Data Clean Rooms

Guidance and Recommended Practices

Version 1.0

Released July 5, 2023

Please email support@iabtechlab.com for questions. This document is available online at *https://iabtechlab.com/datacleanrooms/*

About this document

Most organizations engaging in digital marketing and advertising must share data across their internal departments and with other organizations, that frequently includes personal data of individuals. In the race for adoption of privacy technologies, "Data Clean Rooms (DCR)" have become increasingly sought-after as mechanisms to facilitate data sharing within specific limits. As DCR solutions begin to mature, the industry must come to a consensus on how Data Clean Rooms operate, develop a set of canonical use-cases, and standards around data processes and input-output formats. This document provides a set of common DCR principles, marketing and advertising use cases, and operating recommendations to help organizations better understand, evaluate, and implement a Data Clean Room. Key takeaways for readers:

- What principles do Data Clean Rooms share for advertising use cases?
- What functionality should Data Clean Rooms be expected to enable for advertising?
- What problems does a Data Clean Room solve?
- What are the limitations of Data Clean Rooms and what problems are best addressed with other tools and techniques?
- What Privacy-Enhancing Technologies (PETs) are commonly used by Data Clean Rooms?
- How do Data Clean Rooms fit into the broader ad tech ecosystem?

This document is intended to be an informational guide to understanding DCR technology and functions and help organizations understand and evaluate different DCR offerings against their own privacy policy and requirements.

This document is developed by the IAB Tech Lab Rearc Addressability Working Group. IAB Tech LAb will develop and release DCR standards and specifications as a follow-up to this document to address given use cases and application of DCRs in advertising.

Note: *The use of words or phrases 'Privacy", "Private" "Security", "Control", "Processing", "Personal Data" in this document is generic and does not refer to definitions in any specific regulation e.g. GDPR or CCPA.*

License

Significant Contributors

Devon DeBlasio, *InfoSum*; Chris Watts, *NumberEight*; Justin Langseth, *Snowflake*; Rachel Blum, *Snowflake*; Brian May, *dstillery*; Edik Mitelman, *AppsFlyer*; Andrew Knox, *Decentriq*, Ted Flanagan, *Habu*; Justin Li, *Optable*; William Syms, *Optable*; Jay Rakhe', *Tapad/Experian*; Matt Zambelli, *TransUnion (Neustar)*; Amanda Martin, *Goodway Group*; Jesus Mascias, *Freewheel*; Dan Morris, *Databricks*; Spencer Janyk, *Google*

IAB Tech Lab Lead

Shailley Singh, EVP Product & COO, IAB Tech Lab

Miguel Morales, Director Addressability & Privacy Enhancing Technologies (PETs)

About IAB Tech Lab

The IAB Technology Laboratory is a nonprofit research and development consortium charged with producing and helping companies implement global industry technical standards and solutions. The goal of the Tech Lab is to reduce friction associated with the digital advertising and marketing supply chain while contributing to the safe growth of an industry.

The IAB Tech Lab spearheads the development of technical standards, creates and maintains a code library to assist in rapid, cost-effective implementation of IAB standards, and establishes a test platform for companies to evaluate the compatibility of their technology solutions with IAB standards, which for 18 years have been the foundation for interoperability and profitable growth in the digital advertising supply chain. Further details about the IAB Technology Lab can be found at https://iabtechlab.com.

Disclaimer

THE STANDARDS, THE SPECIFICATIONS, THE MEASUREMENT GUIDELINES, AND ANY OTHER MATERIALS OR SERVICES PROVIDED TO OR USED BY YOU HEREUNDER (THE "PRODUCTS AND SERVICES") ARE PROVIDED "AS IS" AND "AS AVAILABLE," AND IAB TECHNOLOGY LABORATORY, INC. ("TECH LAB") MAKES NO WARRANTY WITH RESPECT TO THE SAME AND HEREBY DISCLAIMS ANY AND ALL EXPRESS, IMPLIED, OR STATUTORY WARRANTIES, INCLUDING, WITHOUT LIMITATION, ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AVAILABILITY, ERROR-FREE OR UNINTERRUPTED OPERATION, AND ANY WARRANTIES ARISING FROM A COURSE OF DEALING, COURSE OF PERFORMANCE, OR USAGE OF TRADE. TO THE EXTENT THAT TECH LAB MAY NOT AS A MATTER OF APPLICABLE LAW DISCLAIM ANY IMPLIED WARRANTY, THE SCOPE AND DURATION OF SUCH WARRANTY WILL BE THE MINIMUM PERMITTED UNDER SUCH LAW. THE PRODUCTS AND SERVICES DO NOT CONSTITUTE BUSINESS OR LEGAL ADVICE. TECH LAB DOES NOT WARRANT THAT THE PRODUCTS AND SERVICES PROVIDED TO OR USED BY YOU HEREUNDER SHALL CAUSE YOU AND/OR YOUR PRODUCTS OR SERVICES TO BE IN COMPLIANCE WITH ANY APPLICABLE LAWS, REGULATIONS, OR SELF-REGULATORY FRAMEWORKS, AND YOU ARE SOLELY RESPONSIBLE FOR COMPLIANCE WITH THE SAME, INCLUDING, BUT NOT LIMITED TO, DATA PROTECTION LAWS, SUCH AS THE PERSONAL INFORMATION PROTECTION AND ELECTRONIC DOCUMENTS ACT (CANADA), THE DATA PROTECTION DIRECTIVE (EU), THE E-PRIVACY DIRECTIVE (EU), THE GENERAL DATA PROTECTION REGULATION (EU), AND THE E-PRIVACY REGULATION (EU) AS AND WHEN THEY BECOME EFFECTIVE.

