


From theory to practice


Running a PET analysis (without programming) on vantage6





A theoretical scenario

- An international consortium of research organizations working together on aging research.

Apple Innovations 




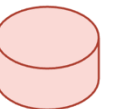
Banana Tech Solution 





Elderberry Enterprises 





Fig Future Tech 



Cherry Digital 



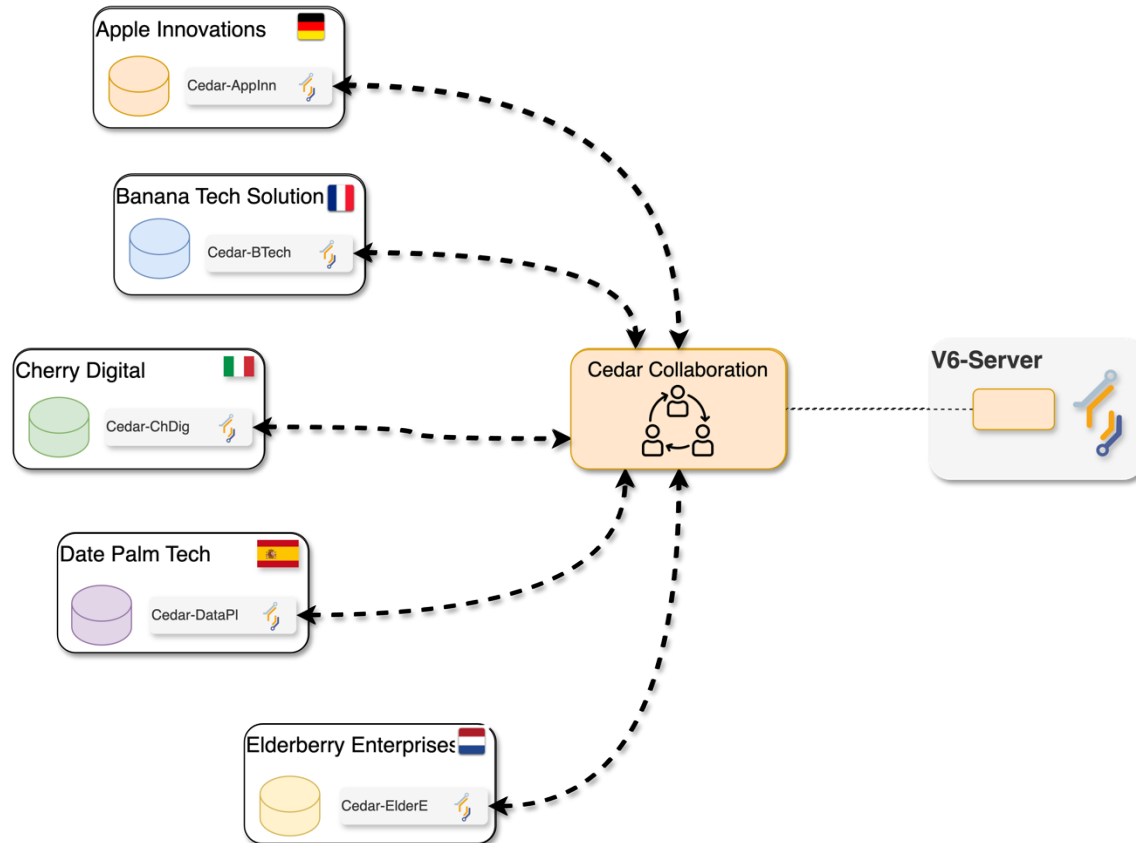
Date Palm Tech 



gender	age	height	weight	isOverweight	ageGroup
M	39	152	108	False	30 - 40
M	8	118	106	False	0 - 10
M	16	161	110	True	10 - 20
M	94	110	115	True	90 - 100
M	47	117	152	True	40 - 50
F	29	127	110	True	20 - 30
M	5	95	65	False	0 - 10
M	39	142	196	False	30 - 40
F	20	189	112	False	20 - 30
F	84	145	116	False	80 - 90



The scenario – checking v6 concepts!



- Six organizations, one collaboration.
- How many nodes need to be set up?

Node 

- Where are these nodes installed?

The scenario – checking v6 concepts!

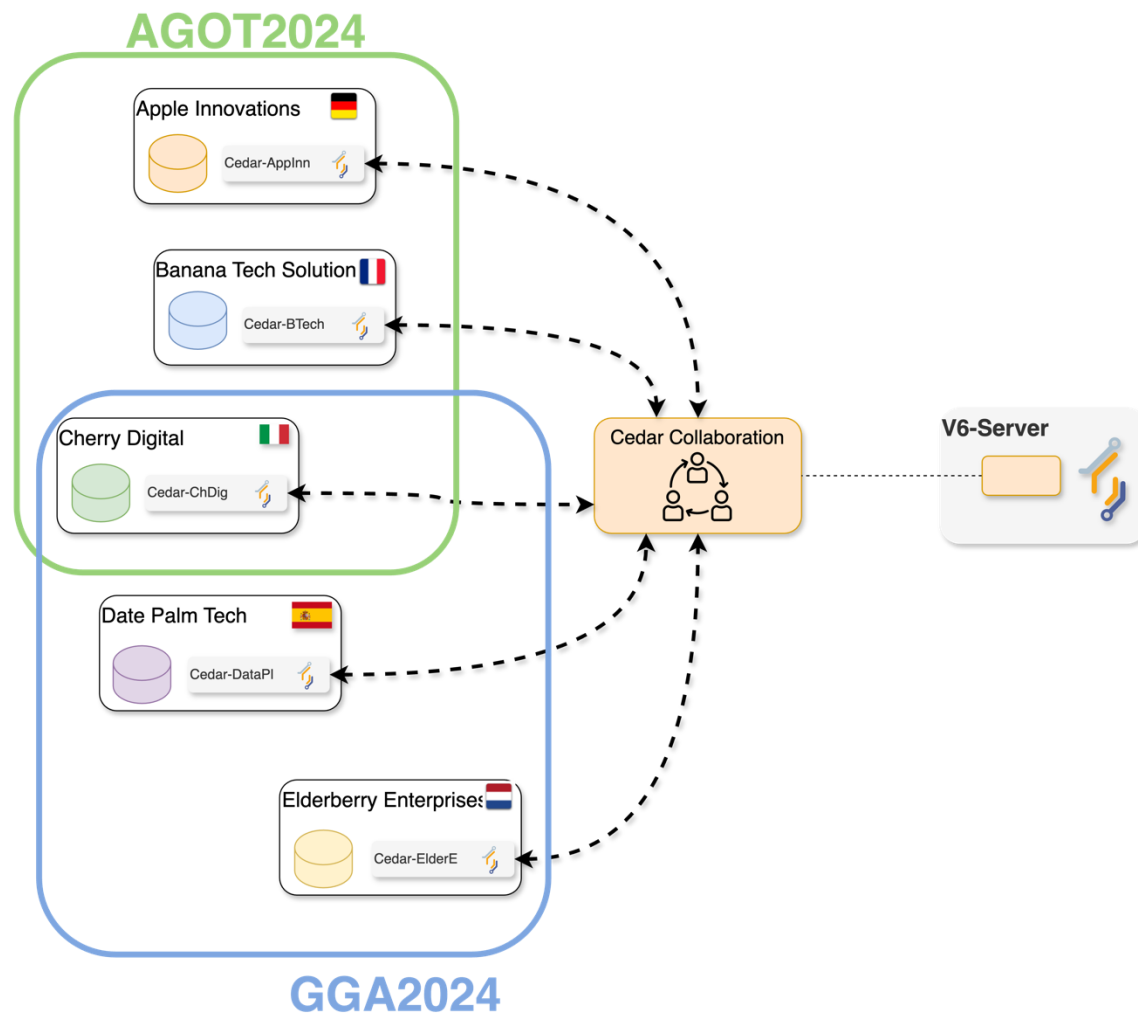
The consortium (from now on also referred as the **collaboration**) will initially focus on two different projects, each one involving different organizations:

- *Age-Related Variations in Overweight Prevalence: A Comparative Analysis Across Gender and Age Groups (**AGOT2024**) : four organization with relevant data.*
- *The Effect of Gender on Height Development Across Various Age Groups (**GGA2024**): three organization with relevant data.*

