



PRIVACY-ENHANCING TECHNOLOGIES

- Anonymization
- Synthetic Data Generation
- Federated Learning
- Explainable AI
- Secure Multiparty Computation



PAMOLA

Privacy AI

DATA PRIVACY STUDIO

TECHNICAL SCOPE DOCUMENT



2/6/2025

WHAT IS PAMOLA?

PAMOLA is a Privacy AI Studio designed for data anonymization, synthetic data generation, and risk analysis (Privacy Enhancing Technologies – PET). It offers modular workflows and tools to ensure secure data handling and compliance with regulations such as GDPR and HIPAA.

What PAMOLA Does

PAMOLA is tailored for corporate environments, enabling users to:

- Upload datasets and perform data profiling.
- Apply privacy-preserving measures (e.g., anonymization, synthetic data generation).
- Evaluate data quality and privacy risks.

This document defines the scope for building the Minimum Viable Product (MVP).

Competitive and Related Solutions

- **Data Management:** OpenML, CKAN
- **Synthetic Data Generation:** Gretel SDV, Mostly.AI, Hazy.AI, Tonic.AI
- **Data Anonymization:** ARX, Informatica Data Masking, IBM Privacy PowerKey, Bizagi

KEY FEATURES

1. Data Anonymization:
Differential Privacy, k-anonymity.

2. Synthetic Data Generation:
PATE-GAN, SDV.

3. Core Metrics: KS-Test, KLD, ..., risk simulation.

4. Pipeline Approach: Modular and configurable workflows.

5. User Experience: Intuitive and scalable UI design.

CORE FEATURES AND MVP SCOPE

To accelerate development, the MVP maximally utilizes existing work and open-source components. Preliminary implementations, such as the DEMO version of PATE-GAN, and frameworks like DataHub for metadata management and MUI/React for the frontend, form the foundation for the PAMOLA MVP.

Feature	Description	Comments	v1 Scope
Data Anonymization	Process datasets using differential privacy, k-anonymity, suppression, and masking techniques.	Utilize PyCanon, Faker,...	Included
Synthetic Data Generation	Create synthetic datasets using PATE-GAN and evaluate quality using statistical metrics.	Built on existing DEMO implementations of PATE-GAN and risk metrics (KS-Test, KLD).	Included
Metrics and Risk Evaluation	Assess dataset utility and simulate attacks like membership inference and linkage attacks.	Built on existing DEMO implementations of quality/risk metrics (KS-Test, KLD, attack simulation) + SciPy, NumPy, Seaborn, Stats models.	Included
Pipeline Management	Step-by-step workflows for anonymization and synthetic data generation.	Utilizes DataHub as a backend for managing datasets, metadata, and pipelines.	Included
Workspaces	Isolated spaces for managing datasets, projects, and models.	Use FastAPI/Sanic or Django	Included
Dashboards	Display data quality, risk metrics, and project progress visually.	Developed using MUI/React for intuitive and modular UI.	Limited

ARCHITECTURE OVERVIEW

PAMOLA's architecture is designed to streamline privacy-focused data workflows by connecting an intuitive user interface, a powerful backend, and advanced data processing capabilities. Each component plays a crucial role in managing datasets, executing tasks, and presenting results effectively.

Core Components

1. Frontend (React UI):

- Centralized dashboards for managing datasets, projects, and metrics.
- Modular widgets (e.g., Quick Notes, Latest Datasets).

