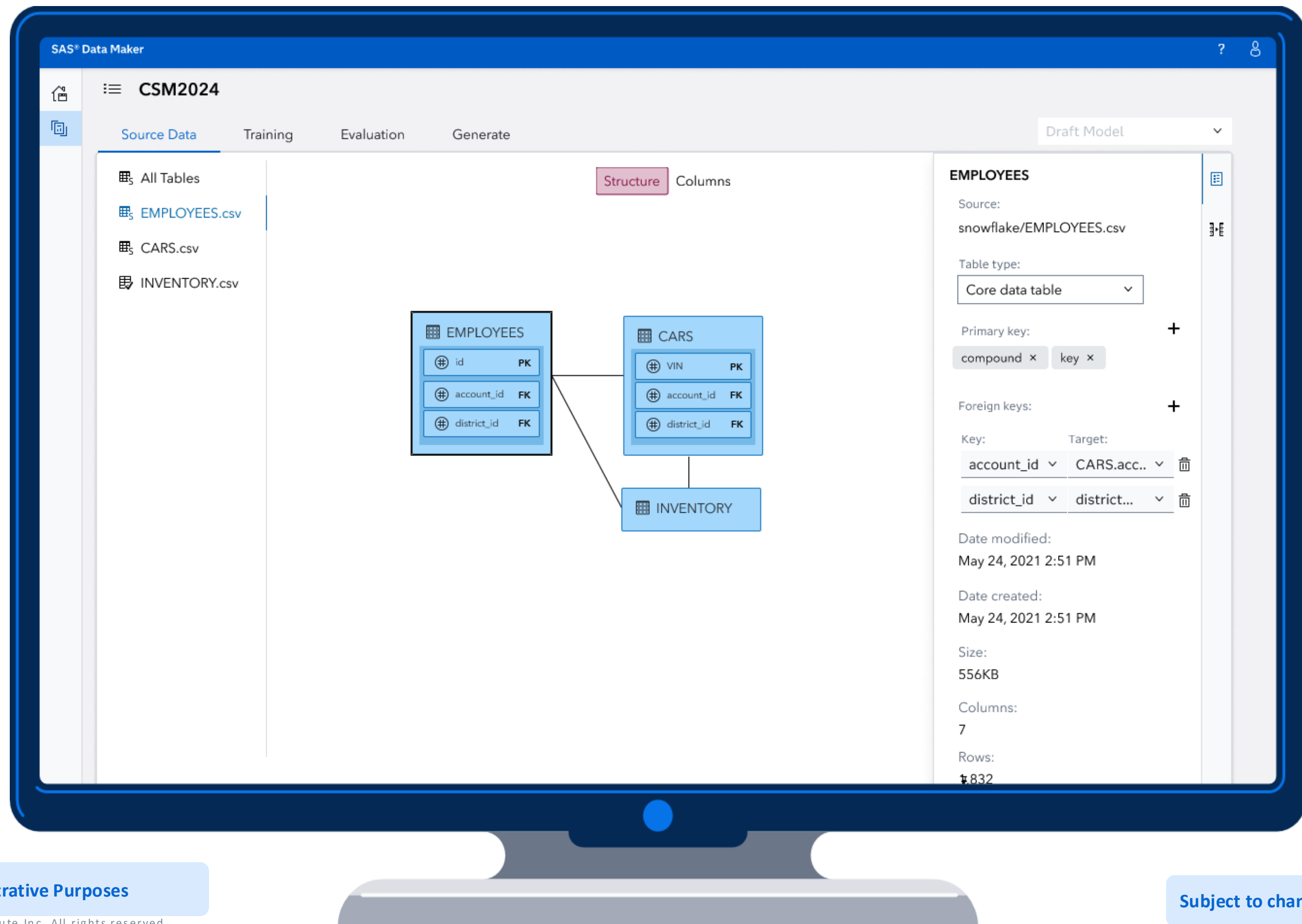
Search Project Name or Source Data Set

☐	Project Name	Source data	Status	Modified	Modified by
☐	User activity log	user_activity_log.dat	✓	Sat Jun 26 2004 23:00:55	cacook
☐	2024 Sales Performance	sales_performance2024.dat	✓	Sat Oct 18 2003 18:43:24	bjmiller
☐	Sensor data stream	sensor_data_streamQ3.dat	✓	Tue Feb 22 2000 08:40:27	supope
☐	Historical Temperature Data	hist_temp_data.dat	✓	Mon Oct 18 2038 10:11:22	tjsimmons
☐	Product review sentiment	product_review.dat	✓	Thu Jan 02 2048 16:12:25	prenoyds
☐	Logistics tracking	logistics_tracking.dat	✓	Sat Jun 26 2004 23:00:55	cacook
☐	Geo Data Locations	geo_location_data.dat	✓	Sat Oct 18 2003 18:43:24	conolan
☐	Real Time Inventory	inventory.dat	✓	Tue Feb 22 2000 08:40:27	wawright
☐	Demographic data	demographic.dat	✓	Mon Oct 18 2038 10:11:22	sacazi
☐	Traffic flow data	traffic_flow_data.dat	✓	Thu Jan 02 2048 16:12:25	hjberry
☐	Credit Card Transactions	cc_transactions.dat	✓	Sat Jun 26 2004 23:00:55	kjharrist
☐	Financial Forecast data	financial_forecast_data.dat	✓	Sat Oct 18 2003 18:43:24	klgarrett
☐	Economic risk indicators	economic_risk_data.dat	✓	Tue Feb 22 2000 08:40:27	bffrankie
☐	Capital flows	capital_flows.dat	✓	Mon Oct 18 2038 10:11:22	aamills
☐	Asset price simulation	asset_price_simulation.dat	✓	Thu Jan 02 2048 16:12:25	hjkarry
