# **The Hidden Challenges of Building a Scalable Data Clean Room**

*Original Research Article*

## **Abstract**

Data clean rooms have emerged as critical infrastructure for enabling privacy-preserving collaborative analytics across heterogeneous data ecosystems. This article presents a comprehensive examination of the architectural and operational barriers organizations face when implementing scalable clean room solutions and offers field-tested architectural patterns and system-level strategies to overcome these challenges. Drawing on over three years of production-grade deployments, the study identifies key implementation bottlenecks related to data schema standardization, multi-cloud security integration, and resource-efficient privacy-preserving computation. The findings are applicable across various industry sectors and provide actionable insights to support secure data collaboration while ensuring regulatory compliance, cost efficiency, and operational scalability. This research addresses critical limitations of current clean room models by proposing concrete technical solutions for cross-cloud data collaboration architectures that accommodate diverse data volumes, complex privacy requirements, and evolving compliance frameworks.

**Keywords:** data clean rooms, privacy-preserving computation, multi-cloud integration, differential privacy, secure multi-party computation, data ingestion, regulatory compliance, data activation

## **1. Introduction**

Data clean rooms have emerged as a pivotal component in the privacy-preserving analytics landscape, enabling inter-organizational data collaboration while maintaining regulatory and ethical data protection standards. As data sharing becomes central to value generation across industries, particularly in finance, healthcare, and marketing, the limitations of conventional architectures have become increasingly evident [13, 21].
 Current data clean room implementation strategies exhibit several critical limitations. First, they suffer from poor frameworks for handling heterogeneous data sources and data types, resulting in poor integration. Second, they are mostly ill-suited for cross-cloud deployments, which create silos and suppress collaborative potential. Third, classical implementations struggle to balance computational capability with high privacy requirements at scale, particularly when handling sensitive data from diverse partners.
Recent analyses have underscored that data collaboration strategies are increasingly driven by the value of partnerships and the analytical utility of data storage infrastructures, rather than their physical location [3]. However, when attempting to deploy production-grade, clean-room environments across heterogeneous cloud providers, organizations frequently encounter a range of architectural and operational bottlenecks. These include delayed data ingestion caused by schema misalignment, fragmented integrations stemming from incompatible cloud-native security frameworks, and computational inefficiencies that deteriorate under parallel multi-party workloads, ultimately impeding scalability as the number of collaborators increases.

Drawing on extensive field deployments and empirical evaluations [4], this study identifies critical failure modes in existing clean room implementations. It proposes a set of targeted architectural interventions to mitigate these limitations. The paper consolidates both technical insights and implementation experiences into a comprehensive methodological framework, enabling system architects and data engineering teams to design agile, secure, and horizontally scalable clean room infrastructures that are adaptable across various industry verticals.

## **2. Securing Analysis Through Privacy-Preserving Computation**

The key to a successful data clean room is its ability to protect sensitive information while it is being analyzed and processed. Existing research has demonstrated that significant hurdles exist in deploying advanced, privacy-preserving technologies at scale [1]. This presents a clear technical challenge that necessitates meticulous architectural consideration to strike a balance between analytical utility and privacy protection.

### **2.1 Computational Performance Constraints**

Privacy-preserving protocols are secure but at computation costs that initial benchmarks grossly underestimate [14, 16, 17]. Recent studies indicate that secure multi-party computation organizations experience a catastrophic performance slowdown when scaling beyond eight data sources [2, 18]. Such constraints necessitate architectural innovations that can provide acceptable performance while still maintaining privacy guarantees.

Beyond raw computational challenges, organizations must also consider how these privacy-preserving technologies integrate across diverse cloud security environments. The performance constraints naturally intersect with the practical security implementation challenges that arise in multi-cloud deployments.

### **2.2 Security Framework Integration Challenges**

Implementing secure computations across AWS, Azure, and GCP environments can generate compatibility issues within different security environments [15]. Consistent with industry research, 73% of organizations attempting multi-cloud clean room deployments attribute integration issues requiring customized development as the reason [11]. The remedy to such problems entails standardizing security layer deployment methods.

