V6 basics

Why using vantage6

Open source

Container orchestration for PETs

Extensible to different data source types

Algorithms in any language
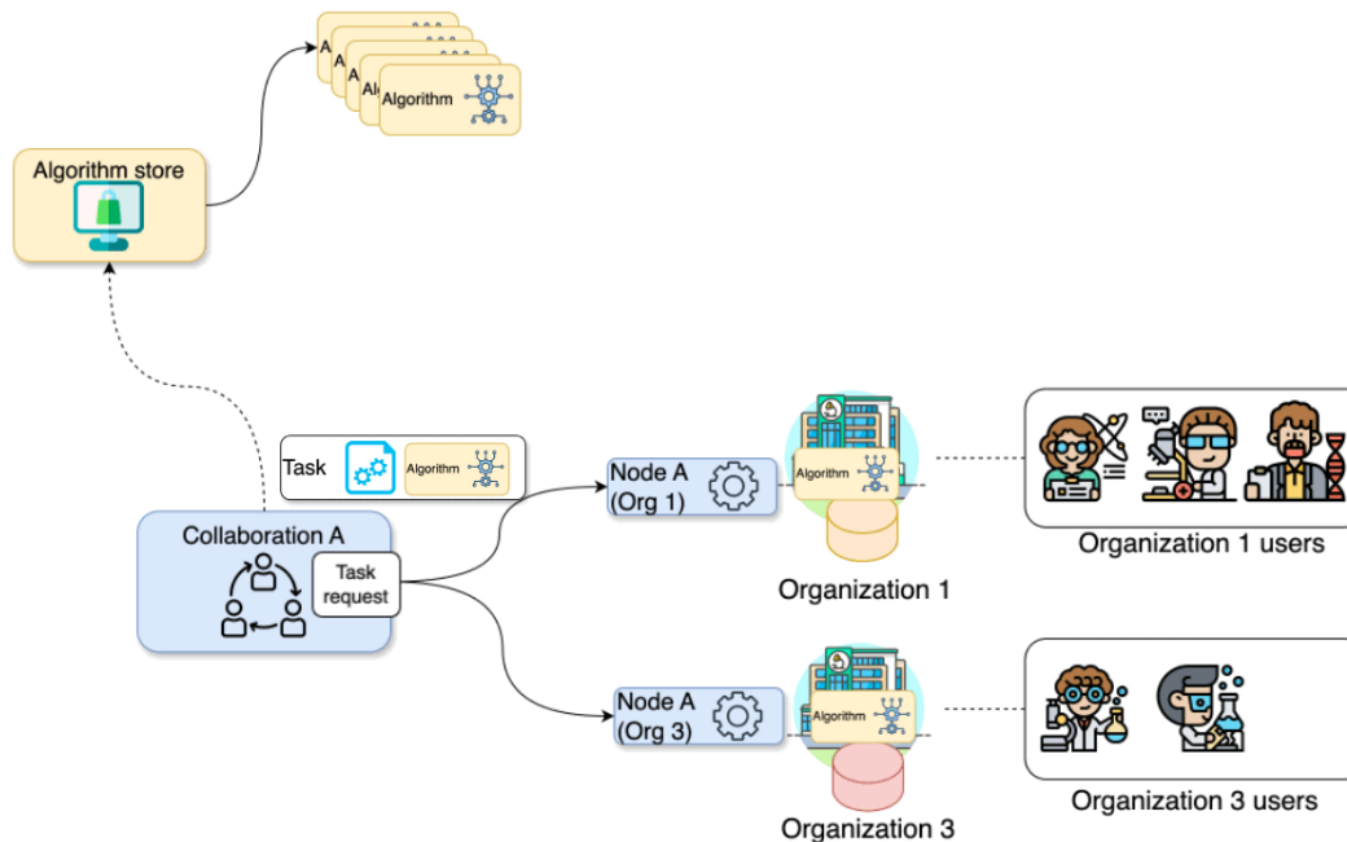
Other applications can connect using the API

Managing collaboration policies

Minimal network requirements

Federated analysis projects with vantage6

vantage6 encompasses a project administration system that allows the user to manage permissions and access to the resources, while assuring the protection of the data.
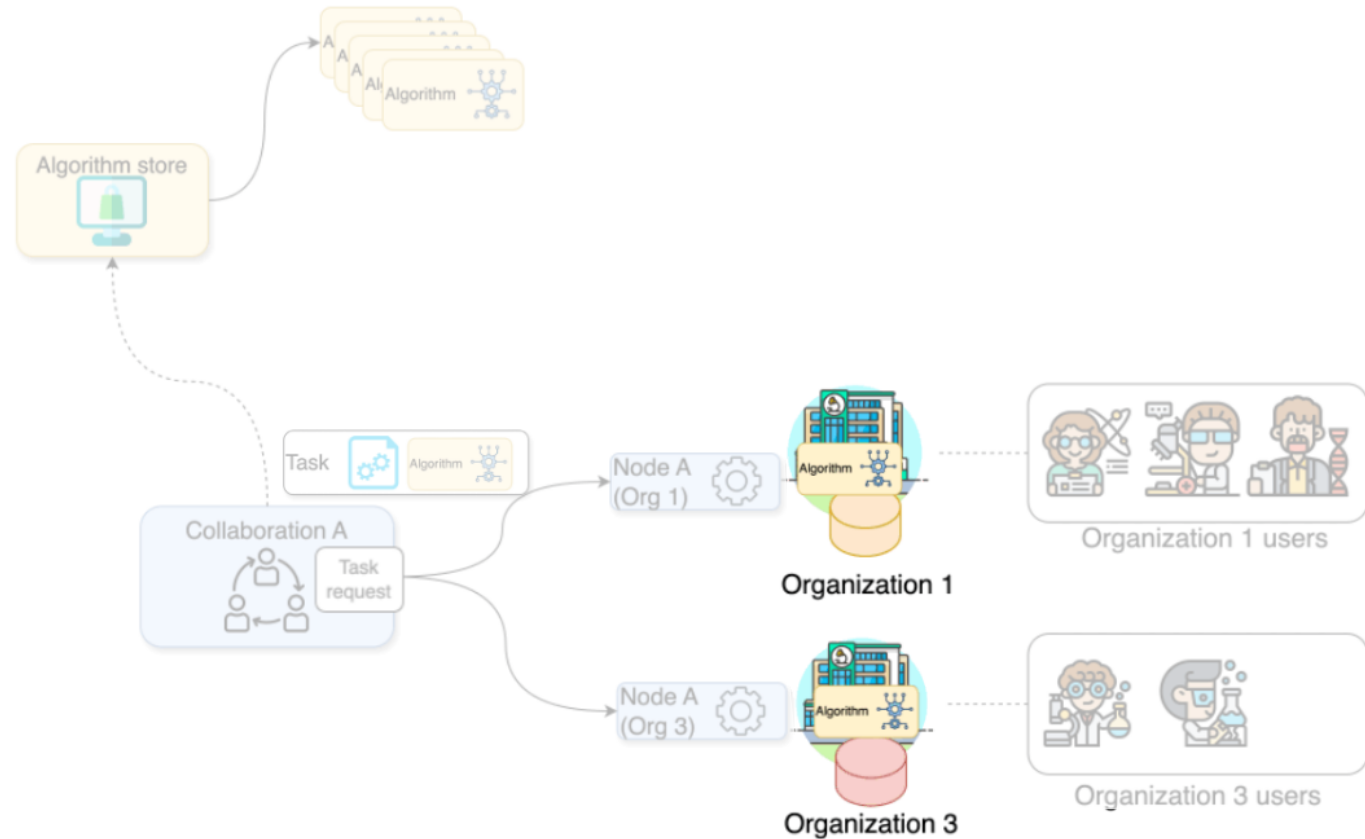
Federated analysis projects with vantage6

Organization

A group of users that share a common goal or interest (e.g., a consortium, an institute, etc.).
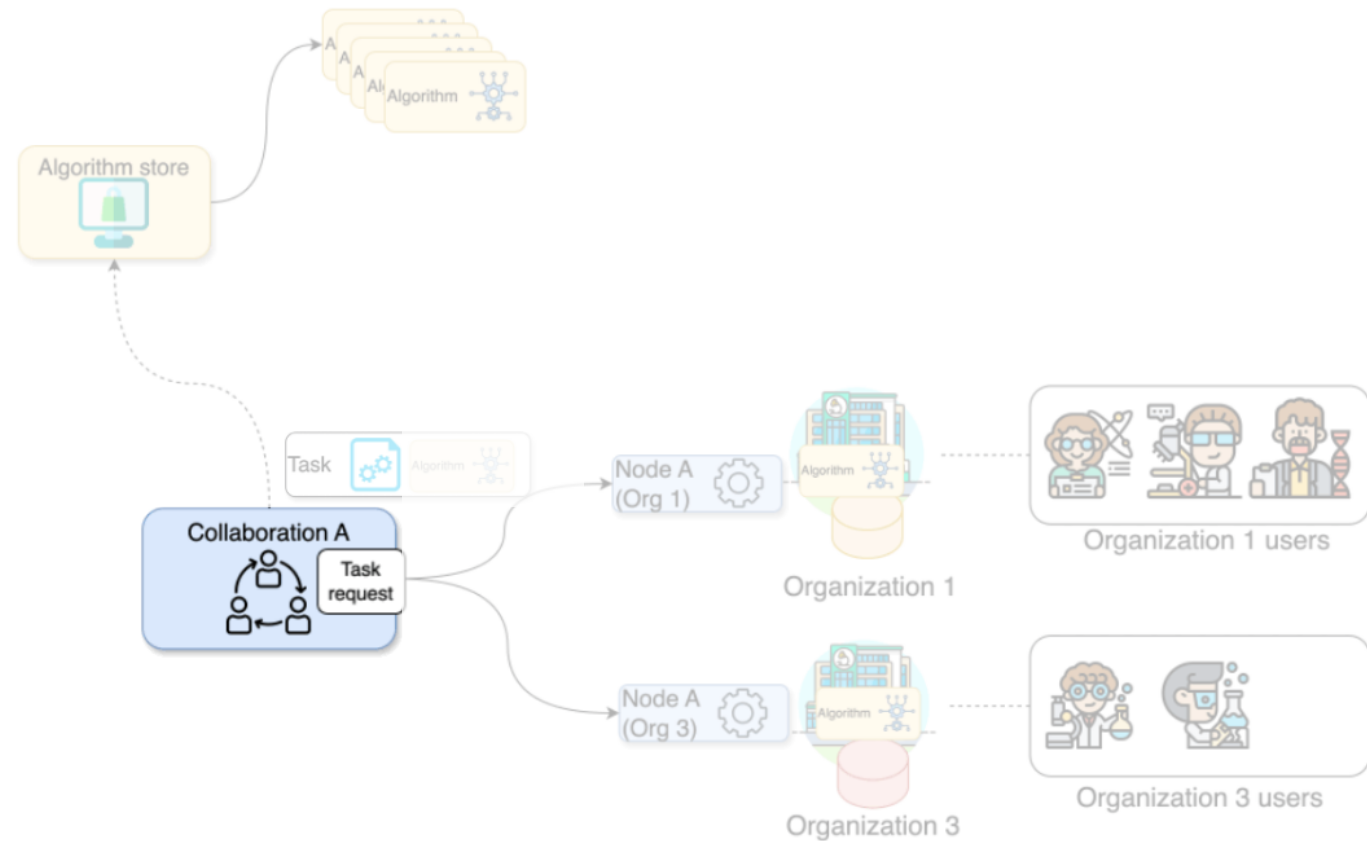
Jointly organized by:

