Synthetic Data Generation PoC

Jan-March 2025

Generalitat de Catalunya

Agenda

1. Generalitat of Catalunya - CTTI
    - Introduction

2. Synthetic Data PoC
    - Overview and Scope
    - Implementation and Evaluation Results
    - Challenges and Lessons Learned

Syntethic Data Generation PoC

1. Generalitat of Catalunya – CTTI

Introduction

CTTI, **Telecommunications and Information Technology Center of Generalitat de Catalunya**, is responsible for designing, building, coordinating and deploying technological projects to provide solutions to the departments and different bodies of the Public Administration.

Currently some of the digital transformation initiatives of CTTI are related with data, namely the implementation of a **centralized data platform called PTD** – Plataforma Transversal de Datos, and in top of that, **an advanced analytics platform to support AI and GenAI use cases.** Another reference project is the development of an integrated platform for the Social services of the government, named **eSocial**.

These initiatives led to the need of having a platform to generate anonymized data, to be used for example in advanced analytics, e.g., AI model training, and for software development integration tests.

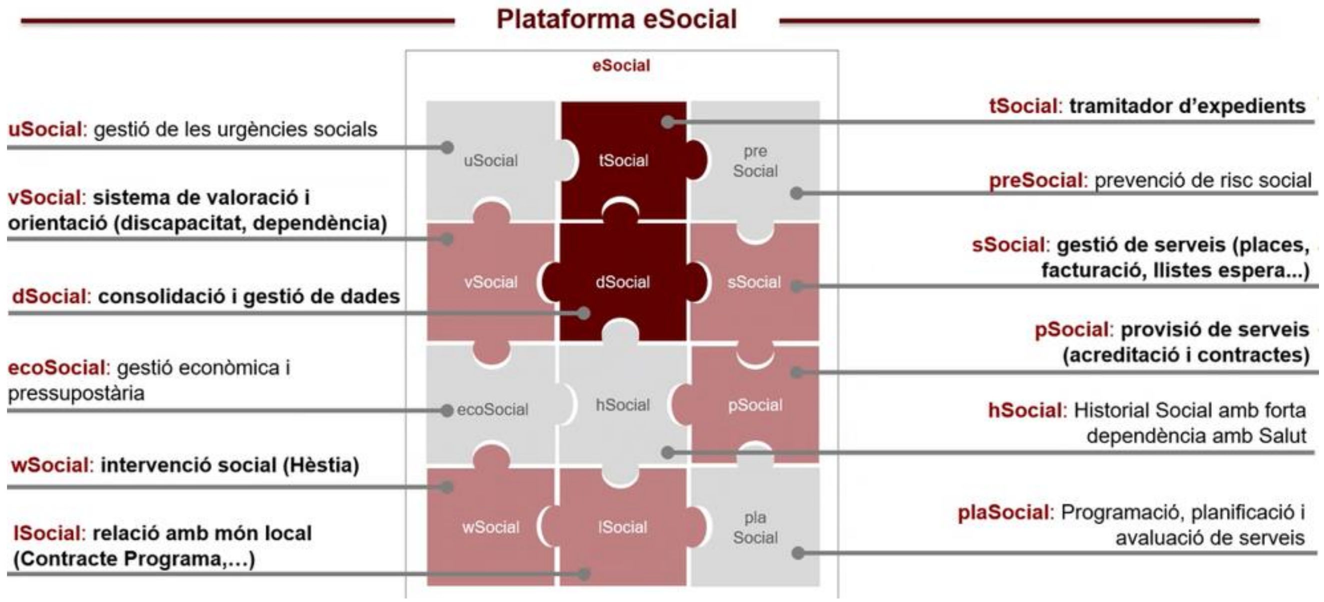
Generalitat de Catalunya
Centre de Telecomunicacions i Tecnologies de la Informació

Syntethic Data Generation PoC

2.1 Overview

Introduction

The **eSocial information ecosystem** is composed by a mosaic of 12 information systems that supports different areas and processes of the Department of Social Rights, and it's being modernized and integrated. For the integration tests exists a need of usage of real data to test all the scenarios, but it's necessary that this data is fully anonymized, keeping the representativeness.



- Management of social emergencies
- Assessment and orientation system (disability, dependency)
- Data consolidation and management
- Economic and budgetary management
- Social intervention
- Relationship with the local community
- Case file processing
- Social risk prevention
- Service management (places, billing, waiting lists...)
- Social history with strong dependence on Health
- Service programming, planning, and evaluation

For this **PoC** was selected data from **tSocial** system (Case file processing) to be fully anonymized.

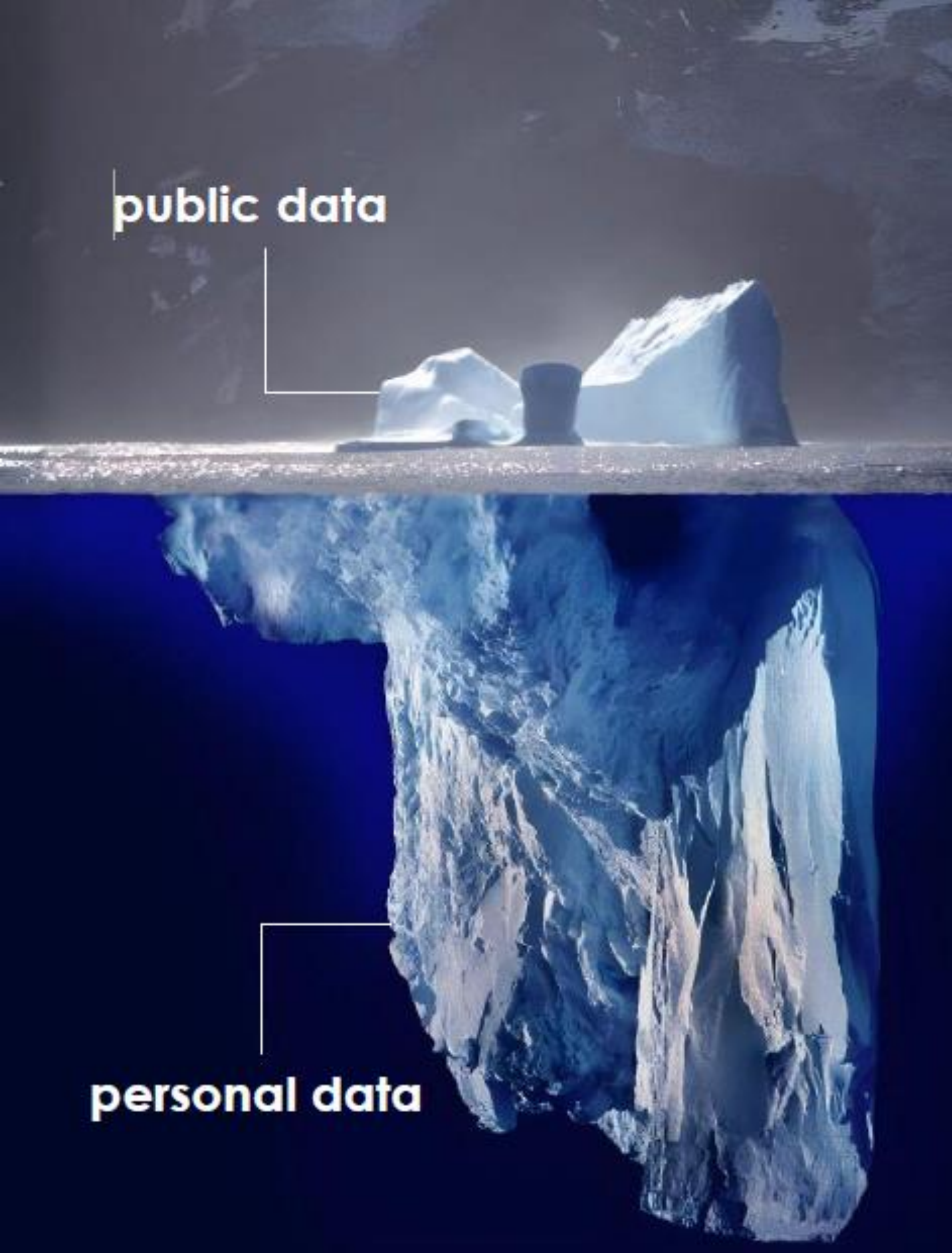
MOSTLY·AI
is the **leader** in this space

