

 integrate.ai

integrate.ai vs Data Clean Rooms Comparison



Contents

01 **Executive Summary**

02 **Overview**

Summary Comparison

03 **Key Differentiation**

3.1 Federated Learning

3.2 Fit for Purpose

3.3 Cross-Platform

04 **Conclusion**



01. Executive Summary

For the purposes of data evaluations and collaborative data experimentation, integrate.ai's Federated Learning based platform is more fit-for-purpose than traditional data clean rooms. Data clean rooms are versatile tools for preserving privacy between two parties, but they lack the pre-built tools necessary to perform scalable and repeatable data evaluations. These types of collaboration jobs would need to be custom-built and configured by the parties for each type of job, which brings with it both execution risk as well as risk of data leakage due to misconfiguration. The integrate.ai platform is built specifically for the evaluation of third party data and collaboration among multiple parties.

There are three key advantages to integrate.ai over the use of traditional data clean rooms in the context of external data evaluation and collaboration:

1. FEDERATED LEARNING

Federated Learning guarantees the non-movement of data for both regulatory (GDPR, HIPAA, data residency, etc.) and contractual purposes, while data clean rooms do not, by virtue of the underlying technologies they use. As such, IAI is better suited for use cases involving regulatory constraints and highly sensitive data, and provides downstream protection from regulatory tightening around data privacy. Furthermore, Federated Learning allows for the elimination and/or shortening of legal and privacy review cycles because record-level data is neither viewable by the collaborating parties nor moving from its safe perimeters by design.

2. FIT FOR PURPOSE, SCALABLE AND REPEATABLE

integrate.ai's platform was designed from the ground up as a platform for scalable data evaluation and experimentation, while data clean rooms are a set of primitives which require customization for each use case. For use cases like pre-sales data evaluations, PoC's, and champion/challenger analyses, IAI offers a productized solution that can be used across multiple partners and use cases, while data clean rooms would require internal resources to "productize" them for each use case and each partner.

