Data Clean Rooms: Enabling Secure and Scalable Data Matching in a Privacy-First World

Poorna Chander Kola

Director of Engineering
Capital One Services LLC

DOI: 10.29322/IJSRP.15.06.2025.p16233
https://dx.doi.org/10.29322/IJSRP.15.06.2025.p16233

Paper Received Date: 15th April 2025
Paper Acceptance Date: 6th June 2025
Paper Publication Date: 28th June 2025

Abstract- In an increasingly privacy-conscious world, the ability to collaborate on data without exposing personally identifiable information (PII) has become a cornerstone of modern data ecosystems. Data clean rooms (DCRs) are emerging as a critical infrastructure for enabling secure, privacy-preserving data analysis and matching across organizational boundaries. This paper explores the concept of data clean rooms, the power of privacy-compliant data matching, architectural patterns, real-world use cases, and the path forward in building scalable, secure data collaboration frameworks.

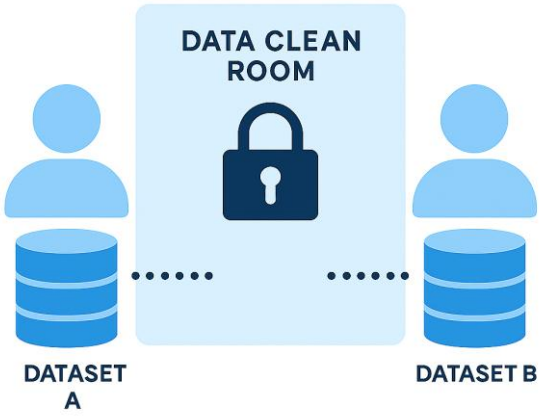
Index Terms- Data clean room, privacy-preserving analytics, identity resolution, secure data matching, differential privacy, data collaboration, multi-party computation, data governance.

I. INTRODUCTION

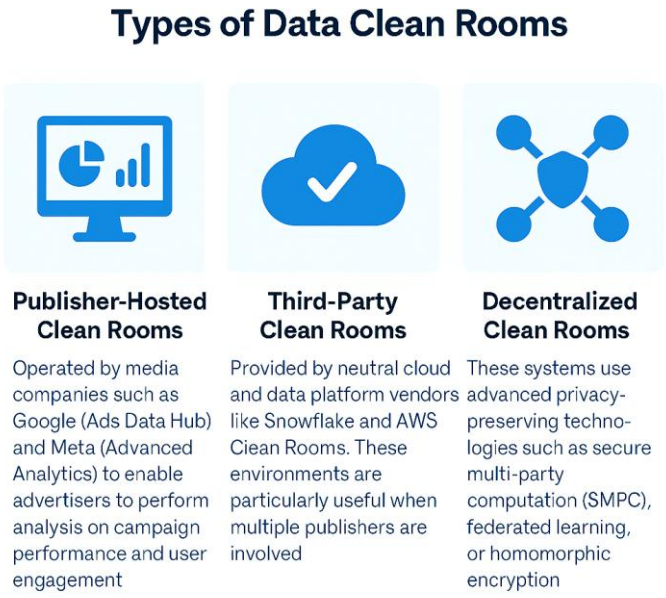
The surge in data regulations such as GDPR and CCPA has drastically reshaped how organizations collect, share, and utilize customer data. While the need for collaboration across data silos grows, so does the importance of preserving individual privacy. Traditional data-sharing mechanisms often fail to address this dual need, creating a gap that data clean rooms are poised to fill. By allowing organizations to analyze and match datasets without revealing raw data, clean rooms offer a future-proof solution to collaborative analytics.

II. Data Clean Rooms

Data clean rooms are secure computing environments where two or more parties can join their datasets for analysis without exposing the underlying data. These platforms rely on privacy-enhancing technologies to ensure that sensitive information is never directly shared or seen by another party.



Types of Data Clean Rooms:



- **Publisher-Hosted Clean Rooms:** These platforms restrict access to user-level data and allow queries only on aggregated outputs, helping maintain privacy while extracting marketing insights. The data never leaves the media company's environment, and advertisers are only allowed to bring their own data in and match it securely.
- **Third-Party Clean Rooms:** Provided by neutral cloud and data platform vendors like Snowflake and AWS Clean Rooms. These environments are particularly useful when multiple brands or institutions want to collaborate without involving a dominant publisher. Third-party clean rooms are more flexible, supporting various data formats, integration models, and compute workloads. They are well-suited for multi-party collaborations and enterprise-wide data governance initiatives.

- **Decentralized Clean Rooms:** These systems use advanced privacy-preserving technologies such as secure multi-party computation (SMPC), federated learning, or homomorphic encryption to enable data matching and analysis across participants without centralizing data. Each party keeps control of its own data, and computation is distributed across nodes. This model is ideal for sensitive domains like healthcare, finance, and government collaborations, where centralization is either a compliance risk or technically infeasible.