Europe
USA
Active

30+
Team members

15+
Nationalities

$31M+
Total funding



95% of the world's valuable data remains locked away and underutilized

At **MOSTLY AI**, we're unlocking this potential by providing privacy safe access for everyone

What is AI generated Synthetic Data?

Synthetic data is an artificial version of your real data

Synthetic data looks and feels like real data, but because it's artificially created it's very flexible

- Bigger or smaller
- Rebalanced
- Augmented
- Imputed

Synthetic data is NOT mock data

Synthetic data is much more sophisticated than mock data

It retains the structure and statistical properties (like correlations) of your real data

You can confidently use it in place of your real data

Synthetic data is safer than legacy data anonymization

Legacy data anonymization can be dangerous

Synthetic data points have no 1:1 relationship to the original data

Synthetic data is a much safer alternative

The Open-source Synthetic Data SDK sets a new standard

→ `pip install mostlyai`

Versatility: Full-Featured SDK

- **Broad Data Support**
  - Mixed-type data (categorical, numerical, geospatial, text, etc.)
  - Single-table, multi-table, and time-series
- **Multiple Model Types**
  - TabularARGN for SOTA tabular performance
  - Fine-tune HuggingFace-based language models
  - Efficient LSTM for text synthesis from scratch
- **Advanced Training Options**
  - GPU/CPU support
  - Differential Privacy
  - Progress Monitoring
- **Automated Quality Assurance**
  - Quality metrics for fidelity and privacy
  - In-depth HTML reports for visual analysis
- **Flexible Sampling**
  - Up-sample to any data volumes
  - Conditional generation by any columns
  - Re-balance underrepresented segments
  - Context-aware data imputation
  - Statistical fairness controls
  - Rule-adherence via temperature
- **Seamless Integration**
  - Connect to external data sources (DBs, cloud storages)
  - Fully permissive open-source license

TABULAR & LANGUAGE

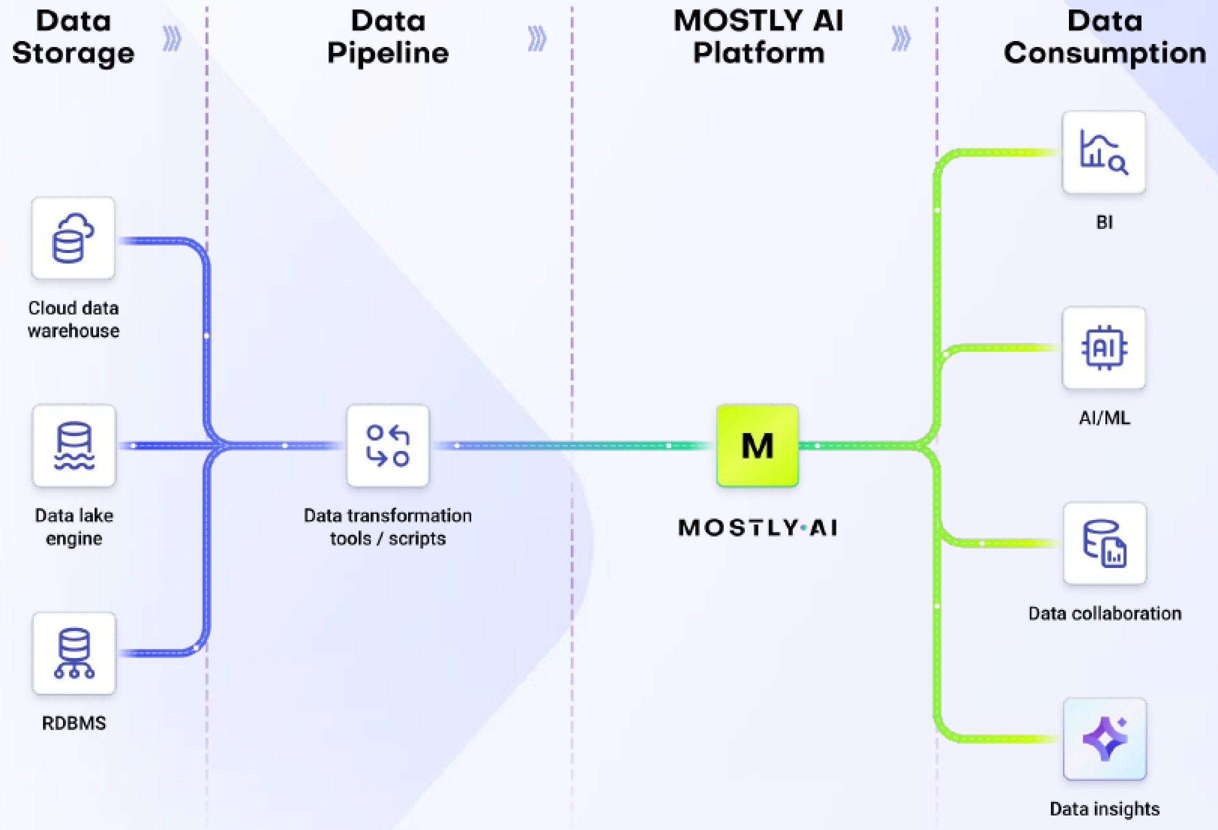
Efficiency: SOTA Accuracy while 10-200x faster than anyone else!