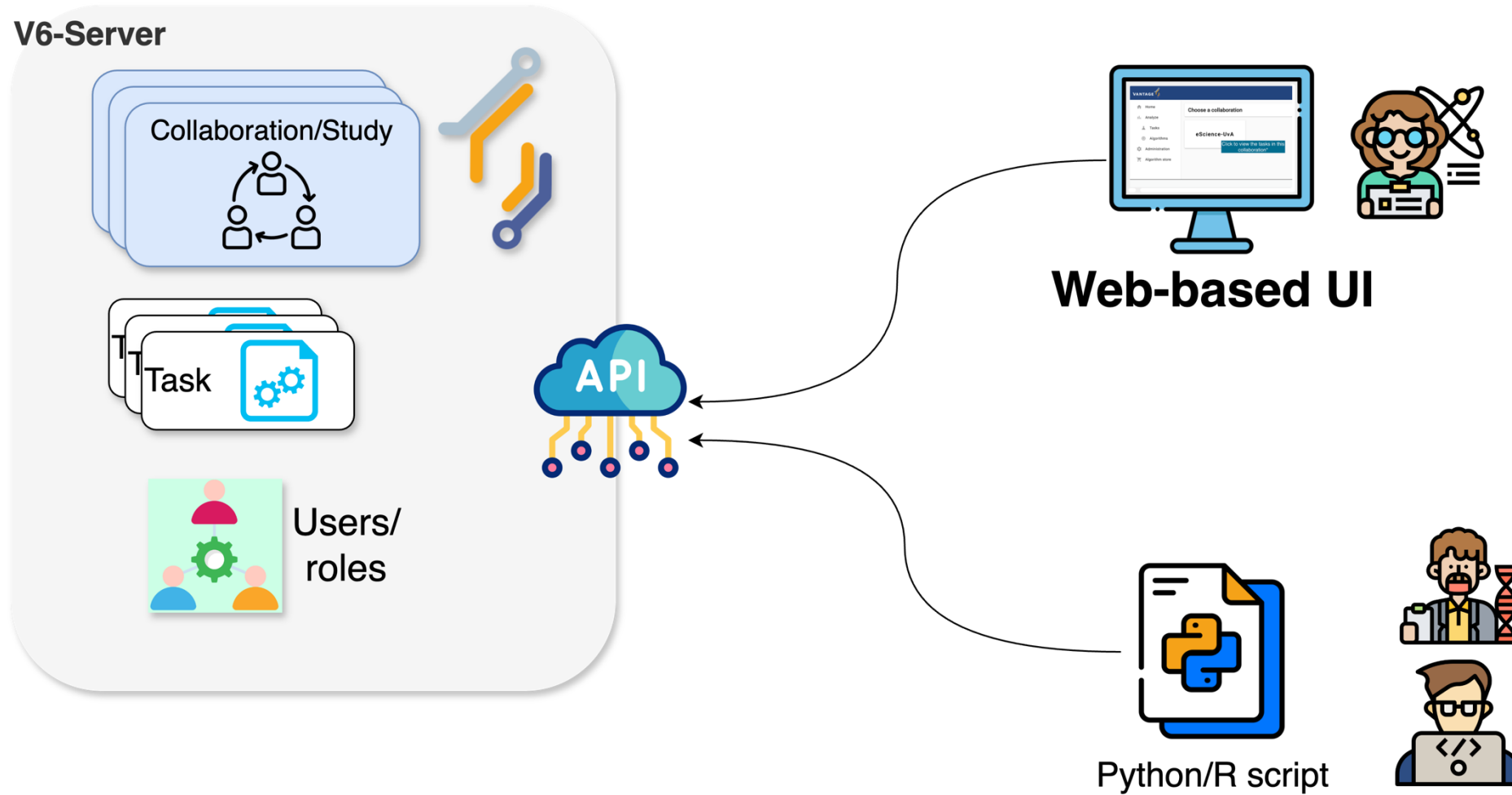
Which V6 concept is applicable here?

The scenario – checking v6 concepts!

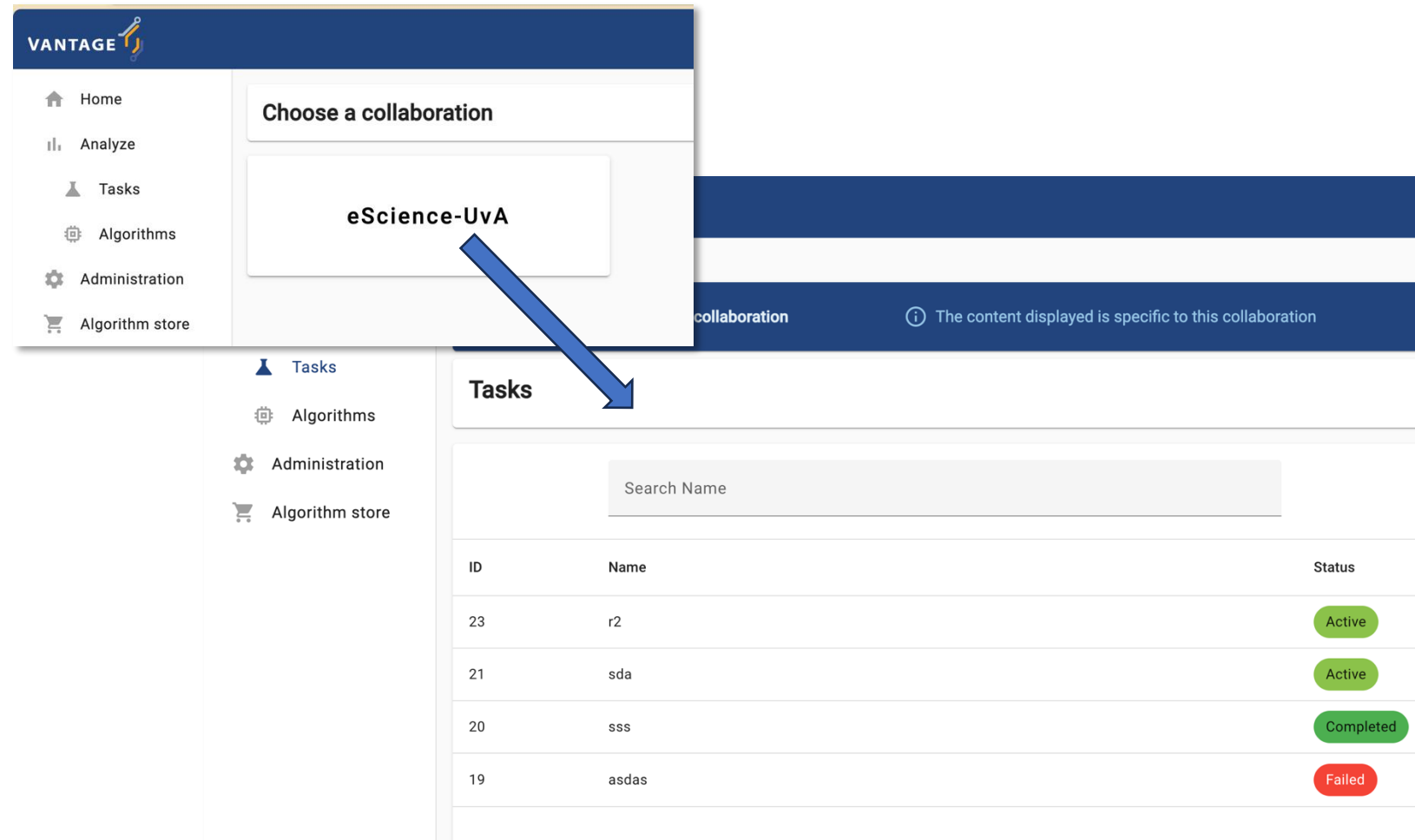
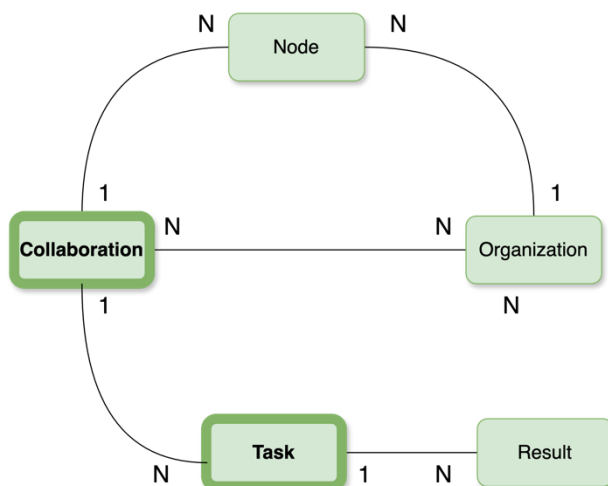
- Two studies: GGA2024, AGOT2024.



