Aridhia Informatics – DARE UK TREvoultion Interim Report

Author: Ross Stiven **Date:** 07/11/2025 **Version:** 1.0

1. Introduction

This interim report details initial work carried out by Aridhia under the DARE UK TREvolution project, specifically a proof of concept (PoC) integration of a secure TRE workspace with a federation component based on the GA4GH Task Execution Service. It provides:

- Overview of components
    - Aridhia DRE Workspace
    - Federated Node
- PoC high level design
- Testing
- PoC Assessment
    - Strengths
    - Weaknesses
- Further work

Overview of Components

The components of the federation PoC are an Aridhia DRE workspace and the Federated Node, an open source software component for data federation.

Aridhia Workspaces:

The Aridhia DRE is a Trusted Research Environment (TRE). It is certified under ISO27001/27701 and HITRUST, and is also compliant with the DARE UK SATRE specification for TREs.

The DRE has two primary components FAIR Data Services, a metadata catalogue, and Workspaces a secure platform for collaborative data analysis. FAIR was not used to implement the PoC.

Each DRE workspace provides users with a number of tools for collaborative research, but only two of its features were required for the PoC:

- A workspace database for hosting data
- A workspace container registry for publishing containerised code.

Remote Data hosting does not impact the availability of FAIR features, users are still able to search for, request and transfer data to a workspace using FAIR where the federated data is hosted.

The Federated Node:

The Federated Node (FN) is an open source software component for federated analysis, derived from the GA4GH TES.

The FN is based on the Common API. The Common API was developed by the International Covid Data Alliance, and is an open standard for a federated data sharing API which includes the TES endpoints. The FN takes this standard and develops it into a deployable software component for executing federated tasks.

A deployed FN must be associated with:

- One or more databases which hold the federated data,
- A container registry from which it retrieves the analytical code.

In the PoC these resources are provided by the DRE workspace.

PoC High Level Design

Each instance of the Aridhia DRE is deployed in a secure cluster within an Azure subscription. For the PoC a Federated Node was deployed in the same subscription as the DRE, but hosted in its own secure K8s cluster.

The network security on the DRE cluster was updated to allow the FN access to the workspace database and Azure Container Registry (ACR), these were then mapped to the FN. The network security of the FN cluster was set up to allow external users to contact the /tasks and endpoint.

Figure 1 below shows the relationship between the PoC components at a high level:

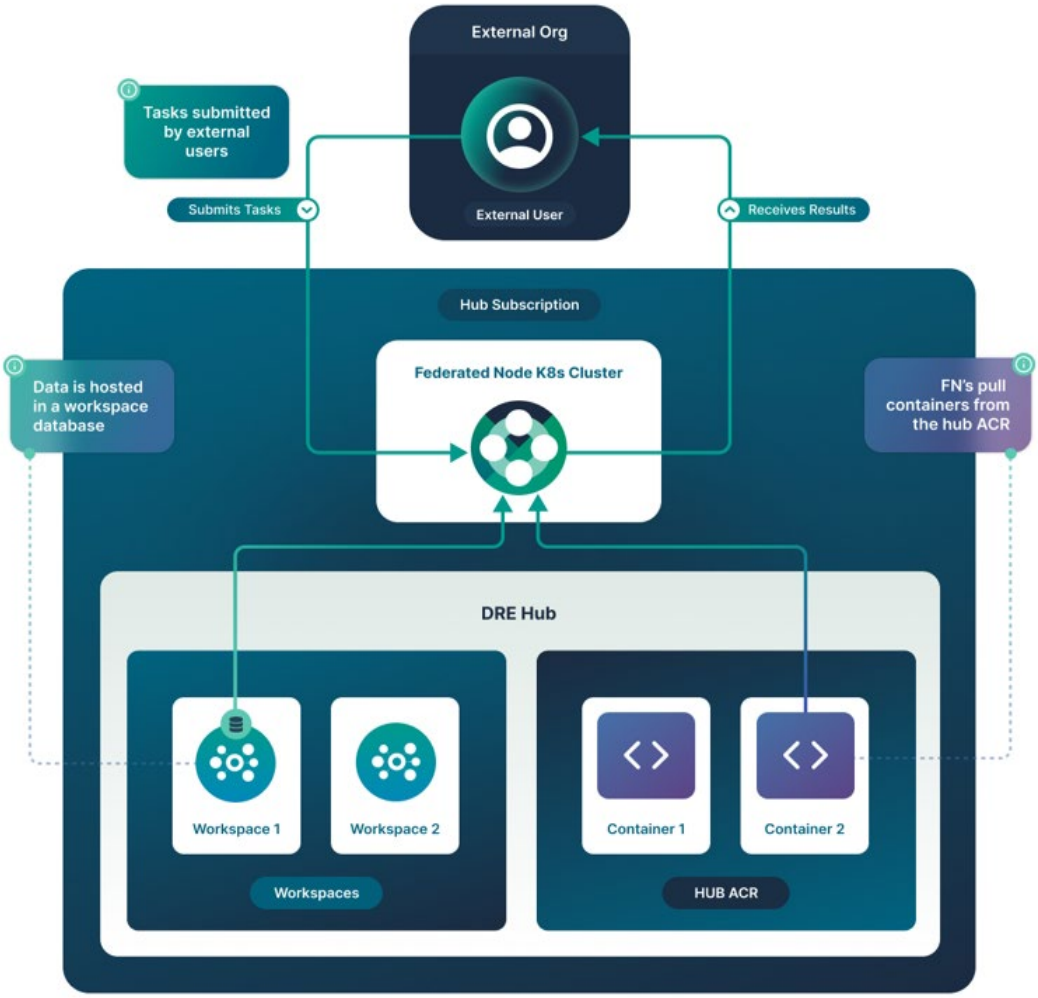

Figure 1 Workspace FN Integration

Testing

A small test dataset was uploaded to the workspace and made available in its database, and some simple analytical code was deployed to the ACR.

An auth token for the FN was generated and provided to a test user with no internal access to the Azure subscription. Using Postman to call the exposed FN endpoints the user was able to successfully submit multiple tasks to the FN, have these run in the secure FN cluster, and then retrieve the results.

This shows that by integrating with the FN it is possible to enable federated queries on data held in a DRE workspace.

Security testing was not carried out during the PoC phase.

PoC Assessment

Strengths:

We believe that the strengths of the current PoC deployment lie in security and clear separation of concerns.

However minimal, opening a secure environment to a federated network represents an increased risk. Deploying the FN in the same subscription as the workspace, but in a different secure cluster helps mitigate this risk.

This separation in deployment also makes that the compute required for federated tasks is provisioned in the stand alone FN cluster, which has two principal benefits:

- Where the federated task requires a large amount of compute, provisioning this in the TRE hub could lead to degradation of service for other users. This is avoided with separate deployments in the same subscription
- Cost recovery is an increasing concern when creating federated networks, and having a dedicated cluster for the FN with its own compute resource makes it easier for the data owner to attribute costs to federated tasks.

Weaknesses:

As noted above the current weaknesses of the PoC relate to scalability. The process of mapping the deployed federated node to workspace resources is a largely manual process that requires direct access to the Azure subscription and in the case of the FN, its secure cluster.

This means that at present a data controller would not be able to connect a workspace to a Federated Node without substantial assistance from the Aridhia service desk and operations team.

Additionally in the PoC no automated safeguards are applied to a workspace which is being used as federated infrastructure. Where this is the case we would recommend that users do not transfer further data into a workspace, but this is not currently programmatically enforced.