Glossary

Addressability	Ability or extent of capability to uniquely identify an individual or a device between data sets of two or more parties in a given context e.g. targeting individuals with advertisements
Audience	Group of people with a common set of characteristics used to create profiles and affinity categories such as demographics, interests, and intents, whom an advertiser wants to show an ad. More specifically, for e.g. this could be a list or group of customers or individuals that is most likely to purchase a given product or service from an advertiser or a list of individuals or households of a well defined set of attributes with common interests.
Audience Activation	A process of connecting advertiser target audience with publisher audience for targeting them through digital advertising channels
Crosswalk	A table that maps identifiers from one identity space to another. For example from Liveramp RAMPID to Experian LUID.
Data minimization	Data minimization refers to collecting, providing or restricting access to only the data that is absolutely necessary for performing a specific task or achieving a given outcome.
First-party data sets	Data acquired by an organization as a result of an individual's interaction with the organization either online on their website or mobile app or connected device or offline in their physical locations or by mail or phone
Identity Partner or Provider	An organization that maintains an individual, household or device level unique identification that can be used to perform a match between two or more organizations' data sets
Machine Learning	A mechanism and technology by which a computer can be trained to use existing data and learn how to perform a specific task
Multi-touch attribution (MTA)	MTA is a method of attributing credit to different touch points in a customer's interaction (for e.g different media channels where the customer viewed or engaged with an advertisement) with the advertiser that resulted in a customer action (for e.g. purchase of goods or services).

Onboarding	Process of setup and configuration to bring an organization's data or processes into a well defined system e.g. a Data Clean Room.
Personalization	Mechanism by which products, services, content and experiences (including but not limited to advertisements) can be delivered to an individual according to the characteristics or attributes of that individual's demography, interests, behavior, location or other expressed intent and information about the individual
PETs	Privacy enhancing technologies (PETs) are technology solutions such as differential privacy, secure multi party compute, confidential computing, and federated learning to accomplish complex data processing functions for sharing and analysis without revealing the individual, household or device level personal information to parties that do not already have them. These technologies aim to safeguard personal information and data from unauthorized access, use, and disclosure.
Post cookie	A common and popular term to describe the state of addressability after the loss of traditional identifiers
Purpose limitation	Purpose limitation is a way to constrain a task or an operation within a very specific outcome or use
ROAS	Return on Ad Spend (ROAS) is a measure of total returns from an ad campaign arrived at by calculating the total revenue earned and direct expenses. It does not include other expenses and does not tell if a paid campaign is profitable for the advertiser
ROI	Return on Investment (ROI) is a measure of overall return on investment arrived at by calculating total profit and all expenses- both direct spend on an ad campaign as well as other expenses. ROI determines how profitable is an ad campaign
Third party	A party to an interaction that has no direct relationship with the individual involved.
Traditional identifiers	Commonly used mechanisms like 3rd party cookie on the browser or Identifier for advertisers on mobile/ device platforms (for e.g. IDFA on iPhone , Android Id on Android devices) to uniquely identify a device or a browser typically used for associating a user or a household.

Table of Contents

1. Introduction: Why Data Clean Rooms?

Diminishing availability of traditional identifiers for advertising like third party cookies and mobile advertising identifiers as well as privacy and data sovereignty requirements are becoming a top issue for organizations that must execute advertising campaigns or measure their results. The primary alternative to the use of traditional identifiers is joining first-party data sets. The benefits and opportunities of being able to join these first party "data silos" are varied and needed for several use cases in digital advertising and marketing. However, the clear benefits to organizations of sharing their data needs to be balanced against the technical challenges they face and the risk they may incur due to ever growing need for data governance and compliance with privacy regulations. Organizations leveraging privacy technologies are increasingly looking to address the following business and consumer challenges when sharing data with their business partners:

- **Consumer trust:** Consumer awareness of privacy and use of their data for advertising is at an all-time high due to media focus on data misuse and leakage stories, mainstream documentaries about digital technology businesses and consumer privacy, and increasingly aggressive privacy pop ups many organizations believe are mandated by new regulations.
- **Loss of scale in addressability:** Diminishing availability of traditional identifiers for advertising and continued expansion of privacy regulation are motivating organizations to find new ways to maintain addressability for two reasons
    - Maintain personalization of connected consumer experiences.
    - Improve marketing performance of advertising campaigns driven by data driven use cases
- **Risk to business and reputation:** Traditional and legacy data-sharing solutions risk continual exposure of customer's data by sharing directly with the ecosystem. Besides compliance with regulations, it requires operating in an environment that allows businesses to manage and extend privacy permissions and controls of their customers' data when collaborating with their partners.