☐	Neural Network Weights	neural_network_weights.dat	✓	Sat Jun 26 2004 23:00:55	cacook
☐	Classification dataset	classification.dat	✓	Sat Oct 18 2003 18:43:24	eferrick

>>

Screenshot For Illustrative Purposes

sas



Screenshot For Illustrative Purposes

Copyright © SAS Institute Inc. All rights reserved.

Subject to change





Steps

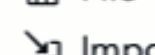


SAS Steps : Shared

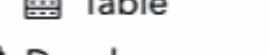
Data (Input and Output)



Export



File

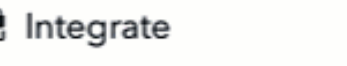


Import Files



Table

Develop



Python Program



SAS Program

Transform Data

Integrate

Enrichment

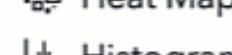


Geocode Data



Verify with Loqate

Visualize Data



Bar Chart



Bar-Line Chart



Box Plot

Bubble Map

Bubble Plot

Choropleth Map

Heat Map

Histogram

Line Chart

Pie Chart

Scatter Map

Start Page

Synthetic Data Generation - HMEQ.flw x +



Run



Cancel



Add ▾

View ▾

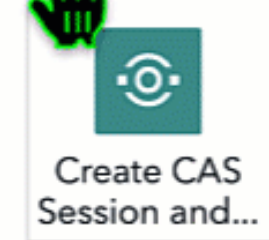


Flow

Generated Code

Submitted Code and Results

Connect to CAS (if required)



Baseline



Generate Synthetic Data using Generative Adversarial Networks (GAN)

Generate Synthetic Data using Synthetic Minority Oversampling Technique (SMOTE)

Example SMOTE application to boost a class

Assess Generated Data (GAN)

