Data Federation in a Trusted Research Environment

Introduction: PHEMS - A Case Study

PHEMS (short for "Paediatric Hospitals as European drivers for multi-party computation and synthetic data generation capabilities across clinical specialities and data types") is a Europe-wide consortium of paediatric hospitals that:

[...] aims to revolutionize the way health data is managed and utilized across Europe. This project is particularly focused on addressing the challenges posed by privacy concerns and the complexity of data sharing due to varying interpretations of the EU General Data Protection Regulation (GDPR). By developing a decentralized and open health data ecosystem, PHEMS strives to facilitate easier access to health data, thereby advancing federated health data analysis and creating services for generating shareable synthetic datasets.

The consortium has two high-level use cases:

Use Case 1	Use Case 2
Hospitals in the PHEMS network to share benchmarking data for clinical outcomes	Hospitals in the PHEMS network pool their data to train machine learning (ML) models

Both use cases need to be met in a way that complies with GDPR and the national data protection legislation each partner is subject to. GDPR alone creates a substantial barrier to direct sharing of clinical data across research projects, and the different national data protection frameworks that the PHEMS partners operate under further complicate this situation.

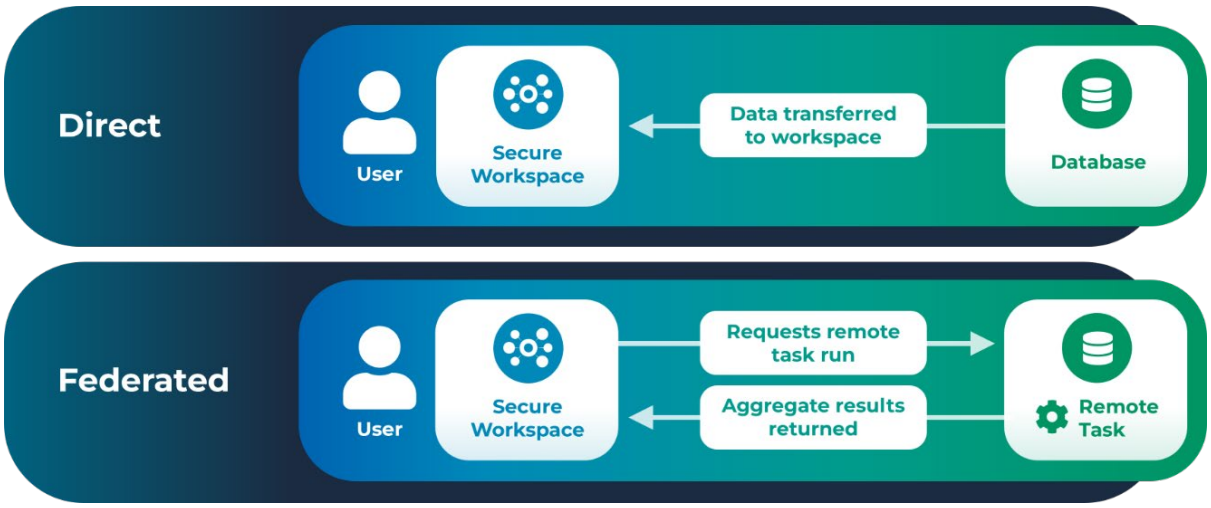
One possible technical solution to these issues is the creation of a federated data sharing network to allow federated analysis, and as a technical partner of the consortium, Aridhia is providing a data federation capability that will allow these use cases to be met in a way that is fully compliant with all relevant data protection legislation.

What is Federated Analysis?

Typically, when a project is granted access to clinical data, a copy of the data is transferred to a secure trusted research environment where researchers can view the full dataset, view record level data, and perform their analysis on it directly.