Organizations that collect, store and manage their own customer data need new ways to increase the effectiveness of their data-driven strategies without having to expose their customer's data to their business partners. These new market conditions have driven organizations to leverage privacy technologies in their operations and protect their customer data while enabling data driven marketing and advertising.

Data Clean Room (DCR) has emerged as a critical solution to enable safe collaboration between advertisers, media owners, technology platforms, and data providers:

- DCRs can be purpose limited for the use case and enable development of new and effective strategies for advertisers to connect with their customers
- DCR can help with adherence to the privacy policies of the organizations
- DCRs enable media owners to maximize the value of their audience in a post cookie world with loss of traditional web and device identifiers
- DCRs enable businesses to extract greater value and operationalize their first-party data without exposing or sharing proprietary data with other parties
- DCR can be used to ensure that data is not altered or manipulated during the analysis process

Note: *While DCRs can help achieve these outcomes, organizations should still assess DCRs to ensure that their particular needs are met. The rest of this document outlines ways of starting to do that.*

2. What is a Data Clean Room?

A data clean room is a secure collaboration environment which allows two or more participants to leverage data assets for specific, mutually agreed upon uses, while guaranteeing enforcement of strict data access limitations for e.g, not revealing or exposing the personal data of their customers to other parties. DCRs can be designed to serve an array of purposes and deploy different mechanisms, for e.g. performing a specific computation for determining matching of audience data between two parties. Below is a list of capabilities that could be offered as part of a DCR solution:

- **Data Isolation:**A DCR allows the parties to isolate from one another or the DCR Provider itself from their raw data. In other words, raw or plaintext data cannot be observed or learned by any participant or DCR Provider unless the participants agree. A DCR ensures that a party retains full control over how their own data is used or available to other participants in a DCR environment.

- **Privacy Enhancing Technologies (PETs):** DCRs integrate technologies to minimize data movement, risk of exposure of personal data, and misuse of data for re-identification of individuals. These are best accomplished by applying Privacy Enhancing Technologies (PETs) like encryption and double blinding while storing input and output data, confidential computing, use of differential privacy in running queries, injecting data noise, maintaining k- anonymity thresholds etc. A DCR must deploy a combination of one or more PETs to accomplish the privacy needs of participating organizations.

- **Privacy control mechanisms:** A DCR implements the principle of least privilege, which is, that a user only has access to specific data and resources required to complete a task and no more, by applying the following controls:
    - Limiting the number of queries allowed, even when differential privacy is being used
    - Limiting the time for which access to compute operations are allowed, expiring the data access after a certain time window
    - Limiting the type or complexity of queries that can be executed
    - Restricting reuse of one data set with other participants
    - Requiring rebuild of input data sets for each operation.
    - Apply statistical noise on query results

    - Limit the outputs or granularity to only those necessary insights that are required for the task

- **Access Controls:** DCRs provide permissions and scoped access controls to define, monitor, and control who can perform what specific action, for what purpose, at what granularity, for how long.

Besides, it is important to understand what is NOT a DCR. Some examples of data sharing and analysis operations that would not be considered a DCR are:

- Standard joins where either party can see the individual records
- Queries that output individual level data
- Queries from whose output individuals can be identified even in the presence of some safeguards
- Non-private data sharing procedures such as exchanging plain text data between parties
- Applying only one PETs usually is not enough

3. What can you do with a Data Clean Room?

As per the IAB State of Data Report 2023, “64% of companies leveraging privacy preserving technologies are using DCRs”. Most of this use is to address privacy concerns like protecting personal data and enforcing privacy policy compliance while sharing data for different purposes. DCR technology has potential to support several advertising use cases that require joining two or more data sets with matching keys. Below is a representative list of use cases that can be accomplished using a DCR.

Addressable Identity and Audience activation

Among many others, one key goal for an advertiser is to find audiences i.e. individuals or households across different media channels so they can target advertisements that are relevant to that audience. In addition they need to perform this at scale with accuracy. Similarly a publisher or media owner goal is to provide enough information about their users to understand the audience an advertiser can target on their media properties.