How to conduct an analysis?



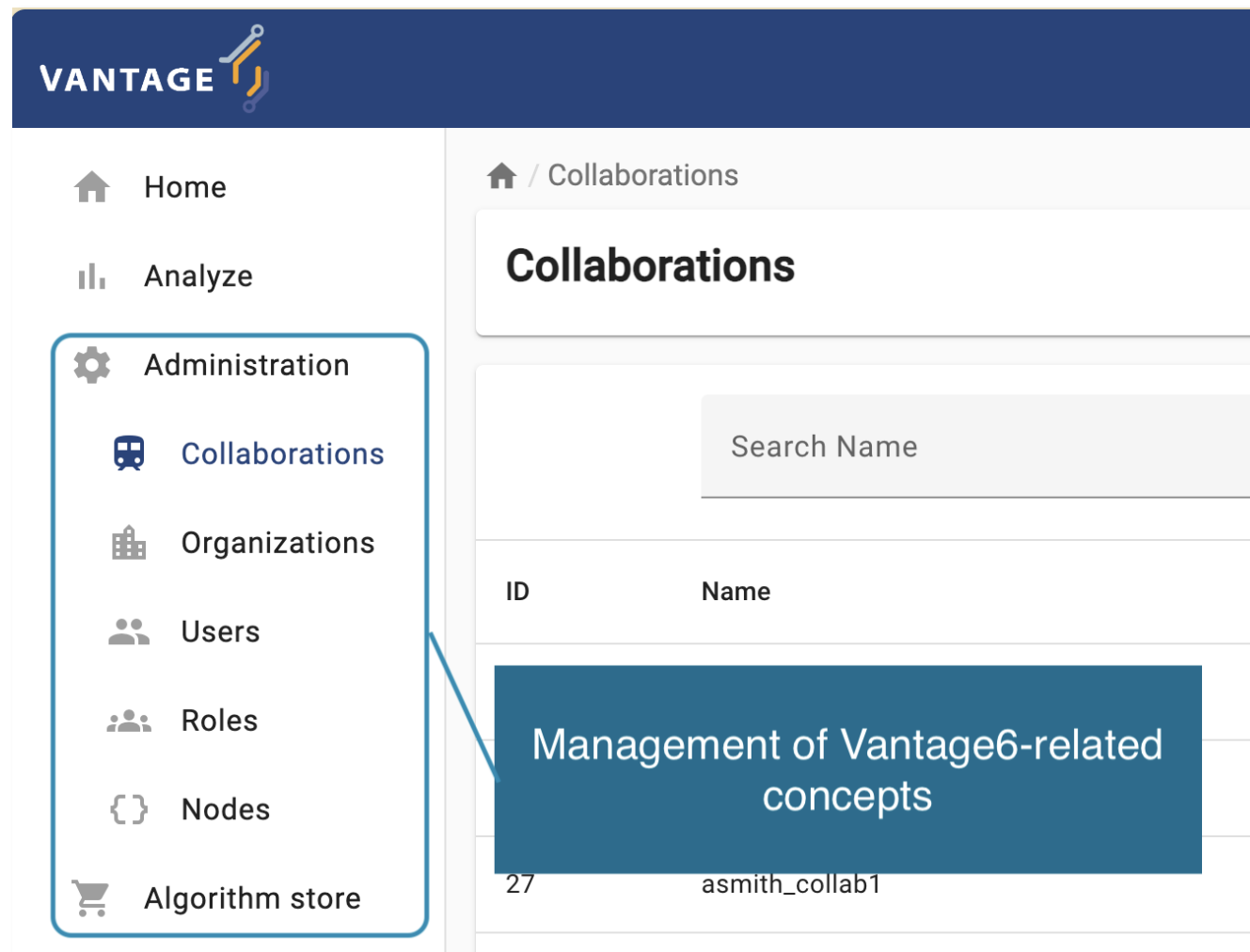
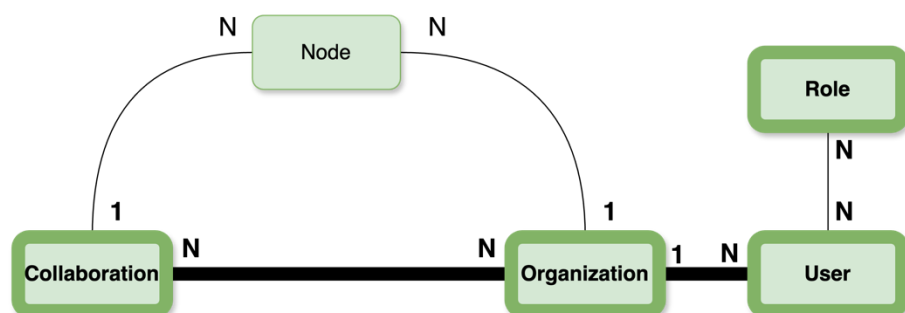
In this episode: web-based UI



The screenshot displays the VANTAGE web-based UI. The top navigation bar includes the VANTAGE logo and a sidebar menu with options: Home, Analyze, Tasks, Algorithms, Administration, and Algorithm store. The main content area is titled 'Choose a collaboration' and features a list of collaborations, with 'eScience-UvA' highlighted. A blue arrow points from this collaboration to the 'Tasks' section below. The 'Tasks' section includes a search bar and a table of tasks.

ID	Name	Status
23	r2	Active
21	sda	Active
20	sss	Completed
19	asdas	Failed

Web-based UI



The screenshot displays the VANTAGE web-based UI. The top navigation bar includes the VANTAGE logo and a home icon. The left sidebar contains a menu with the following items: Home, Analyze, Administration (highlighted with a blue box), Collaborations, Organizations, Users, Roles, Nodes, and Algorithm store. The main content area shows the 'Collaborations' page, which includes a search bar labeled 'Search Name' and a table with columns 'ID' and 'Name'. A blue callout box points to the 'Collaborations' menu item and the table, containing the text 'Management of Vantage6-related concepts'.

ID	Name
27	asmith_collab1

Getting familiar with the UI

CHALLENGE 1

Log into the UI using the information provided and navigate to the administration page. Then, update your email, first name, and last name.