2. App Server (Python/Flask/FastAPI):

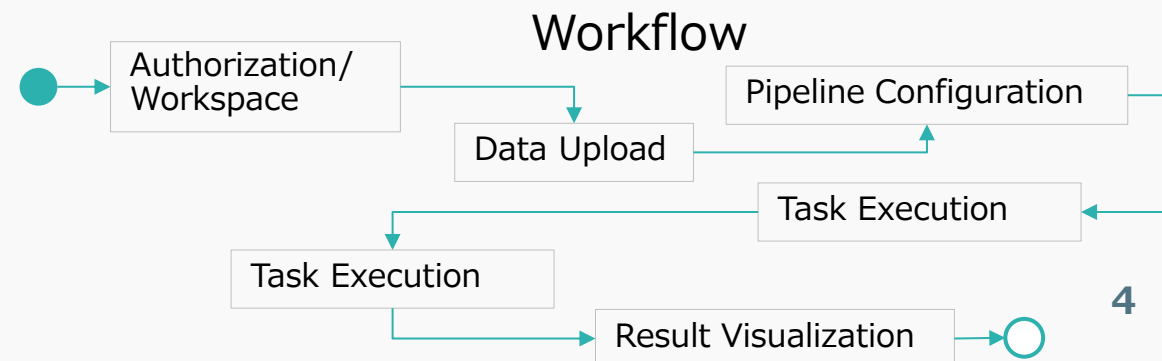
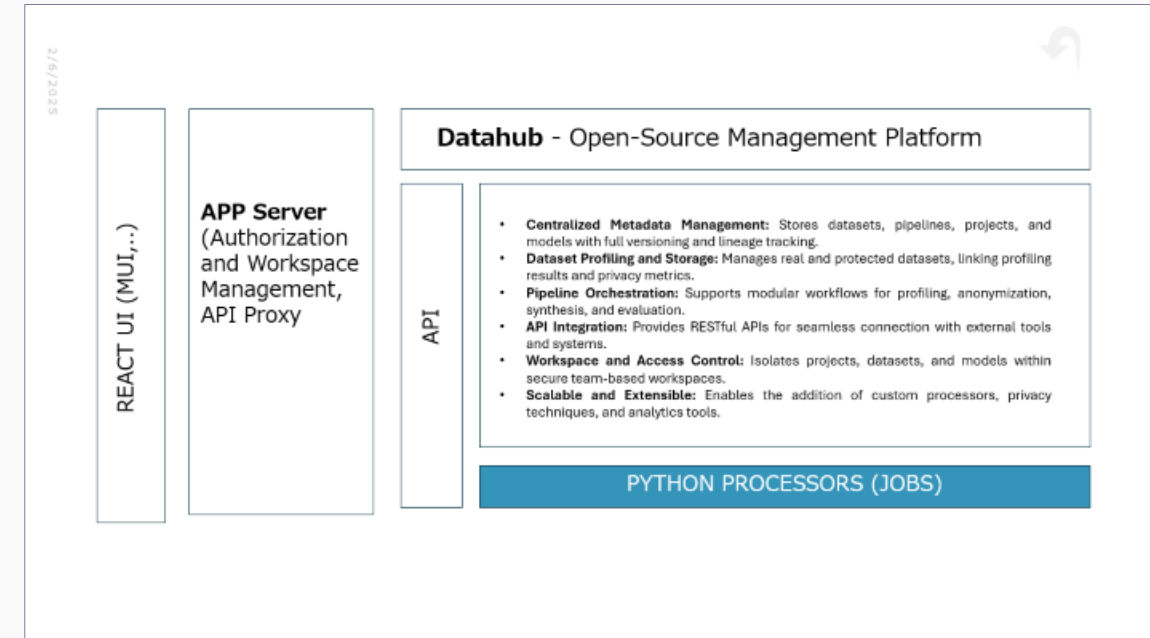
- Orchestrates user requests, pipelines, and integrations.
- Handles API endpoints for seamless frontend communication.

3. DataHub (Backend):

- Manages metadata, datasets, and pipeline configurations.
- Tracks data transformations and stores results securely.

