DECENTRIQ

EXECUTIVE WHITEPAPER BRIEF

The enterprise SaaS platform for secure data collaboration

Combine, share, and analyze sensitive data with anyone - enabled by Confidential Computing

For access to our full Technical Whitepaper, please request at the following link:

https://www.decentriq.com/request/whitepapers/technical

Introduction

Decentriq's mission is to unlock the value of data assets by removing the barriers that hinder data collaboration and innovation.

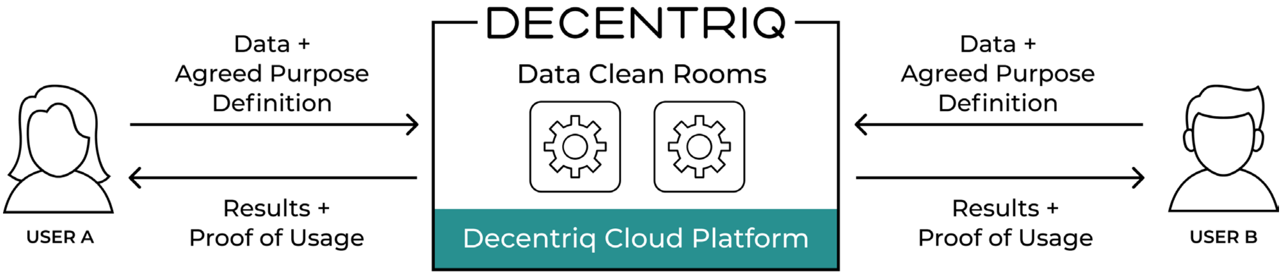
Companies forgo revenue and new business opportunities on a daily basis due to the risks of working with sensitive data. Successful organizations understand the value of their information assets and collaboration with others. However, organizations seeking to collaborate with data rightfully worry about the associated legal, reputational, and financial risks of intentional or unintentional misuse of the data sets. These risks inhibit innovation, collaboration, and partnerships around sensitive data and result in lost value. Data-centric industries such as marketing, healthcare, financial services, and academic research[1] each miss out on billions of dollars of opportunity annually[2] because they lack a secure means for data collaboration.

Decentriq's mission is to **unlock the value of data assets by removing the barriers that hinder data collaboration and innovation**. Decentriq is a cloud-based enterprise SaaS platform offering Data Clean Rooms that enable its users to work with and collaborate on data assets with minimal risk. As the first data collaboration platform where users do not have to trust each other, the platform operator, or the cloud provider, Decentriq mitigates the risks of collaborating on sensitive data sets and helps organizations unlock the full potential of their data.

Decentriq is a cloud service software platform that provides a unique promise: **data can be used for collaboration without being shared, data can only be used for a specific limited purpose,** and **compliance can be remotely verified**. This is true even when the data is used in collaboration with other companies. The original owner retains full control over the data during the entire lifecycle and can prove that it was only used for an explicit and limited purpose. This provides tangible benefits to organizations that use Decentriq's platform:

1. The platform drastically reduces the risk that a Data Analyst mishandles or leaks data.
2. Participants can extend the scope for which they can compliantly process personal data under GDPR and other regulations by eliminating the need to share personal data, reducing the risk of misuse.
3. Participants can collaborate with other Data Owners without having to trust anyone with their data, enabling the joint use of business secrets even between competitors.

To provide these fully verifiable data protection guarantees, Decentriq implements **Confidential Computing**. While data can traditionally be encrypted at-rest (on disk) and in-transit (over



Data may only be used in accordance with purpose definition

network), Confidential Computing also encrypts information in memory while it is in-use. The platform is able to do this with the support of new **trusted execution environment** hardware technologies like Intel Software Guard Extensions (Intel SGX), AMD Secure Encrypted Virtualization with Secure Nested Paging (AMD SEV-SNP), and other emerging technology. These technologies provide support to ensure that the program code and data are **isolated** into an **enclave** that cannot be accessed or modified. They also provide a way to verify what code is running inside the enclave. This **remote attestation** allows Data Owners and analysts to be certain about exactly which code can run over which data. These components come together to provide a **Data Clean Room** that enables collaboration without data sharing. See Section 4 Security Guarantees for more detail about how the technology and security guarantees work together to provide this confidentiality.