MOSTLY·AI

MOSTLY AI Synthetic Data

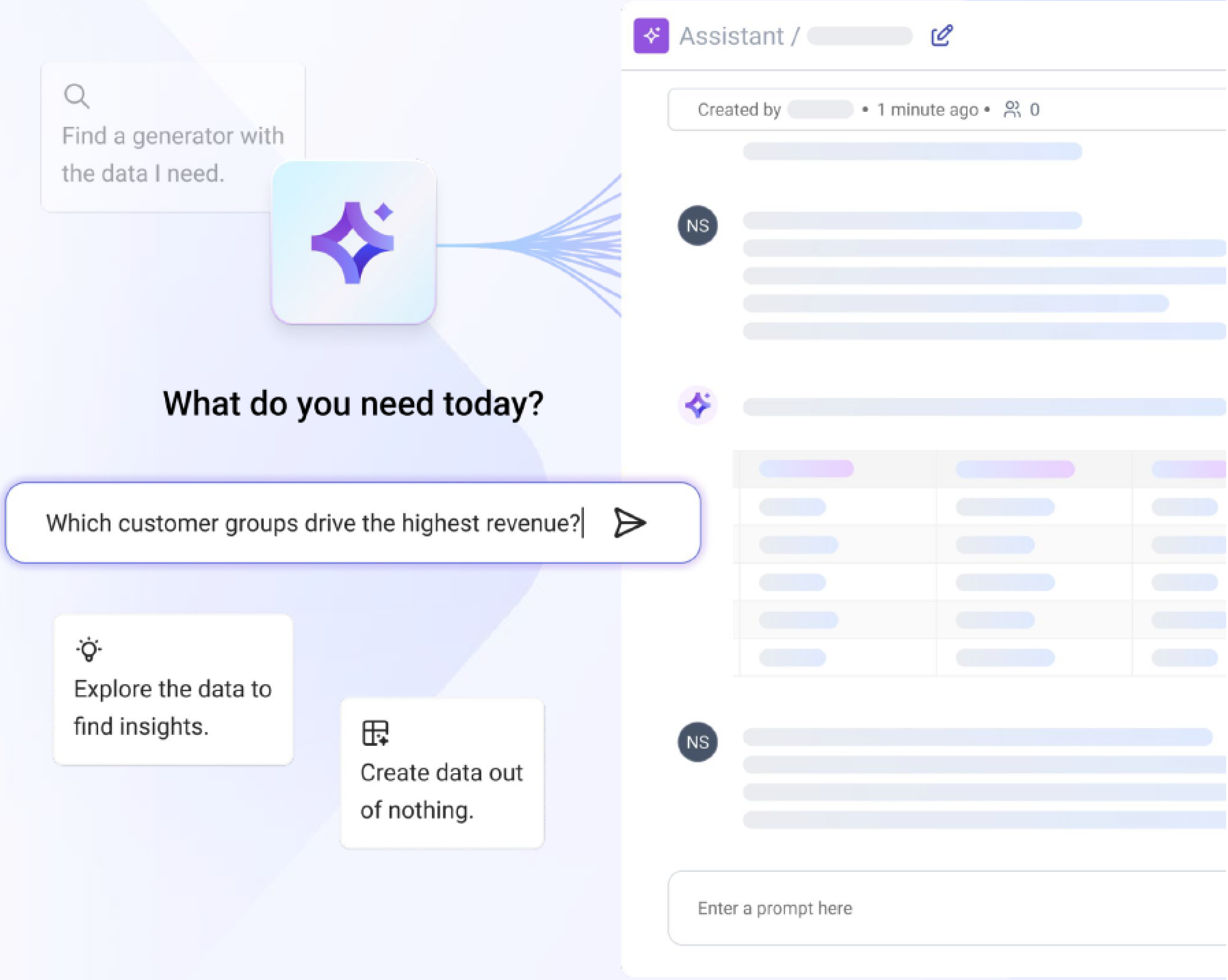
Access + Insights

- **Connect MOSTLY AI to** your **data**
- **Create a Generator** with that data as an input and share it with others
- **Data Consumers** discover available Generators on the Platform
- And **use Generators** to flexibly **create synthetic data** for their use cases
- Unlike real data, synthetic data has **no privacy limitations**



MOSTLY AI Assistant
Access + Insights

- A **natural language interface** on top of **privacy-safe** synthetic data as the **key** to **data democratization**
- Conduct **exploratory data analysis** (EDA) by simply talking to your data
- Create **charts** and **visualizations**
- **Train ML models** on your data and leverage explainable AI (XAI) to understand model decisions



MOSTLY AI enables a wide range of use cases

Data Sharing

- Generate **anonymous** synthetic data to **share** your data easily

AI/ML development

- Speed up your AI/ML development initiatives and **get to value faster**

Customize Language Models

- **Fine-tune LLMs** with privacy-preserving synthetic text

Self-service analytics

- Use the MOSTLY AI Assistant to **understand** your **data** and generate insights

Software Testing & QA

- Leverage synthetic data for development and testing in **non-production environments**