Challenge: To accomplish this goal, advertisers and publishers use their first party data - data they have collected from their customers or individuals who have interacted with their brand or website or an app. Traditional processes require sharing and transferring of first party data including personal data with a third party or intermediary to determine matching audience and activating the audience. Without privacy protections, it risks leakage of their customers’ or users’ data to third parties.

Data Clean Rooms provide the environment and services with PETs and other controls that prevents exposing one party’s data to another party while enabling audience matching to enable following advertising functions:

- Onboarding: Enable businesses to select and transform first-party data assets into addressable IDs for targeting and measurement.
- Identity resolution/bridging: Use an identity bridge partner to generate a direct match between two or more datasets with or without a shared key.
- Match extension: Use one or multiple identity bridge partners to generate a direct match between two or more datasets to extend the overlap of shared customer records with additional key values.
- Audience discovery: Identify most relevant audience segments and activate directly with media owners or platforms.
- Audience expansion: Use first and second party data to expand and enrich audience profiles.

- Audience targeting: Plan and execute audience-based campaigns across media providers safely and without a reliance on third-party cookies or other perishable IDs.
- Lookalike Modeling - identify and target an audience based on conversion and attribution analysis.

Some specific use cases may be:

> *As a marketer I would like to identify the overlap of customers across external partners to understand audience reach, and accurately estimate the outcomes of an ad campaign or compare two or more partners based on the match rates.*
>
> *As an Advertiser, I would like to build relevant audiences and segments using customer demographic, behavioral, or other consumer-level insight data from another vendor and push select audience segments to an activation platform or destination of choice.*

Customer insights and Data enrichment

Organizations are always looking to learn more about their customers so they can connect with them better and provide better products as well as develop advertisements that are relevant to their customers. This requires combining data from both internal sources e.g. data captured by different departments through customer interactions and external sources e.g. another organization that can provide more information about a customer or customer groups. Organizations need to enrich the data about their first-party data using demographic, psychographic, behavioral, or other consumer attributes in order to understand their customers better.

Challenge: Data enrichment with internal sources as well as external collaboration partners requires data to be shared or transferred between parties risking loss of control and ownership of data. Enrichment services often are done in black box manner with limited transparency and cannot be undone further risking loss of competitive advantage.

Data Clean Rooms provide a controlled environment where organizations can ensure their privacy policies are implemented and they retain control over their customer data and no personal data is shared with collaborating partners.

- Single customer view: Break down internal data silos to improve business performance without sharing any data.

- Data enrichment: Amplify the performance of your data with direct access to exclusive consumer intelligence from a network of quality partners.
- Audience and measurement enrichment: Extend the performance of audience planning, activation, and measurement with a broader access to relevant consumer attributes and data points.
- Apply Machine Learning and advanced modeling to gain better insights about the customer.

A common use case may be:

> *As a marketer, I would like to enrich my first-party data using demographic, psychographic, behavioral, or other consumer attribute data provided by internal or external data collaborator(s) to better understand my customers and optimize audience performance.*

Note: *Data enrichment is typically only applied to facilitate a predefined use case within a DCR environment instead of being directly shared or appended to a collaborating partner's dataset. Due to privacy concerns, a DCR operator usually restricts the export of net-new data insights e.g cross media profiling in order to limit data leakage.*

Optimization and Measurement

It is critical for organizations to understand Return on Investment (ROI) or Return on Ad Spend (ROAS) to justify advertising spend as well as optimize media planning and define campaign objectives. This requires matching campaign exposure data like impressions, campaign events like clicks or other user interaction, user action on advertiser landing page or other campaign objectives that is derived from reports originating at multiple sources e.g. publisher property (website, app etc.), publisher ad server, advertiser ad server, verification service, advertiser property (website, app etc.) and other sources that may be deployed by advertiser for campaign measurement.

Challenge: Current measurement calculations require impression/exposure data from different sources to be joined and matched with conversion/sales data from advertiser's data sources and often with one or more parties requiring personal data to be shared and transferred directly to another party. This risks exposure of personal data as well as advertiser's business information.

Data Clean Rooms provide organizations with matching capabilities to quantify effectiveness and performance of marketing and advertising campaigns without requiring any data to be shared, commingled, or transferred to other parties.

- Incremental lift measurement: A solution that calculates the incremental impact of a campaign on metrics such as sales without exposing actual individual customer purchase data.
- Reach and frequency: Calculate the number of individuals or audience that saw an advertisement and at what frequency over a period of time.
- Campaign and audience verification: Identify if a campaign reached the right audience on the right channel at the right time.
- Attribution analysis: Perform real-time or offline attribution actions to identify which touch point across the various channels and campaigns was responsible for the conversion without exposing user level conversion/engagement data.
- ROI/ROAS analysis: Apply actual or predictive models of calculating return across the media mix to optimize campaign performance and budget allocation based on aggregated-level insights only without sharing individual conversions, purchases or other activities.