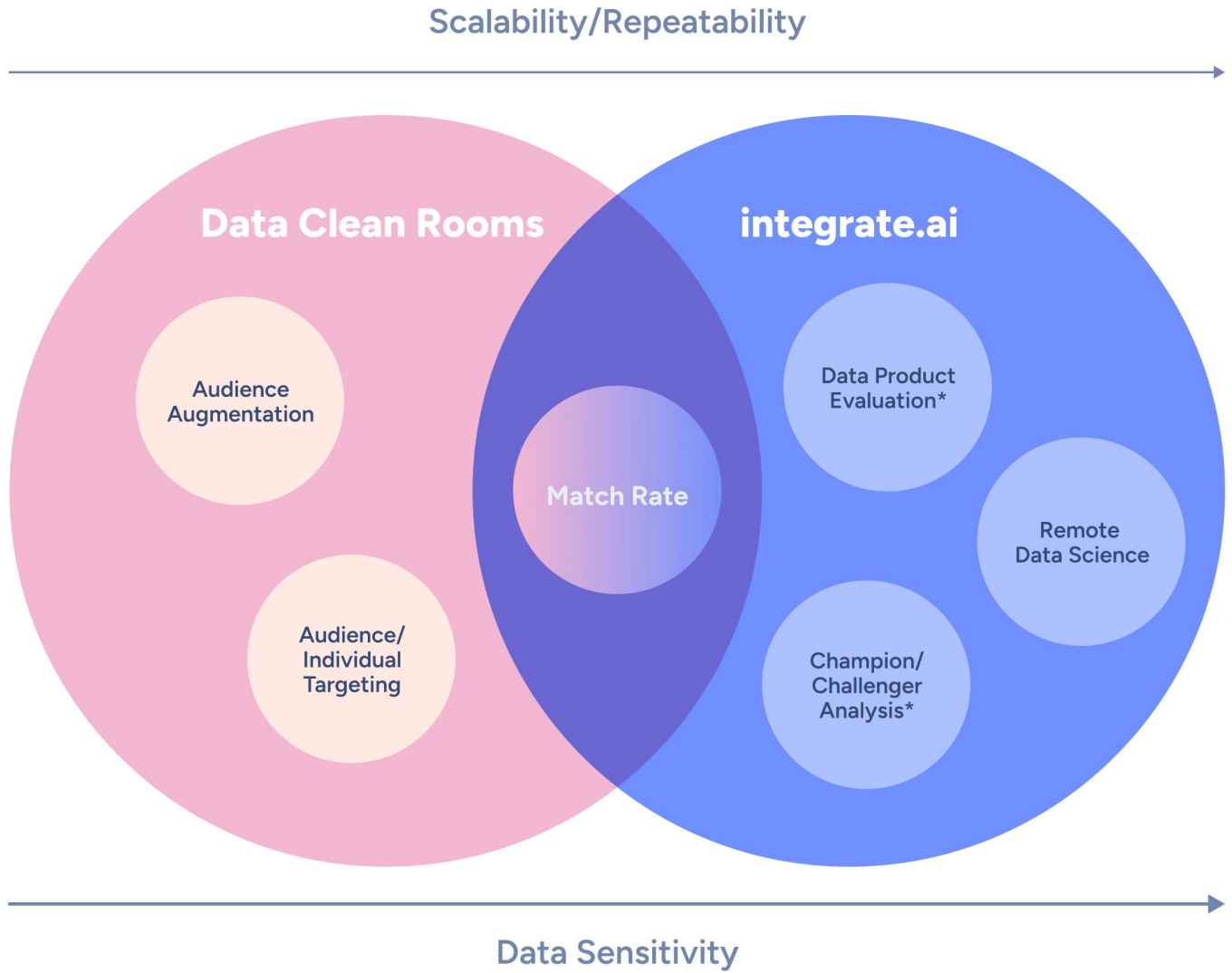
3. CROSS-COMPATIBILITY

With integrate.ai, customers can invite third party guests into a collaborative workspace at no cost. In contrast, data clean rooms require all parties to be customers of the provider and sometimes require aggressive prepaid compute plans (e.g., Databricks, Snowflake). This creates a significant barrier to conducting scalable and repeatable data evaluations across multiple collaborators using a traditional data clean room. This restriction is by design, because the primary business model of data clean room providers is to achieve lock-in to their compute environment, whereas IAI is incentivized to promote cross-platform collaboration. Furthermore, most data clean rooms do not support multi-region collaboration. This means that not only do all parties need to be customers of a data clean room, but there are also restrictions on the regions that can participate in data clean rooms with each other.

GENERAL PURPOSE DATA PLATFORMS VS DATA CLEAN ROOMS

In many cases the data clean room vendors are also vendors of general purpose data platforms (e.g., Databricks, Snowflake, etc). Integrate.ai seamlessly interoperates and integrates with these data platforms and the comparisons in this document are solely focused on the data clean room products.

Summary Comparison



*Insurance-oriented use cases where target variables may be sensitive, such as claims or policy data



02. Overview

integrate.ai

integrate.ai enables businesses to safely collaborate on data and data science jobs to unblock, enhance, and accelerate a variety of internal and external engagements. It incorporates pre-packaged tools and configurations to support specific use cases (detailed below). It's built upon Federated Learning as a means to ensure that data does not move outside one's environment, thus reducing the legal and governance burdens associated with collaborations.

CORE PROBLEM IT SOLVES

Avoid the legal and technical hurdles associated with moving data within and across organizations while still enabling a broad range of data science workloads and collaborations.

KEY USE CASES

- Increase the velocity of vendor / data evaluation
- Data Science across internal data silos (e.g., across data residency regions, subsidiaries, or business units)
- Ongoing Data Science experimentation with strategic vendors/ customers/partners

KEY TECHNOLOGIES

- Federated Learning
- Differential Privacy
- Homomorphic Encryption

Data Clean Rooms

Data clean rooms allow businesses to collaborate on sensitive data in a secure environment.

CORE PROBLEM IT SOLVES

Collaborate on data with external parties with strict governance over the usage and visibility of the data and results.