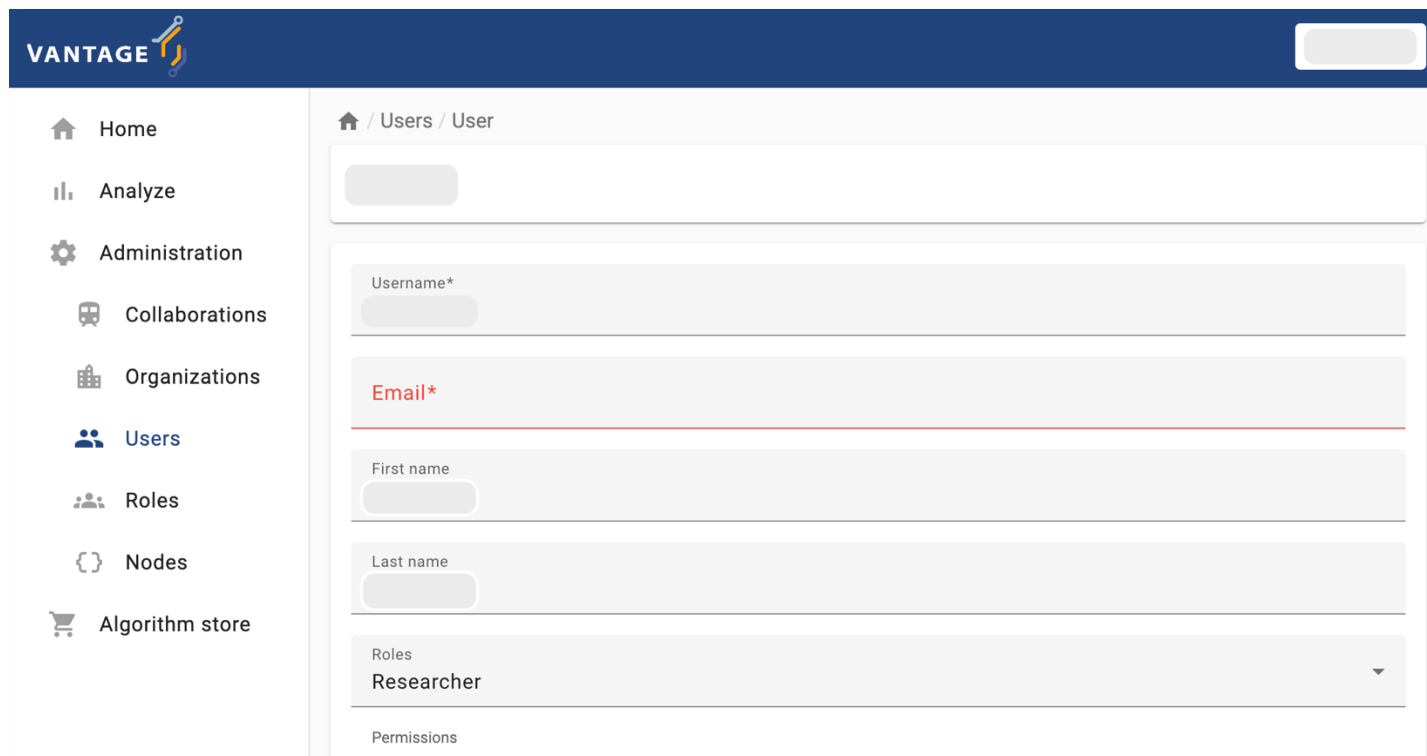


HINT


Use the **Users** option on the Administration panel on the left.

Getting familiar with the UI

 SOLUTION



The screenshot displays the Vantage web application interface. The top navigation bar is dark blue with the VANTAGE logo and a user profile icon. A left sidebar contains a menu with icons and labels: Home, Analyze, Administration, Collaborations, Organizations, Users (highlighted), Roles, Nodes, and Algorithm store. The main content area shows the breadcrumb path 'Home / Users / User' and a form for creating a new user. The form includes input fields for Username*, Email* (with a red error line), First name, and Last name. Below these is a dropdown menu for Roles, currently set to 'Researcher', and a section for Permissions.

VANTAGE 

Home / Users / User

Username*

Email*

First name

Last name

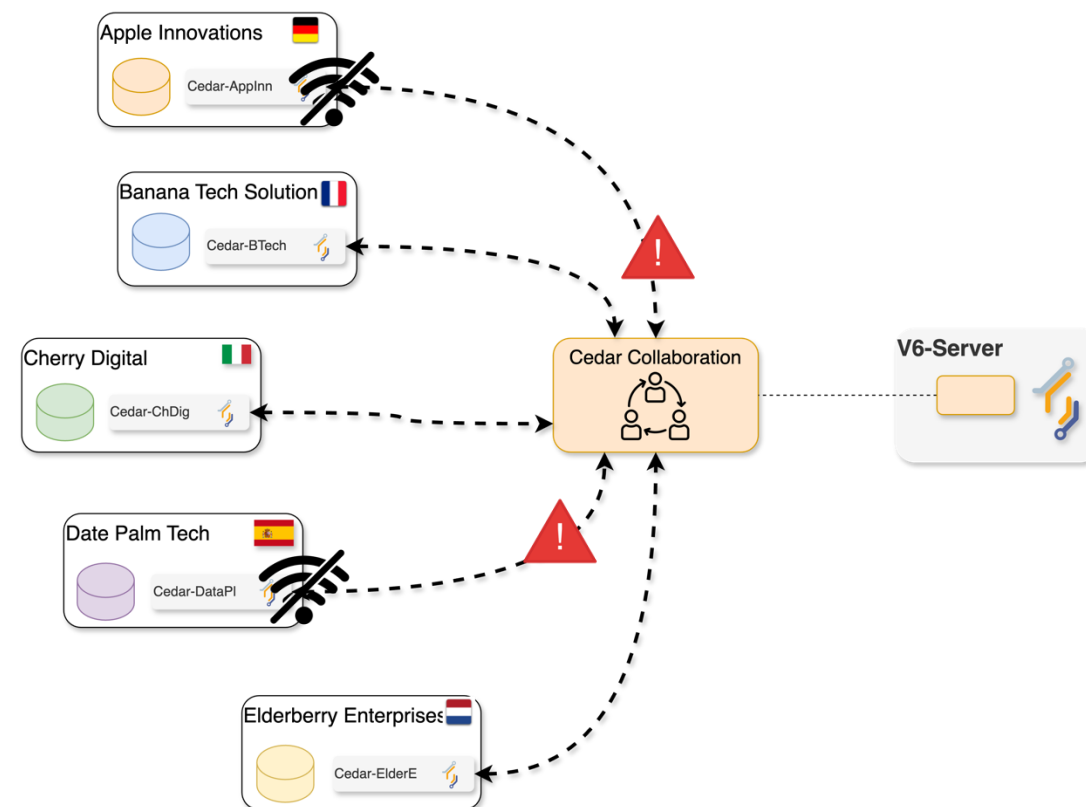
Roles
Researcher

Permissions

Running a PET: federated analysis

Are the nodes on my collaboration or study ready for a federated analysis?

- Node within a collaboration: autonomously managed by its corresponding organization.
- Some could be offline.
- UI is handy for checking this!



Running a PET: federated analysis

CHALLENGE 2

With your researcher credentials, explore the collaboration you have access to. Check which organizations are part of it and if they are online. Then, check which organizations were assigned to each study: **AGOT2024, GGA2924**.

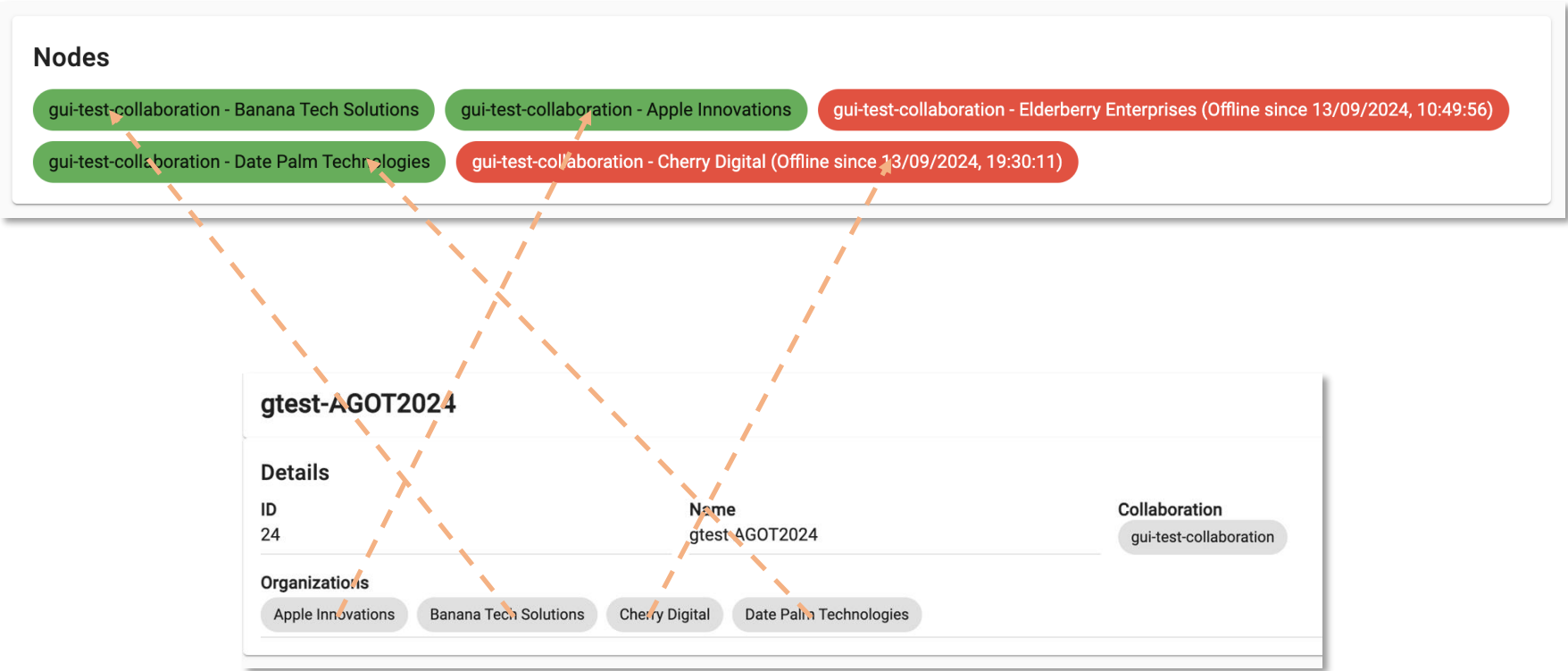
1. Which study is ready for a federated analysis?
2. If you need to perform an analysis for the study that is **not** ready, which organization you would need to contact to fix this situation?

