
# PrivacyGo Data Clean Room

## *A Secure and Private Platform for data collaboration via TEEs*

Dayeol Lee, Mingshen Sun and Vini Jaiswal

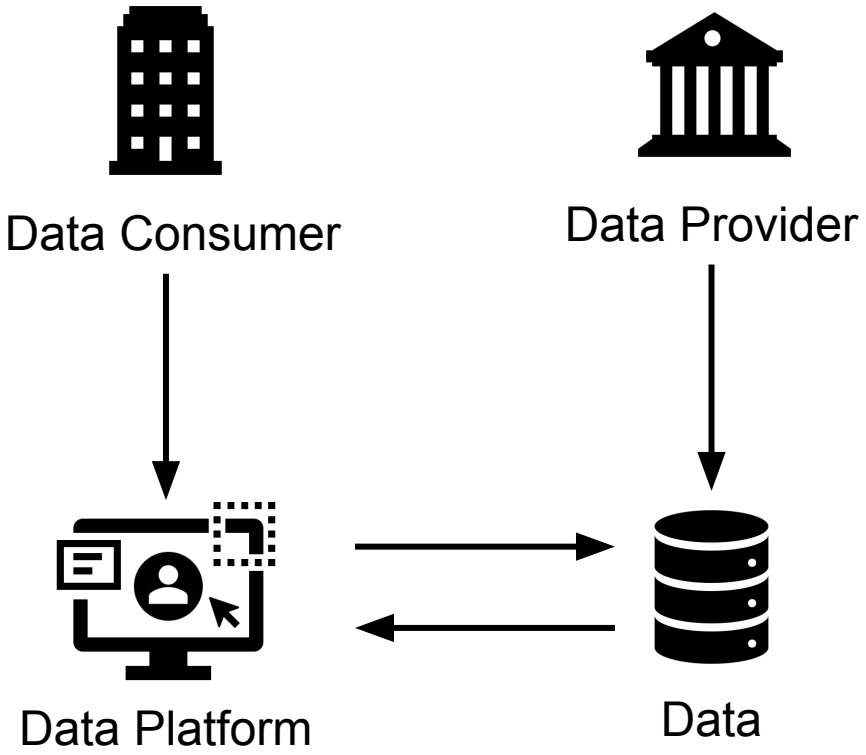
# Agenda

- Problem that we solving for
- Existing Solutions and our Approach
- PGDCR Architecture
- Why PGDCR was built and its use cases
- Why LF CCC?
- Project Status & Growth Plans