This level of data protection is Decentriq's unique value proposition – neither Decentriq, nor the cloud provider, nor the other Data Clean Room participants can see the data or change its purpose. This is in sharp contrast to other Data Clean Rooms and data analytics platforms which rely on traditional security methods. Without Confidential Computing there are no technical protections for the data, such that users can't verify the legitimacy of the server code and platform admins could easily extract data. This means that participants in such systems have to rely on unverifiable access controls and legal contract terms.

As the first data collaboration platform where users do not have to trust the platform operator or each other, Decentriq mitigates the risks of collaborating on sensitive data sets and helps organizations unlock the full potential of their data.

Using Decentriq

Decentriq makes it easy to use secure Data Clean Rooms with robust functionality

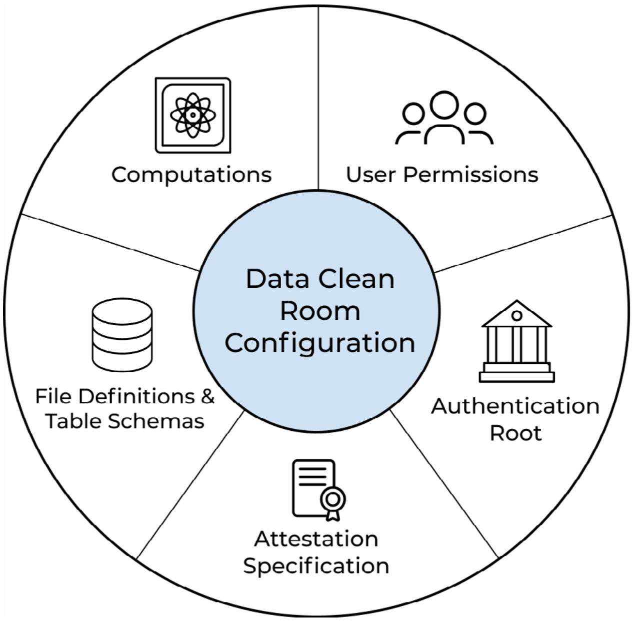
The Decentriq platform is flexible: it natively supports the SQL, python, and R programming languages, including most popular machine learning, statistics, and visualization libraries and packages. It also supports arbitrarily structured data such as JSON, images, and trained ML models in python and R using the standard techniques for those languages. Several features such as automatic shareable summary statistics, synthetic data generation, and result previews make it particularly easy to work with structured datasets.

Users interact with the platform through a Web UI or several Software Development Kits (SDKs) wrapping the Decentriq API. The web-based UI provides an intuitive interface for configuring and interacting with the platform. Decentriq currently supports Python and JavaScript (also used by the web-UI) SDKs which enable participants to build automated integrations and custom applications on top of the platform. Both interaction methods allow users to encrypt data on the client side, perform remote attestation, and interact with the platform workflow (configure environment, submit data, trigger execution, and retrieve results).

Decentriq relies on hardware attestation to provide the Confidential Computing guarantees, but abstracts away hardware vendor-specific logic. Decentriq currently supports securing data using both Intel SGX and AMD SEV-SNP, see Section 3.3 Confidential Computing Technologies.

Decentriq's SaaS offering is mainly hosted on Microsoft Azure Confidential Computing instances. For certain platform services and security requirements, Decentriq-owned hardware located in a secure datacenter in Switzerland is being utilized. Decentriq has the capability to easily extend hosting support to additional cloud providers as they begin offering the hardware support necessary for Confidential Computing.