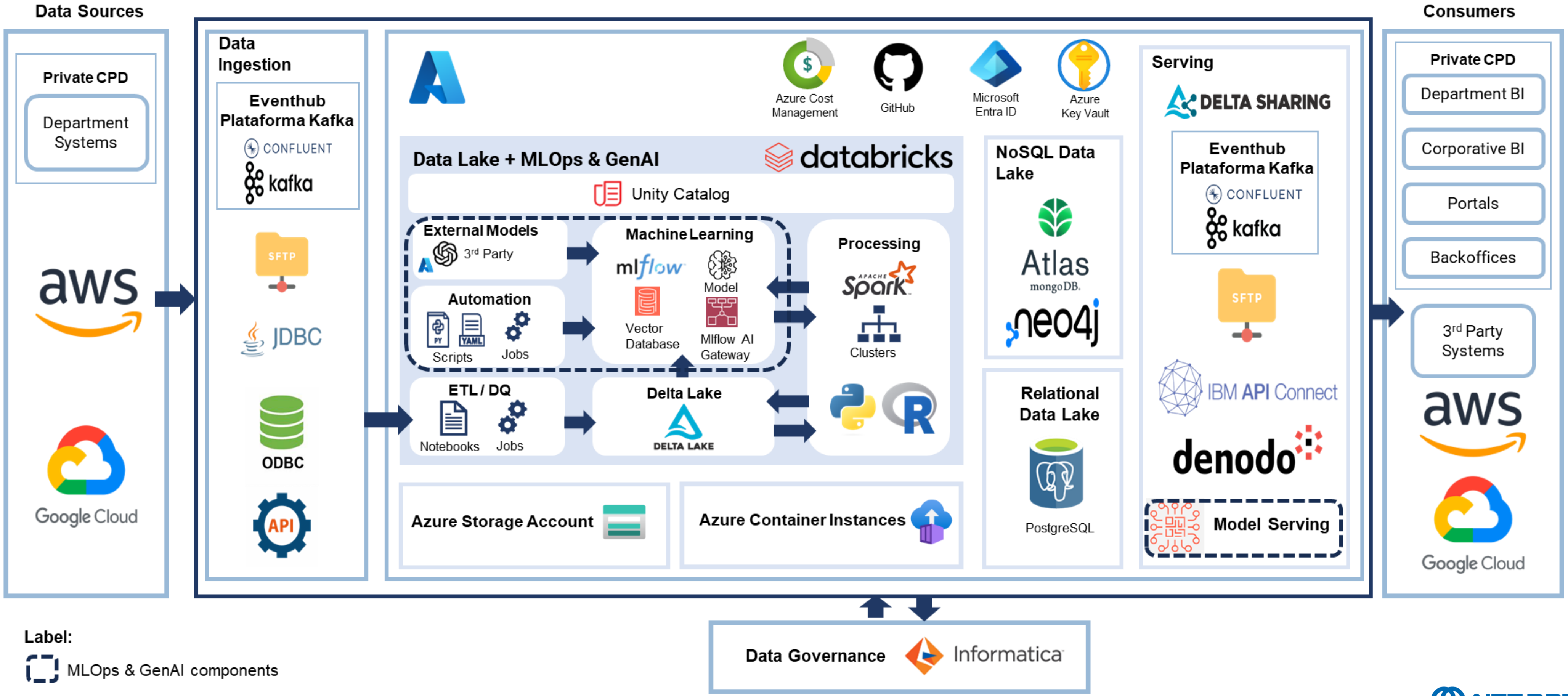
MOSTLY·AI

Syntethic Data Generation PoC

2.2 Implementation and Results

Introduction

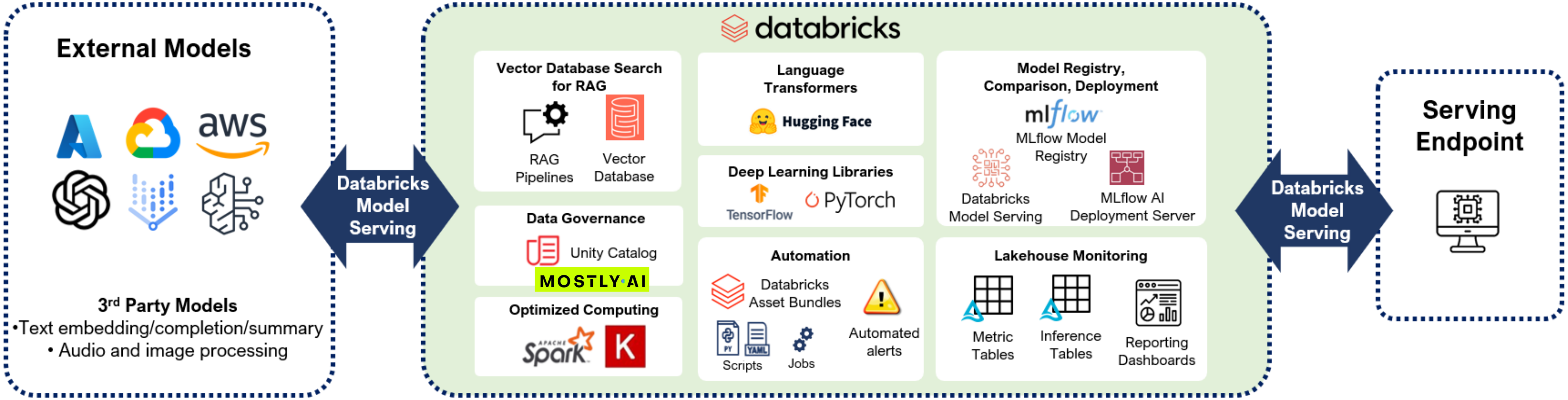
The **PoC** was implemented in the **development environment of the brand-new data platform of CTTI** composed of several components for data ingestion, processing and storage, and for advanced analytics (AI and GenAI), leveraging Databricks ecosystem implemented.



NTT DATA

Introduction

The **Mostly.AI SDK (Python SDK)** was installed in Databricks development environment, adding capacities of generation of synthetic data, in the Data Governance module.

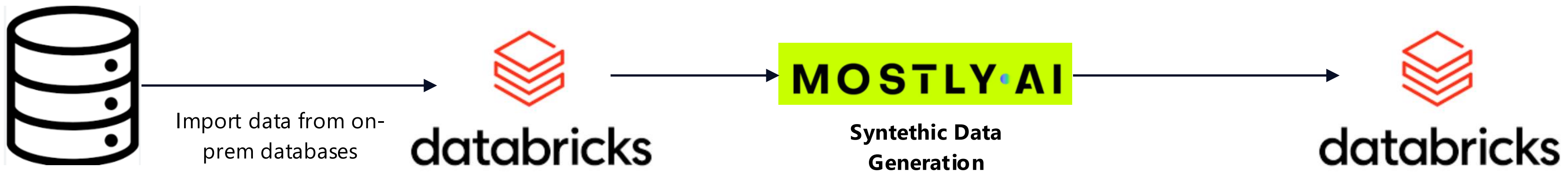


Scope of the PoC

Generate synthetic data to be used for software development integration tests in a pre-production environment, **for the modernization of the platform eSocial** of the Generalitat of Catalunya.

Approach

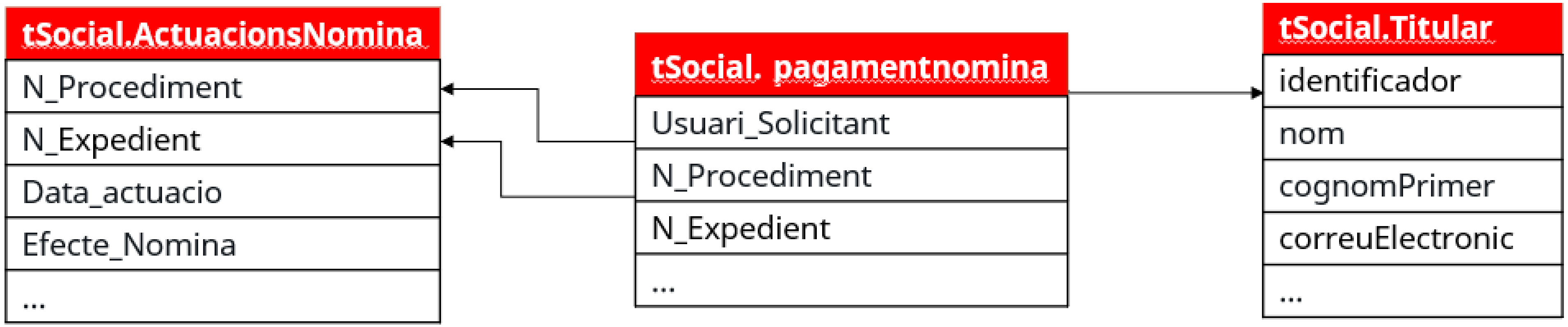
- ✓ Generate a synthetic dataset that maintains the characteristics of the original data, the coherence between the data, and the relationship between tables.
- ✓ All data/attributes in the dataset will be replaced by synthetic data – full privacy protection.
- ✓ Verify that the use of Mostly.AI meets the needs for generating synthetic data across cross-database tables.
- ✓ In this PoC 3 tables will be used, loaded in Databricks, from on-prem database tSocial.