HINT

Use the **Collaborations** option on the Administration panel on the left. A hand-drawn diagram of your collaboration and studies will be handy for the next steps!

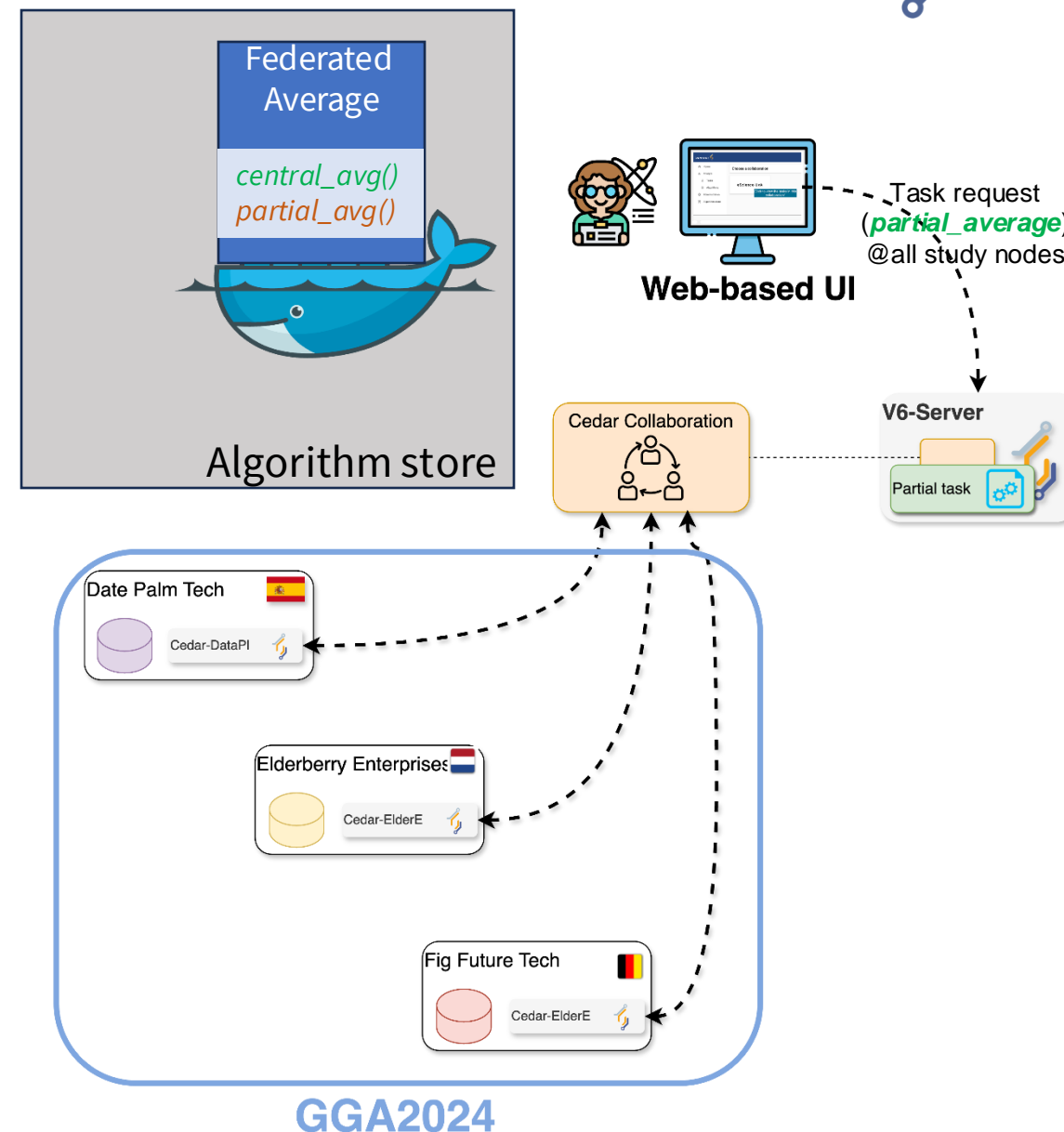
Getting familiar with the UI

 **SOLUTION**



Running a PET: federated analysis

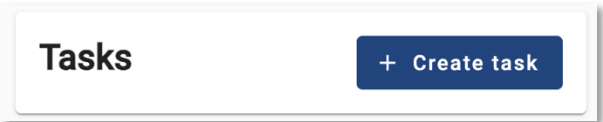
- Creating a new task:
 - You will run the **partial_average** function, of the **Federated Average algorithm**, on **all** the nodes of a study (the one with no offline nodes).



Running a federated algorithm

⚡ CHALLENGE 3

Create your first task!



- Use the *study* with **no** offline nodes.
- **Give your own name to the task so you can find it later.**
- Choose the Average algorithm.
- Select the **partial_average** function and all the organizations within the study.
- Choose the **default** database.
- Choose a numerical variable as the input.

The task you just requested should be listed with a 'pending' status. Once finished, download the JSON results and open them on a text editor.

- What does the content of these files mean?
Why the results are formatted like that?

gender	age	height	weight	isOverweight	ageGroup
M	39	152	108	False	30 - 40
M	8	118	106	False	0 - 10
M	16	161	110	True	10 - 20
M	94	110	115	True	90 - 100
M	47	117	152	True	40 - 50
F	29	127	110	True	20 - 30
M	5	95	65	False	0 - 10
M	39	142	196	False	30 - 40
F	20	189	112	False	20 - 30
F	84	145	116	False	80 - 90

⚡ HINT

The data on 'default' databases on all the nodes looks like this 🙌.

Running a federated algorithm

SOLUTION

Create task

1 Select study

Your collaboration d
Select the study you

Study*
AGOT2024

Next

Home / Tasks / Create task

Selected collaboration **workshop-collab** ⓘ The conten
sk for.

Create task

1 Select algorithm

Select which algorithm you

Algorithm*
Average

Provide a suitable name fo

Name*

Optionally, describe your ta

Description

Back Next

3 Select function

Select which function you want to run

Function*
partial_average (federated)

Select the organizations that will carry out a partial function

Organization(s) to run this function*

☐ Apple

☐ Banana

☐ Cherry

☐ Elderberry

Submit

4 Select databases

Select the database(s) you want to use. The options in the dropdown

Main database

Database*

default (csv)

kaplan_meier (csv)