The Data Clean Room Abstraction


Computations
User Permissions
Data Clean Room Configuration
File Definitions & Table Schemas
Authentication Root
Attestation Specification

Decentriq enables users to collaborate on data with the guarantee that their contributed data sets can only be used for the explicitly specified intended purpose. The platform uses the metaphor of a **Data Clean Room**, a highly controlled environment free from outside influence, to describe a specific act of collaboration.

The key element that ties a clean room together is a data structure called the **Data Clean Room Configuration**. This data structure contains the essential terms of the collaboration:

1. Table Schemas and File Definitions - what data will be included and how it will be structured.
2. Computations - what operations will be performed on the data.
3. User Permissions - who is responsible for contributing data (Data Owner) and who can see the results (Data Analyst).
4. Authentication Root - The root certificate authority used for user authentication.
5. Attestation Specification - other technical details required to ensure that the configuration is unambiguous and complete. This includes the verification policy for the hardware.

The Data Clean Room Configuration ties these parameters together into a verifiable format which allows each participant to review and understand the scope of collaboration. The Decentriq platform uses cryptographic protocols to indelibly bind these parameters to the code that executes them. This is the “contract” for the collaboration – readable by humans, but enforceable by the platform.

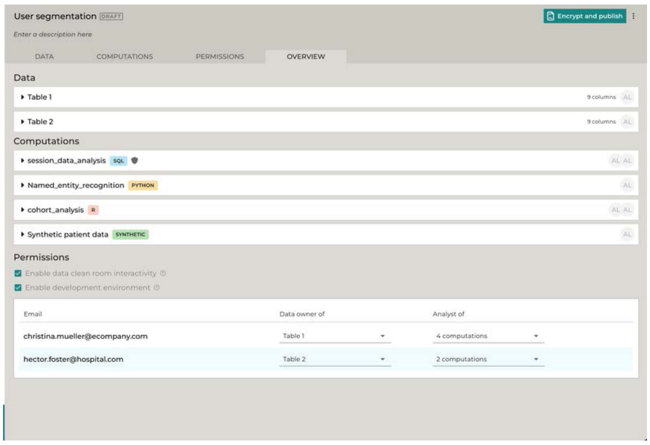
Any user may create a Data Clean Room by creating a Data Clean Room Configuration using a Decentriq client (UI or API) and inviting other participants. The Data Clean Room is then provisioned and can be used by invited participants who accept the configuration. A participant that can contribute a data source is a **Data Owner**, while a participant that can see the output of a computation is a **Data Analyst**. It is common for a single participant to be both a Data Owner and a Data Analyst.

At time of creation, a Data Clean Room can be set to be ‘interactive’. If this setting is active, any participant may propose amendments to the Data Clean Room Configuration which take effect if all affected Data Owners explicitly approve them. See the section below on Interactivity for more details.

Since the Data Clean Room Configuration is such a central concept to the platform, a detailed technical description of each component appears in Section 3.1 Data Clean Room Configuration.

Platform Features and Benefits

Web UI and API Access



User segmentation DRAFT

Enter a description here

Data

- Table 1 — 9 columns AL
- Table 2 — 9 columns AL

Computations

- session_data_analysis SQL — AL AL
- Named_entity_recognition PYTHON — AL
- cohort_analysis R — AL AL
- Synthetic patient data SYNTHETIC — AL

Permissions

☑ Enable data clean room interactivity
☑ Enable development environment

Email	Data owner of	Analyst of
christina.mueller@ecompany.com	Table 1	4 computations
hector.foster@hospital.com	Table 2	2 computations

Users can interact with the Decentriq platform either using an intuitive web-based interface or by using the Python and JavaScript SDKs that expose the platform's functionality via easy-to-use programming constructs. This functionality is described in detail in the platform documentation [3].