4. Python Processors:

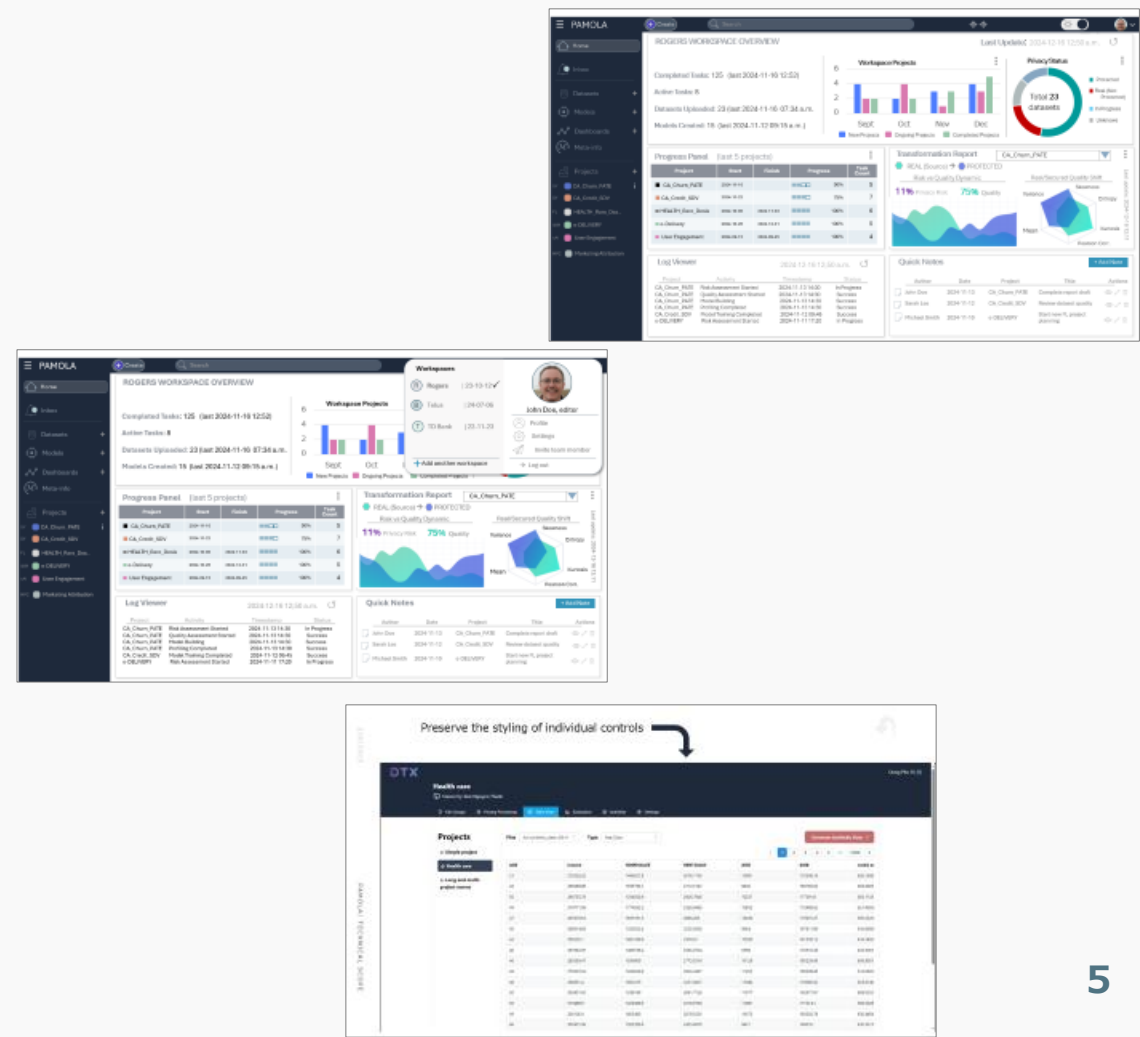
- Executes tasks like anonymization, synthetic data generation, and metric calculation.
- Supports modular addition of custom algorithms.



PAMOLA's UI is designed to prioritize projects and workflows over individual data types, ensuring a seamless user experience. Data is organized into Workspaces, grouping datasets, projects, and models under a unified access point for team collaboration. Tasks are visualized as pipeline steps, providing clarity on progress and dependencies.

Home Screen Structure

- 1. Top Panel:** Search bar, user profile, and workspace selection for quick navigation.
- 2. Left Sidebar:** Navigation for key sections: Workspaces, Datasets, Projects, Models, Dashboards, and Notifications.
- 3. Central Widgets:**
 - General Statistics:** Displays workspace-level metrics, such as completed tasks and uploaded datasets.
 - Active Projects:** Lists ongoing pipelines with progress bars and quick actions (e.g., Stop, Retry).
 - Latest Datasets:** Highlights the most recent datasets with key attributes (e.g., row count, protection status).
 - Quick Notes:** Simple reminders linked to projects or datasets for effective task management.