Federated analysis projects with vantage6

Collaboration

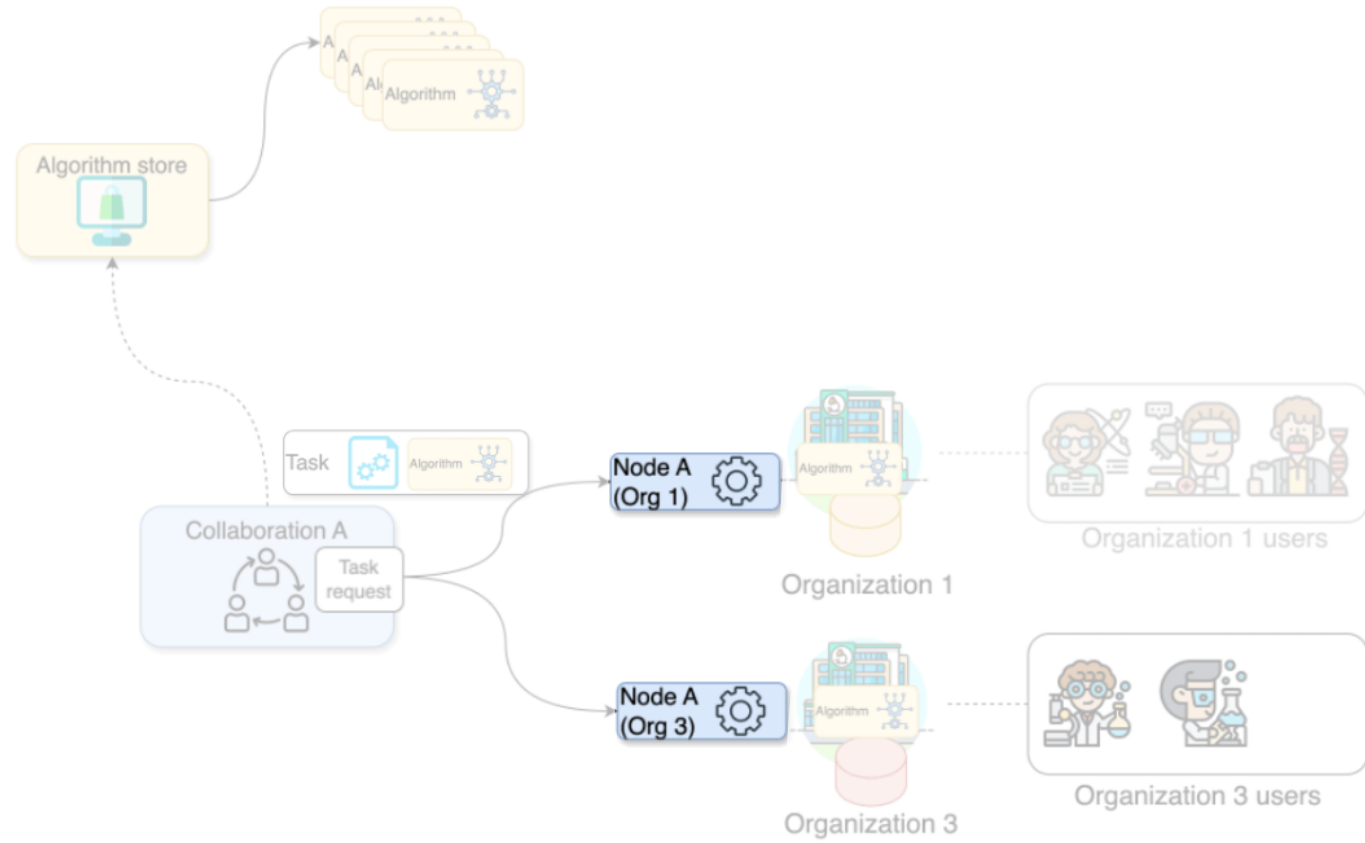*One or more organizations working together towards a shared objective.*

Federated analysis projects with vantage6

Node

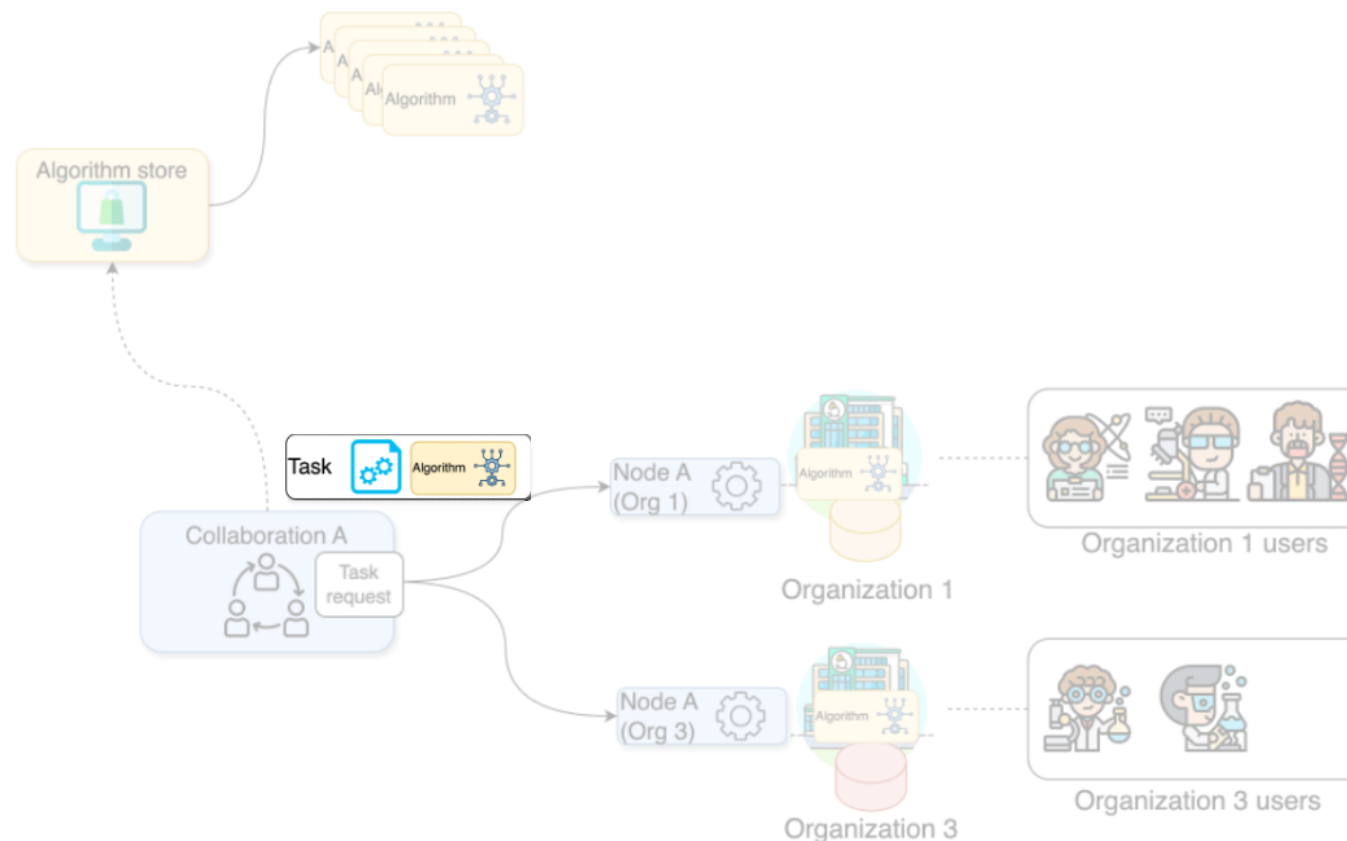*Component that accesses the organization data and executes algorithms on it.*

Federated analysis projects with vantage6

Task

A request for the execution of an algorithm. It is handled by the corresponding organizations' nodes.
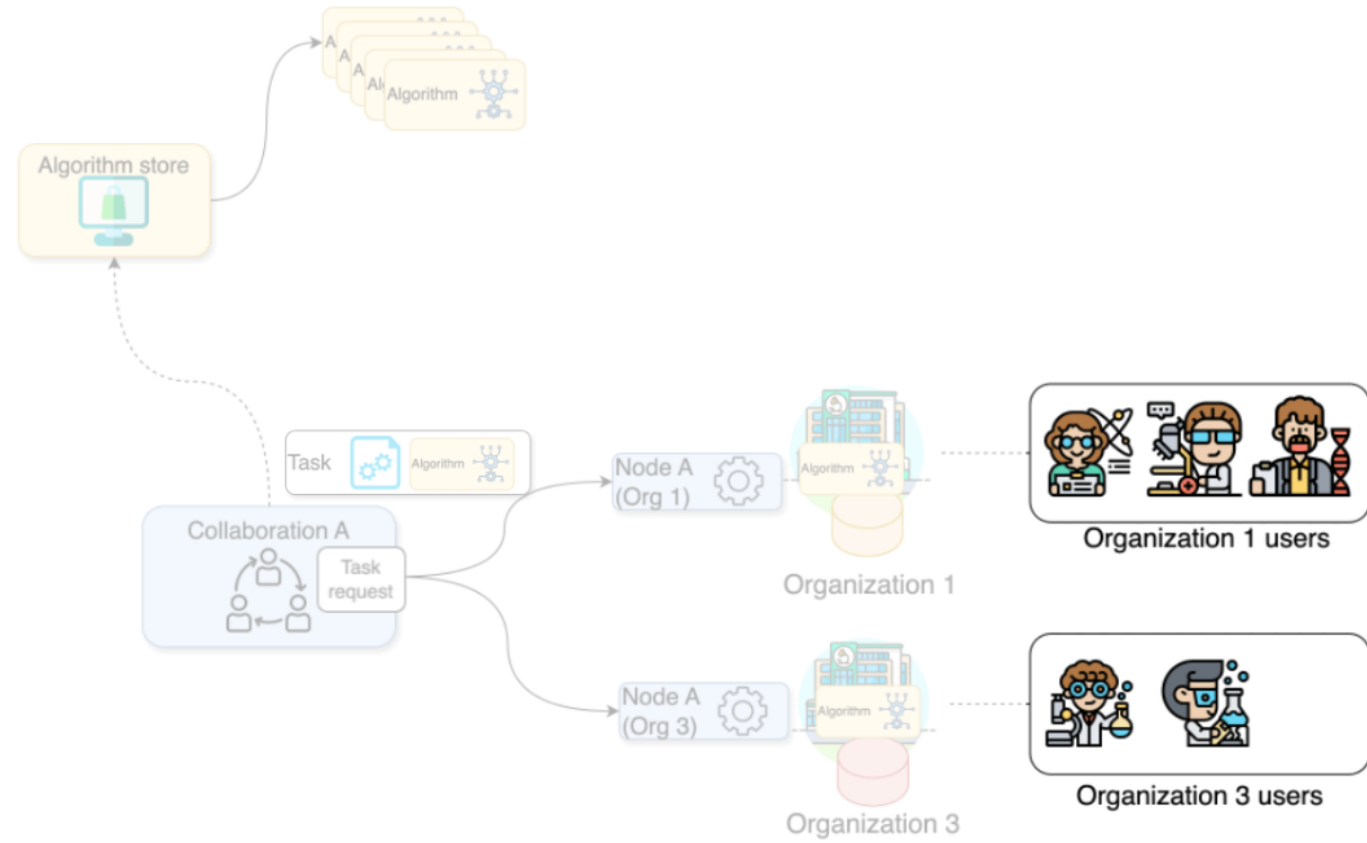
Federated analysis projects with vantage6

