PoV

Data Clean Rooms

by Christeena Uzhuthuval

Introduction

> " *Privacy is not an option, and it shouldn't be the price we accept for just getting on the Internet.*
>
> *- Gary Kovacs* "

In today's world, whenever we wish to buy a product, we first go online and explore everything about it. Haven't you often researched products before buying them? Later, when you visit some websites, boom! There is a deluge of ads related to the product you searched for. This is all thanks to cookies set on our web browser by the sites we visit. These cookies collect our information based on clicks, locations, etc., and then use this data for targeted advertising. Accepting cookies on websites could also lead to stolen credit cards/identity thefts, etc. Apple's Safari and Mozilla's Firefox already block third-party cookies by default. Google Chrome has long planned the phase-out to support third-party cookies by 2024; others are most likely to follow suit. Similarly, under GDPR and other regulations, organizations cannot legally process any individual's personally identifiable information (PII) without the subject's explicit consent.

We are moving towards a cookie-less future with stricter adherence to data privacy regulations. First-party data sets are a powerful alternative that marketing and advertising teams can leverage. Matching this data helps derive insights for customer engagement, collaboration, and running marketing campaigns. This is the essence of data clean rooms—a privacy-enhanced mechanism for collaboration without using traditional identifiers.

What is a data clean room?

A data clean room is a concept where two or more parties or organizations can come together and share data securely and compliantly. With fine-grained controls, the data exchange is done without compromising any PII data. If PII data needs to be shared, it is anonymized before being processed and stored so that users cannot be identified at an individual level. In some cases, parties to the clean room may not have a common identifier for joining the data. In such scenarios, a third-party identity resolution provider can provide a unique identification across parties, which will help them merge their data. For example, using the unique identity given by the identity resolution provider, parties in the clean room can join Mary.John@yahoo.com with Mary.J@gmail.com without access to any PII.

This concept, also known as walled gardens, has been around for some time, especially for running campaigns under the guise of Google's Ads Data Hub, Amazon Marketing Cloud, and many other products. Its most common use cases range from marketing and advertising to audience segmentation, attribution modeling, inventory management, and demand forecasting. Data clean rooms also allow you to find your audience overlap, aggregated profile enrichment, campaign definitions, and measurement. You can thus monetize your data without compromising on security.

Multiple Media Players that have their data on cloud and collaborate with each other without moving data but by accessing a single copy of data

Brand / Advertiser

Publisher / Partner

Figure 1: Data clean room

How to implement a data clean room

To implement a data clean room, we can take advantage of native cloud features like data sharing and data masking. For this paper, I have considered Snowflake as the data warehouse where my data resides, which I would like to monetize through clean rooms. However, this concept can be implemented across any cloud environment or hyperscalers with similar features.

To implement clean rooms on Snowflake, we can make use of the following natively available features:

Secure Data Sharing—A feature for real-time data sharing

Row Access Policies—Policies that allow viewing data to which access has been granted

Masking Policies—Policies that will hide or mask PII data

Tasks, Streams and Notifications—For real-time scheduling, monitoring, and alerting

Parties to the clean room are termed producers, i.e., dataset owners and consumers, who receive the data. A producer can create a data share, which is a Snowflake object, and add his dataset or identity graph to this share. He can further apply restricted access controls, ensuring the data won't be visible unless certain conditions are satisfied. Next, he can define business rules to decide what data should be accessible. For example, he may set a rule that the consumer can count the number of customers per city but cannot view the actual customer name/ID. Once all these controls are in place, he can share this data.

At the consumer end, the consumer is notified via email that there is an incoming Snowflake data share. He can view the shared items and place a request for viewing the data.

The request goes to the producer for approval. Based on whether the access controls and conditions are met or not, the request will be approved or rejected. Once the request is approved in the system, the consumer can see a live copy of the data.