A common use case may be:

As a marketer, I would like to connect the impression/exposure data with conversion/sales data to calculate the performance of a single or multi-channel ad campaign to better understand and optimize future ad campaigns as well as overall media mix.

4. How does a Data Clean Room work?

A DCR shares characteristics common to any data sharing operation like data preparation, data normalization, data matching or querying and data outputs. What differentiates a DCR from other types of data collaboration is the emphasis on:

- Security: Combination of access control, governance protocols and system design to maintain the confidentiality, integrity, and availability of data used in a collaboration.
- Privacy: deployment of technology and controls to uphold privacy expectations of Data Contributors, for e.g. not revealing or exposing personal data of a party's data to other parties involved in the operation.

Data Clean Room Roles

In this section we will describe different DCR roles and operations.

a. **Data Contributor:** A party that provides their data assets to a DCR is a data contributor. Data contributors must ensure that the data they contribute and maintain has been acquired by them after following all applicable regulatory and policy requirements and is updated as such as long as the data is part of a DCR. Data contributors must implement controls and mechanisms to deploy the outputs from the DCR operations to ensure privacy expectations are met during the end use.

b. **DCR Provider:** A party that provides the technology, interface and environment to perform computations and/or extract insights by querying the connected datasets. A DCR provider may be an independent third party that provides this service to Data Contributors or one of the Data Contributors may also be a DCR Provider.

c. **Data Consumer:** A party that uses the DCR to run queries, extract outputs and insights from the data provided by Data Contributors. A Data Consumer may or may not contribute data or provide the DCR environment.

d. **Data Services Provider:** A party that may provide enhanced or value added services for data management or computation to Data Contributors or DCR or Data Consumers, e.g. a Data Services Provider may offer predictive modeling algorithms or a secure compute service or activation service in a DCR to its users.

Data Clean Room Operations

In order to accomplish a secure and private DCR collaboration, there are several operations that need to be performed

Operation	Description	Responsible Party
Data Connection	DCR must provide a secure (for e.g. https or secure ftp) way for Data Contributors to connect their data to the DCR and define the format and structure for e.g. data types, data fields etc. to ensure Data contributors can properly send data to DCR. For e.g. upload file formats or APIs.	DCR Provider
Data Transformation	Data Collaboration may require assembling the data in a form and shape that is ready for joining with other data sets and querying. This is done by **Transformation**:A process for Converting and organizing the data in a consistent format to ensure uniformity and remove redundancies and improve integrity of data. It can include changing the structure, format and values of data as required by data connection requirements or the collaboration use case or by other data contributors	DCR Contributor, DCR Provider, Data Services Provider
Data Processing	DCRs may provide two types of processing modes- centralized or federated. In **centralized** mode the data from all contributors is co-mingled in the DCR environment. In **federated** mode the data can stay in contributor's environment and DCR enables data connections and mechanisms to perform the processing in Data Contributor's environment or in a neutral DCR environment	DCR Provider, Data Services Provider
Data Preparation and Protection	DCR provides the capability to protect and secure the personal data by converting them to irreversible anonymized values. This can be done inside the DCR environment if you fully trust the provider of the environment, or prior to submitting data to the DCR based upon agreed technologies and mechanisms. Some common mechanisms are: ● Salted hash ● Encryption	DCR Provider, Data Contributor, Data Services Provider

Operation	Description	Responsible Party
	• Commutative Encryption There may be other techniques that may be provided by the DCR to achieve the data protection requirements.	
DCR Environment and Interface	A DCR may provide either a User Interface (UI) in the form of an application for e.g. a web application or a script/API access for interaction with the DCR. **UI** requires limited or no technical knowledge and a marketer, analyst or a planner can use UI features to run queries in the background and view or download the results. However it does require knowledge and understanding of the data. **Script/API** based environments require technical knowledge, and often requires a data scientist, data engineer, or advanced analyst that can execute complex script-based queries or relies on third-party resources to execute.	DCR Provider
Data Computation	To enable collaboration between parties, a DCR may offer different join types and matching types to enable understanding of data sets from different contributors as well as other more complex data processing capabilities. There are two **Join types:** 1. Party-to-party join - connection and applied computation between two parties (Include detail from “private-join” proposal) 2. Multi-party join - connection and applied computation to 3 or more datasets simultaneously. Common **Matching types are:** • **Intersection**: Determines the volume (count) of records shared between multiple parties’ datasets, without exposing any other information. • **Union**: Determines the total volume (count) of records that exist in totality across	DCR Provider, Data Services Provider