# Problem that we solving for

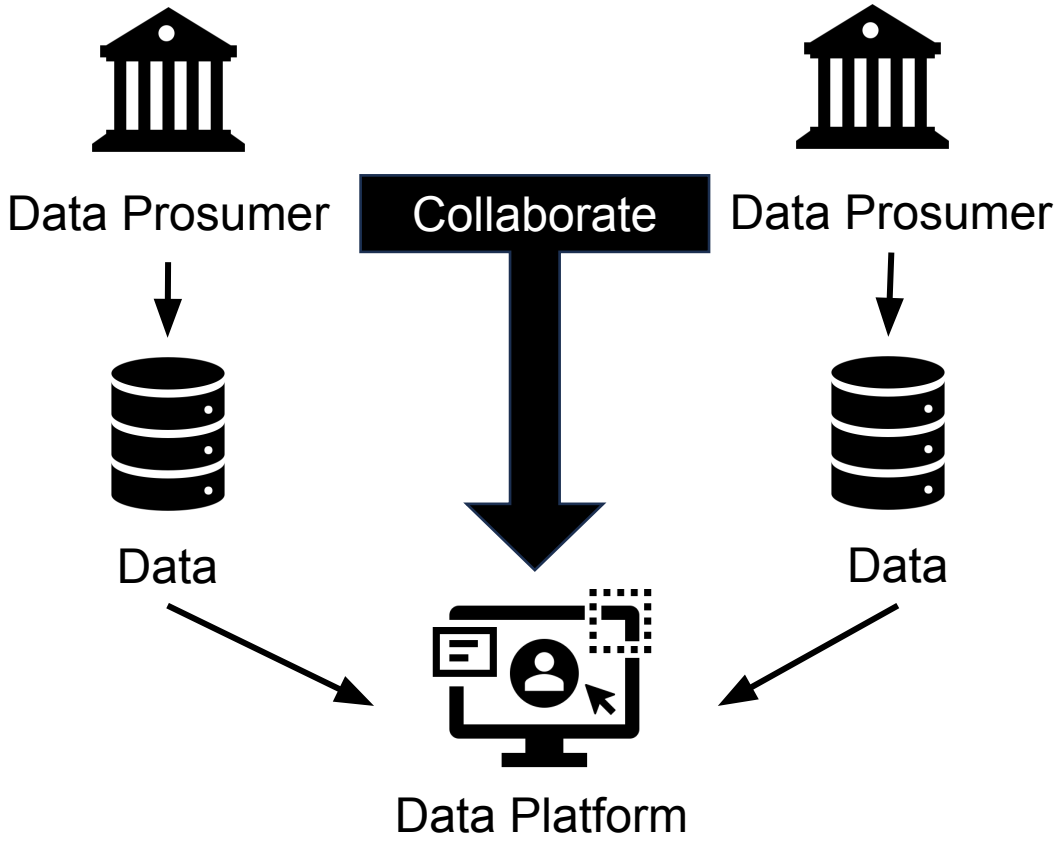
# One-Way Data Collaboration

- Makes data from **data provider** available to **data consumer**
- Allows data provider to **control policies** (e.g., permission, retention)
- Provides data consumer with tools to utilize data (e.g., interactive data query interface)



# Multi-Way Data Collaboration

- **Multiple** data providers and data consumers
- Each party can be data provider and consumer at the same time, i.e., a **prosumer**
- Platform provides the prosumers with tools to **collaborate**



# Privacy & Security Issues with Data Collaboration

**Security**: How can we share data securely?

- Confidentiality and integrity of data and computation

**Privacy**: How can we enforce different privacy policies?

- Purpose limitation: any data should be collected and processed for specific purposes
- Data retention: data should persist only for a defined period of time