After having created a user account, users can login to the web-based UI and create new Data Clean Rooms or interact with existing ones. When creating new Data Clean Rooms, the UI provides the user with an interface in which the user defines all the input datasets, as well as their data schema in case of structured datasets. Users can also specify any number of computations they want to run within the Data Clean Room. Code to be executed can be entered directly into a code editor widget that supports syntax highlighting for the chosen computation type (e.g. R, Python, or SQL).

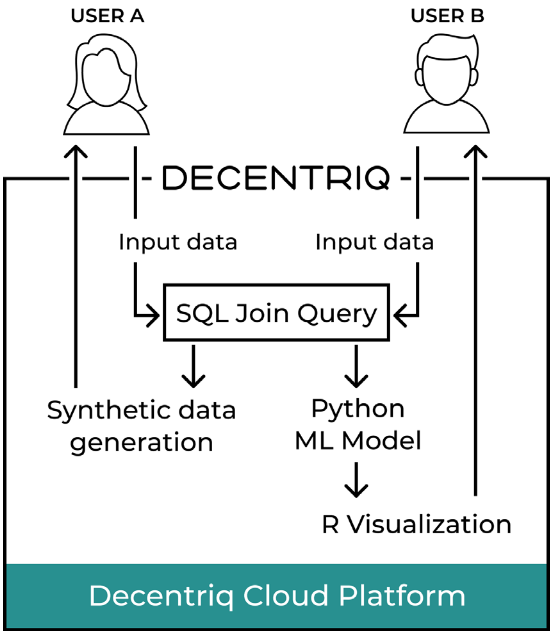
All of the actions that can be performed in the web UI can also be performed programmatically using the Decentriq SDKs. The SDKs offer object oriented programming (OOP) abstractions for the protocol buffer-based API and for processes such as the definition of a Data Clean Room. Additional helper functions for uploading data, triggering computations, and retrieving their results are also provided.

In contrast to the web UI, where computations are always configured to use the latest versions of available enclaves, users of the SDKs can define the exact versions of enclaves that are accepted in the remote attestation process. This way, users can be certain that only enclaves built from audited source code will be able to process any user data.

Using the SDKs, the Decentriq platform can be integrated with almost any existing infrastructure and can enable automated data processing workflows powered by Confidential Computing technology.

Data Pipelines

The Decentriq platform supports multi-stage data pipelines. While defining computations in a Data Clean Room, users can specify the relationship between computations and which other participants can read the output. Data can only be read by the downstream compute node.



Interactivity

The logic that controls what changes to an existing Data Room Configuration are possible is encapsulated in a **Governance Protocol**, an immutable property that is part of every interactive Data Clean Room. By default, Data Clean Rooms are immutable. Once published, changes such as adding new computation nodes, tables, or participants are not possible unless the interactivity function is enabled.

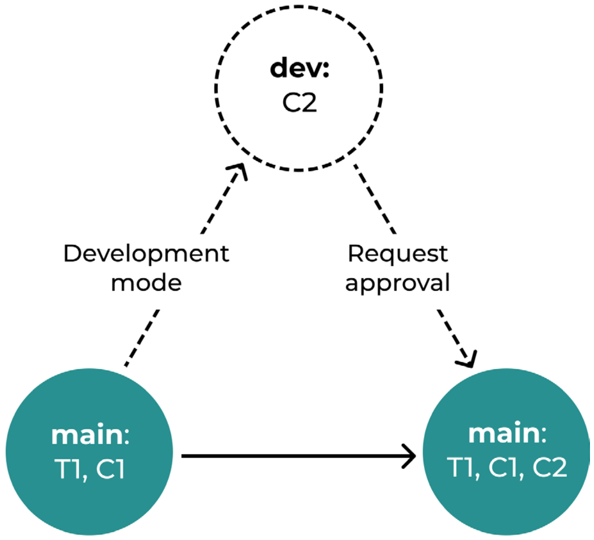
It is possible to set a Data Clean Room to **Interactive Mode** in the Governance Protocol. This allows participants to submit change requests to the current configuration of a Data Clean Room, such as adding new computations. Requests need to be approved by all participants whose data the new computation depends on.