Where direct sharing of data is not possible (as in PHEMS) federated analysis may be an appropriate solution. With federated analysis, researchers never have direct access to the dataset, cannot view record-level data, and therefore cannot perform direct

analysis. Instead, they are granted permission to send approved analytics to the data, and then receive aggregate results when this analysis has been run:



The Federated Node

The Federated Node (FN) is a software component for running federated tasks, and is based on three existing open-source projects:

- The Common API
- Keycloak
- nginx

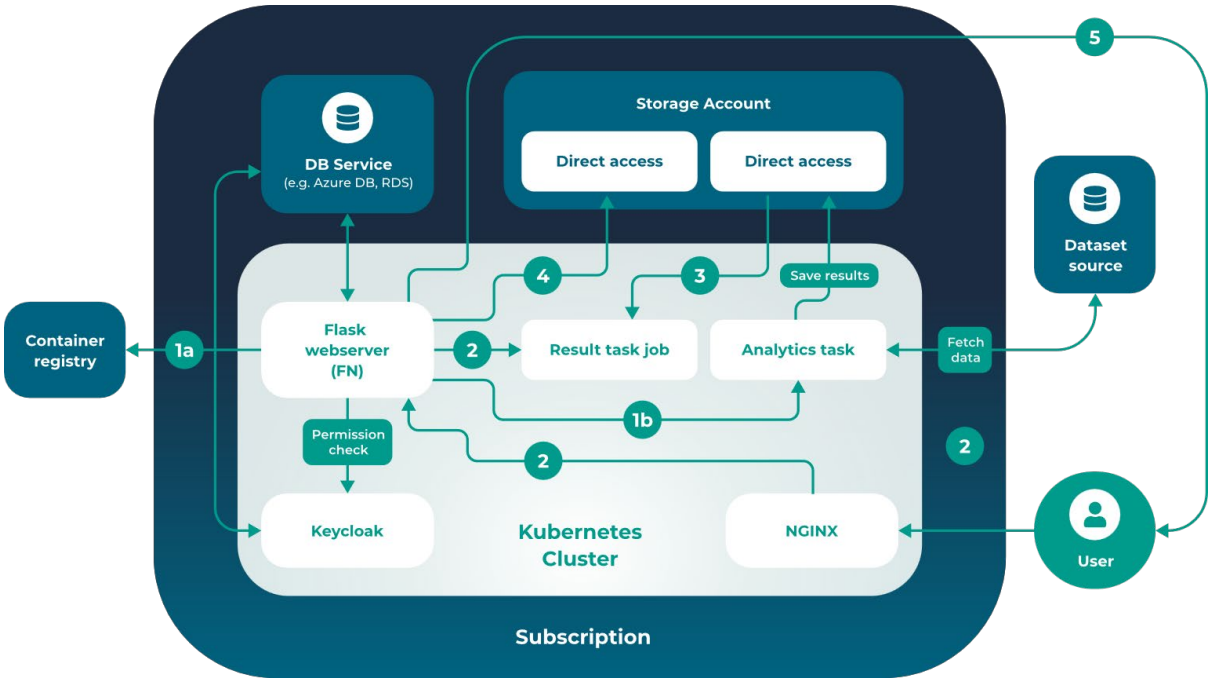
The Common API was developed in 2021 as an open standard for data platforms to participate in data sharing networks. It specifies a set of endpoints that provide a framework for organisations that wish to collaborate on federated data sharing and analysis. It provides the structure of the Federated Node API.

Keycloak is used for token and user management, and nginx is used as a reverse proxy, to route incoming requests.

Federated Node deployments are lightweight and use common technologies. All Federated Nodes are deployed to a Kubernetes cluster and require a Postgres database for storing user credentials.

A deployed Federated Node also needs to be associated with a Docker container registry. This is used to store the remote tasks that are run against the data. This architecture gives the data owner full control over what code is run against their data, as only scripts stored in the associated container registry can be used, and only authenticated users have the ability to initiate federated tasks.

Diagram showing how the Federated Node processes a federated task:



1a	Before creating the task pod, the FN checks if the docker image needed can be found in any of the docker container registries associated with the FN.
1b	The task pod is created, and the results are saved in the storage account.
2	On /results calls, if the task pod is on completed status, a job is created.
3	The job's pod will have the 2 storage environments mounted. It fetches the tasks result folder and zips it.
4	The webserver reads the zip contents from the live job pod and saves it in its own storage account environment.
5	The resulting archive is returned to the end user.

The Federated Node was made available in open source in November 2024, and the initial release was comprehensively tested by an independent contractor who identified no significant concerns. We anticipate that subsequent major versions will also be tested in a representative environment.

The Federated Node is available under the GNU GPL v3 license. Releasing under this license means that the Federated Node will be free to use, and that other projects can modify and distribute the source code as they need, while ensuring that any subsequent projects based on the FN must also be open-source.

