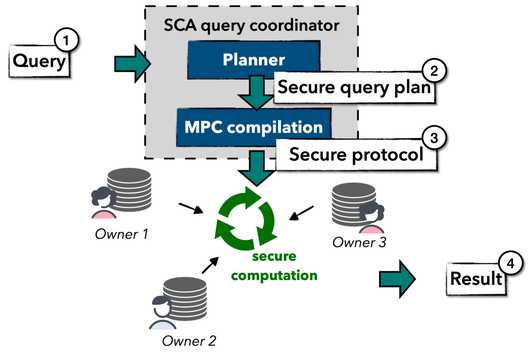
CICI: UCSS: SciPDF: Usable Private Data Federation for Secure Scientific Collaboration (Award# 2419821)

Chenghong Wang (Indiana University), Johes Bater (Tufts University)

Abstract

Private Data Federations (PDF) are powerful tools to address privacy barriers in collaborative scientific research involving sensitive data. However, significant usability gaps have hindered their widespread adoption in practical scientific workflows. SciPDF democratizes the complex PDF pipeline by creating a science-native system that allows the scientific community to easily use advanced PDF features without needing security expertise.

What is private data federation?

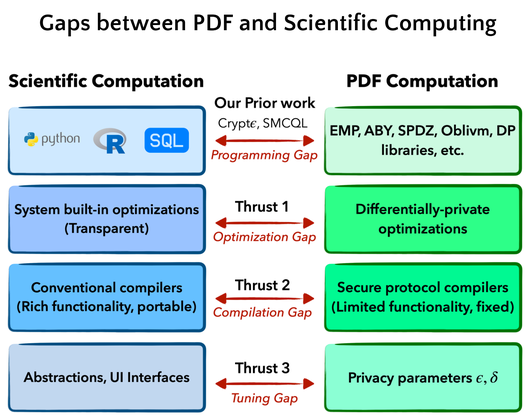


Simply put, PDF is a secure database that enables multiple data owners to pool their private data and perform collaborative queries without directly sharing their raw data.

Computing lifecycles:

1. A vetted analyst (e.g., a scientific researcher) submits an analytical query to a secure collaboration coordinator.
2. The coordinator optimizes the query execution and compiles it into a secure multi-party computation (MPC) protocol spanning multiple data owners.
3. The data owners jointly execute the MPC protocol over their respective private data.
4. The resulting output is encrypted and can only be decrypted by the vetted analyst.

Research objectives



Research Thrusts

Thrust 1. Self-sustaining PDF optimizer. We plan to design a self-sustaining query optimizer that fully automates the PDF optimization.

Thrust 2. Full-fledged, portable PDF compiler. We will develop a portable PDF compiler capable of automatically translating high-level logical queries into executable PDF protocols using various secure computing primitives.

Thrust 3. Policy-controlled system tuning. Our main objective is to enable non-expert admins to configure the PDF system using high-level policies. To achieve this, we will design tools that translate formal privacy definitions into user-digestible privacy semantics in scientific contexts.

Thrust 4. Benchmark and evaluation. A thorough evaluation in collaboration with domain scientists from Yale, Cornell, Emory and IU medical schools.

Current outcomes

SPECIAL System (VLDB 2024)

SPECIAL (Synopsis-Assisted Secure Collaborative Analytics) is a PDF system designed to automatically perform query optimization.

Intellectual Merit (IM): SPECIAL is the first system to leverage private synopses for query optimization in PDF, enabling both privacy-preserving and automatic performance tuning—an unprecedented advancement in secure collaborative analytics.

Broader Impacts (BI): We are exploring integration with medical databases at Indiana University, the design is also completely open-sourced at https://github.com/lovingmage/SPECIAL

Picachv System (Usenix 2025).

Picachv is a system tool that enables data administrators to formally verify data use policies and automatically ensure their compliance and enforcement within private data federations. Our vision is that such a tool can significantly reduce human efforts in future policy compliance practices.

Intellectual Merit (IM): Picachv is the first system to enable formally verified data use policy enforcement for secure data analytics by operating directly over relational algebra query plans. It introduces a novel design where both the semantics of queries and the data use policies are mechanized in the Coq proof assistant, providing machine-checked guarantees of correctness.

Broader Impacts (BI): We are currently exploring extensions of Picachv to modern data lakehouse environments, such as those built on Apache Iceberg or Delta Lake, to bring formally verified policy enforcement to large-scale, real-world analytics infrastructures. The raw implementation of Picachv is open-sourced at https://github.com/picachv

Next steps

Year 2.

1. We are currently developing a new query acceleration tool designed to automatically speed up PDF computations that perform poorly on modern CPUs.
2. We will extend the Picachv system to support policy-driven system configurations and incorporate privacy visualization tools.
3. We will design and implement compilers that can efficient compile PDF logical queries to different protocols (e.g., MPC, trust hardware, etc.)

Year 3.

1. We will explore automated configuration of PDF systems, moving beyond manual interfaces for administrators. We will make these configurations not only automatic but also explainable.
2. We will develop full-fledged benchmarks as a general-purpose performance evaluation framework for PDF systems, enabling comprehensive testing of both our designs and existing state-of-the-art solutions.
3. To maximize broader impacts, we will pursue real-world deployment opportunities, including collaborations with our partners in medical schools to apply these tools in healthcare settings.

Acknowledgements

NSF
Yale
EMORY UNIVERSITY SCHOOL OF MEDICINE
Weill Cornell Medicine