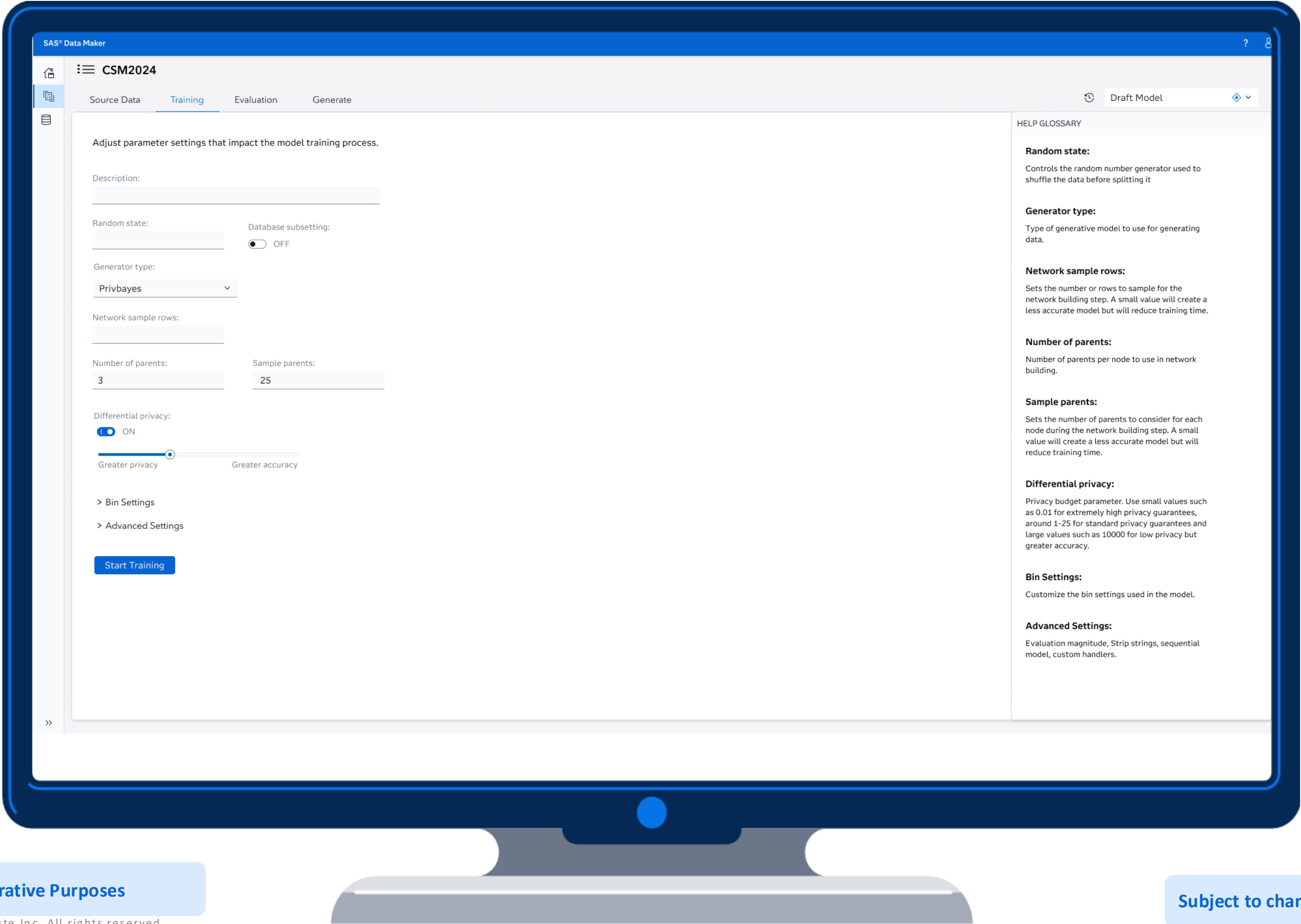
Assess Generated Data (SMOTE)