Core Features: - Data anonymization and encryption - Access controls and audit trails - Privacy-preserving computation (e.g., SMPC, homomorphic encryption)

III. The Power of Data Matching

At the heart of data clean rooms is the ability to match identities or entities across datasets in a secure way. This matching unlocks powerful business insights, particularly in:

- **Marketing Attribution:** One of the most powerful applications of data matching in clean rooms is linking digital ad impressions to real-world outcomes, such as in-store purchases or call center interactions. This enables advertisers to understand the effectiveness of their campaigns across channels without compromising user privacy. By joining hashed identifiers from advertisers with conversion data from publishers in a clean room, granular insights into consumer behavior can be obtained without accessing individual-level data.
- **Audience Expansion:** Clean rooms allow brands to securely share and enrich audience segments with second- or third-party data. Using secure entity resolution, marketers can identify common user traits and build lookalike models to reach new high-potential audiences. These lookalike segments are generated based on aggregated insights, ensuring no raw user data is exchanged.

- **Fraud Detection:** Financial institutions and digital platforms can collaborate within clean rooms to detect anomalies and suspicious behavior patterns across datasets. By securely linking behavior data (e.g., device fingerprinting, IP activity, login behavior), organizations can flag potential fraud without revealing proprietary or sensitive information. This enables a broader defense against systemic fraud across institutions, while maintaining compliance with data sharing policies.

Matching Techniques:

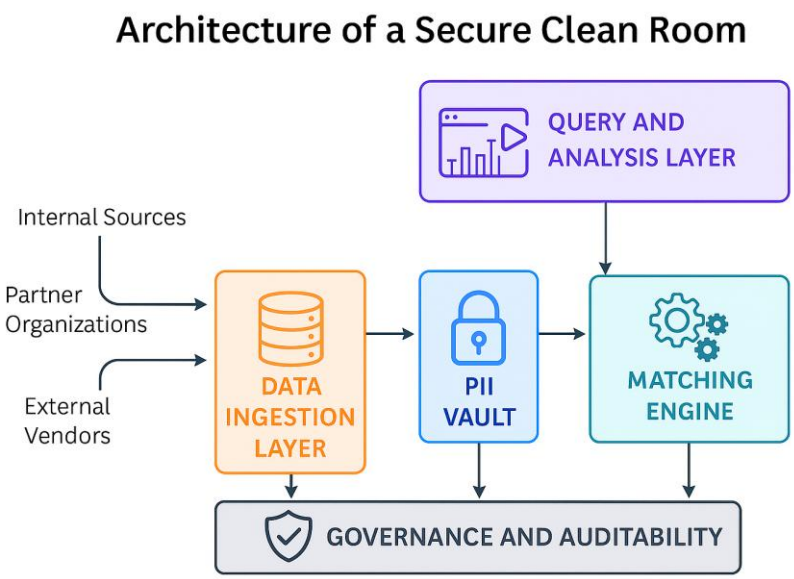
- **Deterministic Matching:** This technique relies on exact matches of unique identifiers that have been pseudonymized or hashed, such as email addresses, phone numbers, customer IDs, or device IDs. Since the identifiers are derived from a known source and consistently formatted, deterministic matching provides high accuracy and low ambiguity. It is commonly used when collaborating parties share a common identity space or have prior consent to link user records across datasets. For instance, a retailer and a brand might hash their customer emails using the same algorithm to identify overlapping users in a clean room without ever revealing the plain text data.
- **Probabilistic Matching:** This method uses statistical techniques, machine learning models, and fuzzy logic to infer likely matches between datasets when deterministic identifiers are unavailable or incomplete. It evaluates a combination of attributes such as name, location, IP address, browser fingerprints, transaction history, and behavioral patterns. These variables are assigned weights, and a match confidence score is computed to determine the likelihood that two records refer to the same individual. While less precise than deterministic matching, it significantly increases coverage and is often the only option in fragmented or anonymous data environments.

Modern clean rooms leverage machine learning models to enhance matching accuracy while maintaining privacy thresholds, often using confidence scoring systems.

IV. Architecture of a Secure Clean Room

A robust clean room architecture includes:

- **Data Ingestion Layer:** Responsible for receiving data from multiple sources—internal systems, partner organizations, or external vendors—and preparing it for secure processing. This involves schema validation, format normalization, deduplication, and encryption of data at the field or file level. Often, ETL pipelines such as Apache Spark, AWS Glue, or cloud-native services are used here. The ingestion layer ensures that no raw PII is exposed or transferred without encryption, applying tokenization or hashing early in the pipeline.