User

A person that belongs to one organization who can create tasks for one or more organizations within a collaboration.
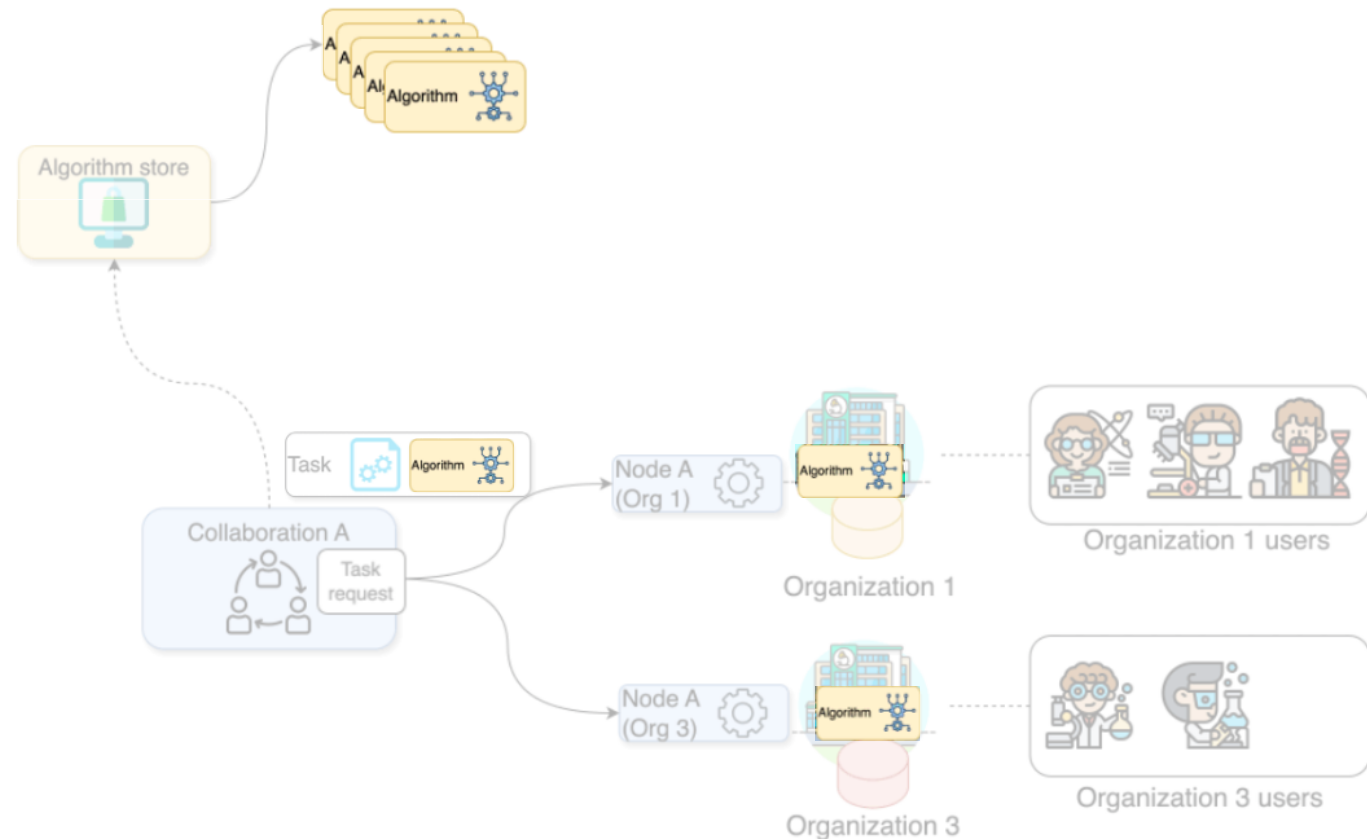
Federated analysis projects with vantage6

Algorithm

A computational model or process which can be securely distributed to nodes for execution.

