# PrivacyGo Data Clean Room

An Open-Source TEE-based Data Clean Room for Secure Trusted Research & AI Collaboration


Dayeol Lee†
Confidential Computing
TikTok
San Jose California USA
dayeol.lee@tiktok.com

Mingshen Sun
Confidential Computing
TikTok
San Jose California USA
mingshen.sun@tiktok.com

Vini Jaiswal
Open Source Operations
TikTok
San Jose California USA
vini.js1303@gmail.com

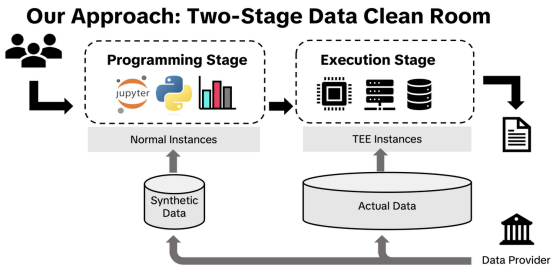


## ABSTRACT

Data collaboration is not a new concept, and numerous data collaboration frameworks already exist. However, different frameworks try to apply different privacy-enhancing technologies (PETs), which have different strengths and weaknesses. Traditional data protection methods, such as encryption at rest and in transit, offer limited protection when data is being processed. SQL policy and differential privacy are two common solutions, but they have some limitations to verify and collaborate on data before releasing it. Applications and data remain vulnerable to attacks during runtime, regardless of infrastructure access privileges. This leaves organizations storing and processing sensitive and regulated data exposed to potential security breaches. Additionally, without remote attestation, it becomes difficult to verify the integrity and authenticity of the computing environment, raising concerns about the security of data in use. Based on this situation, we designed a two-stage data clean room that combines different technologies to balance usability, accuracy, and privacy.


## What is PrivacyGo Data Clean Room

PrivacyGo Data Clean Room (PGDCR) is an open-source project for easily building and deploying data collaboration framework to the cloud using trusted execution environments (TEEs). PGDCR achieves this by combining different PETs in different stages.




In the programming stage, the platform allows data consumers to explore the dataset while providing interactive data usage and protecting data providers' privacy, in that data provider decides the protection mechanism. In the execution stage, the workload runs in an isolated environment, and data providers can manage the data, code, and output space. By using attestation, the data providers can control which program can access their data. TEE also assures the data scientists, the integrity of their program and legitimacy of the output from executions by providing JWT-based attestation report that can be publicly verified for authenticity.

## Use cases of PrivacyGo Data Clean Room

The system built on top of cloud infrastructure allows for multi-cloud usage and provides benefits such as transition of trust, integrated code output, and monitoring. Use cases include providing transparency to researchers and enabling data analytics for marketing purposes. Some of the potential use cases of the PGDCR include:

- **Trusted Research Environments (TREs)**: Some data may be valuable to various research on public health, economic impact, and many other fields. TREs are a secure environment where authorized/vetted researchers and organizations can access the data. The data provider can choose to use PGDCR to build their TRE.
- **Advertisement and Marketing**: Advertisement is a popular use case of data collaboration frameworks. PGDCR can be used for lookalike segment analysis for advertisers, or ad tracking with private user data.
- **Machine Learning**: PGDCR can be useful for machine learning involving private data or models. For example, a private model provider can provide their model for fine-tuning, but do not reveal the actual model in the Programming Stage.

## Availability

The project is open-source[1] as of June 6, 2024[2] and available on GitHub[3]. Currently, it only supports one-way collaboration, uses Google Cloud Platform as the backend and currently supports the computation on CPU. The data provisioning, policy and attestation is manual for the current initial version. Project's growth plans include expanding to multi-user collaboration, platform extensibility to support multiple backends, bringing automation to the data provisioning, policy and attestation and computation to be supported on both CPU/GPU.

[1] https://developers.tiktok.com/blog/privacygo-data-clean-room-open-source
[2] https://www.confidentialcomputingsummit.com/session/a-secure-and-private-platform-for-transparent-research-access-via-trusted-execution-environments
[3] https://github.com/tiktok-privacy-innovation/PrivacyGo-DataCleanRoom