Scope of the PoC

For this **PoC** was selected **3 tables** from **tSocial** system (approximately **7.5M data points**) to be fully anonymized:

- *ActuacionsNomina;*
- *PagamentNomina*
- *Titular*

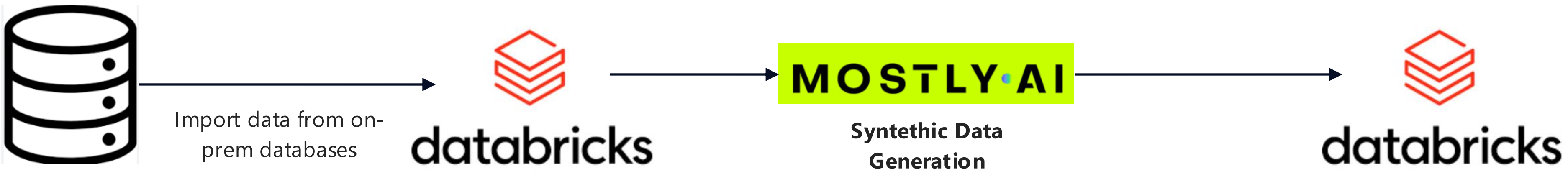


# columns: 4	**# columns: 12**	**# columns: 39**
# rows: 190.142	**# rows: 278.373**	**# rows: 86.569**
Size: 1.1 MB	**Size: 1.69 MB**	**Size: 6.2 MB**

Mostly.AI process

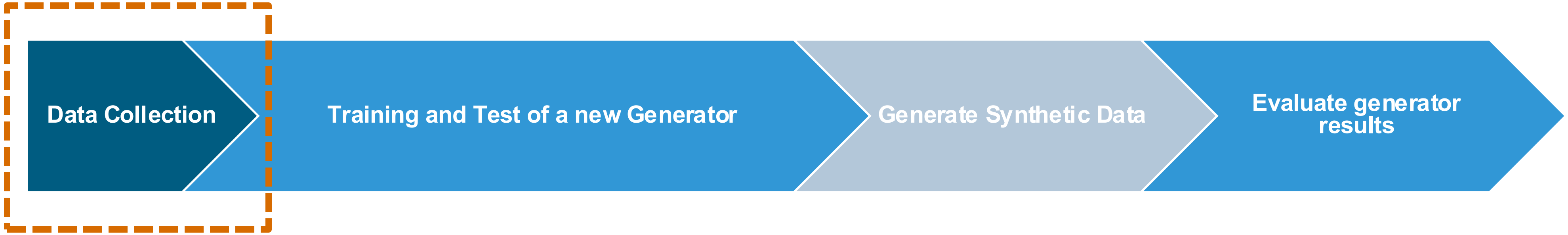
The process to generate synthetic data with Mostly.AI involves the following steps, that can be done using the **Synthetic Data SDK** (open source) or **App**.

- **Data Collection**
- **Training and Test of a new Generator**
- **Generate Synthetic Data**
- **Evaluate generator results**



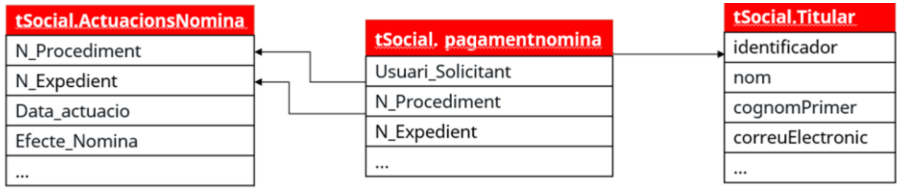
A **Mostly.AI Generator** bundles the training of Generative AI Models and the definition of metadata about tabular data, including table schemas, table relationships and data types. They leverage advanced AI techniques such as Transformers, GANs, Variational Autoenconders and Autoregressive Networks to ensure accuracy, privacy and flexibility.

Mostly.AI Implementation Details



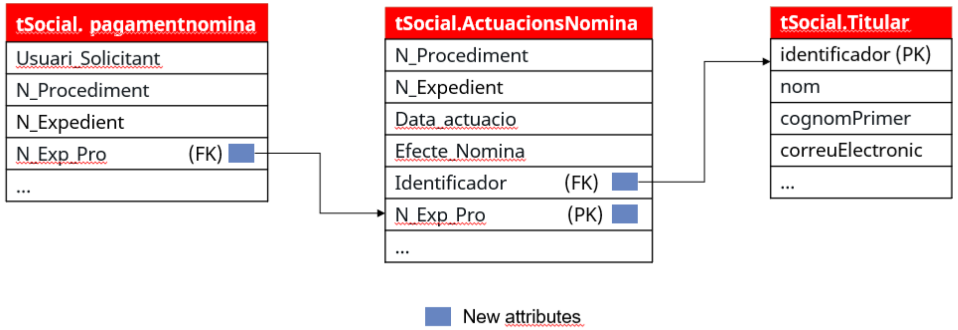

```
▶ ✓ 2 days ago (<1s)                                                    7

from pyspark.sql.functions import countDistinct, lit, concat

df_rel_exp_pro_titular = (df_pagamentnomina.groupBy(
    df_pagamentnomina.N_Expedient,
    df_pagamentnomina.N_Procediment,
    df_pagamentnomina.Usuari_solicitant
).agg(
    countDistinct("N_Expedient", "N_Procediment", "Usuari_solicitant").alias("distinct_user_count")
    )
.withColumn("N_Exp_Pro", concat(df_pagamentnomina.N_Expedient, lit("_"), df_pagamentnomina.N_Procediment))
.drop("distinct_user_count")
)

#Add N_Exp_Pro and Usuari_solicitant to dataframe ActuacionsNomina
df_actuacionsnomina_processed = (
    df_actuacionsnomina
        .withColumn("N_Exp_Pro", concat(df_actuacionsnomina.N_Expedient, lit("_"), df_actuacionsnomina.N_Procediment))
        .join(df_rel_exp_pro_titular, on=["N_Exp_Pro"], how="left")
        .select(df_actuacionsnomina["*"], df_rel_exp_pro_titular["N_Exp_Pro"], df_rel_exp_pro_titular["Usuari_solicitant"].alias("identificador"))
)

#Add N_Exp_Pro to dataframe PagamentDomina
df_pagamentnomina_processed = (
    df_pagamentnomina
        .withColumn("N_Exp_Pro", concat(df_pagamentnomina.N_Expedient, lit("_"), df_pagamentnomina.N_Procediment))
)

▸ df_actuacionsnomina_processed: pyspark.sql.dataframe.DataFrame = [N_Expedient: string, N_Procediment: string ... 4 more fields]
▸ df_pagamentnomina_processed: pyspark.sql.dataframe.DataFrame = [Usuari_solicitant: string, N_Expedient: string ... 11 more fields]
▸ df_rel_exp_pro_titular: pyspark.sql.dataframe.DataFrame = [N_Expedient: string, N_Procediment: string ... 2 more fields]
```