- **PII Vault:** A secure enclave or isolated container designed to manage sensitive identifiers such as names, email addresses, phone numbers, and government-issued IDs. Access to the vault is tightly restricted and governed by fine-grained access controls. It supports secure identity resolution by generating pseudonymous tokens or encrypted values (e.g., via HMAC or SHA-256) that can be used in subsequent matching processes without leaking raw information. These tokens are consistent within a session or across data partners, enabling linkability while preserving privacy.
- **Matching Engine:** This is the core computational unit that performs identity resolution by comparing encrypted or hashed identifiers across datasets. It supports both deterministic and probabilistic matching approaches and may include confidence scoring, duplicate suppression, and clustering algorithms. Some matching engines are enhanced with ML-based techniques for improved scalability and accuracy in complex, high-volume environments. Matching operations are typically executed within privacy-aware computation frameworks to prevent any leakage of intermediary data.
- **Query and Analysis Layer:** Once the data is matched and anonymized, this layer enables analysts and data scientists to query the combined datasets. It ensures that outputs are compliant with pre-defined privacy thresholds, such as minimum aggregation levels or noise-injection for differential privacy. This layer may include SQL engines, dashboarding tools, or APIs with policy-driven query approval workflows to prevent data misuse.
- **Governance and Auditability:** A comprehensive governance framework ensures that all activities within the clean room are transparent and accountable. This includes enforcing role-based access controls, maintaining immutable audit logs of data access and transformations, applying policy checks, and integrating with external consent management systems. Governance also encompasses legal and compliance reviews to certify that all clean room operations meet regulatory standards (e.g., GDPR, HIPAA).

Example: Using AWS services: Clean room could integrate AWS Glue for ETL, S3 for encrypted storage, Lambda for orchestration, and AWS Clean Rooms for matching and analysis.

V. Real-World Use Cases

- **Retail:** A brand and a retailer can join forces in a clean room to analyze purchase behavior, basket size, product preferences, and channel performance. This allows the retailer to share sales insights while the brand contributes advertising and demographic data. By combining anonymized datasets, they can jointly assess the impact of promotions or product placements without disclosing individual customer identities.
- **Healthcare:** Hospitals, research labs, and pharmaceutical companies can collaborate in clean rooms to evaluate the effectiveness of drugs and treatment protocols across broad populations. By using de-identified health records and aggregating outcomes, researchers can draw robust conclusions while remaining compliant with regulations such as HIPAA and GDPR. For example, clean rooms enable multi-institutional COVID-19 studies without sharing sensitive patient-level data.
- **Finance:** Banks, credit bureaus, and fintech platforms can use clean rooms to identify trends in credit defaults, identity theft, or suspicious transactions that span multiple institutions. Sharing behavioral and transaction data in a privacy-safe manner enables coordinated fraud detection and risk mitigation strategies. This collaborative intelligence is especially valuable in combating cross-bank synthetic identity fraud.

- **Advertising:** Clean rooms empower advertisers to understand cross-platform campaign effectiveness. Publishers provide impression and engagement data, while advertisers upload purchase and loyalty data. The clean room environment allows secure matching to determine which impressions led to conversions, enabling granular attribution modeling and return on ad spend (ROAS) optimization—all while keeping individual user data shielded.

VI. Challenges and Future Directions

Despite their promise, data clean rooms face challenges:

- **Interoperability:** One of the most pressing challenges in the adoption of clean rooms is the lack of standardization. Different platforms use varying formats for tokenization, encryption, and data schemas, making it difficult for participants to integrate clean rooms across multiple vendors. Additionally, the absence of a universal API or interoperability layer limits the portability of clean room workloads. This creates friction in multi-party collaborations and hinders ecosystem-wide scalability.
- **Scalability:** Cryptographic computations used in clean rooms, such as secure multi-party computation (SMPC) and homomorphic encryption, are resource-intensive and can introduce significant latency. As data volumes grow and use cases become more complex, maintaining performance while ensuring privacy becomes a key bottleneck. Real-time or near-real-time analytics is especially hard to achieve in high-scale environments without compromising security or incurring excessive compute costs.
- **Usability:** Many clean room platforms are built with technical users in mind, requiring advanced knowledge of cryptography, data engineering, or statistical modeling. Non-technical stakeholders such as marketers, product managers, or legal teams often struggle to navigate the setup, analysis, and governance of clean rooms. A lack of intuitive interfaces, guided workflows, and no-code tools limits broader organizational adoption, thus reducing the potential impact of these solutions.
- Future Trends: - Automated clean room configuration and policy templates - Cross-cloud interoperability via open standards (e.g., Clean Room APIs) - AI-enhanced matching algorithms with privacy-aware learning

VII. Conclusion

As data collaboration becomes vital for innovation, clean rooms offer a scalable and secure solution to match and analyze data across domains without compromising privacy. Their adoption will only grow as organizations seek to comply with privacy regulations while unlocking data's full value. Investing in clean room technology and best practices is not just a compliance decision but a strategic imperative in the era of responsible data stewardship.

REFERENCES

[1] https://cloud.google.com/bigquery/docs/data-clean-rooms
[2] https://www.snowflake.com/en/blog/data-clean-rooms-privacy-collaboration

http://dx.doi.org/10.29322/IJSRP.X.X.2018.pXXXX www.ijsrp.org
International Journal of Scientific and Research Publications, Volume X, Issue X, Month 2018 3 ISSN 2250-3153

AUTHORS

First Author – Poorna Chander Kola, Director of Engineering, Capital One Service LLC, poornachander.kola@gmail.com

http://dx.doi.org/10.29322/IJSRP.X.X.2018.pXXXX www.ijsrp.org