Operation	Description	Responsible Party
	datasets, without exposing any other information. ● **Exclusion**: Determines the total volume (count) of records that are NOT shared between parties, without exposing any other information. **Other Data processing** capabilities: DCR may also offer advanced compute and querying capabilities once the join and matching have been completed. For example, encrypted predictive machine learning models can run on the matched data set to either train or execute logic to generate insights and outcomes. Instead of the matching output, the insights and outcomes generated by predictive modeling can then be extracted as outputs of the DCR compute. The DCR Provider should be transparent about features used for data modeling, e.g. look-a-like segments so that the DCR consumer is aware of attributes used in models and they can remain compliant under laws for e.g. Equal Credit Opportunity Act (ECOA) that prohibit certain demographic elements usage for fair lending	
Data Output	DCR provides the ability to use the results of queries Data Contributors perform and insights extracted from connected datasets to be exported in well defined transformations, structures and formats to ensure the privacy expectations agreed upon by Data Contributors involved. The data outputs may be aggregate or at the individual user level. ● Aggregate outputs ○ Insights ■ Customer overlap analysis ■ Consumer segmentation ■ Lookalike modeling ■ Audience expansion ○ Measurement	DCR Provider, Data Services Provider

Operation	Description	Responsible Party
	■ Frequency/lift analysis ■ Reach and frequency ■ Audience verification ■ Attribution ● User level output (media activation/serving) ○ Direct activation ■ Emerging media, Connected Television (CTV), streaming audio, gaming, and retail media ■ Walled gardens ○ Indirect or open activation ■ Private marketplace and direct premium digital publishers ■ Longtail media inventory over open bidding programmatic channels through integrated partners	

5. What protections do Data Clean Rooms provide?

As mentioned in the previous section, DCRs emphasize security of the data assets and privacy of the personal data in the data assets of the Data Contributors. To accomplish this a DCR must integrate privacy enhancing technologies to prevent parties from learning an individual's personal data and establish data governance controls to ensure the data is protected in all stages of the collaboration.

Data privacy

Data Clean Rooms must leverage one or more of these privacy-enhancing technologies to ensure the privacy of personal data in the data set as necessary to satisfy the privacy outcomes (e.g. mathematical guarantee of privacy provided by Differential Privacy) required by the parties involved in the DCR operations:

- **Encryption** is the technology to convert or scramble plain text data into an unintelligible form which can only be understood by decrypting or reversing the encryption. It prevents non DCR participants from learning about the Data Contributors' data.
- **Homomorphic Encryption** is a type of encryption which allows a party to perform computation on data while it is still encrypted.
- **Commutative Encryption:** if the double encryption using two different keys produces the ciphertext that can be correctly decrypted using the keys in arbitrary order, then it is commutative encryption. It is a way to enhance privacy as it requires two keys from two different parties to decrypt the ciphertext providing an additional layer of protection
- **Secure Multi-party Compute** is a technology where multiple parties perform a computation keeping their data private from each other and yet infer the overall results and insights.
- **Private Set Intersection** is a cryptography technology that allows two parties with datasets to learn about the intersection between those two datasets without exposing the data other than the intersection.
- **Federated Learning** is a machine learning technology that allows multiple parties to locally perform the part of a computation relevant to their data, which is then aggregated to infer the overall results and insights.
- **Trusted Execution Environments,** commonly referred to as **Confidential Computing,** is a hardware technology that isolates a computation from the host system to keep the data and state private from all parties and yet infer overall results and insights.
- **Synthetic data** is a technique in which artificially generated data that is similar to the original data reduces the risk of individual personal data from being exposed or re-engineered.

- **Pseudonymization** is a private technique that obfuscates a user identifier in a data set by replacing it with an unintelligible number. It is typically done by applying hash functions, salted hash functions, encryption, double blinding i.e. encrypting twice.
- **Noise Injection** is a technique to add random irrelevant data into a dataset to obfuscate an individual's data or make an individual's data statistically irrelevant
- **Differential privacy** is a mathematical technique to rigorously guarantee a specific level of privacy for an operation by injecting noise.
- **K-Anonymity** or cohort sizing is a property of an anonymized data set which makes it much more difficult that an individual can be re-identified i.e. each person cannot be distinguished from at least k-1 other persons. It requires determining a minimum cohort size based on the characteristics of a data set.

Data governance

DCR must implement a combination of security and trust controls to ensure Data Contributors can maintain ownership over their datasets and are in full control of how their data is accessed and used with precise granularity.

- **Scoped access and permission controls**: Allow a company to define the rules for each secure data clean room.
- **Auditing and Logging:** Maintain metadata for usage and requests, monitoring for security purposes. This should be transparently visible to all Data Contributors
- **Data Residency Control:** Ability to control and manage the physical/geographic location of Data Contributor's datasets
- **Data rights management:** Allow Data Contributors to manage and update the datasets based on the permission updates of data usage either by the individuals in the datasets or by the Data Contributor's policies.
    - Right of access - can they get a copy of their data?
    - Right of rectification - can they correct their data, or at least the parts that they entered themselves?
    - Right of erasure - can they request that their data is destroyed or revoked?
    - Right to restrict processing - can they request that only part of their data is processed or that only specific types of processing can occur?
- **Consent Orchestration:** Ability to execute or process consent management requests prescribed by the Data Contributors.