Algorithm store

A A Algorithm

Task — Algorithm

Collaboration A — Task request

Node A (Org 1) — Algorithm

Organization 1

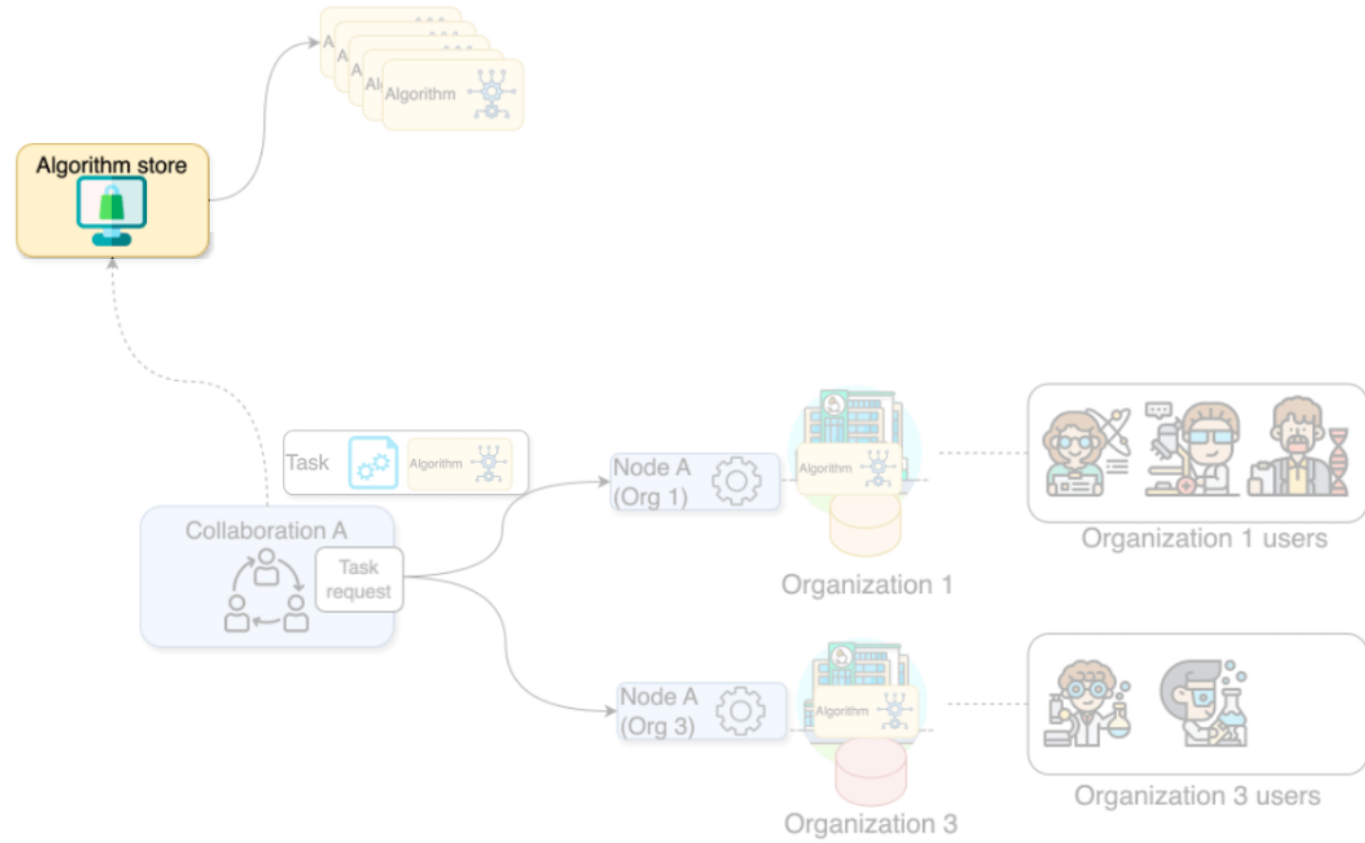Organization 1 users

Node A (Org 3) — Algorithm

Organization 3

Organization 3 users

Federated analysis projects with vantage6

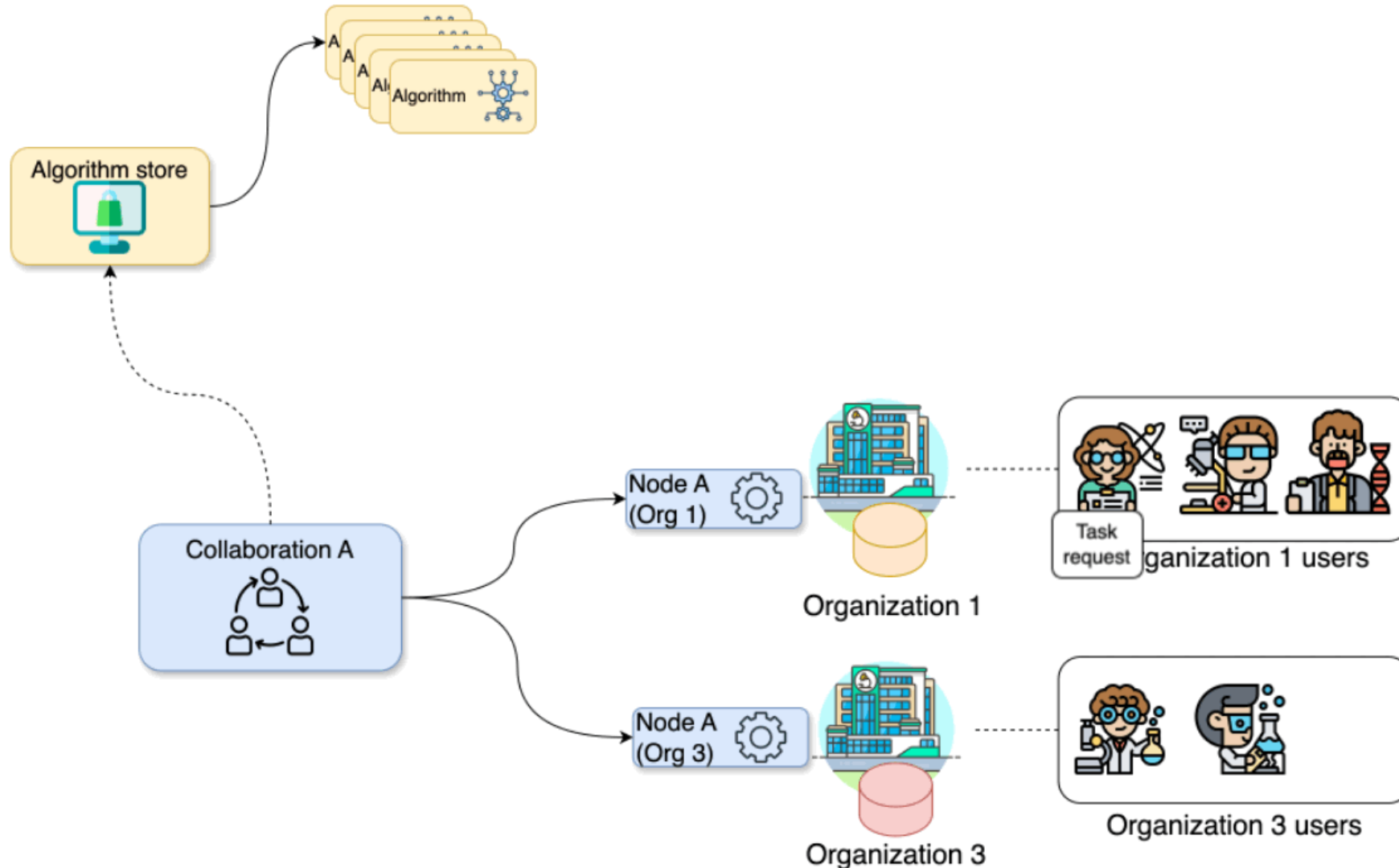## Algorithm store

A repository for trusted algorithms within a certain project.

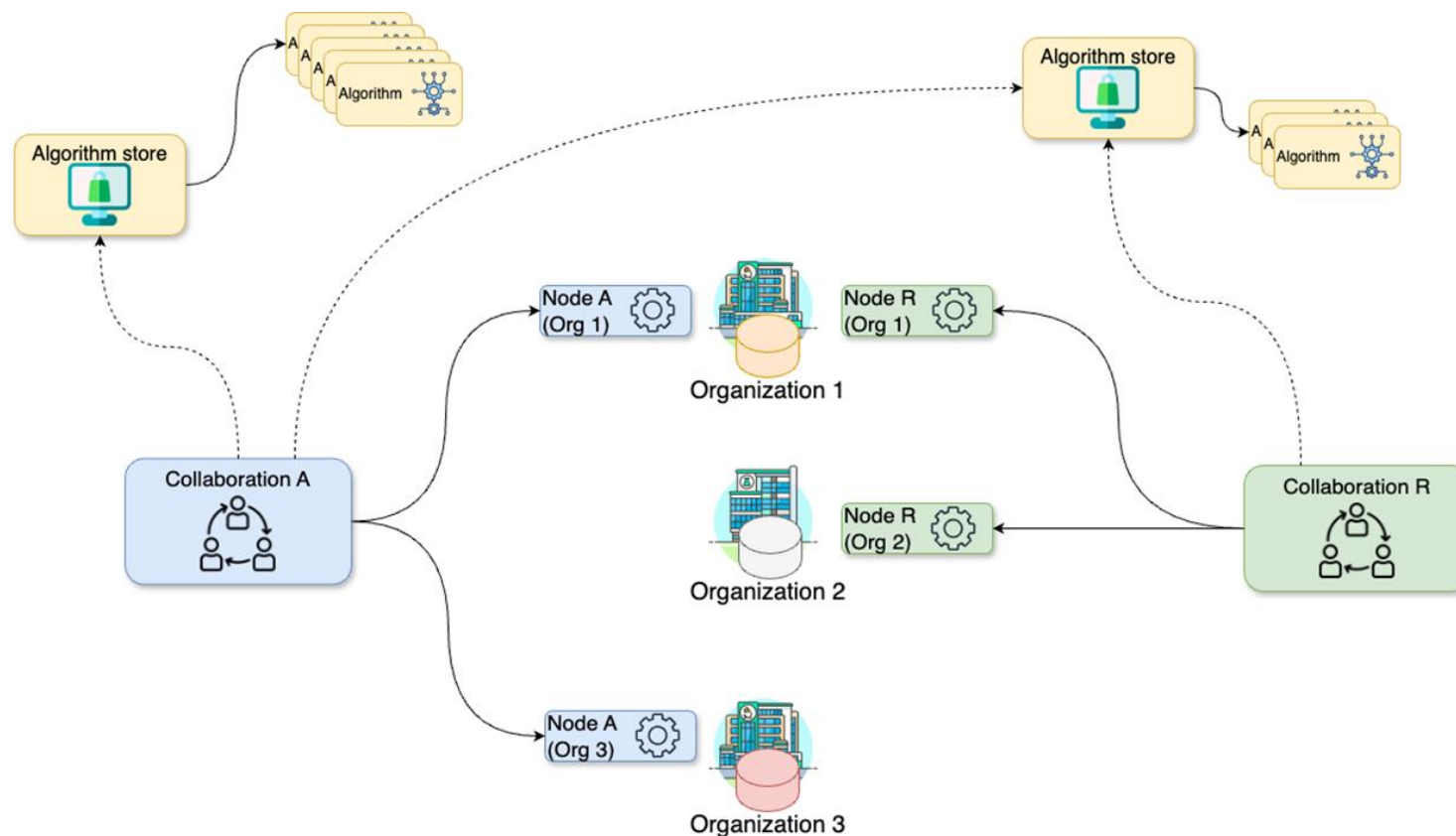# Federated analysis projects with vantage6

Federated analysis projects with vantage6
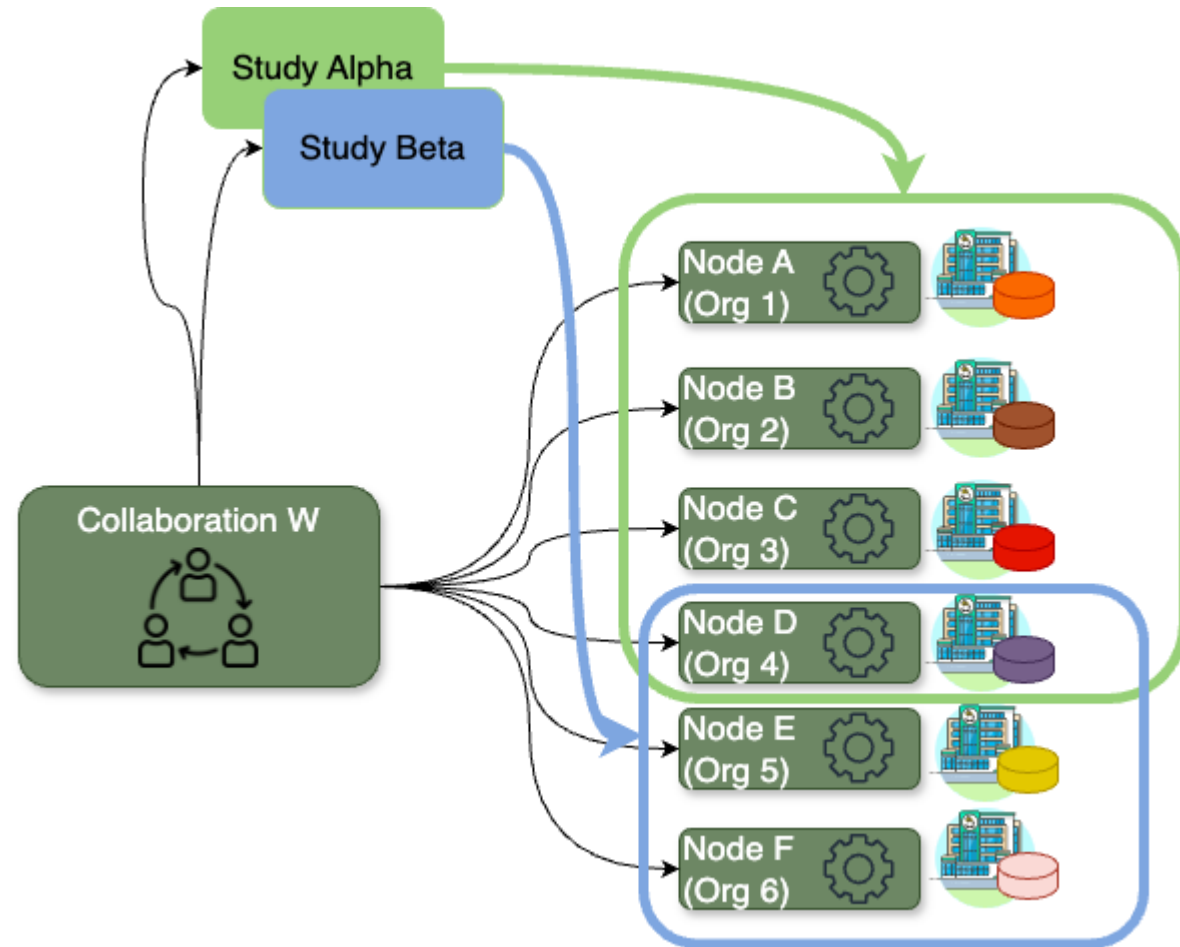

Expanding the scenario: Organizations can take part to multiple collaborations

Federated analysis projects with vantage6

Design your own project structure

Let's consider a scenario where you, on behalf of your research institute, want to conduct a new study on a particular illness across three major academic hospitals in the Netherlands: VUmc in Amsterdam, Maastricht UMC+, and UMC Utrecht, as these have valuable data related to the illness. Consider the following:

- Your research institute has an existing collaboration (with a different purpose, not related with yours) with UMC Utrecht and UMC Groningen. Hence, there is a vantage6 node already running on your institution for the said collaboration.

- You will be conducting this study with a colleague from your institute named Daphne. Both of you are already registered on the organization but without access to the existing collaborations.

How would the concepts described above map to your potential use case?

1. **Which organizations will you need to add to your collaboration?**
2. **How many new nodes would you need to set up and on which organizations?**
3. **How many users would be created?**

Solution

1. **Which organizations will you need to add to your collaboration?**

   In this case the organizations would be the academic hospitals as well as your own organization: VUmc, Maastricht UMC+, UMC Utrecht and your research institute. Note that UMC Utrecht must be added to the new collaboration despite being already part of an existing one.