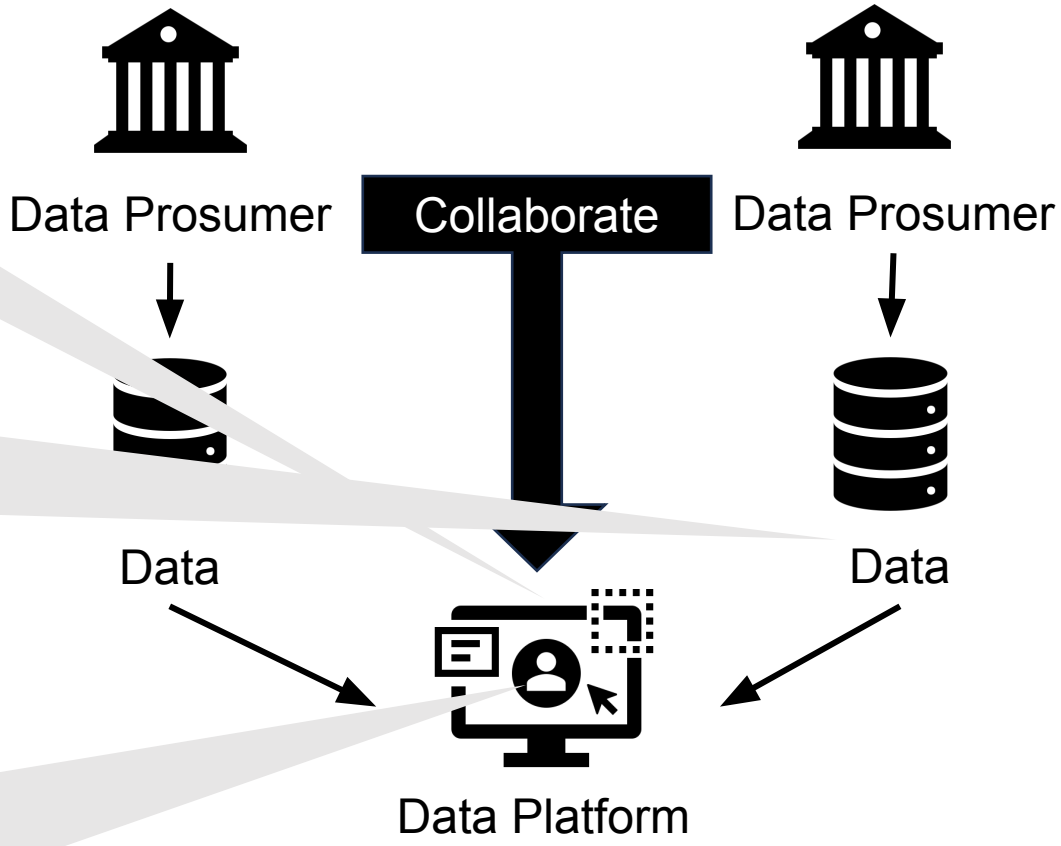
# Our Goals

- **Usability**: provide an interactive tool to utilize the data
- **Security**: protect confidentiality and integrity of data in use, and provide strong access control
- **Privacy**: make it possible to enforce privacy policies such as purpose limitation and data retention
- **Accuracy**: provide accurate results on real data, as well as an evidence of execution
- **Deployment**: make it easy to deploy to the cloud

# Existing Solutions

# Existing Industry Solutions for Security & Privacy

| | |
|---|---|
| SQL Policy-based Data Clean Room | Rely on SQL/data platforms provided by 3rd party, who is free of conflict of interest |
| Differential Privacy | Preprocess data or add noise to the aggregated SQL results to limit information leakage |
| Trusted Execution Environment | Use remote attestation to co-verify the code before releasing data; contain data in an isolated environment during execution |



# Technical Difficulties of Existing Solutions

| | Interactive Tools | Accurate Results | Privacy Protection | Privacy Policy | What You Trust |
|---|---|---|---|---|---|
| SQL Policy-based Data Clean Room | Yes | Yes | Hard | Hard | 3rd Party |
| Differential Privacy | Yes | No | Yes | N/A | DP Algorithm, Curator |
| Trusted Execution Environment | No | Yes | Yes | Yes | TEE |

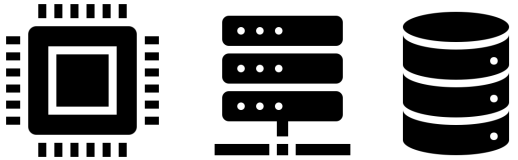
# Our Approach:
# Two-Stage Data Clean Room

# Different Need at Each Stage

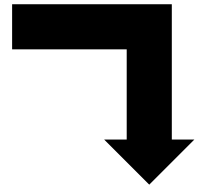
## Programming Stage


- Smaller Data/Compute
- Interactive
- Hard to Control Data
- Higher Privacy Risk