Table 1 illustrates that these technologies exhibit varying trade-offs between security and performance.



*Table 1.* Comparative analysis of privacy-preserving technologies based on implementations. *The chart illustrates the comparative performance of major technologies across security strength, computational efficiency, and implementation complexity, encompassing differential privacy, secure multi-party computation, and federated learning.*

## **3. Streamlining Multi-Source Data Ingestion**

Data ingestion is often the weakest link in clean room deployments. Ingesting data from several dozen external sources without compromising privacy and compliance has led to significant project delays in real-world implementations.

### **3.1 Resolving Data Format Inconsistencies**

A particularly challenging retail project highlighted this issue —data originated from historic mainframe dumps to newer flows of JSON, presenting formatting issues for typical analysis environments. Recent industry surveys indicate that standardizing data formats is a top bottleneck in 68% of clean room deployments [4].

Effective approaches for addressing format inconsistencies include implementing schema enforcement using Apache Avro, which substantially reduces integration problems, and re-engineering data flows with tools like Apache NiFi to normalize schemas before clean-room ingestion. These strategies have demonstrated significant reductions in integration timelines and marked improvements in overall data quality, streamlining the entire ingestion process across varied data sources.

While resolving format inconsistencies addresses the structural challenges of data ingestion, even properly formatted data can contain quality issues that undermine clean room operations. Beyond standardization, organizations must implement robust quality control mechanisms to ensure the integrity of the data they collect and process.

### **3.2 Proactive Data Quality Assurance**

Detection of data inconsistencies during ingestion prevents cascading issues in subsequent analysis phases. Data quality issues have been cited as the cause of about 40% of clean room failures [5]. Adherence to strict quality assurance processes is a key driver of success.

Implementation strategies for quality assurance should integrate continuous data profiling that analyzes incoming data structures in real-time, helping teams identify anomalies before they enter processing pipelines. This approach, complemented by the incorporation of automated validation protocols, ensures conformity with required specifications and significantly minimizes analytical errors downstream. Together, these proactive measures form a comprehensive quality framework that stabilizes the entire clean room operation.

## **4. Navigating Multi-Cloud Data Architectures**

Modern data clean rooms often span multiple cloud environments, introducing significant challenges in ensuring data consistency, security, and availability across these platforms. As organizations become increasingly inclined toward multi-cloud strategies, clean room deployment must support diverse infrastructure needs without introducing friction-based functionality.

### **4.1 Creating Seamless Cross-Cloud Integration**

Modern organizations utilize multiple cloud providers for various functionalities, resulting in disparate data environments that significantly complicate clean room operations. A significant number of organizations report challenges with inconsistent workflows and data management across multi-cloud environments, which can impact performance and scalability [6]. Our field deployments across various industries have identified several critical integration challenges, including data fragmentation across clouds, incompatible security models, complex cross-boundary metadata management, and network configuration variations causing performance bottlenecks [12]. To address these challenges, successful implementations have developed comprehensive architectural solutions that combine container orchestration-based (Kubernetes) cloud-agnostic infrastructure for platform-independence, synchronized data layers with event-driven updates to maintain consistency, federated identity and access management systems for unified access controls, and standardized API gateways providing a single data access interface that abstracts underlying complexity [7]. As Table 2 illustrates, an effective architecture must integrate these specialized components to enable secure and seamless operations across diverse cloud environments, while maintaining both performance and security standards.



*Table 2.* Key components of multi-cloud data clean room architecture. *The table illustrates the essential architectural elements for implementing a clean room across diverse cloud environments, including data ingestion layers, privacy-preserving computation modules, and secure activation frameworks.*

While establishing seamless integration across cloud environments forms the foundation of multi-cloud clean rooms, the efficient allocation and management of computational resources across these integrated environments presents its own distinct set of challenges. Organizations must not only connect these environments but also optimize how processing power is distributed and utilized within them to balance performance requirements with cost considerations.