KEY USE CASES

- Category management for retail and consumer goods
- Personalization with expanded interests for retailers
- Know Your Customer (KYC) in banking
- Audience overlap exploration for media and entertainment

KEY TECHNOLOGIES

- Access Control
- Data Mapping
- Anonymization
- Encryption



03. Key Differentiation

3.1 FEDERATED LEARNING

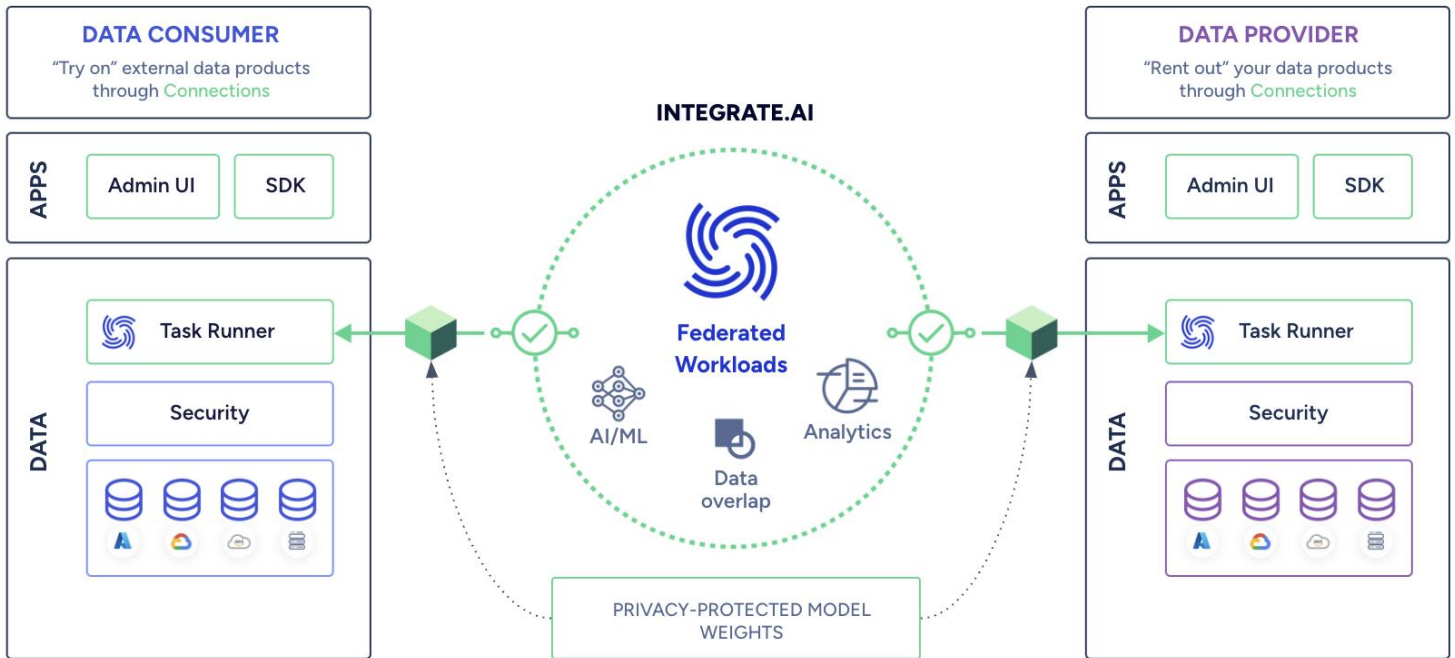
integrate.ai

integrate.ai is built upon Federated Learning, which provides the following data governance guarantee:

Raw data never moves out of the respective environments of each data custodian; integrate.ai never has direct access to data assets, nor do the collaborating parties

The artifacts that are shared between parties in a collaboration are privacy-protected aggregate artifacts, such as model weights, and cannot be reconstructed back to the original data. As such, the integrate.ai platform does not require the parties to trust one another or rely exclusively on a contract around the usage of the data.

In addition to the underlying trustless architecture of federated learning, Integrate.ai also offers governance controls which enable participants to control what jobs collaborators can perform and how the output artifacts are managed and shared (e.g., trained models, analytical results).