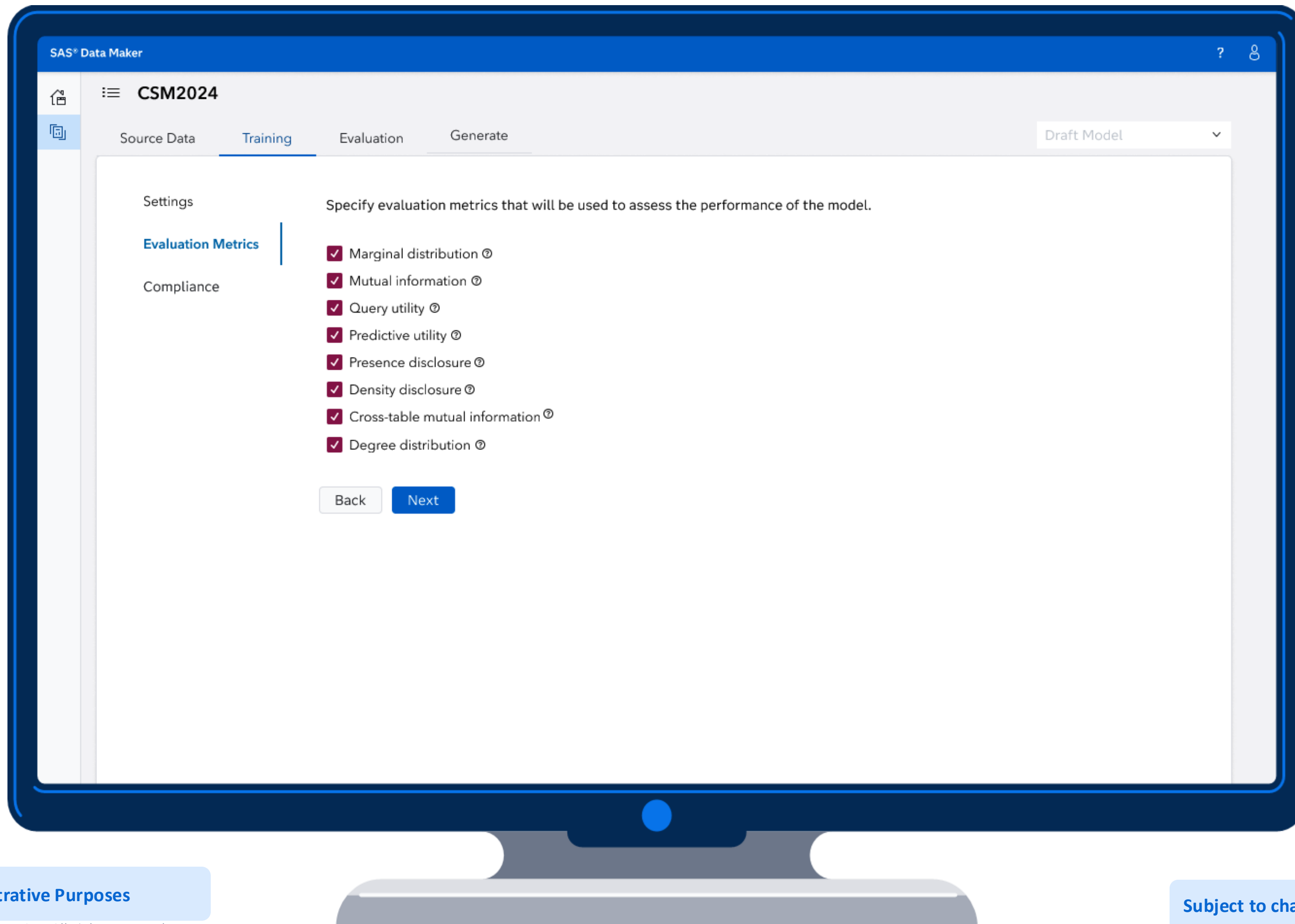
Persist table

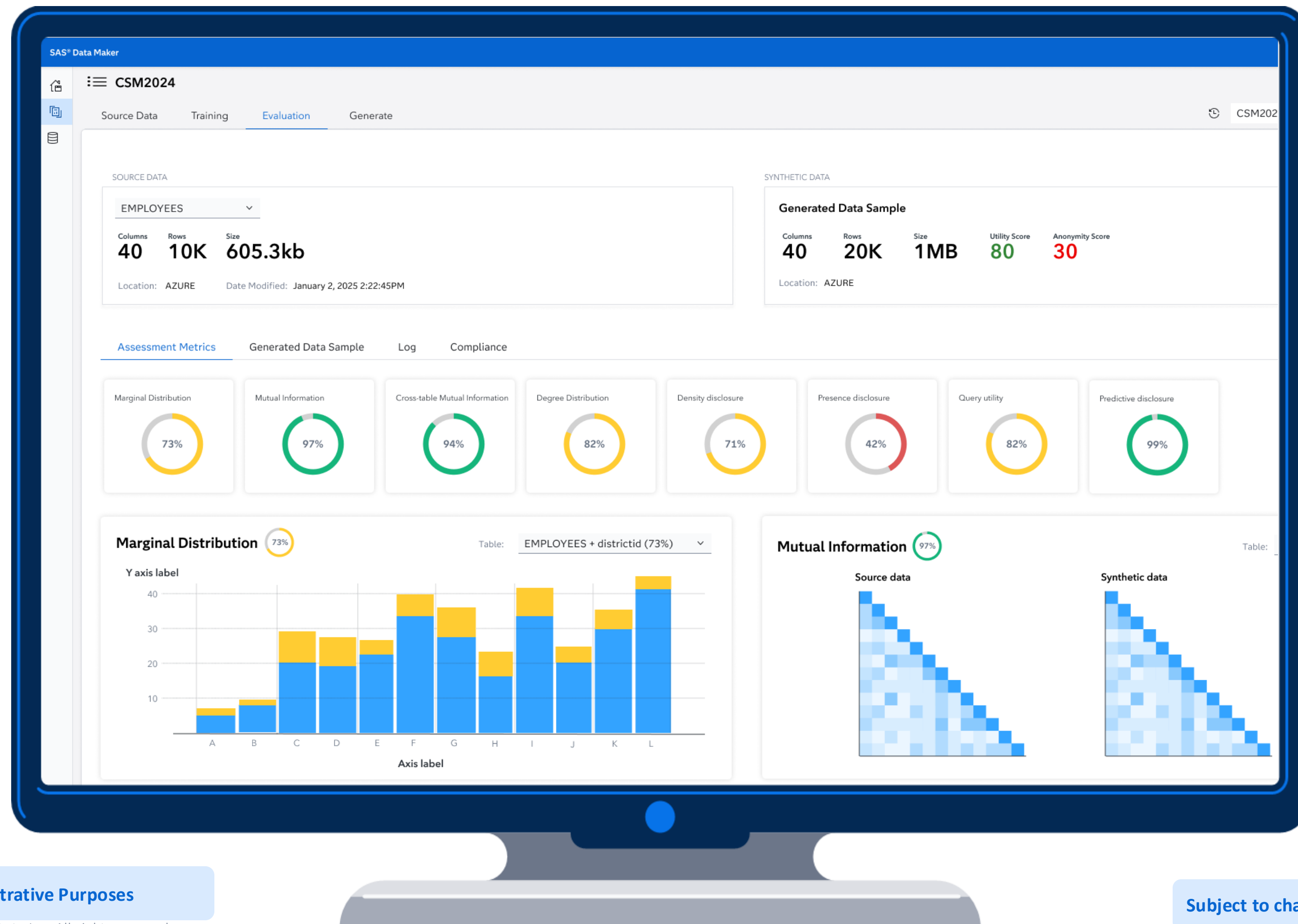
Create Analytical Base Table (ABT)

Create Analytical Base Table (ABT) based on SMOTE

Wind Down (if required)