## Execution Stage

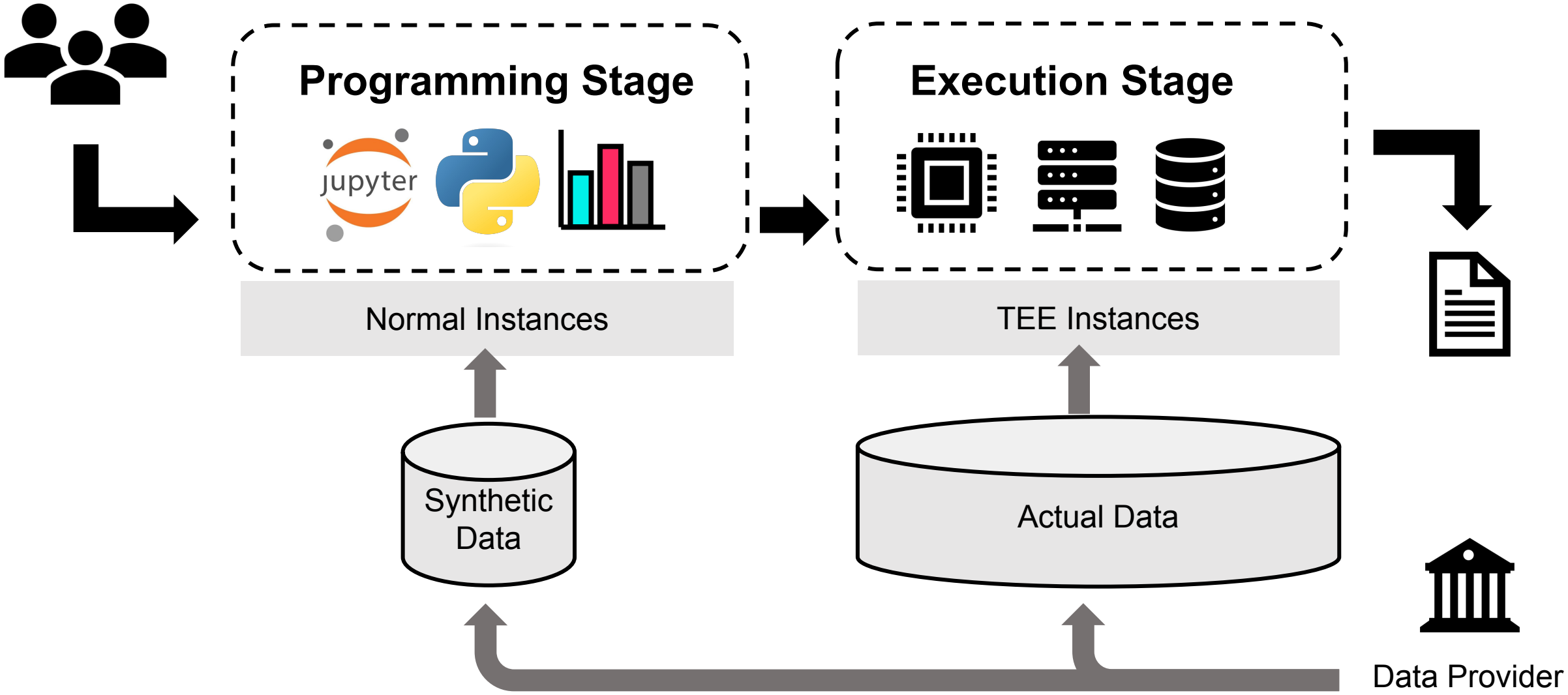


- Larger Data/Compute
- One-Time Execution
- Easier to Control Data
- Lower Privacy Risk