TYPICAL PIPELINES FOR KEY TASKS

PAMOLA utilizes modular pipelines to structure workflows for privacy-preserving data operations. Below are examples of typical pipelines for synthetic data generation using deep learning models and data anonymization with k-Anonymity, l-Diversity, and t-Closeness.

Pipeline 1: Data Anonymization

1.Pre-Processing:

- **Data Profiling:** Identify direct identifiers, quasi-identifiers, and sensitive attributes.
- **Data Cleaning:** Standardize formats and remove irrelevant data.

2.Anonymization Techniques:

- **Suppression:** Remove specific rows or columns.
- **Masking:** Replace identifiers with pseudonyms or random tokens.
- **Generalization:** Replace specific values with broader categories (e.g., age ranges).
- **Noise Addition:** Apply differential privacy methods to add noise to data.

3.Post-Processing:

- **Validation:** Ensure anonymized data meets k-Anonymity, l-Diversity, or t-Closeness thresholds.
- **Data Output:** Create anonymized datasets for analysis or sharing.

4.Evaluation:

- **Quality Metrics:** Analyze data utility using metrics like distribution overlap or classification accuracy.
- **Risk Metrics:** Assess residual risks (e.g., re-identification potential).

Pipeline 2: Synthetic Data Generation

1.Pre-Processing:

- **Data Profiling:** Identify sensitive attributes and correlations.
- **Data Cleaning:** Handle missing values, outliers, and inconsistencies.
- **Data Encoding:** Convert categorical data to numerical formats suitable for model training.

2.Data Synthesis:

- **Model Selection:** Choose models like PATE-GAN or SDV based on the dataset structure.
- **Training:** Train the synthetic data generation model on pre-processed data.
- **Synthetic Data Generation:** Generate synthetic datasets that mimic the statistical properties of the real data.

3.Post-Processing:

- **Decoding:** Reconvert encoded features back to original formats.
- **Sanitization:** Ensure no direct identifiers are included in the synthetic dataset.

4.Evaluation:

- **Quality Metrics:** Use KS-Test, KLD, Pearson Correlation to compare distributions.
- **Risk Metrics:** Simulate attacks (e.g., membership inference) to evaluate privacy.
- **Utility Testing:** Test model performance on synthetic data versus real data.

PAMOLA LIVE DEMO

Discover how PAMOLA revolutionizes privacy-preserving AI with advanced synthetic data generation, federated learning, and risk assessment tools. Explore the live demo on the REALM platform and see how organizations can securely process and analyze data while ensuring compliance and confidentiality.

2/6/2025

PAMOLA: TECHNICAL SCOPE

The image displays a collage of the PAMOLA user interface. On the left, a sidebar menu lists various sections: Home, Inbox, Data, Models, Dashboards, and Meta-info. The main area shows a dashboard with a '321321 Overview' section, a 'Privacy status' gauge, and a 'Progress (last 5 projects)' table. A user profile dropdown is visible, showing the user is 'Owner User' with 'Owner Role'. Overlaid on the right is a circular graphic showing hands interacting with a tablet displaying a network diagram. To the right of this is a 'Sign In' form with fields for Email (super_admin@pamola.com) and Password, a 'Remember me' checkbox, a 'Forgot password?' link, and a 'Login' button. Below the login button is a link to 'Create an account' for new users.

New here?
Sign up to showcase the amazing features of PAMOLA
Sign Up

PAMOLA
Privacy AI

Sign In
Please sign-in to your account and start the adventure

Email
super_admin@pamola.com

Password

☐ Remember me [Forgot password?](#)