Submit

5 Define parameters

Provide parameter values for the selected algorithm function

Name of the column to calculate average for

column_name*

Back

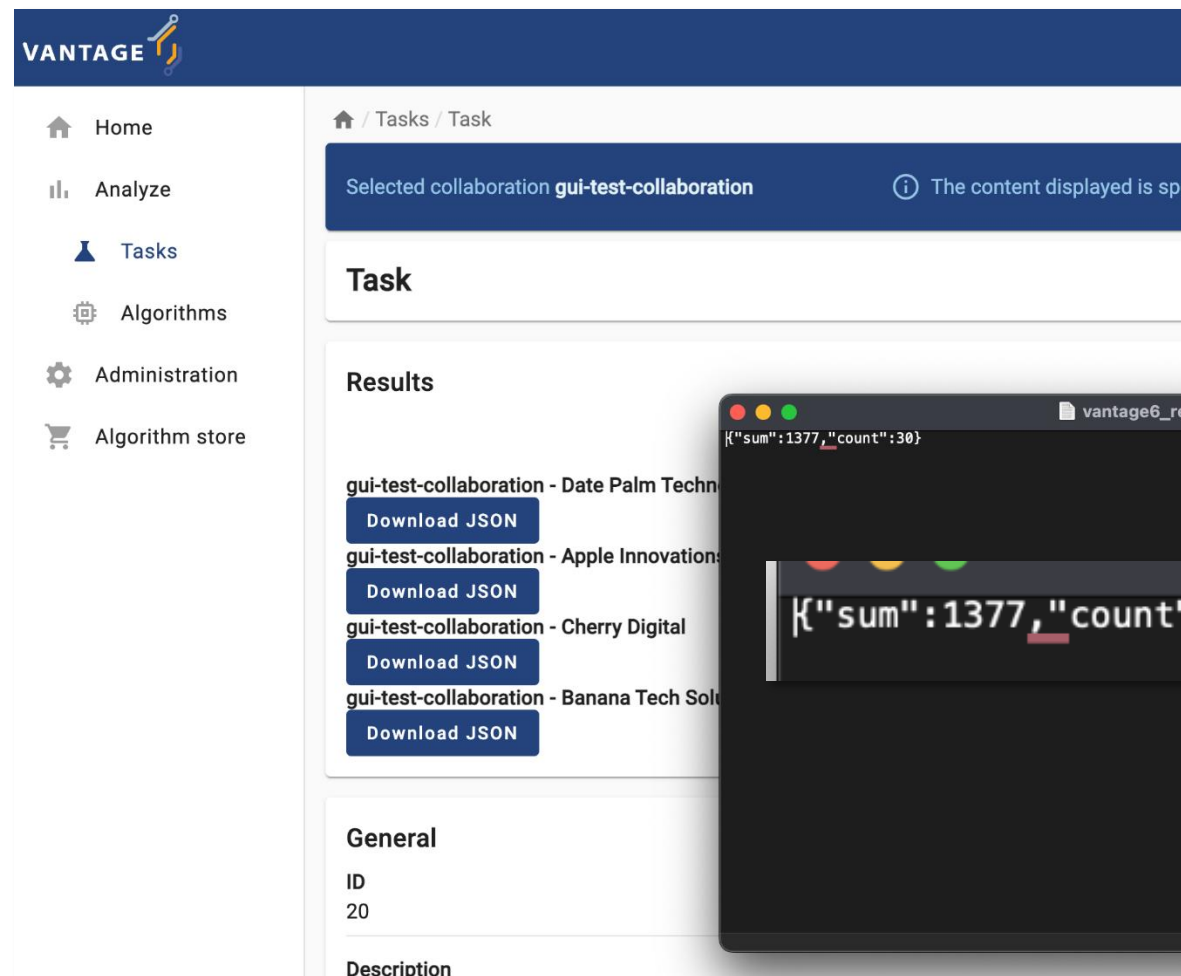
Submit

Getting familiar with the UI

⚡ SOLUTION

The **partial_average** returns the two values needed by the central function of the ‘federated average’ algorithm, as described in Chapter 2: the number of records within the database, and their sum.

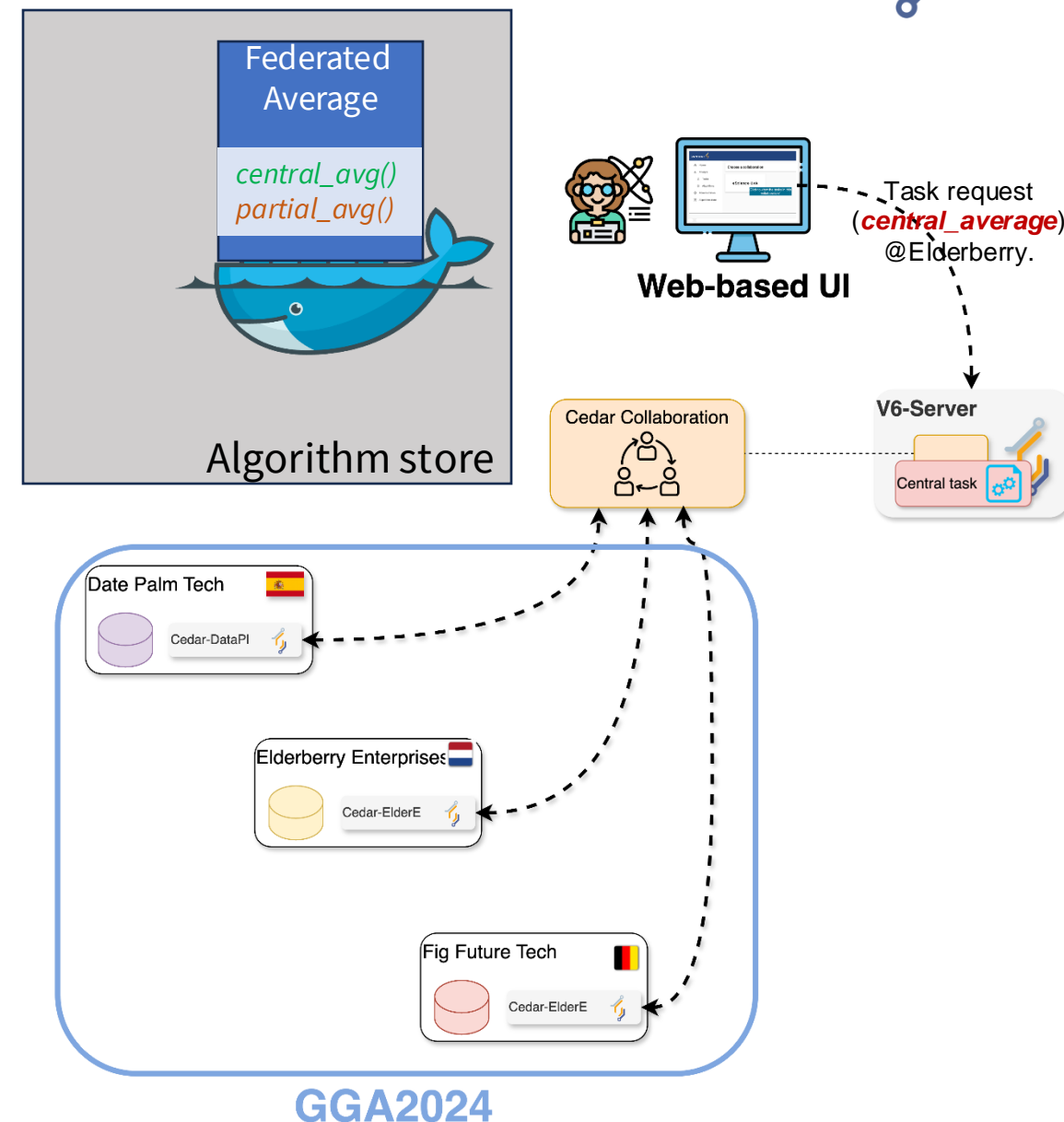
The algorithm is ‘encoding’ these values as a JSON document, so they can be used later (e.g., by another function or program).



The screenshot displays the VANTAGE web application interface. On the left is a navigation menu with options: Home, Analyze, Tasks, Algorithms, Administration, and Algorithm store. The main content area shows the 'Task' details for 'gui-test-collaboration'. Under the 'Results' section, there is a list of four entries, each with a 'Download JSON' button. The entries are: 'gui-test-collaboration - Date Palm Techn...', 'gui-test-collaboration - Apple Innovations', 'gui-test-collaboration - Cherry Digital', and 'gui-test-collaboration - Banana Tech Sol...'. Below the results is a 'General' section showing 'ID 20'. A 'Description' section is partially visible at the bottom. Overlaid on the right side of the screenshot is a dark rectangular box containing a JSON string: `{"sum":1377,"count":30}`. Above this box, a smaller window titled 'vantage6_res' also displays the same JSON string.

Running a PET: federated analysis

- Let's create yet another task.
 - This time you will run the **central_average** function, of the **Federated Average algorithm**, on **one** of the study nodes (once again, on the study with **no** offline nodes).



Running a federated algorithm

CHALLENGE 4

Let's see what the central function does under the hood!

- Use the study with **no** offline nodes.
- Choose the **Average algorithm**.
- Select the **central_average** function and one of the organizations within the study.
- Choose the default database.
- Choose a numerical variable as the input.

Before submitting the request: why does the V6 UI let you select only a single organization for the central function?