# PGDCR Architecture

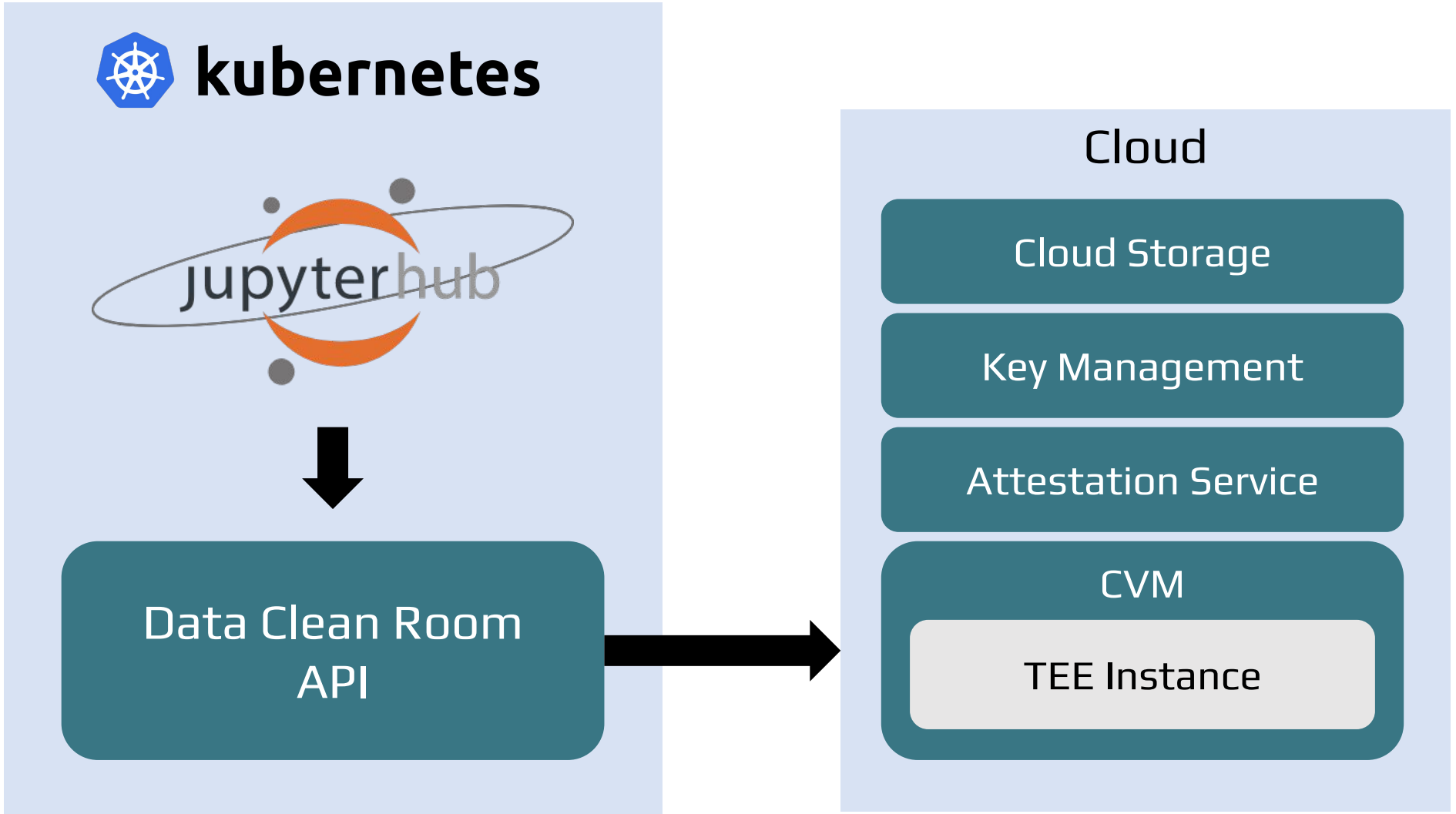
# PGDCR Architecture: Two-Stage Data Clean Room



# Benefits of Two-Stage Data Clean Room w/ TEE

- Data Provider Decides a Protection Mechanism
  - Data at Programming Stage: random data, DP synthetic data, or public data
  - Code/Output Filtering at Execution Stage: can implement coarse-grained policy, instead of per-query policy
- Trusted Execution Environment
  - Provides transition of trust in multi-way data collaboration settings
  - Integrity of code and output
  - Attestation report can be used as a proof of execution
- Accurate Results in Execution Stage
  - Full data access is securely enabled via TEE