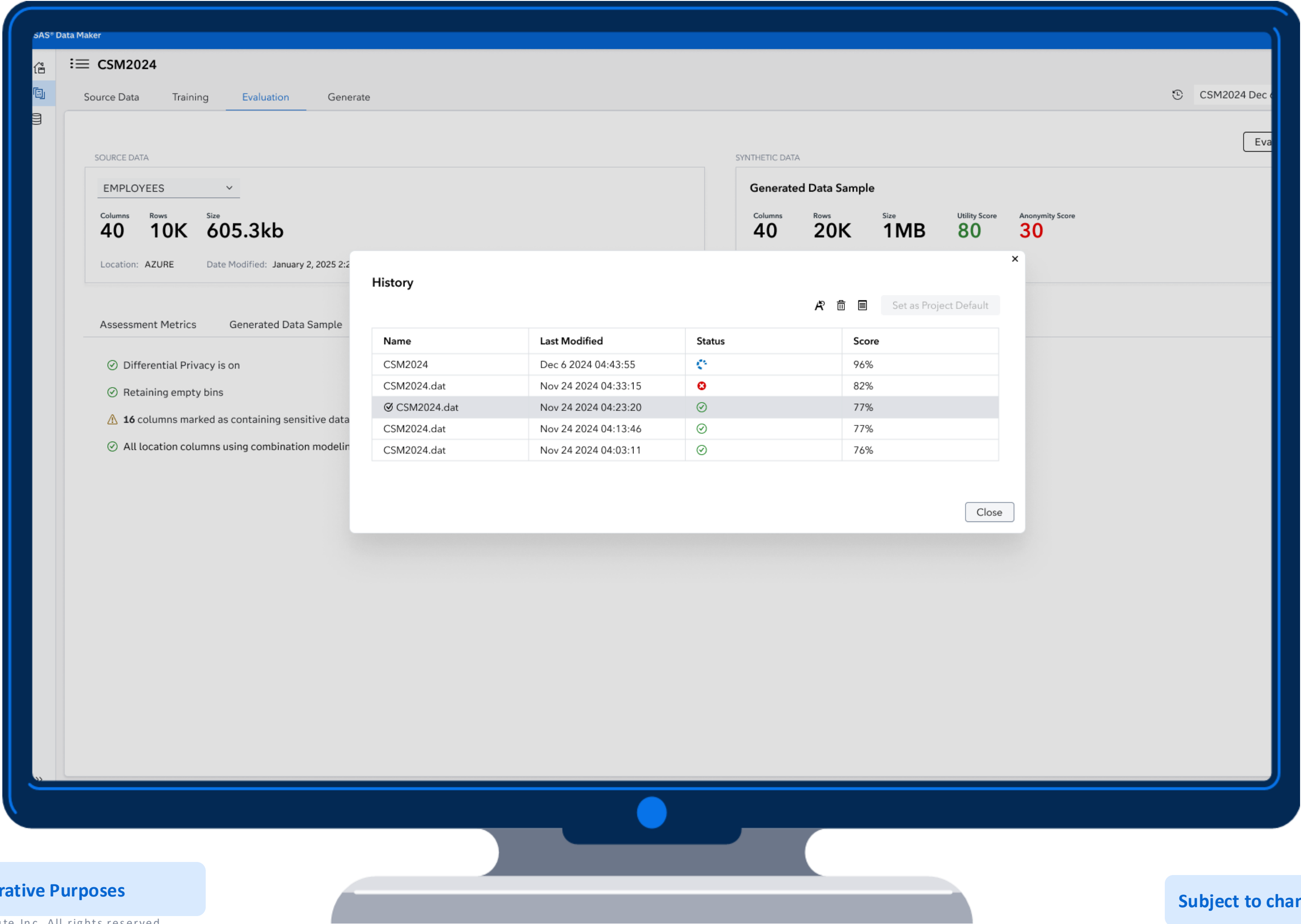


Screenshot For Illustrative Purposes

Copyright © SAS Institute Inc. All rights reserved.

Subject to change

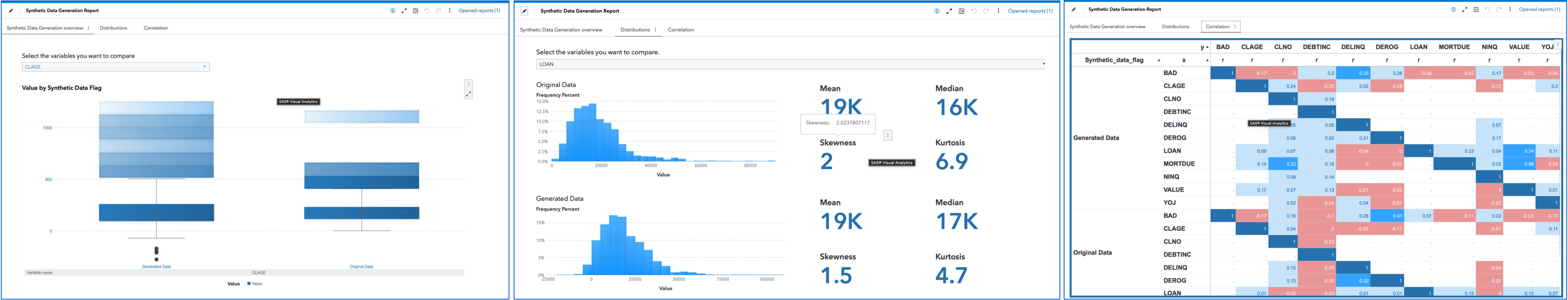




Screenshot For Illustrative Purposes

Post-generation tasks

I'll still poke it to see if it's the right data



Assessment

- Is my generated data the same as original data?