Keep an eye on the Tasks section, and see how the processes are created:

- Why does one of the nodes show up as a Main process and also as a Child one?

HINT

As soon your task is created, go to the task list and open it. See how 'child processes' are created.

Algorithm pseudo-code

FUNCTION *partial_average*(*dataset*, *column_name*)

```
1  INFO("Extracting column", column_name)
2  column_values = dataset[column_name]
3
4  INFO("Computing partials")
5  local_sum = SUM(column_values)
6  local_count = SIZE(column_values)
7
8  RETURN {"sum": local_sum, "count": local_count}
```

END FUNCTION

- Why does one of the nodes show up as a Main process and also as a Child one?

Actual python code?

<https://bit.ly/v6-federated-avg>

FUNCTION *central_average*(*client*, *column_name*)

```
1  study_orgs = client.study.list()
2
3
4  INFO("Requesting partial computation")
5  task = client.task.create(
6      function = "partial_average"
7      column = column_name
8      organizations = study_orgs
9  )
10
11  INFO("Waiting for results")
12  results = client.wait_for_results(task)
13
14  global_sum = 0
15  global_count = 0
16
17  FOR EACH output IN results
18      global_sum = global_sum + output.sum
19      global_count = global_count + output.count
20
21  average = global_sum/global_count
22
23  RETURN {"average": average}
```

END FUNCTION

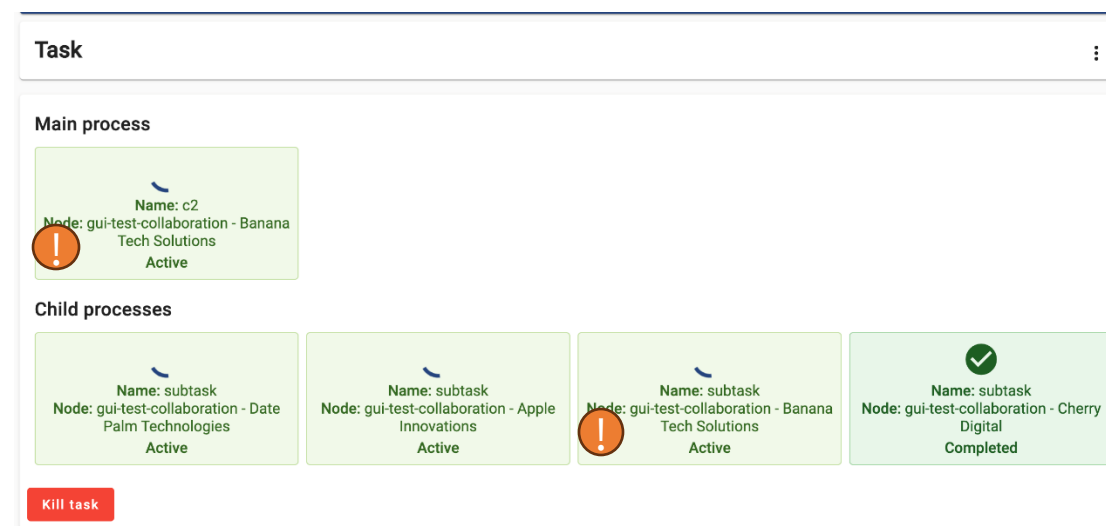
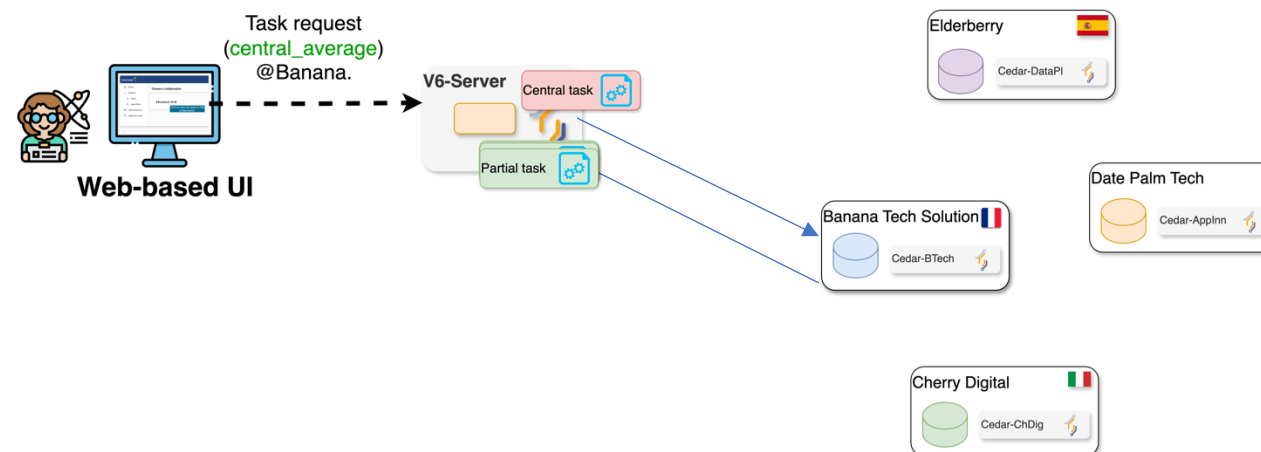
 **HINT**

Getting familiar with the UI

⚡ SOLUTION

The **Task** with the **central_average** function, when executed by a node, requests **all** the nodes within the **Study** to run the **partial_average** one. This particular node is also part of the study, so in the end it gets his own request!

The **central_average** was designed to just consolidate the results of the **partial_average**, sent by the other nodes. Hence, it doesn't need direct access any dataset.



Double-checking core concepts...

FUNCTION **partial_average**(dataset, column_name)

```
1 INFO("Extracting column", column_name)
2 column_values = dataset[column_name]
3
4 INFO("Computing partials")
5 local_sum = SUM(column_values)
6 local_count = SIZE(column_values)
7
8 RETURN {"sum": local_sum, "count": local_count}
```

END FUNCTION

- Why does the **central_average** function, unlike **partial_average**, not get any dataset as an input?
- Why does the **partial_average** function, unlike **central_average**, not get any client as an input?

FUNCTION **central_average**(client, column_name)

```
1 study_orgs = client.study.list()
2
3
4 INFO("Requesting partial computation")
5 task = client.task.create(
6     function = "partial_average"
7     column = column_name
8     organizations = study_orgs
9 )
10
11 INFO("Waiting for results")
12 results = client.wait_for_results(task)
13
14 global_sum = 0
15 global_count = 0
16
17 FOR EACH output IN results
18     global_sum = global_sum + output.sum
19     global_count = global_count + output.count
20
21 average = global_sum/global_count
22
23 RETURN {"average": average}
```