    - Ensure all Data Contributors provide evidence of consent, where applicable/required (e.g., the IAB TCF V2).
    - Provide ability to read privacy signals (such as the IAB Techlab's GPP), where present, and allow Data Contributors and Data Consumers to verify their right to process the data is either explicitly allowed, or not explicitly prevented through such signals.
    - Provide controls to the Data Contributor to set usage and permissions for each individual in the data set including ability to remove an individual from the data set
    - Prevention of "matching further down the line"
- **Data Management:** Ability to ensure prompt removal and destruction of Data Contributor data, as well as any DCR results, when the DCR is no longer required.

Note: *While DCRs provide privacy technologies and data governance tools, it is the responsibility of the Data Contributor to ensure that their datasets have the required compliance with applicable privacy regulations (e.g. GDPR, CCPA etc.) and that the Data Contributor will continuously use the governance tools to apply the most updated regulatory compliance to their datasets*

6. What are the limitations and constraints of a Data Clean Room

While DCRs provide privacy safe methods of collaborating to share and infer data between business partners in the advertising ecosystem, there are certain constraints and limitations that must be considered when operating in a DCR.

- Some level of counterparty trust and due diligence is still required when collaborating within a data clean room:
    - Parties involved in DCRs have important and often differing legal responsibilities based upon the applicable privacy law and the way in which the parties have chosen to structure their relationships. Such parties should consult with their legal counsel about specific steps they should undertake to ensure compliance
    - All parties must operate with trust that other collaborating parties are contributing complete and accurate (not spoofed/doctored/inflated) data
    - In many cases, the design of the clean room requires all parties to trust the DCR Provider and cloud provider. In some cases, all parties must trust the same cloud provider.
    - In many cases, all parties must develop trust that counterparties do not attempt to use sophisticated or novel attacks against the privacy controls or PETs (logging can help address this) by incorporating it in contracts and performing due diligence and monitoring.

Note: *A Data Contributor of personal data should have the option to be aware of the identity of all other data contributors with whom their data is commingled, all Data Consumers who may extract insights from their data and all Data Service Providers who may use their data for queries, modeling or other services inside a DCR*.

- A **common matching mechanism** or set of agreed criteria to be used for matching must exist between the (two or more) data sets being matched. If one does not exist it can be generated by working with an identity partner or third-party to facilitate a match/overlap when available or permissible. For e.g an identifier like an email is a commonly used matching criteria.
- **Match rates**, depending on data clean room functionality and methodology may be different than expected or experienced with another identity provider. This is due to the first-party nature of the underlying datasets and can be impacted by the specific match keys used to facilitate

a match, the level of granularity of those keys (individual or household), and the matching methodology used (often deterministic).

Note: *match rate is not often a single indicator of performance, addressable scale, or accuracy. Due to the first-party nature of the DCR collaboration, it may actually be on the contrary, where higher accuracy results in greater performance (ROAS, ROI) with a lower first-party match rate*

- **Data cleaning and normalization** may be more difficult because one or both stakeholders may not have access to all the raw data (though the clean room providers may offer normalization features).
- **Advanced measurement analysis** such as multi-touch attribution (MTA) and other attribution-like calculations are dependent on multiple-factors that are outside the control of a DCR Provider. Specifically for MTA, which requires vast amounts of media exposure data from multiple sources to be connected and analyzed, may be restricted or limited due to the privacy-safe and collaborative nature of a data clean room.
- **Data enrichment** where net new insight or intelligence is appended directly to an underlying raw dataset is limited or not possible via the DCR environment due to the privacy-first nature of each individual collaboration. This approach may also violate the privacy and data governance principles of the DCR Provider or one or more of the collaborating Data Contributors.

7. How to select a Data Clean Room that's right for you

A DCR approach to data collaboration requires investment in time and resources from Data Contributors. It is necessary to properly evaluate a DCR Provider for various criteria like privacy, security, scale, usage etc, Some common criteria to consider are

Scale

- Compute capabilities: How many datasets, of what size and what level of complexity can a DCR connect and run computation against
- Activation channels: Does the data clean room allow for real-time activation via integrations with activation channels? Which activation channels are supported? Do you have to plan activation outside of DCR?
- Number of Data Contributors or other participants that can interact in the same available in a DCR

Speed

- How quickly can you configure and set up a DCR solution?
- How long does it take to grant permissions and join datasets? How long do computations take to run, and can this be dialed up and down dynamically?
- How fast can insights be gleaned and executed? Can insights be gleaned and activated from connected datasets within a DCR?
- Does a DCR support datasets that are continuously changing?
- Does it support low-latency queries?
- Can data be joined and queries get executed over the joined dataset without the need to upload it into the DCR?