Original tables

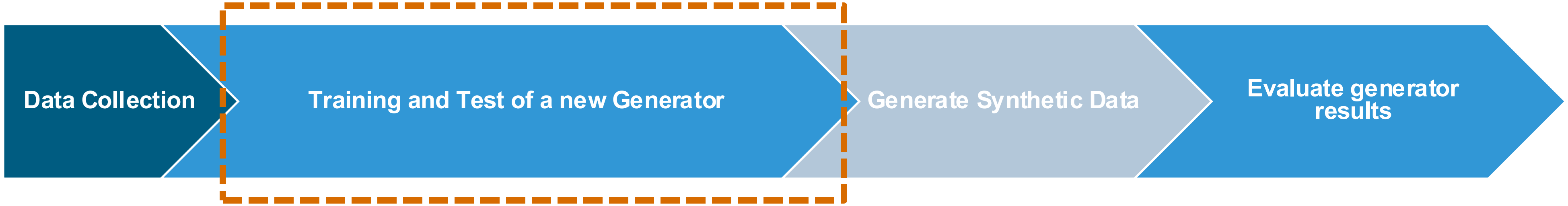


tSocial.ActuacionsNomina	tSocial. pagamentnomina	tSocial.Titular
# columns: 4	**# columns: 12**	**# columns: 39**
# rows: 190.142	**# rows: 278.373**	**# rows: 86.569**
Size: 1.1 MB	**Size: 1.69 MB**	**Size: 6.2 MB**

Processed tables (nested schema approach)



Mostly.AI Implementation Details



1. Each **table** requires **detailed configuration** to enable the generator to learn its structure

```
actuacionsnomina_table_config = {
    "name": "actuacionsnomina",
    "data": df_actuacionsnomina_processed,
    "tabular_model_configuration": {
        "max_training_time": 90
    },
    "primary_key": "N_Exp_Pro",
    "foreign_keys": [{"column": "identificador", "referenced_table": "titular", "is_context": True}],
    "columns": [{"name" : "N_Expedient", "model_encoding_type": ModelEncodingType.tabular_character},
                {"name" : "N_Procediment", "model_encoding_type": ModelEncodingType.tabular_character},
                {"name" : "Data_actuacio", "model_encoding_type": ModelEncodingType.tabular_datetime},
                {"name" : "Efecte_Nomina", "model_encoding_type": ModelEncodingType.tabular_categorical},
                {"name" : "N_Exp_Pro", "model_encoding_type": ModelEncodingType.auto},
                {"name" : "identificador", "model_encoding_type": ModelEncodingType.auto}]
}
```

2. After table configuration, they are **assembled into a multi-table generator configuration**.

```
generator_config = {
    "name": "Multi-table Generator",
    "tables": [titular_table_config, pagamentnomina_table_config, actuacionsnomina_table_config],
}
```

3. Once configuration is complete, the **generator model can be trained**.

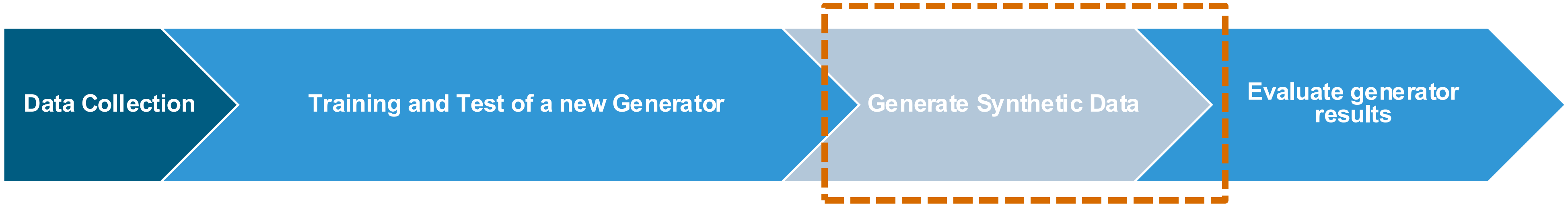
```
# train a synthetic data generator
generator = mostly.train(name="ctti", config=generator_config)
print(f"{generator.id} {generator.name} - {generator.accuracy}")
```

(5) Spark Jobs

```
Created generator 7eb2ad63-777d-4474-9363-df5f64f73b7a
Started generator training
```

NTT DATA

Mostly.AI Implementation Details



After successfully training a generator model, the next phase of the workflow involves **generating the synthetic data and storing it in Databricks tables**.

Generator importing and synthetic data generation

23 hours ago (43m) 16

```
g = mostly.generators.import_from_file('/Volumes/admin_govern_sta_des/tmp_mostlyai/generator/generator-7eb2ad63.zip')
df_samples_test = mostly.generate(g, size=86_000)
```