### **4.2 Optimizing Computational Resource Allocation**

The size and location of compute resources significantly impact both performance and operational costs in multi-cloud clean room implementations. Research indicates that poor resource allocation can account for up to 45% of operational costs for large clean room deployments [7], highlighting the critical need for optimization that balances processing requirements with data sovereignty and privacy requirements. Real-world implementations have revealed several clean room-specific resource challenges, including inter-cloud latency and data gravity issues, complexity in cost management across varying pricing models, computational locality limitations imposed by privacy-preserving measures, and workloads with varied computational demands and unpredictable patterns [7]. To address these challenges, successful production deployments have implemented advanced resource optimization strategies including workload-aware dynamic allocation, event-driven computational frameworks that activate resources only when needed, distributed edge processing to minimize data movement, intelligent caching strategy to reduce redundant computations, and continuous monitoring systems that enable ongoing refinement of allocation approaches [7]. These strategies, as illustrated in Figure 1, have a significant impact on both the operational efficiency and cost management of multi-cloud deployments.



*Figure 1.* **Computational resource optimization strategies for multi-cloud clean rooms.** *This visualization demonstrates approaches to resource allocation across distributed cloud environments, highlighting methods for balancing performance requirements with cost considerations while maintaining privacy guarantees.*

Effective optimization relies on establishing appropriate performance measures that balance business objectives with technical efficiency, including key performance indicators such as query response time distribution, data transfer efficiency, resource utilization density, and cost-per-query metrics. These KPIs form the foundation for continuous improvement in resource allocation, enabling organizations to develop economical clean room facilities that operate within controlled expenses while maintaining necessary privacy safeguards.

## **5. Enabling Secure Data Activation and Export**

Once processed in the clean room, data must be securely exported to external applications or systems without compromising privacy protections or regulatory compliance. Secure activation workflows have been identified as a challenge for the majority of organizations [8]. Successful activation is the key final step to achieving business value from clean room deployments.

### **5.1 Implementing Protected Data Export Mechanisms**

Facilitating access to privacy-controlled exports during export phases is necessary to guarantee security and compliance with the law [20, 22]. From our analysis, insecure export controls represent a critical vulnerability in otherwise secure clean room environments. Institutions must have robust export controls in place that maintain privacy guarantees for the entire data life cycle. Our experience in implementing comprehensive security measures reveals that successful export security regimes require multiple integrated components working in concert, including role-based access controls with granular export authorization that permit only authorized users to export specific data elements, rigorous output validation processes that scan exported data for potential privacy violations, comprehensive audit trails that maintain detailed records of all export activities, standardized export formats with embedded privacy metadata, and digital rights management integration that provides persistent controls for protection against unauthorized use [8]. Several design patterns have proven particularly effective in protecting the export process without compromising usability, including progressive disclosure mechanisms that provide graduated levels of access, purpose-specific export templates for common use cases, privacy budget management systems to reason about cumulative privacy impacts, and approval workflow automation based on predefined policies that balance security with operational efficiency [8]. These patterns form the foundation for designing export mechanisms that effectively balance stringent privacy requirements with operational effectiveness.

While securing the export process is critical, organizations must also consider how exported data will be activated and utilized in downstream systems to realize business value. Beyond the mechanics of secure export, the challenge extends to ensuring that privacy-protected data can be effectively integrated with and leveraged by the broader analytics ecosystem.

### **5.2 Streamlining Downstream Data Activation**

Processed data must be adequately exposed for future use in applications such as machine learning or business intelligence systems. Research on data integration reveals that activation bottlenecks can hinder insight generation in most deployments [9]. Removing such bottlenecks presents the first opportunity to maximize the return on investment from clean room deployments. Field deployment across various industry domains has uncovered several typical activation problems, including privacy-preserving data flow control, output format incompatibilities, process efficiency and automation deficits, governance rule propagation, and contextual metadata persistence [9]. Several advanced methods have demonstrated remarkable effectiveness in maximizing activation potential without compromising essential privacy protections, including automated privacy analysis workflows, pre-configured integration adapters for common destination systems, comprehensive activation pipeline choreography, federated activation patterns, and strategic insertion of privacy-enhancing technologies into activation workflows [9]. As shown in Table 3, different use cases require tailored activation approaches to strike a balance between utility and privacy protection.