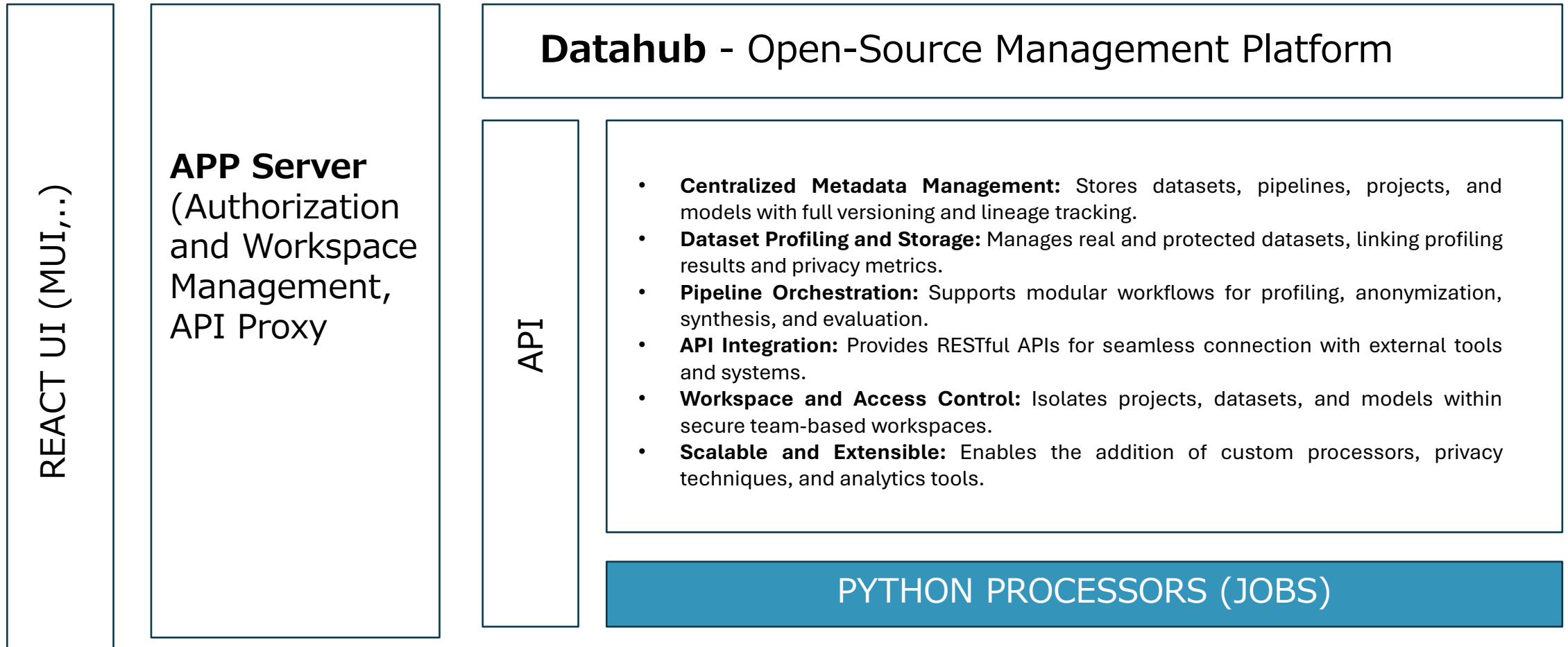
Login

New on our platform? [Create an account](#)

<https://pamola.realmdata.io/>

THANK YOU





🏠

Home

🔔

Inbox

🗄️

Datasets

+

🧠

Models

+

📊

Dashboards

+

🔗

Meta-info

👤

Projects

+

SY

CA_Churn_PATE

⋮

SY

CA_Credit_SDV

FL

HEALTH_Rare_Des...

ANM

e-DELIVERY

xAI

User Engagement

MPC

Marketing Attribution

ROGERS WORKSPACE OVERVIEW

Last Update:

2024-12-16 12:50 a.m.

↺

Completed Tasks: 125

(last 2024-11-16 12:52)

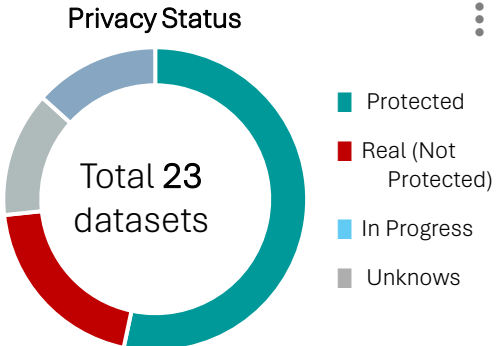
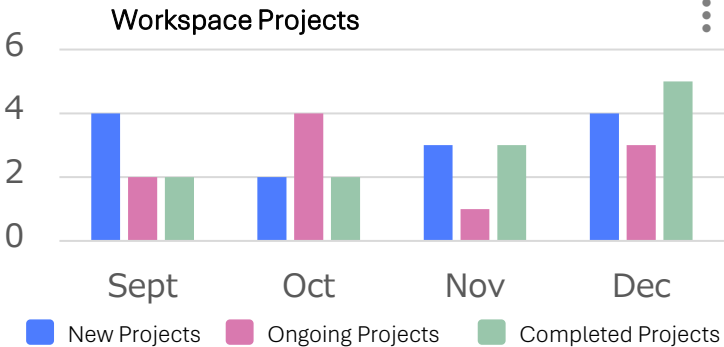
Active Tasks: 8