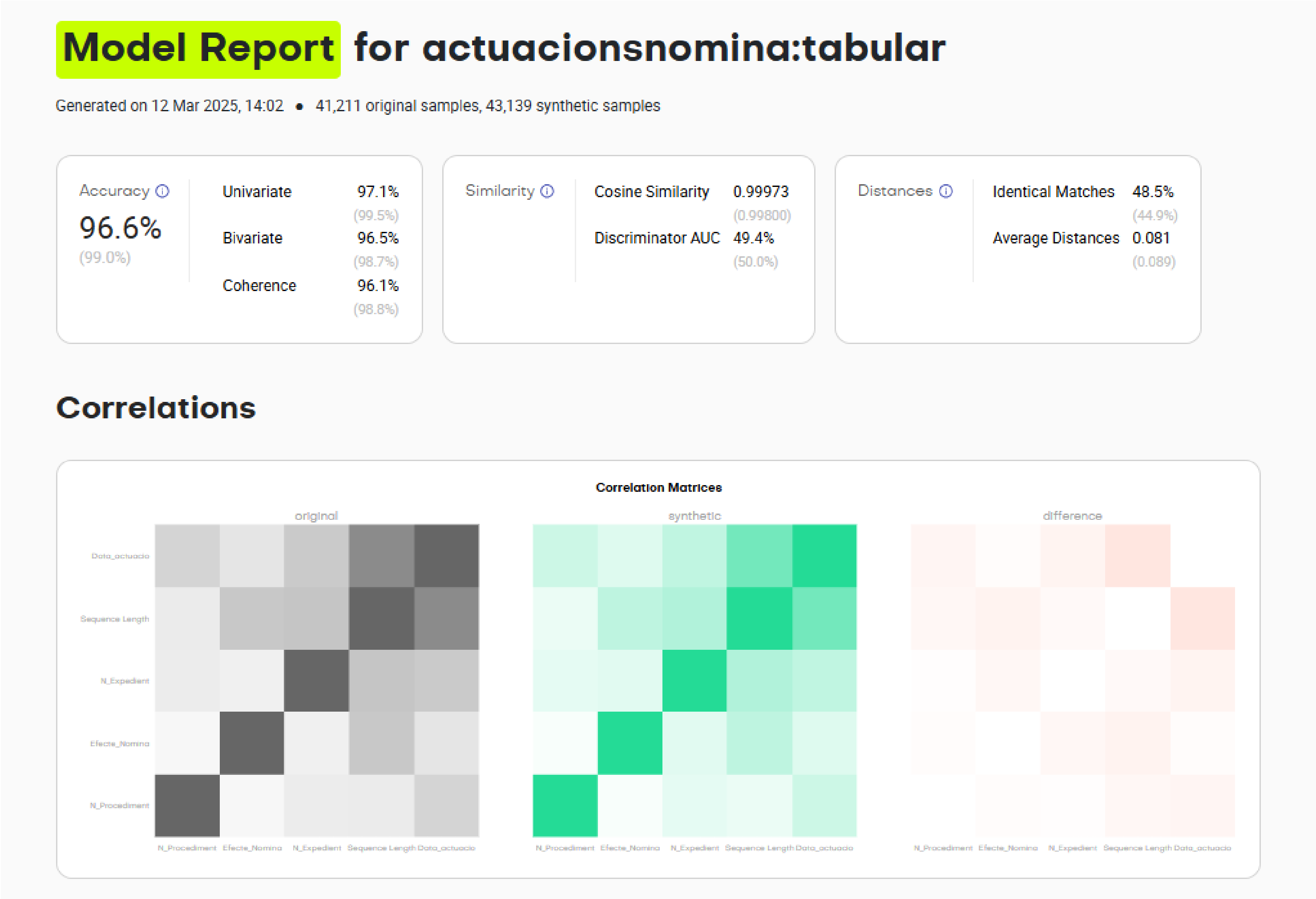
```
Imported generator 337ec6f6-8e45-4307-a73b-68771eb8b396

Created synthetic dataset b5281170-097b-405e-8aeb-b061fe21ffba with generator 337ec6f6-8e45-4307-a73b-68771eb8b396

Started synthetic dataset generation
```

Dataframe conversion and saving to Databricks Unity Catalog

```
df_syntetic_titular_spark = spark.createDataFrame(df_syntetic_titular)
df_syntetic_actuacionsnomina_spark = spark.createDataFrame(df_syntetic_actuacionsnomina)
df_syntetic_pagamentnomina_spark = spark.createDataFrame(df_syntetic_pagamentnomina)

df_syntetic_titular_spark.write.mode("overwrite").saveAsTable("admin_govern_sta_des.tmp_mostlyai.syntetic_titular_preprocessed")
df_syntetic_actuacionsnomina_spark.write.mode("overwrite").saveAsTable("admin_govern_sta_des.tmp_mostlyai.
syntetic_actuacionsnomina_preprocessed")
df_syntetic_pagamentnomina_spark.write.mode("overwrite").saveAsTable("admin_govern_sta_des.tmp_mostlyai.
syntetic_pagamentnomina_preprocessed")
```

Evaluation Metrics

In **Mostly.AI** each trained generator is evaluated with an auto-generated report.



Quality metrics dimensions:

- Accuracy

Accuracy of synthetic data is assessed by comparing the distributions of the synthetic and the original data

- Similarity

Explains how similar the synthetic data is to the training data. It’s expected these similarities to be close.

- Distances

Synthetic data shall be as close to the original training samples, as it is close to original holdout samples, which serve us as a reference.

Evaluation Metrics

Results for the table **actuacionsnomina**

Model Report for actuacionsnomina:tabular

Generated on 12 Mar 2025, 14:02 • 41,211 original samples, 43,139 synthetic samples

Accuracy

96.6%
(99.0%)

Univariate	97.1% (99.5%)
Bivariate	96.5% (98.7%)
Coherence	96.1% (98.8%)

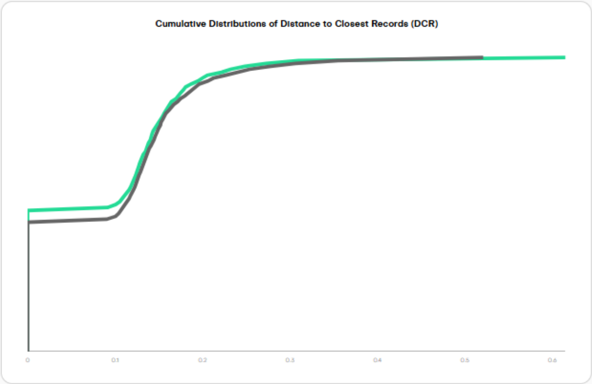
Similarity

Cosine Similarity	0.99973 (0.99800)
Discriminator AUC	49.4% (50.0%)

Distances

Identical Matches	48.5% (44.9%)
Average Distances	0.081 (0.089)

Column	Univariate	Bivariate	Coherence
N_Procediment	99.2%	98.3%	98.5%
N_Expedient	98.6%	97.6%	96.9%
Efecte_Nomina	98.1%	97.7%	96.8%
Sequence Length	96.3%	95.7%	-
Data_actuacio	93.7%	92.8%	92.3%
Total	**97.1%**	**96.5%**	**96.1%**



Evaluation Metrics

Results for the table **pagamentnomina**

Model Report for pagamentnomina:tabular

Generated on 12 Mar 2025, 13:16 • 43,441 original samples, 43,441 synthetic samples

Accuracy
94.0%
(99.1%)

Univariate	95.4%	(99.4%)
Bivariate	92.7%	(98.6%)
Coherence	94.1%	(99.1%)