The steps described above are a simplified way of implementing data clean rooms using Snowflake in an application.
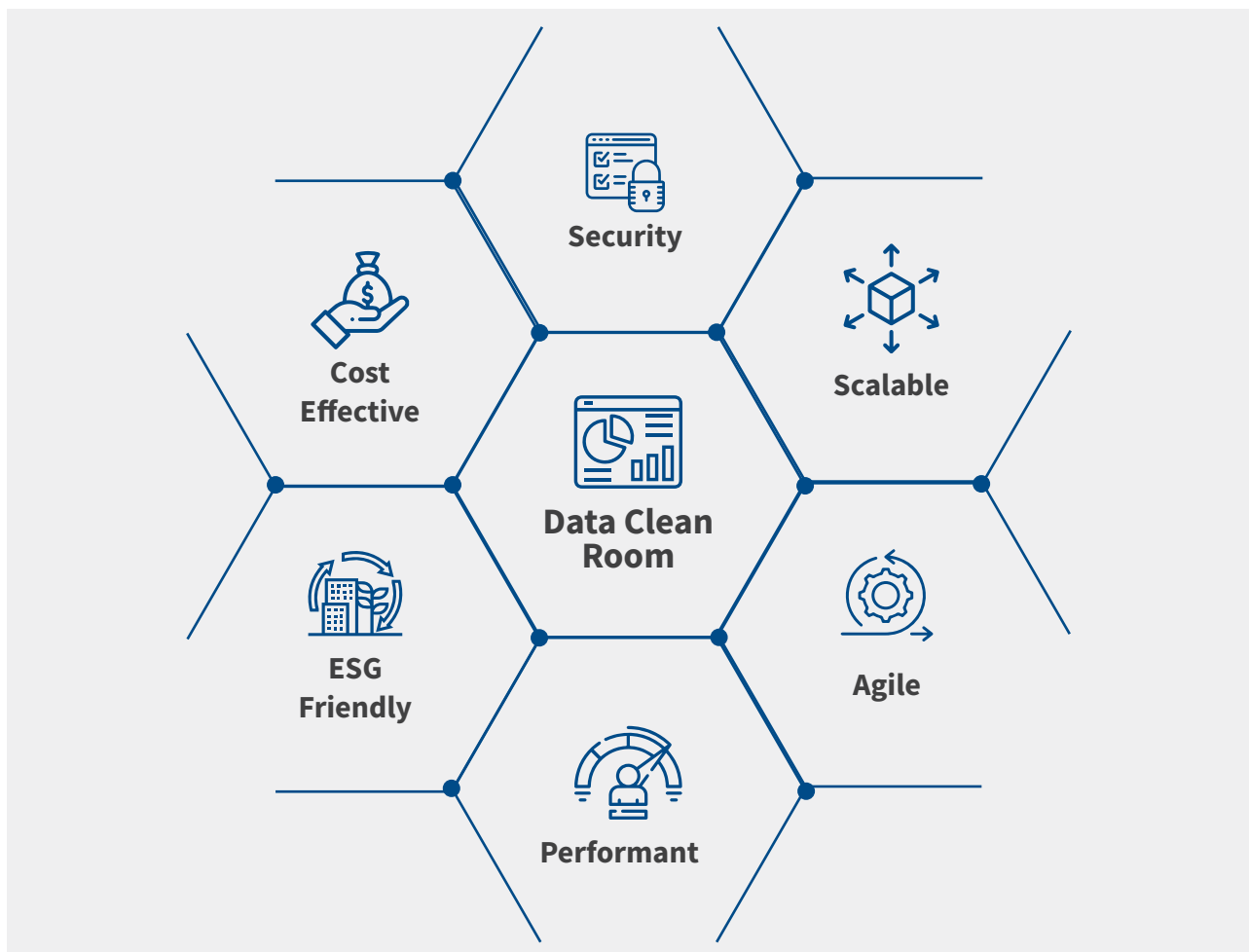

Figure 2: Data clean room implemented on Snowflake

Common use cases

Consider a fictitious media publisher, C-media, who wants to increase its advertising revenue. C-media intends to share its proprietary viewer identity graph for advertisers to find audiences. At the same time, it will want to control the dataset by applying business rules that state what attributes should be viewable or selectable by the advertiser. Suppose C-media shares a surrogate ID or a hashed email for its customer base with the advertiser. The advertiser/brand will join its dataset on the surrogate ID and get the common customers. The brand can now determine first-level addressability. It can also perform matching at the segmentation level with the viewer behavior with respect to the media property at an aggregated level. This gives the brand an idea about the viewing behavior of its common customers. It can target their segments at C-media during their respective show times without knowing which customer watches which shows. Once these customers are activated or shown the ad at the media end, C-media can share this list with the brand. The brand can then determine the attribution of people visiting its store to the ones who saw the ad at C-media. It can thus measure the impact of the marketing campaign.

Let us see some use cases in different domains where data clean rooms can play a significant role.

Health and insurance

In the medical domain, finding candidates for vaccination trials involves putting newspaper or online ads. Checking the suitability of registered applicants by procuring their health information is a tedious process.

Now, with data clean rooms, researchers can join the customer profiles with diagnostic lab and hospital data to understand patient disease history and suitability of blood types. Further, they can join these profiles with insurance data to understand the money aspect, pharma data to understand the prescriptions given, etc. Narrowing down candidates for vaccination trials becomes very quick thanks to secure data sharing without compromising PII.