# Zero-to-One Cloud Deployment with Terraform



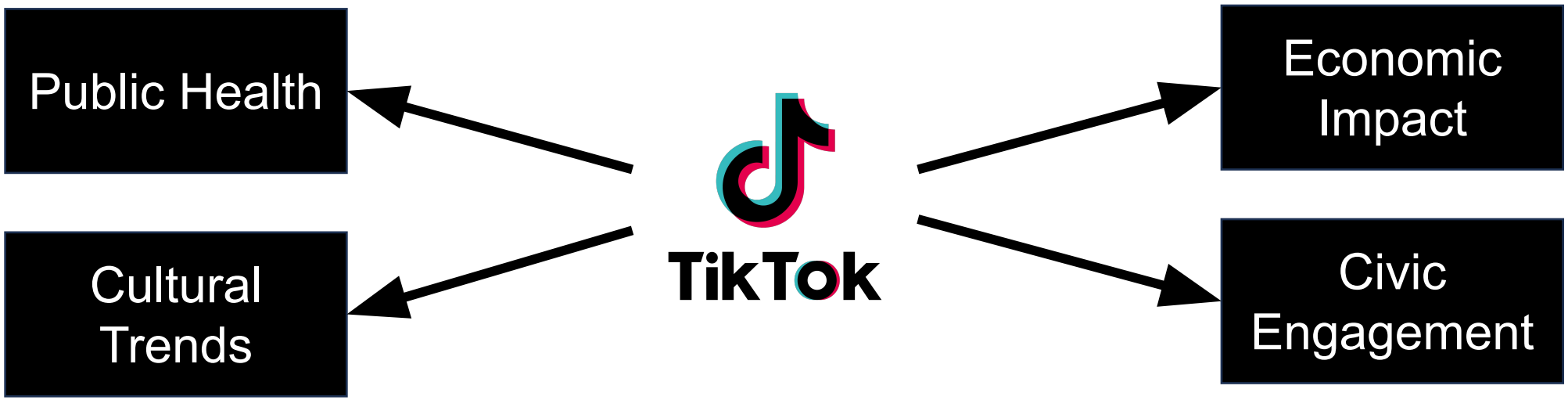
# Demo

# Why PGDCR was built and its use cases

# TikTok

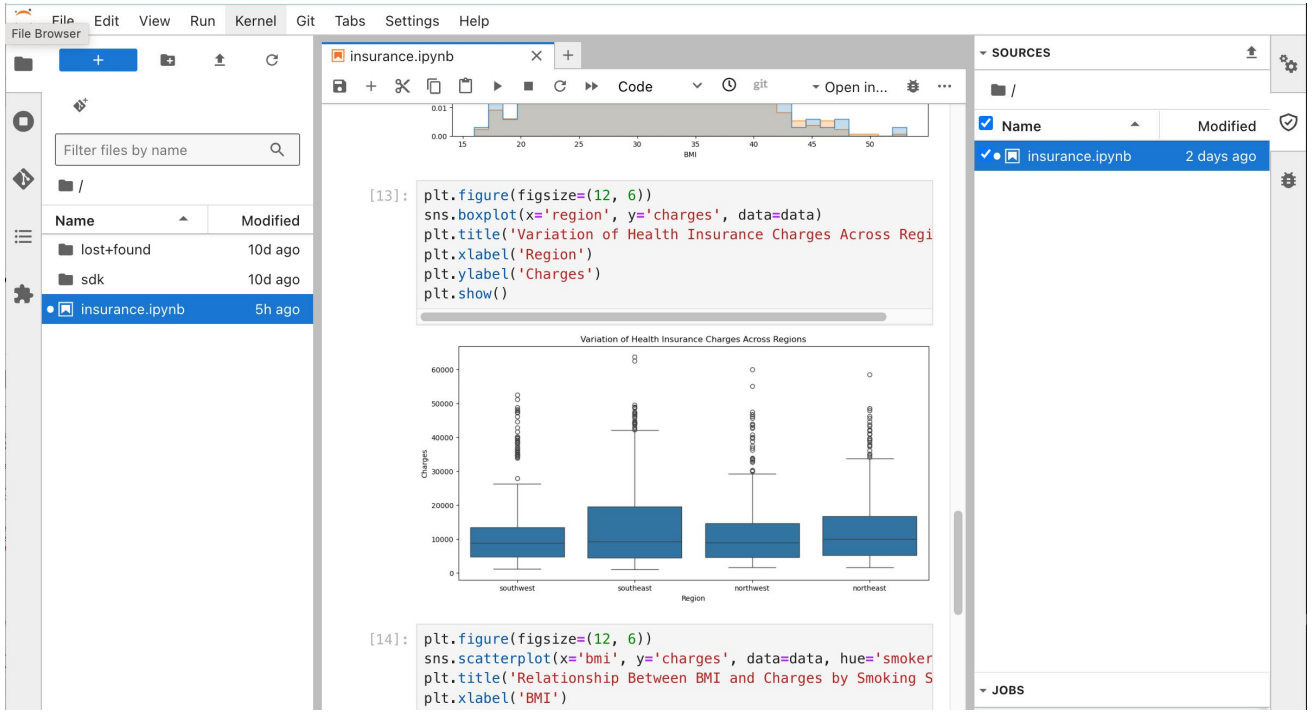
# 1 Billion+ Users

# Providing Transparency to Researchers



# TikTok Research Tools

- Virtual Compute Environment (VCE)
  https://developers.tiktok.com/doc/vce-getting-started



# Other Use Cases

- Ads & Marketing
  - Lookalike segment analysis
  - Measurement and conversion tracking
- Machine Learning
  - Inferencing & training with private dataset
  - Inferencing & fine-tuning private model