2. **How many new nodes would you need to set up and on which organizations?**

   One node for every academic hospital, so 3. Note that UMC Utrecht needs a new node despite already having one, as the existing one is for a different collaboration.

3. **How many users would be created?**

   There is no need to create new users, as these are already registered on the organization. Note that the users are linked only to the organization, not to the nodes.
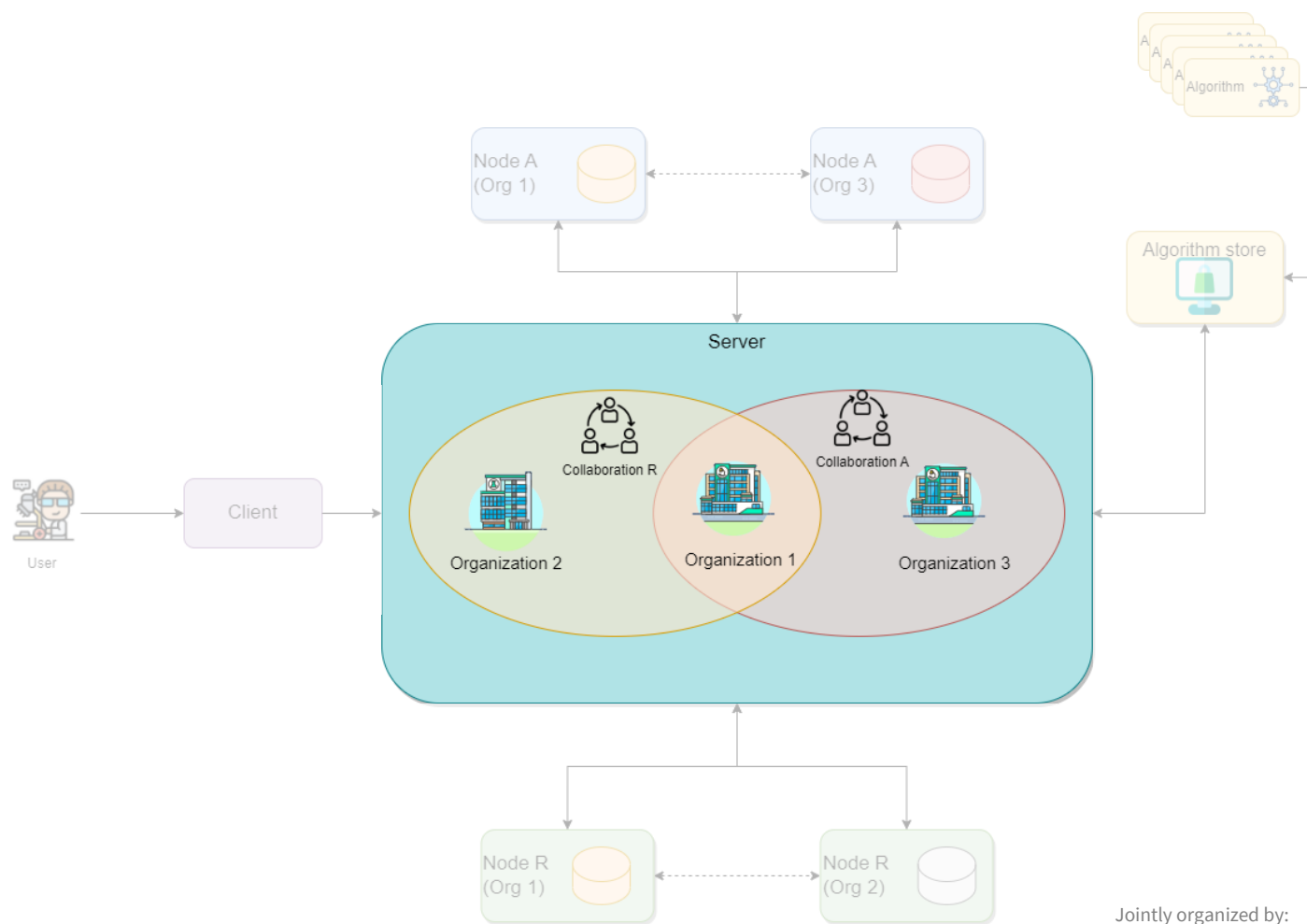
The vantage6 infrastructure
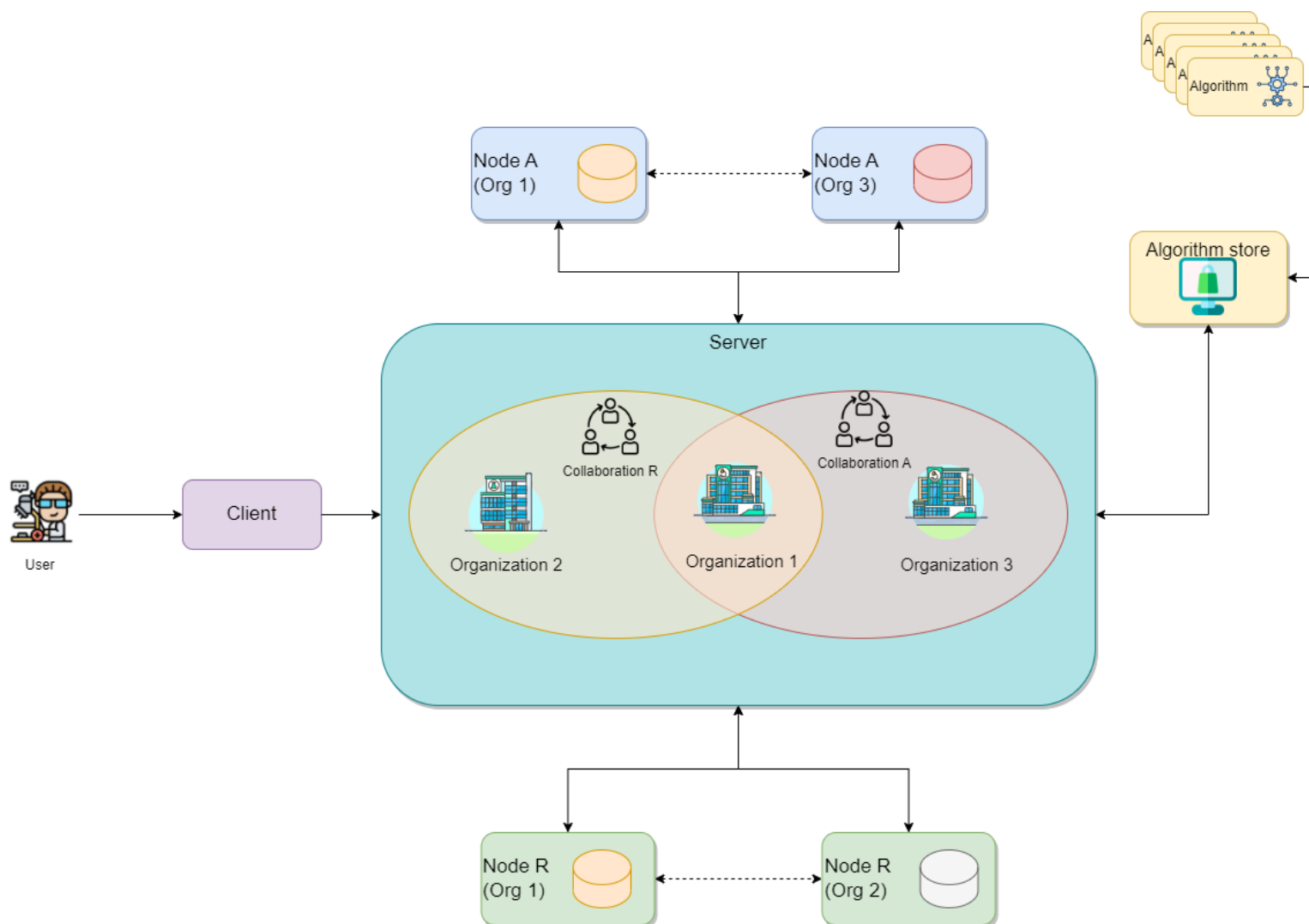
The vantage6 infrastructure

Server

Communication hub between clients and nodes. It also handles authentication and authorization to the system.
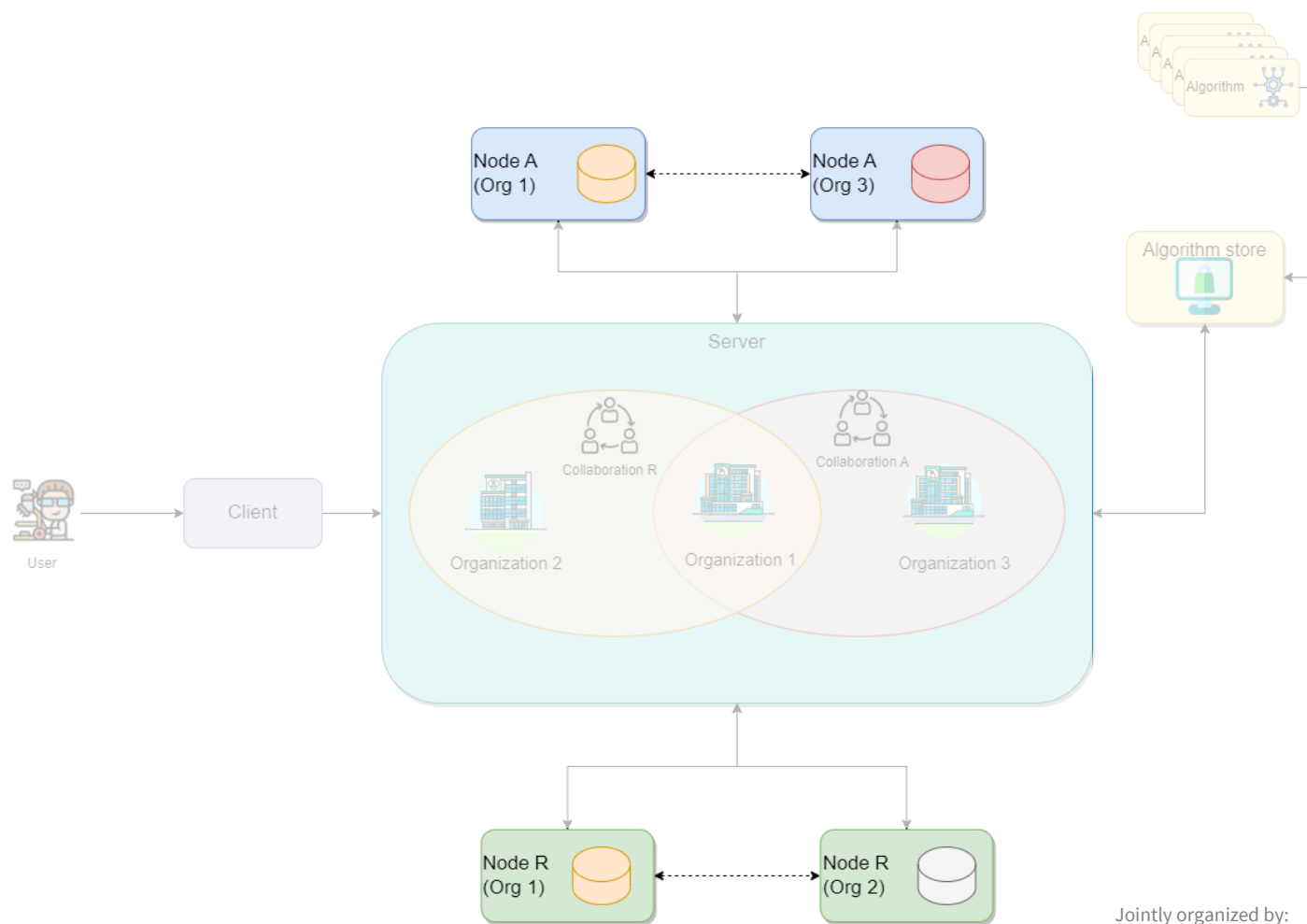
The vantage6 infrastructure

Data station

Hosts the data and the vantage6 node. The vantage6 node executes the allowed algorithms on the local data and sends the results back to the server.