Media & Entertainment(M&E):

Media publishers can increase advertising revenue by monetizing how viewers engage with their media properties. Through clean rooms, M&E publishers can securely share first-party viewer behavior data that they collect and enrich. This data can be combined with datasets available through other parties or from the data marketplace.

Suppose the store decides to target the sports section after getting an overlap between the media publisher and the store. It will send a purchase order to the media publisher to show a sports shoe ad during the basketball shows. The media publisher will know the number of people who watched the ad and will send this data securely to the store owner. Once customers visit the shoe store, the store can overlap their information with the activated audiences at the media end. This will give the store a multi-touch attribution and measurement of the ad campaign's success.

Retail and consumer product goods

Consider a CPG brand interested in targeted marketing for some product categories in a particular retail store. The retailer wants to increase advertising revenues by sharing customer data, with recency, frequency, and monetary (RFM) segmentation details related to the products the brand is targeting. Based on these RFM insights, the brand may want to target its loyal customers and churning customers of other brands. It will send a purchase order to the retailer to offer these customers a discount code, which they can avail of when they visit the store.

After a few days, the brand can do an overlap to know how many customers visited the store and availed of the discount code. It can also get an idea of how many other brands churning customers it was able to convert at an aggregated level.

Financial services

In financial services, banks can engage in targeted marketing of their campaigns with multiple digital platforms or media publishers.

Consider a bank that wants to push its campaigns to a target list of customers across multiple channels. They want to go ahead and sell bank advisory services to their prospective customers. To achieve this, the bank would start by sharing its prospect database for different media publishers to match and activate the campaign on their properties.

The media publisher would show the bank's ad on their media properties during the prospect's favorite shows. Here, only the media house will know their favorite shows. The bank won't know the shows during which they are being shown the ad.

Once activation is done, the media publisher will securely share the list of activated customers with the bank. The bank will match this list of activated customers with their footfall data in the bank to measure their campaign's success. Thus, they can attribute an ad to the footfall data, i.e., those customers who came to the bank after watching the ad.

Travel, transport, and hospitality

In the travel, transport, and hospitality industry, data clean rooms allow multiple parties to cross-sell related services through one another. For example, a car rental may cross-sell its services to customers booking hotels or a hotel may cross-sell its services to travelers booking flights.

The party that wants to cross-sell will share its customer or prospect database with controls in place. When the other party finds matches, when their customers are booking, they may cross-sell the relevant service. For example, when a user books a hotel, he gets an option to book car rental to the airport with live flight updates. The hotel will cross-sell car rentals only if the customer has been marked as desirable in the shared database.

Human resources

An IT services company may want to verify the background information of its job applicants. For this purpose, it would want to match its job applicant dataset to the past employee's dataset, which was securely shared with the provider company. Once it finds an overlap, it will automatically process those candidates whose stated experience and educational qualifications match or are more than those stated in the previous employee database. A manual verification/check would be required for the remaining candidates only.

How to get started with data clean rooms?

Implementing data clean rooms at scale can cause friction due to a lack of skilled resources, technical know-how, etc. An end-to-end application should cater to many factors like account onboarding to clean rooms, adding new users, controlling access, frictionless collaborations, setting up controls, and enabling multiple parties to work together seamlessly.

LTIMindtree has built an application -'LTIM Data Clean Rooms' using Streamlit Python framework, which serves as an accelerator for implementing clean rooms on Snowflake. This accelerator allows data contributors to seamlessly implement access rules while sharing data. It allows back-and-forth communication to derive joint insights with data consumers.

Conclusion

Due to the approaching end of third-party cookies in Google Chrome, organizations are looking for ways to maintain their advertising processes, such as ad targeting and measurement, while respecting users' privacy at the same time. Data clean rooms are rapidly gaining traction as they enable advertisers to run personalized ads, measure advertising outcomes, and attribute it to specific customer behavior while ensuring compliance and data security. It paves the way for businesses to monetize their first-party data sets without compromising their users' PII data.

References

How Retailers and Brands are Using Data Clean Rooms, Jeffery Bustos, Veronica Luik, LinkedIn, Mar 22, 2023: https://www.linkedin.com/pulse/how-retailers-brands-using-data-clean-rooms-jeffrey-bustos

https://www.ltimindtree.com/data-clean-room/

About the author

Christeena Uzhuthuval

Specialist, Snowflake Centre of Excellence, LTIMindtree

Christeena Uzhuthuval is the Technical Lead for LTIMindtree PolarSled Framework as well as LTIMindtree Data Clean Room Solution Accelerator. She is part of Snowflake CoE, in a specialist role. She has over 11 years of experience in data and analytics and is passionate about data and cloud technologies. In addition to reading about emerging technologies, she also likes to read thriller books, engage in some sports, and listen to music.