Data Clean Rooms also include a **Development Mode** that allows even more permissive interactive features. In Development Mode, participants can create new computation nodes that only depend on data they already have access to read without needing approval from other participants. This allows participants to collaborate and quickly prototype new computations using test data without the checks in place that are necessary when working with real data.

Here is an example workflow for a Data Clean Room in Interactive Mode:

1. A Data Clean Room is created with table T1 where Bob is the Data Owner and computation C1 where Carol is Analyst, and publishes it with interactivity enabled.
2. Carol creates a computation C2 in development mode, having T1 as its dependency.
3. Carol submits a request to integrate C2 into the Data Clean Room.
4. Bob, as the T1 Data Owner, must review and approve the request.
5. Upon approval, the Data Clean Room will have T1, C1 and C2 visible to all authorized participants.

This is how the Data Clean Room Configuration evolves:



Synthetic Data Generation

Decentriq offers a unique value proposition with native synthetic data generation. Synthetic data is data that “looks real”, but is not based on any particular row of the original input data. This feature allows for additional flexibility for the Data Analyst, while maintaining privacy for the Data Owner. With this workflow, Data Analysts can choose to generate a synthetic data copy of the original data provisioned by the Data Owner for it to be used on exploratory analysis using the synthetic data node. Data Analysts can then interactively explore the data and develop their models based on the privacy-preserving synthetic copy. Once the training scripts are ready, Data Analysts can request approval, and only when approved can the analyses run on the real data.

Decentriq generates synthetic data by training a Generative Adversarial Network (GAN) machine learning algorithm on the original data. The trained model can then generate rows that contain correctly formatted values with a similar statistical distribution to the original data. In order to safeguard the integrity of the original data, Decentriq uses two techniques: masking and differential privacy.

Masking replaces data with similarly formatted but entirely artificial values before training the model. This is appropriate for variables that might be personally identifying, such as a postal code or name. Even though a synthetic data generation model is extremely unlikely to output a row that was in the input, it can only output column values it has been trained on. Therefore, if someone observed a “first name” with a value of “Alice”, they could infer that there was at least one person named Alice in the input set. In order to prevent this, the Decentriq platform allows users to mark columns for masking, and specify what format to use for the masked data. In this case, they could mark it as a first name field, which would replace all instances of “Alice” with a different randomly generated artificial value, such as “Jordan”.

Decentriq protects individual data during Synthetic Data Generation by introducing differential privacy through the use of a Private Aggregation of Teacher Ensemble (PATE). In a PATE, the input data is split up into multiple smaller data sets, and a “teacher” model is trained on each subset. The teacher models then vote on model updates, with noisy voting. Decentriq uses this particular approach, known as PATE-GAN, because it performs well with very low privacy budgets. Decentriq uses the OpenDP project’s Smartnoise Synth library for this functionality. Users may specify the privacy budget, but Decentriq recommends most users to use the default value of epsilon=1.

Here is an example workflow for a Data Clean Room using Synthetic Data Generation to simplify training a machine learning model:

1. A Data Owner creates a Data Clean Room containing an SQL-based computation that is used to combine (“join”) two input datasets. The output of this computation, a single joined dataset, is then used as the input for a Synthetic Data Generation model.