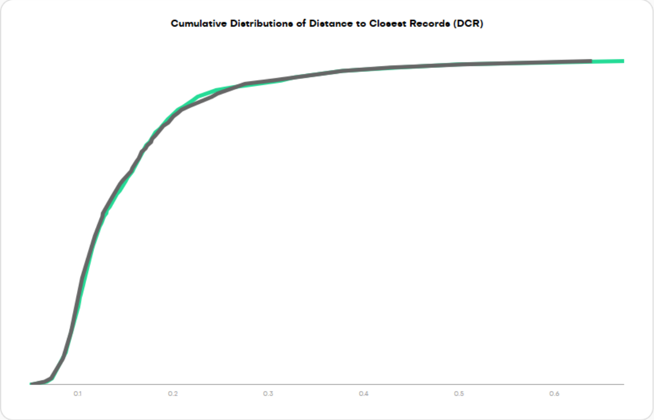
Similarity

Cosine Similarity	0.99769	(0.99930)
Discriminator AUC	93.7%	(51.2%)

Distances

Identical Matches	0.0%	(0.0%)
Average Distances	0.153	(0.152)

Column	Univariate	Bivariate	Coherence
Tipus_Expedient	99.5%	96.0%	99.4%
Ambit_Territorial	99.4%	95.9%	99.3%
Provincia	99.4%	95.9%	99.3%
Tipus_procediment	99.4%	95.9%	98.9%
Municipi	99.2%	95.6%	99.0%
Sequence Length	98.8%	94.3%	-
Data_nomina	97.3%	93.5%	95.2%
Data_pagament	96.7%	92.9%	94.6%



Evaluation Metrics

Results for the table **titular**

Model Report for titular:tabular

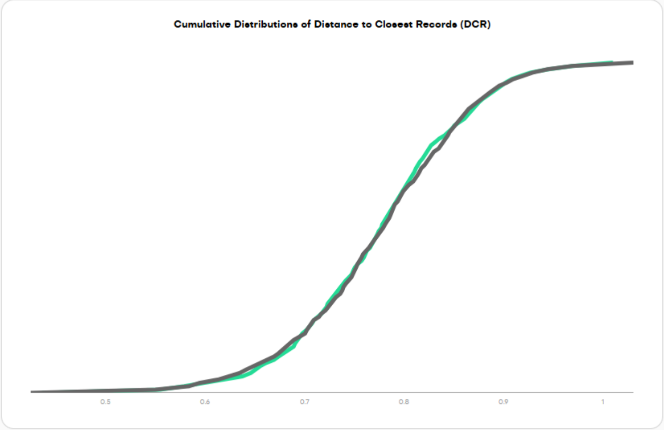
Generated on 12 Mar 2025, 12:35 • 86,569 original samples, 86,569 synthetic samples

Accuracy		
94.5% (99.5%)	Univariate	96.2% (99.7%)
	Bivariate	92.9% (99.3%)

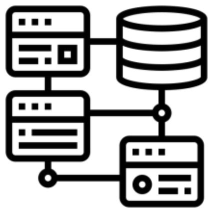
Similarity		
	Cosine Similarity	0.96851 (0.99860)
	Discriminator AUC	80.9% (49.3%)

Distances		
	Identical Matches	0.0% (0.0%)
	Average Distances	0.775 (0.775)

Column	Univariate	Bivariate
telefonFix	100.0%	96.0%
situacioPersona	99.6%	95.9%
residenciaTipusDomicili	99.5%	95.7%
cognomPrimer	99.4%	95.6%
residenciaQualificador	99.3%	95.7%
residenciaBloc	99.2%	95.5%
telefonMobil	99.1%	95.5%
correuElectronic	99.1%	95.6%
identificadorTipus	99.1%	95.5%
notificacioQualificador	99.1%	95.6%
naixementData	98.9%	95.2%
genere	98.8%	95.4%



Challenges



Configure Tables and Their Relationships

Working with related tables, whose relations must be retained, is challenging and involves a proper documentation of the database relational model and clarifications with client stakeholders.

As most of the time the documentation is not updated the effort to have a relational model could be work-intensive and time-consuming as involve several interactions. Could be a blocker in the progress.

Preprocessing Data and Contextual Parental Relationships

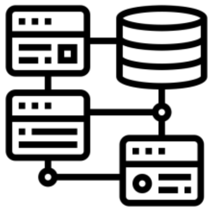
MOSTLY.AI supports retaining correlations between tables, specifically between a child table and its parent table ("context" parent). The challenge lies in deciding which parent table relationship should be considered as the context, as correlations can only be retained for one parent.

Statistical Representativeness of Generated Tables

Ensuring that the newly generated tables are statistically representative of the original data is crucial for the validity and reliability of synthetic data.

Training time & number of training executions

Lessons Learned



Steps to address configuring tables and their relationships:

- Define Primary Keys;
- Establish Foreign Key Relationships;
- Configuration tools.

Steps to address Preprocessing Data and Contextual Parental Relationships

- Identify Contextual Relationships;
- Configure Context Parents.

Steps to address Statistical Representativeness of Generated Tables

- Analyze Original Data;
- Implement Generative Models;
- Validation and Testing.

Training time & number of training executions

- Trade-off between training time and costs

Syntethic Data Generation PoC

Mostly.AI – Running Cost

Running Cost

There are **2 possibilities** of using Mostly.AI for synthetic data generation:

1. usage of the Synthetic Data SDK, under a fully permissive Apache v2 license;
2. usage of the Mostly.AI Platform.

The usage of **the SDK is free** (open-source licensing) and for the **AI Platform there are 4 License options**: Starter, Business, Unlimited and Enterprise.

Package Options	**Starter**	**Business**	**Unlimited**	**Enterprise**
Platform Instances/agents	1	Up to 4	Unlimited	Unlimited
Data Creators	1	Unlimited	Unlimited	Unlimited
Platform License Term	12 months	12 months	12 months	Minimum 3 years
Initial Data Usage term	6 months unlimited	6 months unlimited	12 months unlimited	
Professional Services	n/a	n/a	n/a	Deployment, Onboarding and Training included
Data Credit Cost	$5/€5	$4/€4	n/a	$4/€4
Annual Subscription ($/€)	100.000	150.000	250.000	250.000

Notes:

- Data Usage: 1 credit = 1 million data points (rows x columns x tables)
- Data Credit Cost: After initial data usage term. Packages available for pre-purchased credits.

Running Cost

If the option is to use the **Synthetic Data SDK**, installed on a Databricks instance, for total anonymization with synthetic data of the 3 tables selected from **tSocial** information system, the running cost is presented below.

Licensing	Free
Processing	DBU consumption * 4 DBU/hour – 13 DBU/total
Data Points (=rows x columns x tables)	7.588.860
Runtime	3h10m (Configuration, Training and Generation)
Accuracy	actuacionsnomina: 96.6% pagamentnomina: 94.0% titular: 94.5%
Total Cost Estimate	€ 39,60

Notes:

- Databricks cluster: Standard_D16ads_v5, 1 Driver, 64GB Memory, 16 cores; Unity Catalog;
- Azure Databricks Pricing reference: 3,046 €/DBU-hour


Thank you.