Why Open Source?

The Federated Node is the first product available under the new Aridhia Open Source GitHub organisation.

As a company our goal is to bring large-scale computing and analytics to healthcare and to bridge the gap between clinical research and how that knowledge contributes to day-to-day healthcare delivery. And Aridhia Open Source has been created in that same spirit, reinforcing our commitment to a world with better healthcare outcomes for all patients.

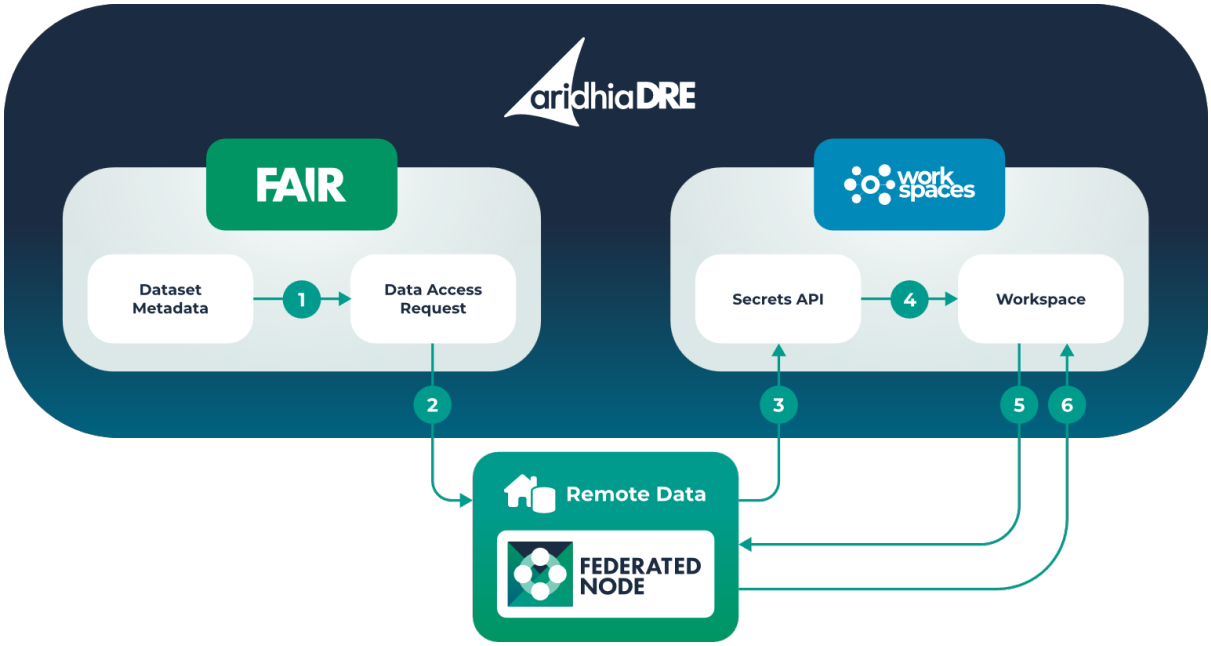
The Federated Node is an appropriate first open-source offering. The successful creation of federated networks depends on the participants adopting common standards, so ease of distribution and adoption is an important consideration for any data federation software, and releasing the FN under a recognised open-source licence should encourage wider usage.

The Aridhia DRE: Federation in a Trusted Research Environment

As detailed above, the Federated Node is a component for running federated tasks, not a complete product, and assumes that users of the FN will integrate it with other components which provide:

- Metadata catalogue
- Data Access Request (DAR) Management
- User interface (UI) for submitting tasks
- UI for viewing results

The following diagram shows the Aridhia DRE integrating with a Federated Node deployed in a remote environment. The DRE provides a metadata catalogue and data access request process through FAIR Data Services, and a secure environment for querying the federated data in Workspaces.



1. Submit data access request
2. Approve data access request
3. Issue credentials for FN
4. Access secure workspace
5. Submit federated query
6. Return results

The data source self-service feature in FAIR Data Services makes it easy for data owners to create a metadata record and associate it with a deployed Federated Node, and to ensure the secure generation and transfer of user auth tokens into Workspaces. In addition to security, querying a Federated Node from a workspace gives the researcher a flexible user interface that facilitates the deployment of custom applications for querying federated datasets and visualising the results.