*Table 3.* Data Export and Activation Strategies by Use Case. *This table outlines appropriate export mechanisms for common clean room scenarios, mapping business requirements to technological approaches while preserving privacy guarantees and regulatory compliance.*

Organizations across sectors have developed customized activation patterns to address their unique requirements; financial services implementations typically emphasize rigorous audit trails and verification mechanisms with multi-stage approval workflows, healthcare organizations often incorporate robust de-identification mechanisms with statistical disclosure controls, while retail and marketing applications frequently leverage purpose-built connectors with built-in anonymization for near real-time activation [8]. Research confirms that context-specific optimization can improve both compliance confidence and operational efficiency [9]. Establishing appropriate metrics for activation effectiveness enables organizations to optimize their clean room implementations, with key performance indicators including time-to-insight, confidence in privacy preservation, activation success rate, and downstream integration efficiency. Effective solutions combine algorithmic anonymization with automated data delivery systems to reduce manual handling and accelerate availability to downstream processes. This establishes a framework for consistent, secure, and efficient data activation, ensuring privacy guarantees are maintained throughout the data lifecycle.

## **6. Conclusion**

Creating truly scalable data clean room infrastructure requires addressing multifaceted challenges in data ingestion, privacy-preserving computation, cross-cloud architecture, and secure data activation. By implementing comprehensive validation systems, real-time data quality monitoring, cloud-agnostic infrastructure, and efficient computational resource management, organizations can develop clean room environments that scale effectively while maintaining privacy and regulatory compliance.

Recent industry analyses highlight the growing importance of these solutions across various sectors, where collaborative data analysis provides substantial business value while operating within increasingly stringent privacy constraints **[10]**. The approaches outlined in this article provide a foundation for organizations seeking to implement robust, scalable clean room environments that support growth while protecting sensitive information.

As privacy regulations continue to evolve and data collaboration becomes increasingly important for gaining a competitive advantage, properly implemented clean room solutions will play a crucial role in enabling secure, compliant data sharing across organizational boundaries. The technical foundations presented in this article provide a framework for organizations to develop effective clean room implementations that strike a balance between privacy requirements and analytical utility.

With these architectural patterns and implementation strategies in place, data clean rooms can fulfill their promise of enabling secure collaborative analytics, allowing organizations to derive insights from combined datasets while preserving individual data privacy and meeting regulatory requirements.

**COMPETING INTERESTS DISCLAIMER:**

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc. have been used during the writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