Datasets Uploaded: 23

(last 2024-11-16 07:34 a.m.)

Models Created: 15

(last 2024-11-12 09:15 a.m.)



Progress Panel (last 5 projects)

Project	Start	Finish	Progress	Task Count
CA_Churn_PATE	2024-11-16		50%	5
CA_Credit_SDV	2024-11-22		75%	7
HEALTH_Rare_Desis	2024-10-05	2024-11-03	100%	6
e-Delivery	2024-10-26	2024-12-01	100%	5
User Engagement	2024-09-15	2024-09-25	100%	4

Transformation Report

CA_Churn_PATE

▼

REAL (Source) → PROTECTED

Risk vs Quality Dynamic

11% Privacy Risk

75% Quality



Real/Secured Quality Shift



Last update: 2024-12-16 12:11

Log Viewer

2024-12-16 12:50 a.m.

↺

Project	Activity	Timestamp	Status
CA_Churn_PATE	Risk Assessment Started	2024-11-13 14:30	In Progress
CA_Churn_PATE	Quality Assessment Started	2024-11-13 14:30	Success
CA_Churn_PATE	Model Building	2024-11-13 14:30	Success
CA_Churn_PATE	Profiling Completed	2024-11-13 14:30	Success
CA_Credit_SDV	Model Training Completed	2024-11-12 09:45	Success
e-DELIVERY	Risk Assessment Started	2024-11-11 17:20	In Progress

Quick Notes

+ Add Note

Author	Date	Project	Title	Actions
John Doe	2024-11-13	CA_Churn_PATE	Complete report draft	
Sarah Lee	2024-11-12	CA_Credit_SDV	Review dataset quality	
Michael Smith	2024-11-10	e-DELIVERY	Start new FL project planning	

Preserve the styling of individual controls



2/6/2025

PAMOLA: TECHNICAL SCOPE

DTX

Dung Pho Tri (T)

Health care

Owner by Tam Nguyen Thanh

File Groups

Privacy Processing

Table View

Evaluation

Activities

Settings

Projects

Simple project

Health care

Long and multi-project names

Files

full-synthetic_data-20k-f1

Type

Real Data

Generate Synthetic Data

<

1

2

3

4

5

...

1000

>

AGE	Income	HOMEVALUE	RENTVALUE	AME	AMB	credit_sc
51	272026.25	1484527.8	3078.1736	10981	175548.19	605.1589
43	289289.88	1585796.1	2712.3162	9820	169766.92	603.0683
50	296755.78	1356030.9	2806.7908	10337	177391.9	605.1135
49	274771.06	1774363.2	3266.0469	10852	172499.62	621.4505
47	267670.53	1655181.5	2886.206	10430	175051.27	592.2325
45	260014.06	1333632.6	2525.9336	9664	187811.89	616.6900
43	295225.1	1681240.9	2974.01	10200	161338.12	614.3963
46	287962.97	1489738.2	2545.4734	9789	176516.39	623.3061
46	283550.47	1590060	2773.5518	10129	185229.45	606.9951
49	279535.56	1456626.8	2926.3447	11252	189309.45	613.8603
48	296851.4	1803379	3207.4067	11046	170085.62	623.9146
50	303437.62	1396140	3041.7126	11577	162871.67	608.8353
50	313498.9	1459389.9	3218.9749	11081	171914.1	586.5028
49	256105.4	1605480	2870.5505	10475	193503.78	630.4804
46	302431.06	1583389.8	2433.4878	9421	169316	635.5611