Data Clean Rooms

All data clean room solutions require the movement of data and do not use Federated Learning. Where Federated Learning is a decentralized computation method, these solutions are using centralized computation methods. This means, data clean rooms require participating parties to securely transfer their data into the data clean room environment to be anonymized and encrypted. Depending on the data clean room product, data transfer can be done via integrations with existing data platforms, secure file transfer protocols or direct database connections. Unlike federated learning, this means that there is no guarantee that the data will stay in its original environment, or within its original region.

This means in cases with sensitive data, one or more of the parties may be in breach of data governance requirements in a way that is unknown and uncontrolled.

This is of particular concern when there are data residency requirements. Considering the potentially unmanaged risk of data movement, even if one of the collaborating parties is willing to accept that risk, it's unlikely that all parties will be willing to do so without detailed legal and infosec review of each individual engagement. Consider how data is transferred for the following data clean room products:

databricks

Databricks Clean Rooms use a protocol called Delta Sharing as the underlying connectivity mechanism. Delta Sharing involves the movement of data from one party's environment to another when performing operations. This movement cannot be directly managed by the data custodian (i.e., data custodians don't control the properties of the receiving environment, such as region or security posture).

snowflake®

The Snowflake Data Clean Room offering has two options for usage, either through a web app or a developer API.

The web app is geared towards non technical users leveraging templates that focus primarily on adtech use cases. The web app involves the movement of data from one party's environment to another when performing operations. It leverages a few standard processing regions, meaning that using this tool can result in your data being processed in a different cloud platform and region than your Snowflake account.

The developer API is more robust and enables a user to build and customize their own data clean room, leveraging the Snowflake Native App Framework. This offering supports a linkage concept, meaning data will not be copied or moved into the data clean room, but only if all participating parties have Snowflake accounts and the participating data exists in the Snowflake data platform.

Both Databricks Clean Rooms and Snowflake Data Clean Rooms provide a layer of governance controls as part of their offering that can be used to limit the usage of data to specific data science tasks. However, these controls must be custom designed by the data custodian to ensure that they are broad enough to allow the desired use case but narrow enough to prevent unauthorized usage. The responsibility falls on the data custodian for each individual collaboration, requiring auditing and custom trust agreements between the parties.

◆ Implications and Benefits of Using [integrate.ai](#)

Without the movement or exposure of data assets, [integrate.ai](#) enables rapid and repeatable cross-organization engagements, avoiding the overhead of detailed legal and privacy scrutiny for each interaction. This can increase the volume and speed of engagements between data vendors and prospective buyers by 10x compared to traditional collaboration approaches.

In contrast, data clean rooms require the movement of data, which necessitates close legal and privacy scrutiny for each engagement to meet governance requirements. This results in a slower pace of engagements and a risk of misconfiguration while the volume and speed of engagements are not meaningfully increased.



03. Key Differentiation

3.2 FIT FOR PURPOSE

integrate.ai

integrate.ai has been designed to serve a specific range of collaboration scenarios in highly regulated industries, namely:

1. Data vendors engaging prospective customers
2. Data consumers engaging prospective vendors
3. Unlocking internal data science teams where data movement is a hurdle (e.g., data residency; cross line of business)
4. Engaging strategic partners in ongoing collaboration (e.g., strategic vendors)

Across these use cases, the integrate.ai platform is fit-for-purpose in the following ways:

SAFETY AND SECURITY

- Use of Federated Learning without the movement of data
- Roles and permissions are pre-configured for the specific engagement scenarios

SCALABILITY AND REPEATABILITY

- Federated Learning offers privacy-by-design, which eliminates the need to set up custom governance controls for each use case to protect the underlying data
- Pre-built tools designed specifically for external data evaluation and common data collaboration scenarios with pre-established safety controls
- Lightweight installation that does not disrupt existing data platform investments (e.g., fully interoperable with Databricks, Snowflake, etc)
 - Full cross-platform compatibility
 - Aligned with data science experiment infrastructure and tools
- 'Guest Access' feature provides a lightweight, streamlined way to on-board collaborators without additional contracting or disrupting their platform investments