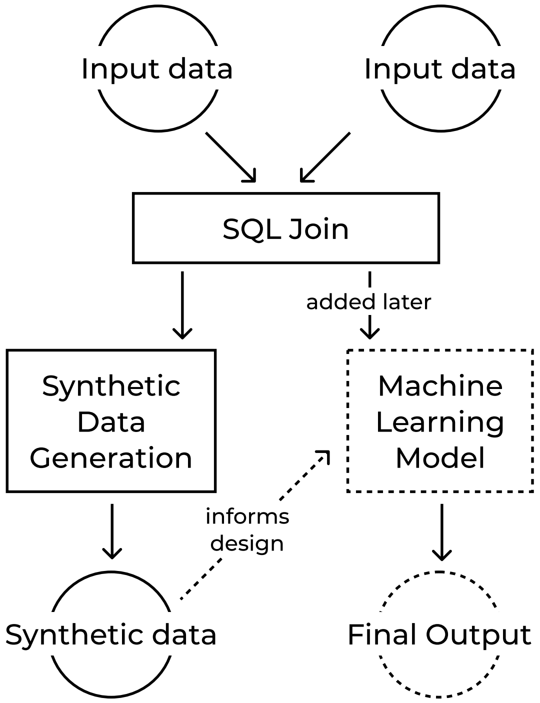
2. The Analyst reads the synthetic data and uses it to explore, troubleshoot, and/or tweak model performance outside of the Data Clean Room.

3. When the result looks acceptable, the Analyst uses the Interactivity feature to replace the Synthetic Data Generation step with the configured machine learning model and runs the model on the original joined dataset.

This has several advantages over a standard synthetic data workflow.

1. **Ability to join data before synthesis**: This means the joined statistics are more accurate, it is operationally easier because different data owners don’t need to coordinate in advance to have consistent join keys, and it has highly attractive privacy properties if the join keys are individually identifying because they can be masked or dropped after the join before synthesis.

1. **Increased accuracy of final model or analysis**: Synthetic data is generated with a model, so training another model on top of it is like making a “copy of a copy”. Even with extremely accurate synthetic data models, it will almost always be



more accurate to train the final model directly on the original data.

1. **Stronger privacy guarantees**: Decentriq's synthetic data models use differential privacy with an extremely strict privacy budget (default epsilon=1). Most synthetic data training paradigms do not offer a privacy guarantee; and of the few that do, almost all use a much weaker privacy budget. This is because they need to then use the synthetic data to train the final model or analysis. However, as the Decentriq platform provides a separate secure and confidential method to train the final model on the real data, it can afford more privacy and less accuracy during exploration as that loss of accuracy will not affect the accuracy of the trained model or final analysis.

Privacy Filter

Decentriq currently offers a simple privacy filter to reduce the risk of accidentally outputting individual data. For SQL query computation nodes, there is an option to suppress any output that has less than a certain number of rows contributing (default: 100 rows). While this is not a strong defense against active attack by itself, it can be used as a tool to cover edge cases and mitigate some forms of accidental disclosure when defining a Data Clean Room.

Decentriq intends to add more powerful privacy filter options in the future, such as differential privacy.

DECENTRIQ

If you're interested in learning more about Decentriq's solution, please request our full Technical Whitepaper here, covering the following topics in more depth:

Request here

Platform and Technology

- Data Clean Room Configuration
- System Architecture
- Confidential Computing Technologies
- Hosting

Security Guarantees

- Threat model
- Guarantees
- Current limitations

Citations

[1] https://www.ouvrirlascience.fr/wp-content/uploads/2019/03/Cost-Benefit-analysis-for-FAIR-research-data_KI0219023ENN_en.pdf

[2] https://www.paladincapgroup.com/confidential-computing-investing-in-decentriq/

[3] https://docs.decentriq.com/

Further Reading

Confidential Computing Consortium	https://confidentialcomputing.io/
Our Blog	https://blog.decentriq.com/
Intel SGX	https://software.intel.com/content/www/us/en/develop/topics/software-guard-extensions.html
Microsoft Azure Confidential Computing	https://azure.microsoft.com/en-us/solutions/confidential-compute/#overview

DECENTRIQ

Founding member of the Confidential Computing Consortium

CONFIDENTIAL COMPUTING CONSORTIUM

Learn more:
decentriq.com

Request a Demo:
hello@decentriq.com