VANTAGE
WORKSHOP

Node A (Org 1)
Node A (Org 3)

Algorithm

Algorithm store

Server

Collaboration R
Collaboration A

Organization 2
Organization 1
Organization 3

User

Client

Node R (Org 1)
Node R (Org 2)

Jointly organized by:
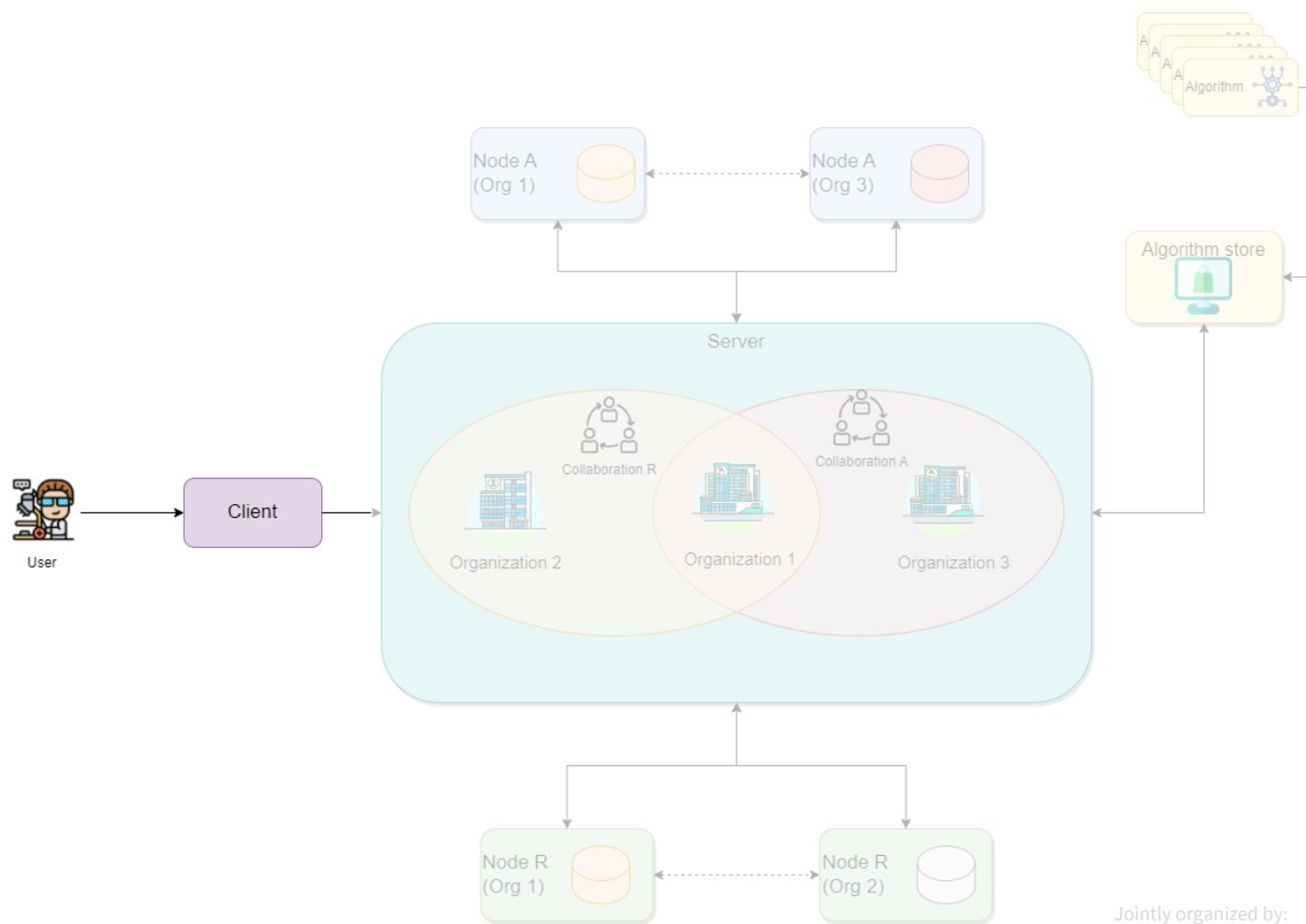
netherlands
eScience center
KNL

The vantage6 infrastructure

Client

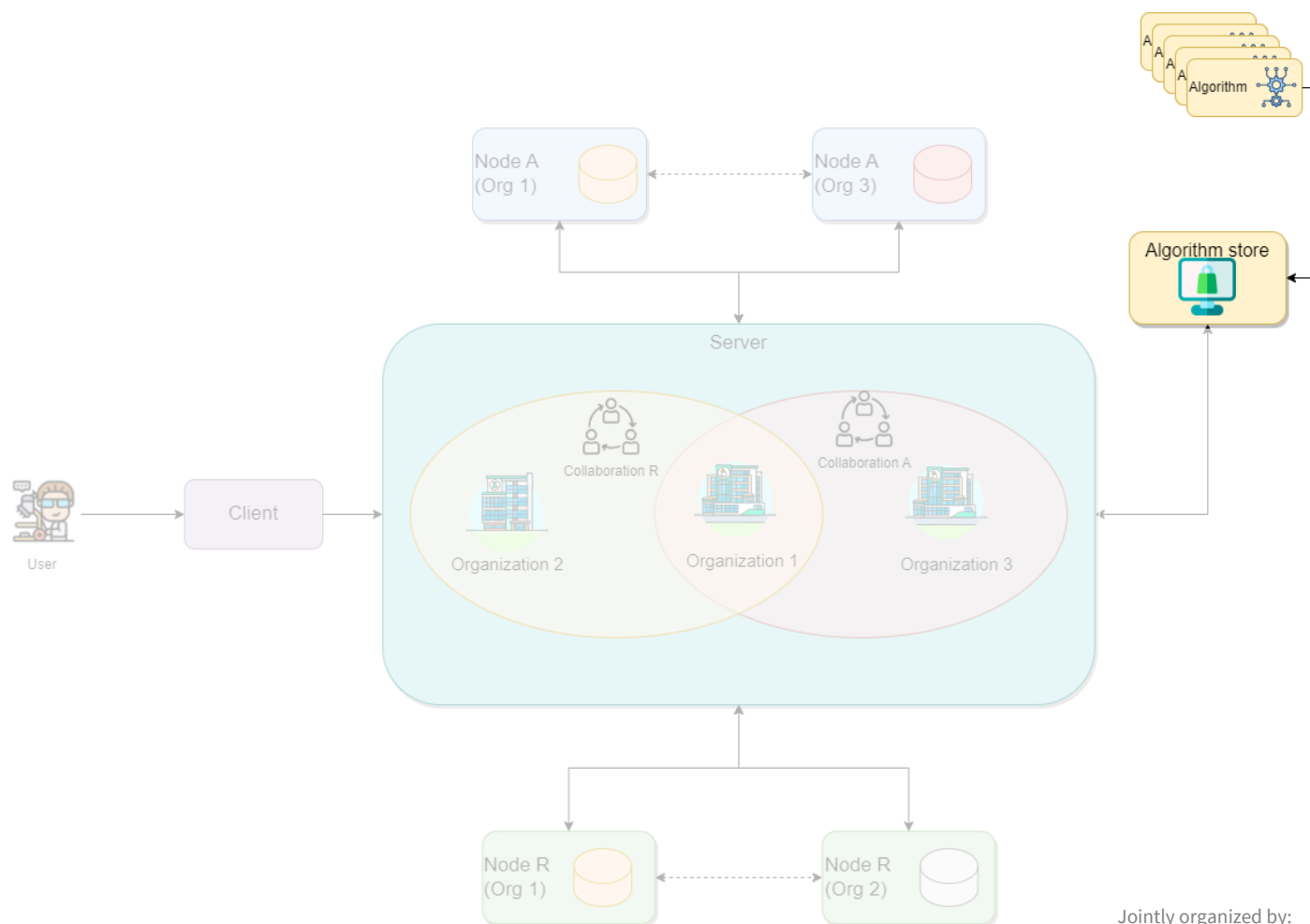*Entity that interacts with the vantage6 server via:*

- *API*
- *User interface*
- *Python client*

The vantage6 infrastructure
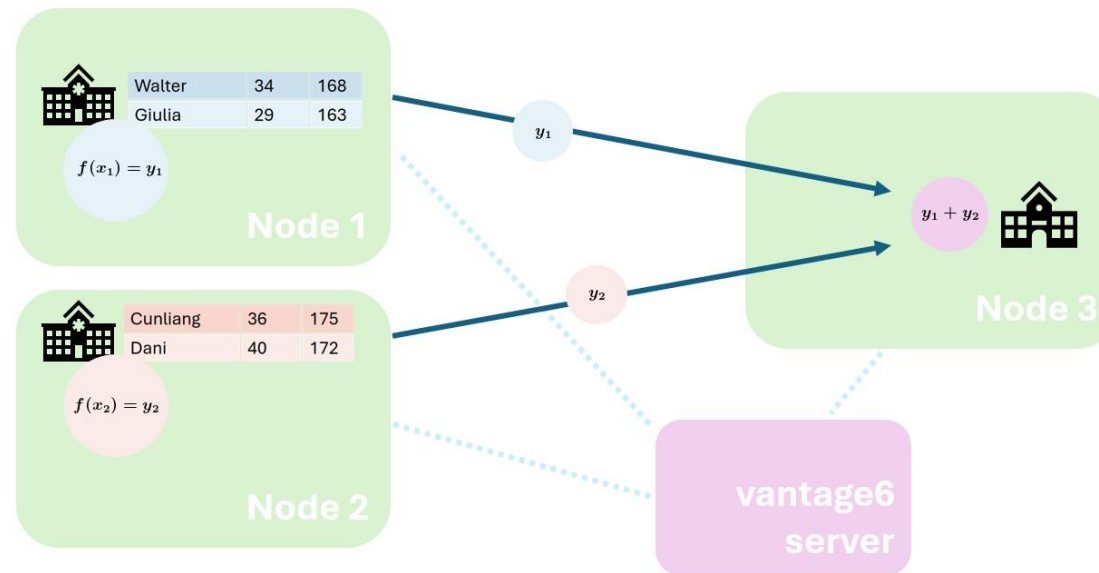
Algorithm store

Repository for trusted algorithms. It allows researchers to explore which algorithms are available and how to run them.