FEATURE RICH

- Landing pages to support the specific mode of collaboration (e.g., storefront for data products)
- 'Projects' make it easy to safely engage multiple parties in a manner that is directly fit to the use case (e.g., insurance carrier can invite multiple vendors to a single project to compare results without exposing the parties or results to the vendors)
- Data science tools, such as 'Shapley Values' and 'SHAP Values' provide unique insights on data set influence and feature importance that are highly valuable in the data evaluation use cases

Data Clean Rooms

Data clean room products, including Databricks Clean Rooms and Snowflake Data Clean Rooms, have largely been designed for adtech / martech use cases that involve two parties collaborating on consumer audience data. This is not fit-for-purpose to the intended use case in the following ways:

SAFETY AND SECURITY

- Data transfer as a requirement for their respective connectivity mechanisms present a legal and infosec risk due to data movement
- Roles and permissions need to be configured with high diligence for every engagement in order to appropriately match the security needs of the use case

SCALABILITY AND REPEATABILITY

- High effort and risk of error to repeat similar engagements - i.e., if an insurance carrier is engaging multiple data vendors on the same project, each individual engagement requires a separate data clean room and associated configuration. In these cases, it is very difficult to compare results across engagements

◆ Implications and Benefits of Using [integrate.ai](#)

integrate.ai has been designed from the ground up to be private- and safe-by-design for data science experimentation between data vendors and data consumers, as well as between strategic collaborators. Beyond the out-of-the box security and privacy, it is also feature-rich to support those collaborations to make them more efficient and effective.

With data clean rooms, there is significant implementation and configuration effort for each individual engagement. Use-case specific features are geared towards advertising / marketing use cases, and infosec, legal, and privacy reviews will need to be performed on each engagement.



03. Key Differentiation

3.3 CROSS-PLATFORM

integrate.ai

integrate.ai has been designed as a non-disruptive extension of existing data platform investments.

It is compatible with all the major cloud providers (AWS, Azure, and Google Cloud), data platforms (e.g., Snowflake, Databricks), and data governance tools (e.g., Unity Catalog).

Participating parties can be on different cloud platforms and only require a lightweight installation of the integrate.ai software. Further, only one party ('host') needs to be a paying customer of integrate.ai while the other parties can be non-paying 'guests' of the 'host.'

Data Clean Rooms

databricks

Databricks Clean Rooms are compatible with the different major cloud providers. However, participating in a Databricks Clean Room requires that all parties have not only an installation of Databricks Clean Rooms, but Delta Sharing capabilities enabled and appropriately configured via one or more additional platform components. This is often a non-trivial IT investment that may be disruptive to existing data platform investments. There is likely also a cost for all participating parties to procure the Databricks Clean Rooms solution.

snowflake®

Snowflake Data Clean Rooms are compatible across major cloud providers. However, participating in a Snowflake Data Clean Room requires that all parties have a baseline installation of Snowflake. A host on Snowflake can add a managed account for a partner that is not on the Snowflake platform (as consumers). The host account needs to have a pre committed "capacity" account, and the managed accounts pay for their participation in the data clean room. This is often a non-trivial IT investment that may be disruptive to existing data platform investments. Additionally, Snowflake Data Clean Rooms have complex regional compatibility requirements. Not all regions are compatible with each other, potentially leading to high costs if a party needs to change regions to participate.

◆ Implications and Benefits of Using **integrate.ai**

integrate.ai serves as an additive platform investment to unlock safe, native collaboration within and across organizations. It is cost effective in terms of leveraging existing investments while invoking no additional dependencies.

Data clean rooms introduce an inherent complexity that often demands a significant reconfiguration of data platform investments (e.g., compute and data storage in new regions), and potentially additional dependent platform investments that are often costly and time consuming.

Conclusion

While data clean rooms can be useful technology primitives to address data sensitivity issues in high trust collaborations, when it comes to scalable and repeatable use cases like vendor evaluations, data product PoC's, remote data science, and strategic partner data sharing, integrate.ai was designed from the ground up to be fit-for-purpose for those jobs. integrate.ai does not compete directly with data clean rooms, but rather is a necessary complement to data platform capabilities for companies that want to scale their external data collaboration capabilities.