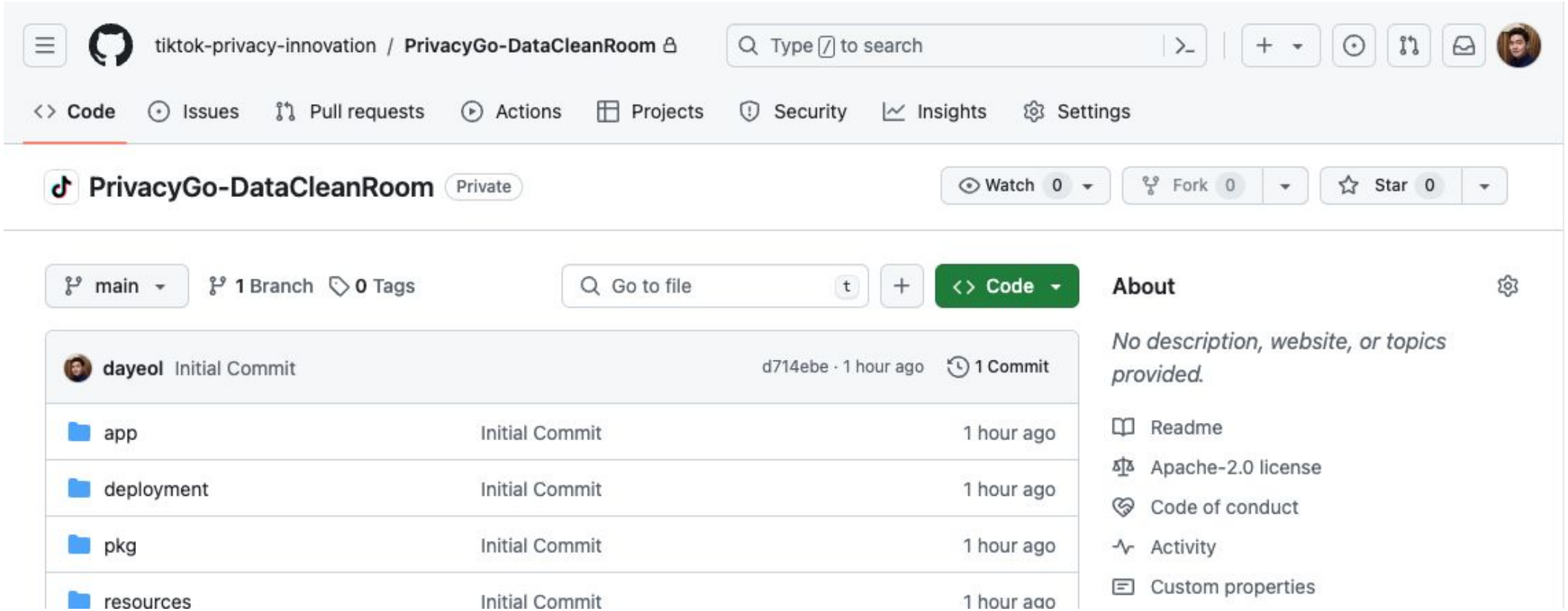
# Project Current Status & Future Growth Plan

| | Current | Future |
|---|---|---|
| Users | One-Way Collaboration | Multi-Way Collaboration |
| Backend | Single Backend | Multiple Backend |
| Data Provisioning | Manual | Automated |
| Policy and Attestation | Manual | Automated |
| Compute | CPU | CPU/GPU |