Figure 2 below shows the current manual steps required to enable federation in a DRE workspace, and the planned improvements to remove or minimise these interventions:

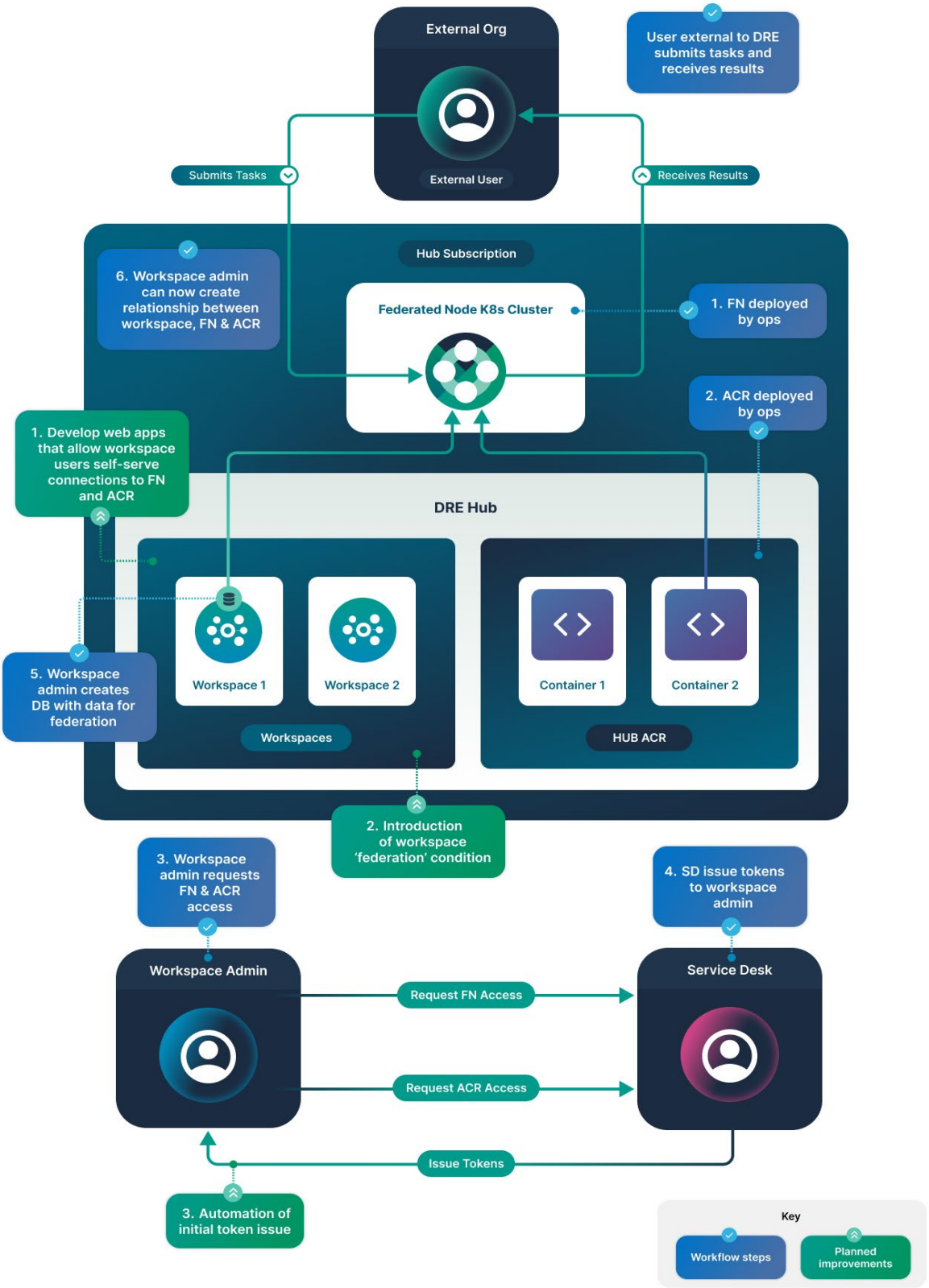

Figure 2 setup steps for FN integration

Further Work

As detailed in figure 2 above, we have identified three possible areas for further work, in current priority order these are:

- Make it possible for the data controller to self-serve connections between a Federated Node and the workspace database and ACR
- Put in place programmatic controls that disable certain workspace features when it is being used as federated infrastructure
- Remove the need for manual intervention during initial set up