How algorithms run in vantage6

Let us consider the federated sum from chapter 1 again

Data sources and the aggregation entity are mapped to nodes.

The vantage6 server is on the side, coordinating the analysis.

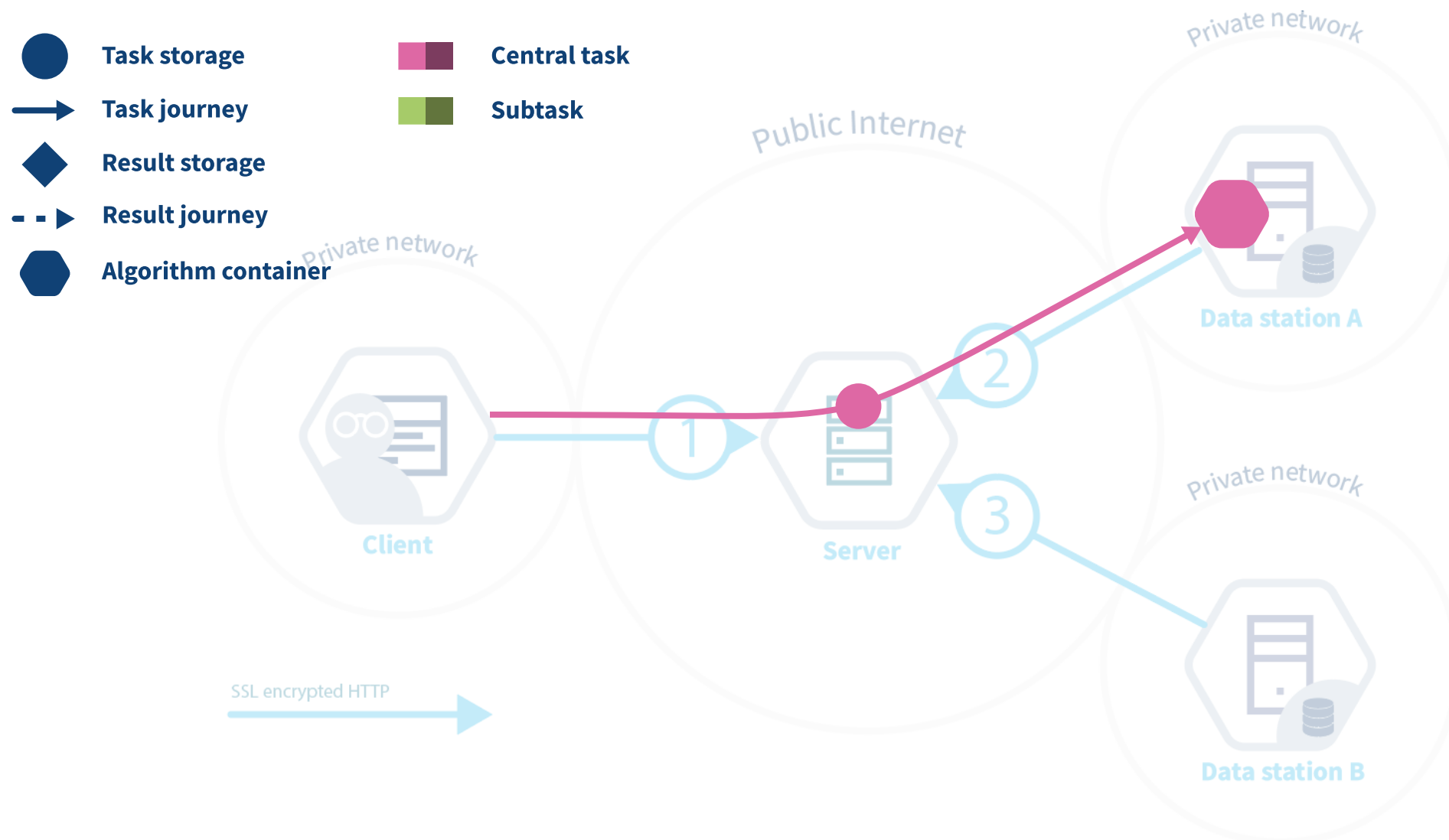# How algorithms run in vantage6

Federated algorithms can be split in a **federated** and a **central** part:
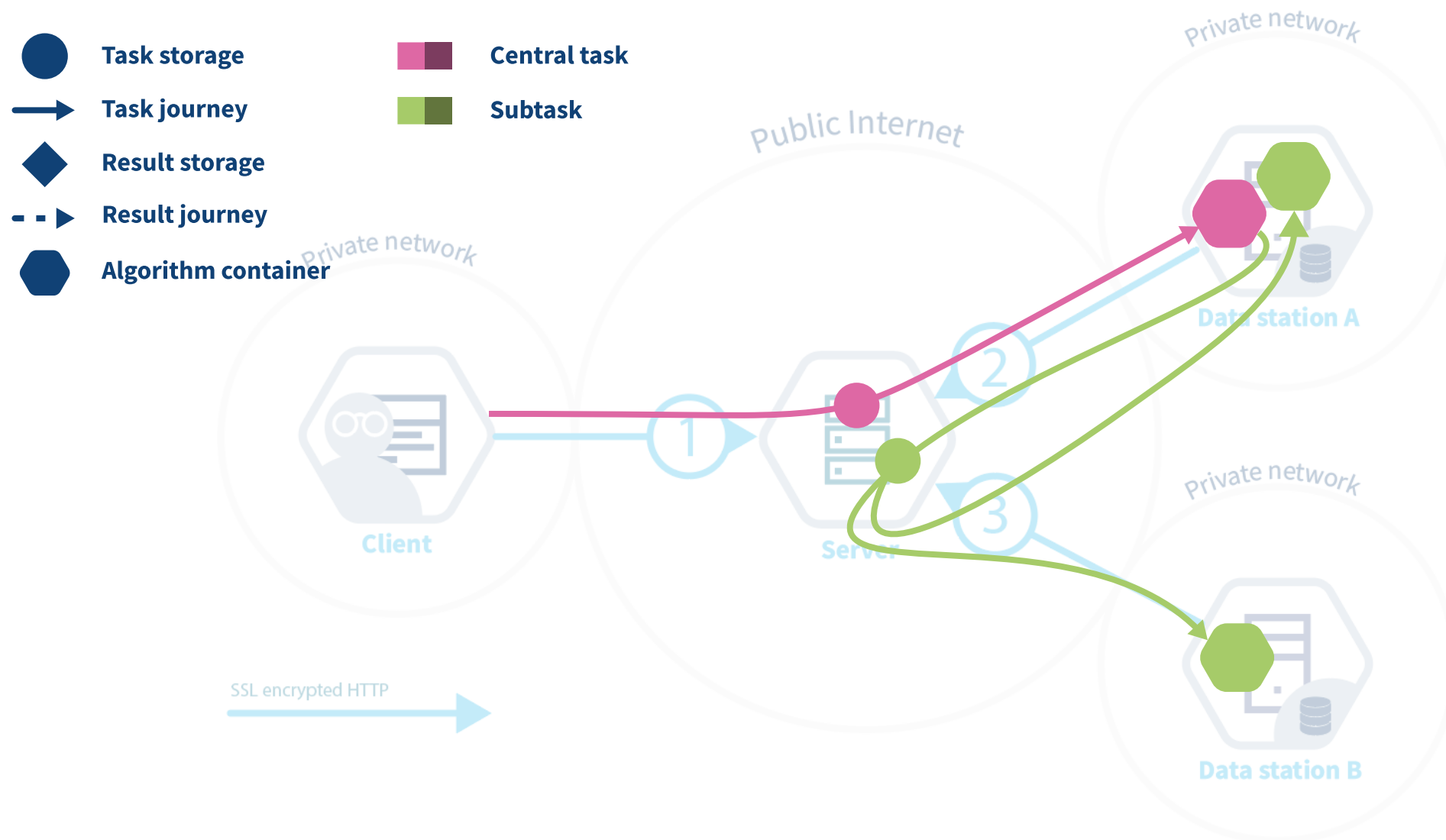
- **Central**: The central part of the algorithm is responsible for orchestration and aggregation of the partial results.