Data Privacy

- Does the DCR provide or support one or more PETs to execute computations that preserve individual privacy with technical protections (e.g. Homomorphic encryption).
- Does the DCR guarantees data privacy at rest, in transit, and in use by leveraging hardware-based confidential computing technologies?
- Does the DCR provide or support technologies that mathematically guarantee a required level of privacy (e.g. Differential privacy)

Defenses against privacy attacks

- Does the DCR protect against the privacy attacks that allow:

    - Re-identification - Re-identification refers to reversing the step of anonymization and tracing an individual record back to its human source
    - Reconstruction - The goal in a reconstruction attack is to determine the non-private identifying attributes for individuals in the dataset.
    - What threat model does the DCR employ? Which parties are trusted, and in what capacity? What are the known failure modes?
    - Do the PETs deployed in a DCR and DCR outputs provide strong privacy and base for mitigation against known privacy attacks,

Common privacy attacks

- Membership inference attacks: A membership inference attack allows an adversary to query a trained machine learning model to predict whether or not a particular example was contained in the model's training dataset. This can expose individual's data if the training model contained personal data
- Outlier injection and profiling attacks: In this type of attack, fake data is inserted in a data set to drive a malicious output. The fake data is considered outlier and there are algorithms that can detect outliers and remove them from the compute
- Dictionary attacks: an attack where a party uses a dataset of all known or possible identifiers (for example phone numbers or IP addresses) instead of their actual dataset in a clean room. This could allow them to profile all or most of the counterparty's data versus just the actual overlap with their dataset.
- Manufactured data join attacks: In this type of attack, adversaries collect auxiliary information about a certain individual from multiple data sources and then combine that data to form a whole picture about their target, which is often an individual's personally identifiable information. For e.g. using 1990 U.S. census data, Stanford researchers showed that they could uniquely identify 87 percent of the U.S. population using only their Zip code, gender, and date of birth
- Forced-error attacks (forcing div/0 errors or join errors) that signal details about the data that should not be exposed

Note: *The above is not a complete but only a representative list and different types of attacks are discovered on an ongoing basis. A DCR may not have control over managing all privacy attacks.*

Cross region privacy restrictions

- Can data remain in the region of origin without copying or transferring the data to perform cross-region collaboration or computation?
- Is data required to be copied from one region to another in order to facilitate a collaboration or computation, if so would this require a transfer of data?
- How many regions does the data clean room facilitator operate and/or have residing data/storage infrastructure?
- Where does computation take place? Does the data need to be moved to use within the DCR?

Simplicity

- Does a DCR provide functionality that maximizes simplicity and ease of use for both data scientists and nontechnical users?
- Does the DCR solution facilitate the ad sales process?
- Does the DCR incorporate data and/or analytics assets, such as pre-defined query templates, to aid in implementation and usage?

Controls and Functionality

- Permissioning and scoped access: Does a DCR provide each data contributor the ability to define their own scoped access and permissions control and at what granularity?
- Functionality: Which programming languages does a DCR support(e.g. ANSI92 SQL, Python, C++, Java, Scala), Can you import/use public and/or private libraries? Can you leverage machine learning within the DCR operations?
- Audit: What level and latency of logging and auditing is provided to the participants of other parties' usage of their data in a DCR?
- Interoperability: Moving data between some DCR platforms may be cumbersome or costly. As a result, DCR users should consider whether potential partners who use different DCRs or have limited data management capabilities will be able to interoperate. Can the solution span across multiple regions, multiple clouds, and/or multiple data platforms.
- Network effect and reputation: Are there publicly referenceable customers using the clean room technology in production at data and organization scales and for use cases similar to yours, who have already approved the solution from a privacy, legal, and scalability perspective? Are there customers that can produce a network effect - will it be easy to onboard data collaboration with existing customers and can I expect to find new data collaborators who are already onboarded

Cost

DCR pricing structure can vary dramatically. You should ask if any of the following fees exist for the solution.

- License or usage fees - how much is a basic license? What does it cover? Are there tiers?
- Participation costs - do all participants or Data Contributors need to pay for licenses, or is it sufficient for only some to have one?
- Data storage costs - how much does data storage cost?
- Data preparation costs - are there additional fees for data preparation?
- Data connection costs - are there additional fees for connecting to specific other databases or other parties?
- Query/ Operations cost - what is the per-query/operation cost? Who pays it?
- Talent/ Skills cost for maintaining privacy technologies like encryption etc.
- Computation cost - is there a fee for computation usage? Is it separate from any fees for queries?
- OPEX (Operating Expenses) cost may be a limitation for smaller firms without a technical team, who are offloading data processing responsibilities to intermediaries. In many cases, platforms may include the data clean room OPEX costs in the cost of media. However, in other cases, marketers and/or Intermediaries may pass this cost along or transfer to ecosystem stakeholders which can become prohibitive for firms with smaller budgets. In due diligence with partners & providers, this is something that should be discussed early in the evaluation conversations.