Leakage (and Leaching)

- Is it still possible to identify a random record?

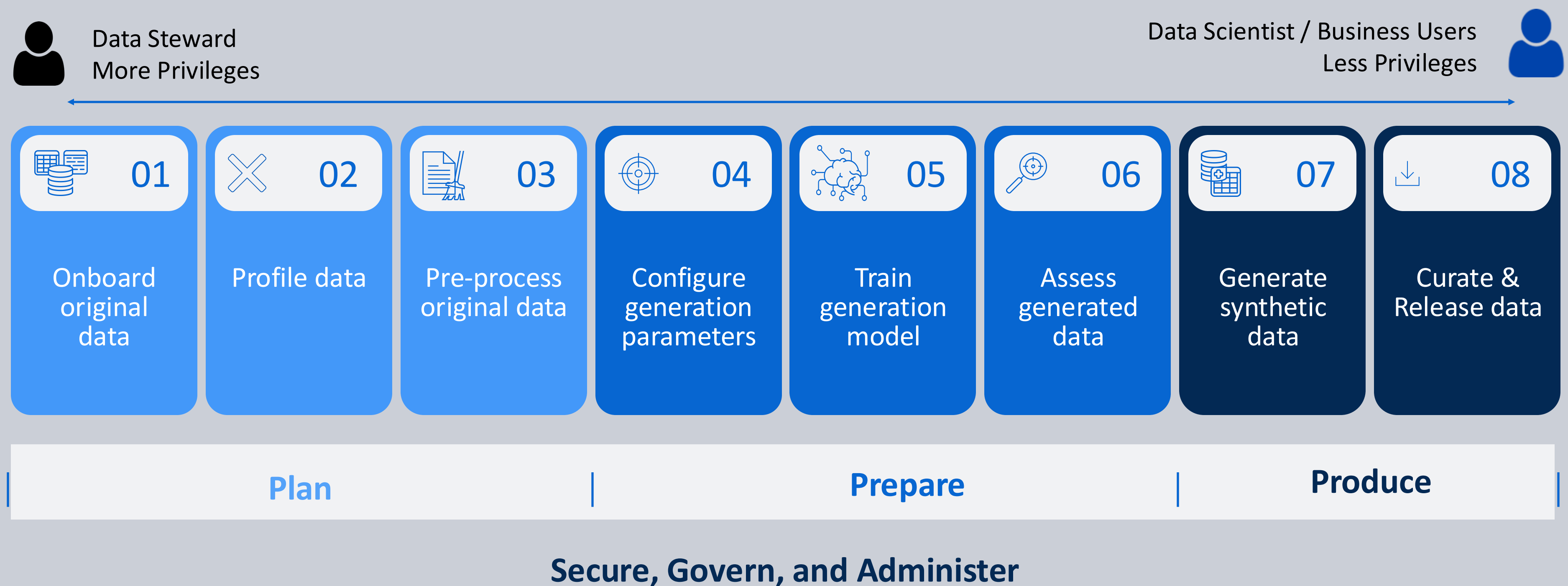
Human-in-the-loop

- Impose limits and business rules



A Process Paradigm for Synthetic Data

Some common steps recommended for synthetic data generation



To Summarise,

A Process Paradigm is required for Synthetic Data Generation

- Analytical tasks are rarely one-switch activities; upstream and downstream considerations require treatment
- User Personas are important considerations while designing a synthetic data generation process
- Governance mechanisms help mitigate risks.
- Security and Protection-first

**Thank You.
Questions?**