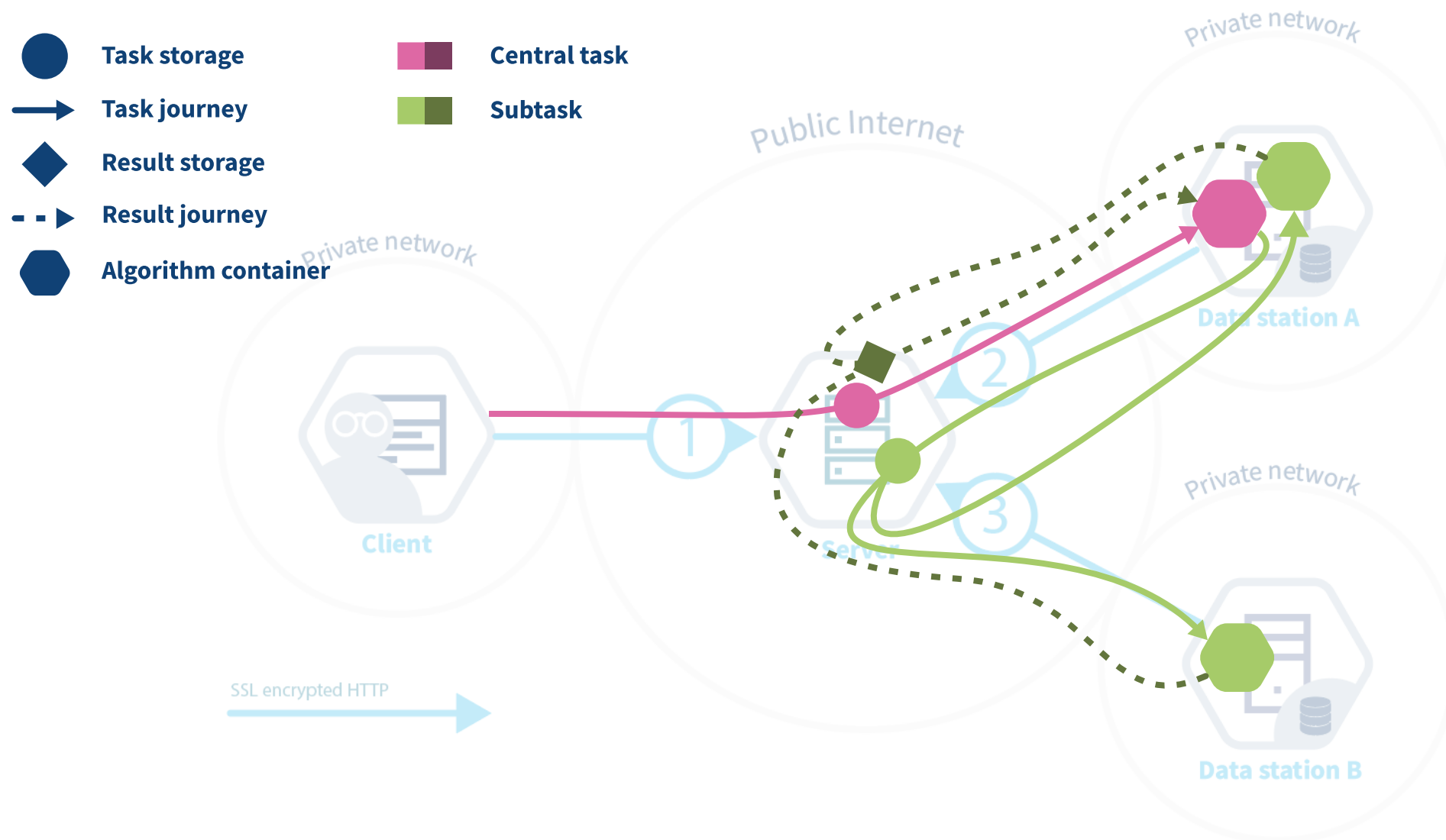- **Federated**: The partial tasks are executing computations on the local privacy sensitive data.
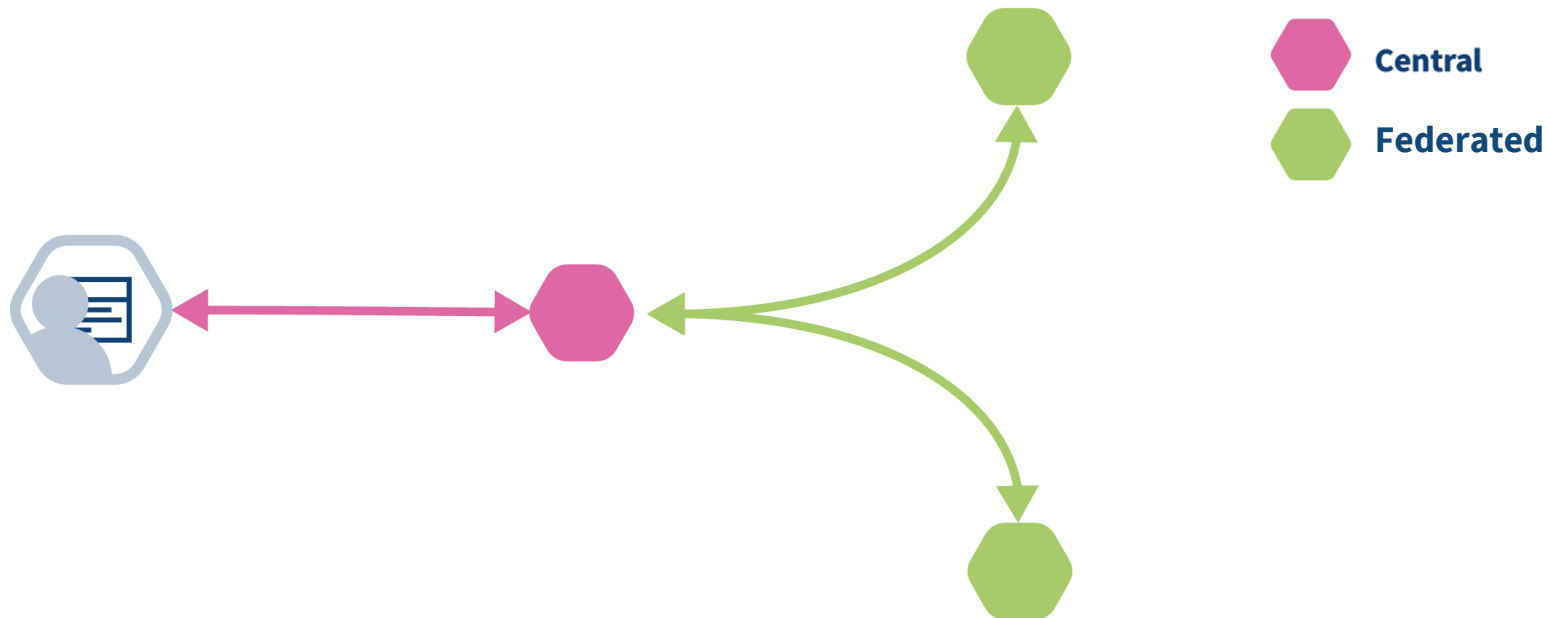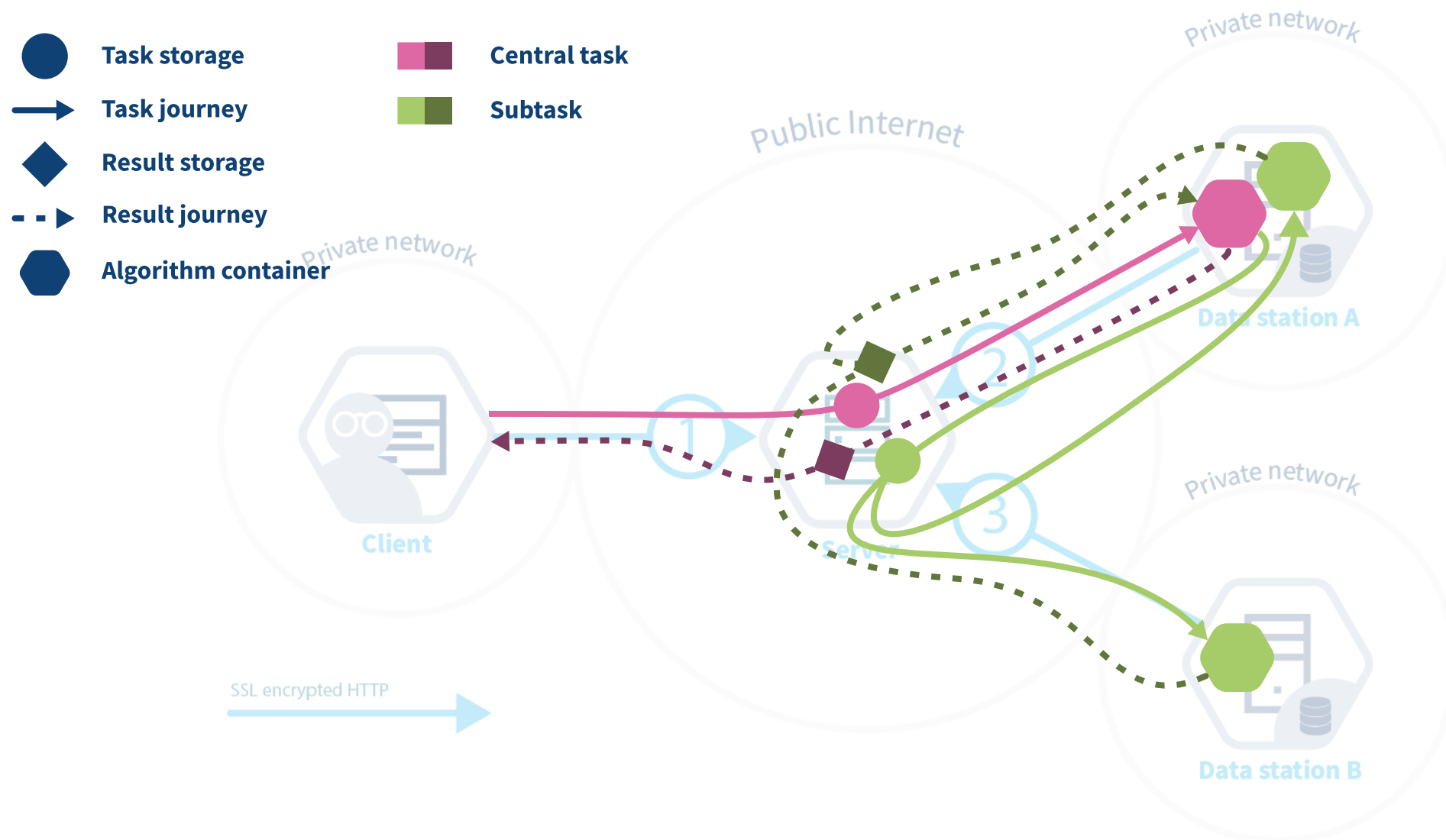
How algorithms run in vantage6

Algorithm projection on infrastructure

Algorithm projection on infrastructure

Algorithm projection on infrastructure

Exercise

⚡ **CHALLENGE**

Two centers A and B have the following data regarding the age of a set of patients:

a=[34,42,28,49]

b=[51,23,44]

Each center has a data station, and we want to compute the overall average age of the patients.

Given that the central average can be computed by summing up all the values and dividing the sum by the number of values, using the following equation:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{38 + 42 + 28 + 49 + 51 + 23 + 44}{7}$$

To make it ready for a federate computation, the equation can be rewritten as the sum of the elements of A plus the sum of the elements of B, divided by the number of elements of A plus the number of elements of B:

$$\bar{x} = \frac{1}{n_a + n_b}\sum_{i=1}^{n_a} a_i + \sum_{i=1}^{n_b} b_i = \frac{1}{4 + 3}[(34 + 42 + 28 + 49) + (51 + 23 + 44)]$$

Can you determine which part of the infrastructure will execute each part of the computation, and which is the result returned by the different parts?

Solution

The Server starts the central task on one of the two nodes (e.g. Data station A).

The node A starts two subtasks, one per node. Node A will run the following computation:

$$S_a = \sum_{i=1}^{n_a} a_i = 34 + 42 + 28 + 49$$

and it will return the following results to the central task:

$S_a = 153, \ n_a = 3$

Node B will run the following computation:

$$S_b = \sum_{i=1}^{n_b} b_i = 51 + 23 + 44$$

and it will return the following results to the central task:

$S_b = 118, \ n_b = 3$

The central task receives S_a and n_a from node A and S_b and n_b from node B, and will run the following computation:

$$\bar{x} = \frac{S_a + S_b}{n_a + n_b} = \frac{153 + 118}{4 + 3} = 38{,}71$$

Future developments of vantage6

- **Policies**: we want to extend the current available policies to a more generic policy framework in which any aspect of the vantage6 platform can be controlled by policies. This will maximize the flexibility of the platform and make it easier to adapt to new use cases.

- **Model repository**: Currently, vantage6 is focused on privacy enhancing techniques. Some of these techniques result in a model that can be used to make predictions. We want to extend vantage6 with a model repository in which these models can be stored, shared and used. This will make it easier to reuse models and to compare the performance of different models.