## **7. References**

1. Dwork, C., & Roth, A. (2014). "The Algorithmic Foundations of Differential Privacy." Journal of Privacy and Confidentiality, 9(3-4), 211-407.
2. Evans, D., Kolesnikov, V., & Rosulek, M. (2018). "A Pragmatic Introduction to Secure Multi-Party Computation." Foundations and Trends in Privacy and Security, 2(2-3), 70-246.
3. Decentriq. (2024). "What is a data clean room, and what can you do with one?" Retrieved from<https://www.decentriq.com/article/what-is-a-data-clean-room>
4. IDC Research Group. (2024). "Data Clean Rooms: Secure and Private Data Collaboration." IDC Market Analysis.
5. DataCo. (2022). "Advanced data clean rooms to meet the needs of regulation-compliant global collaboration." Retrieved from<https://www.linkedin.com/pulse/advanced-data-clean-rooms-meet-needs-regulation-compliant->
6. HashiCorp**.** (2021). *State of Cloud Strategy Survey Report*. Retrieved fromhttps://www.hashicorp.com/state-of-cloud/2021
7. Johnson, O. B., Olamijuwon, J., Cadet, E., Osundare, O. S., & Samira, Z. (2024). Designing Multi-Cloud Architecture Models for Enterprise Scalability and Cost Reduction. Open Access Research Journal of Engineering and Technology, 7(2), 101–113.
8. IAB Tech Lab. (2023). Data Clean Room Guidance. IAB Technology Laboratory.​" Retrieved from
https://iabtechlab.com/wp-content/uploads/2023/02/FINAL-DRAFT-PUBLIC-COMMENT-Data-Clean-Room-Guidance-IAB-Tech-Lab.pdf
9. Kaissis, G., Makowski, M., Riegler, M., & Braren, R. (2020). End-to-end privacy-preserving deep learning on multi-institutional medical data. *Nature Machine Intelligence*, 2(6), 305–311. <https://doi.org/10.1038/s42256-021-00337-8>​[tum.de+5](https://portal.fis.tum.de/en/publications/end-to-end-privacy-preserving-deep-learning-on-multi-institutiona?utm_source=chatgpt.com)
10. Federal Trade Commission. (2024). "Data Clean Rooms: Separating Fact from Fiction." Retrieved from<https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/11/data-clean-rooms-separating-fact-fiction>
11. AdMonsters. (2022). "What Is a Data Clean Room?" Retrieved from<https://www.admonsters.com/what-is-a-data-clean-room/>
12. Kavis, M. J. (2020)**.** *Architecting the Multi-Cloud: Design Patterns and Best Practices*. In: *Proceedings of the IEEE International Conference on Cloud Engineering (IC2E)*.
13. Campbell, M. (2022). Privacy-preserving computation: doomed to succeed. *Computer*, *55*(8), 95-99.
14. Tuyishime, E., Balan, T. C., Cotfas, P. A., Cotfas, D. T., & Rekeraho, A. (2023). Enhancing cloud security—proactive threat monitoring and detection using a siem-based approach. *Applied Sciences*, *13*(22), 12359.
15. Herbrich, T. (2022). Data clean rooms. *Computer Law Review International*, *23*(4), 109-120.
16. **Cheng, Y., Li, W., Qin, S., & Tu, T. (2025).** Differential Privacy Federated Learning Based on Adaptive Adjustment. Computers, Materials & Continua, 82(3), 4777–4795. <https://doi.org/10.32604/cmc.2025.060380>
17. **Shan, F., Mao, S., Lu, Y., & Li, S. (2024).** Differential Privacy Federated Learning: A Comprehensive Review. International Journal of Advanced Computer Science and Applications, 15(7). <https://doi.org/10.14569/IJACSA.2024.0150722>
18. **Fu, J., Hong, Y., Ling, X., Wang, L., Ran, X., Sun, Z., Wang, W. H., Chen, Z., & Cao, Y. (2024).** Differentially Private Federated Learning: A Systematic Review. arXiv preprint arXiv:2405.08299. <https://arxiv.org/abs/2405.08299>
19. **Gu, H., Luo, J., Kang, Y., Yao, Y., Zhu, G., Li, B., Fan, L., & Yang, Q. (2024).** FedAdOb: Privacy-Preserving Federated Deep Learning with Adaptive Obfuscation. arXiv preprint arXiv:2406.01085. <https://arxiv.org/abs/2406.01085>
20. **Li, K., Chen, X., Leng, L., Xu, J., Sun, J., & Rezaei, B. (2025).** Privacy Preserving Conversion Modeling in Data Clean Room. arXiv preprint arXiv:2505.14959. <https://arxiv.org/abs/2505.14959>
21. **Future of Privacy Forum. (2024).** Data Clean Rooms: A Taxonomy & Technical Primer. Future of Privacy Forum Discussion Draft. <https://fpf.org/wp-content/uploads/2024/09/FPF-Data-Clean-Rooms-Discussion-Sept-2024.pdf>
22. **Desai, H. (2025).** Identity Graphs and Privacy-Safe Data Collaboration: A Guide to Modern Data Clean Rooms. International Research Journal of Engineering and Technology (IRJET), 12(2), 390–392.