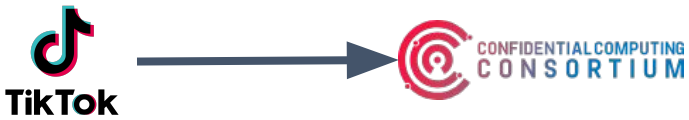
# PGDCR is an Open Source Project

https://github.com/tiktok-privacy-innovation/PrivacyGo-DataCleanRoom

The project was open sourced at CC Summit on June 6, 2024



# Alignment with CCC's Mission




## Why valuable to CCC community?

1. Diversifying the Confidential Computing landscape

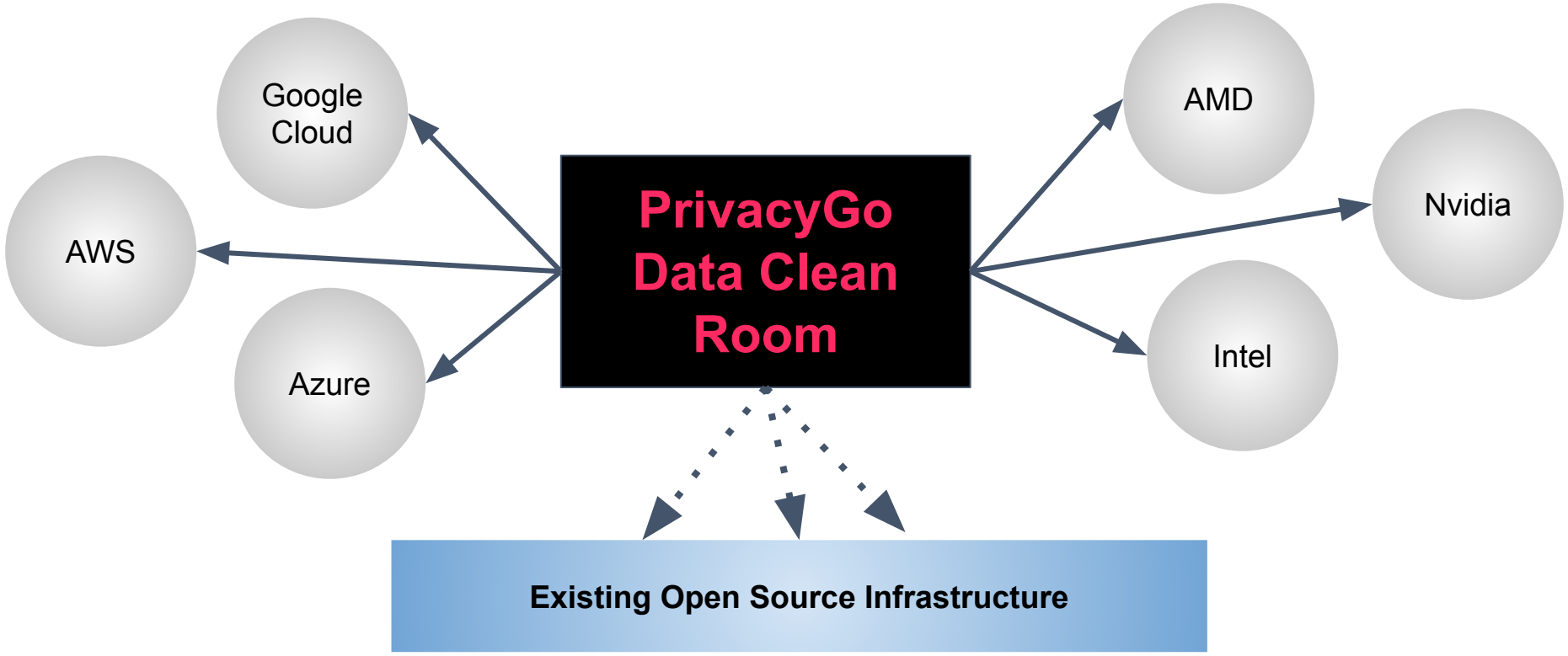
   by providing an open-source solution based on TEE technology
2. Accelerating the CC adoption

   Use case-focused approach by demonstrating how CC can be used to enable advanced use cases of the current industry demands requiring enhanced secure platforms for data collaboration
3. We can drive underlying CC technology to become more mature.

# Alignment with CCC's Mission

## Open Collaboration

Started out using Google cloud and Jupyter based solution. The PGDCR can DCR can utilize existing open source infra, and can be customizable to support multiple backends to build a better platform.



# Find us on online

Learn more on our **website**

Welcome to contribute on **GitHub**

Follow us on **social channels: X | LinkedIn | Medium**

# Q&A

dayeol.lee@tiktok.com
mingshen.sun@tiktok.com
vini.jaiswal@tiktok.com