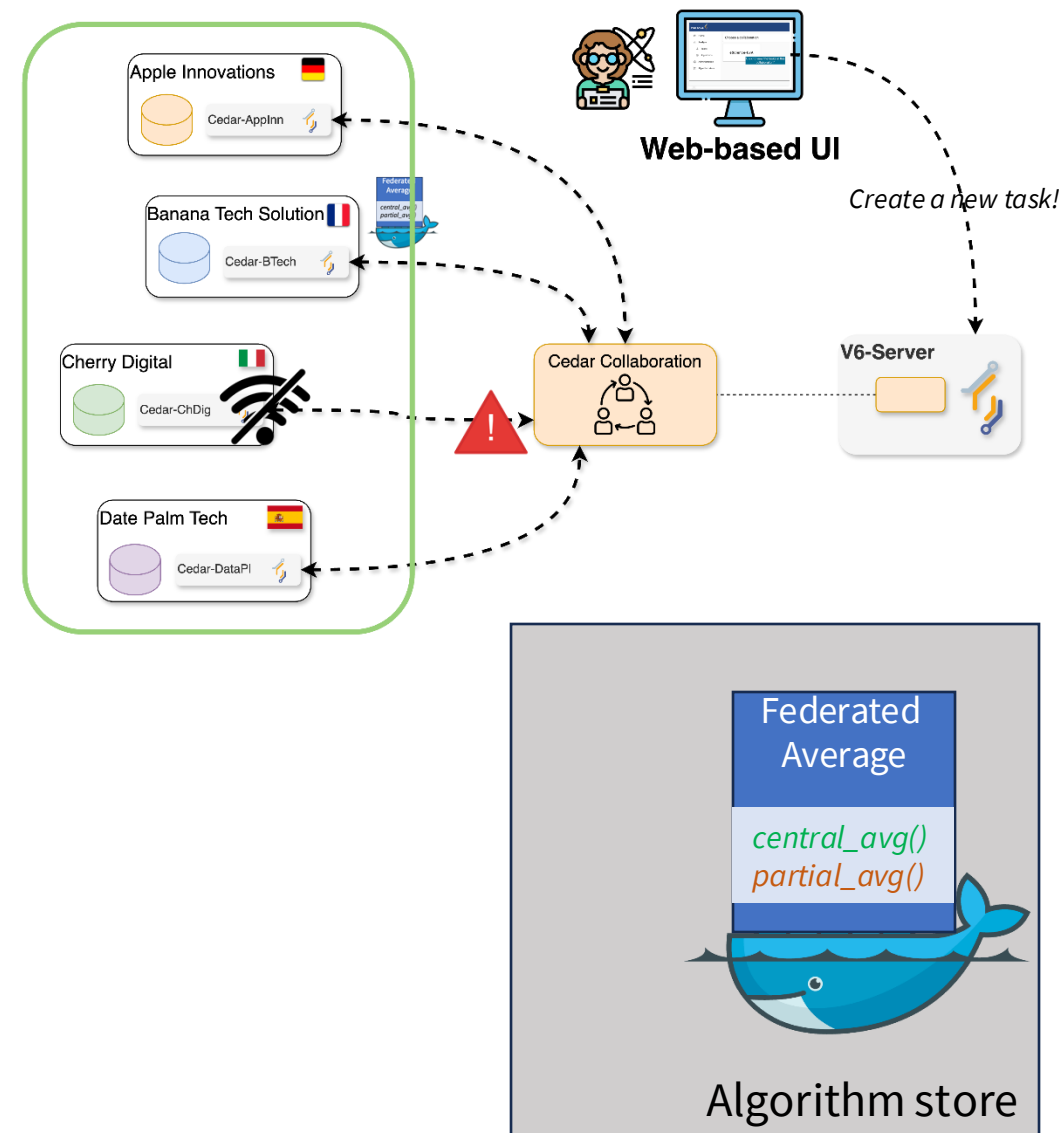
END FUNCTION

Running a PET: federated analysis



Problems on a node?

- What if...
 - You run the **central_average** function (**Federated Average algorithm**), on the study that has an **offline** node?



Running a federated algorithm

⚡ CHALLENGE 5

handling problems through the UI!

- Use the study that has **an offline node**.
- Choose the **Average algorithm**.
- Select the **central_average** function and an **online** organizations within the study.
- Choose the *default* database.
- Choose a numerical variable as the input.

What happened with the task? Why?

FUNCTION **central_average**(**client**, column_name)

```

1  study_orgs = client.study.list()
2
3
4  INFO("Requesting partial computation")
5  task = client.task.create(
6      function = "partial_average"
7      column = column_name
8      organizations = study_orgs
9  )
10
11  INFO("Waiting for results")
12  results = client.wait_for_results(task)
13
14  global_sum = 0
15  global_count = 0
16
17  FOR EACH output IN results
18      global_sum = global_sum + output.sum
19      global_count = global_count + output.count
20
21  average = global_sum/global_count
22
23  RETURN {"average": average}

```

END FUNCTION

⚡ HINT

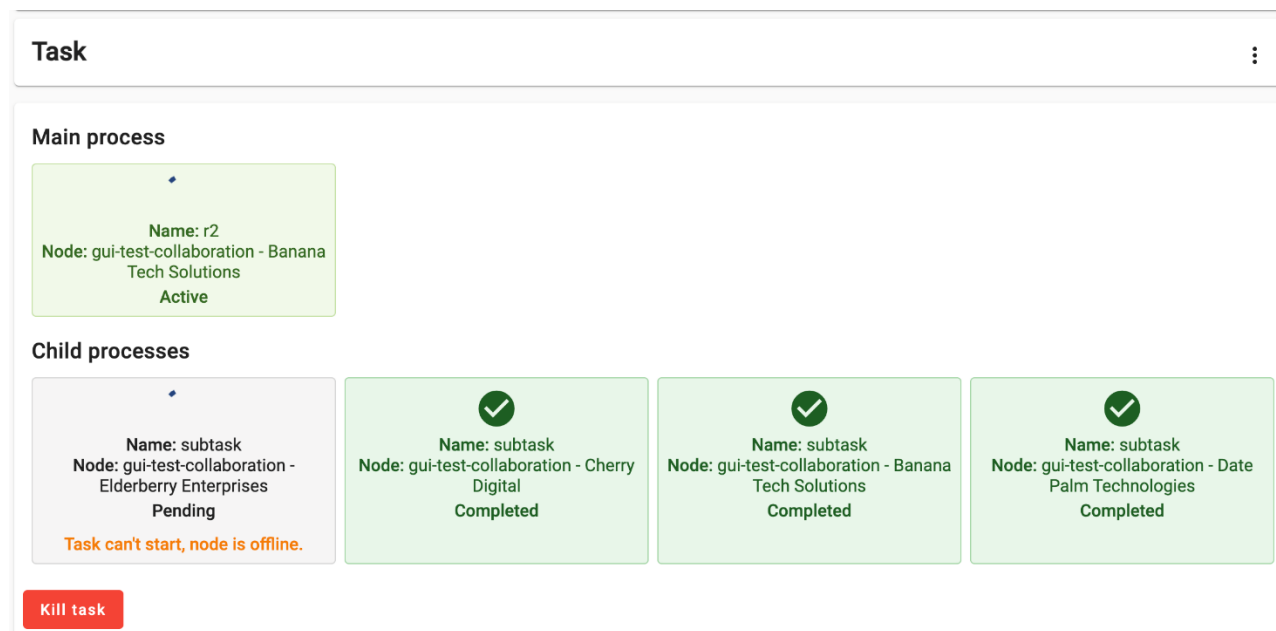
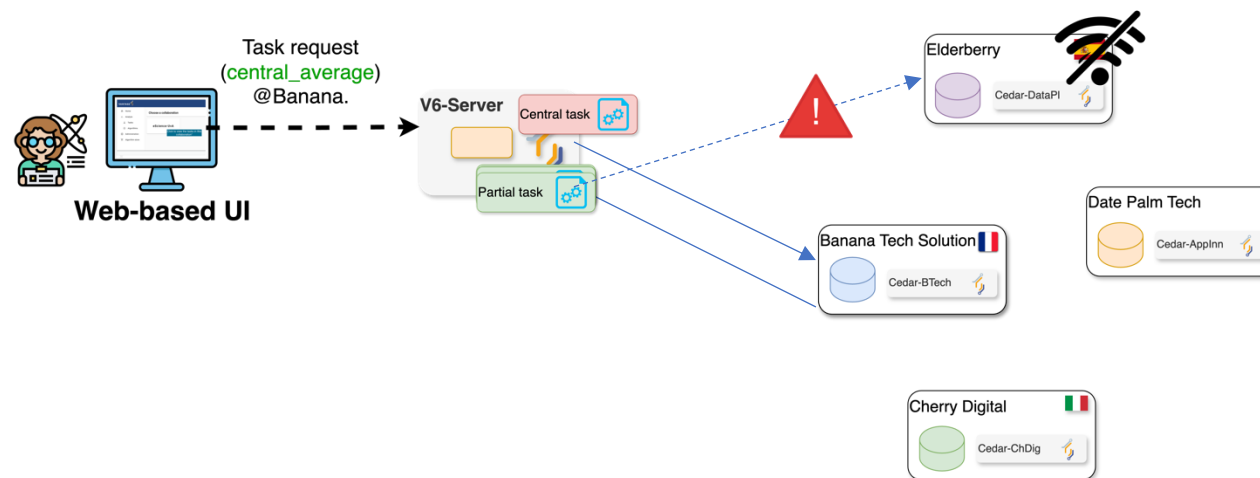
Getting familiar with the UI

⚡ SOLUTION

The algorithm didn't crash.

The Central task requests all the nodes in the study to run the 'partial' function. As the server is unable to transfer this request to the offline node, this child process is kept on hold, until the node is back online.

Consequently, the Main process is also kept on hold, and the process stays with an 'Active' status indefinitely (or until the node is back online).

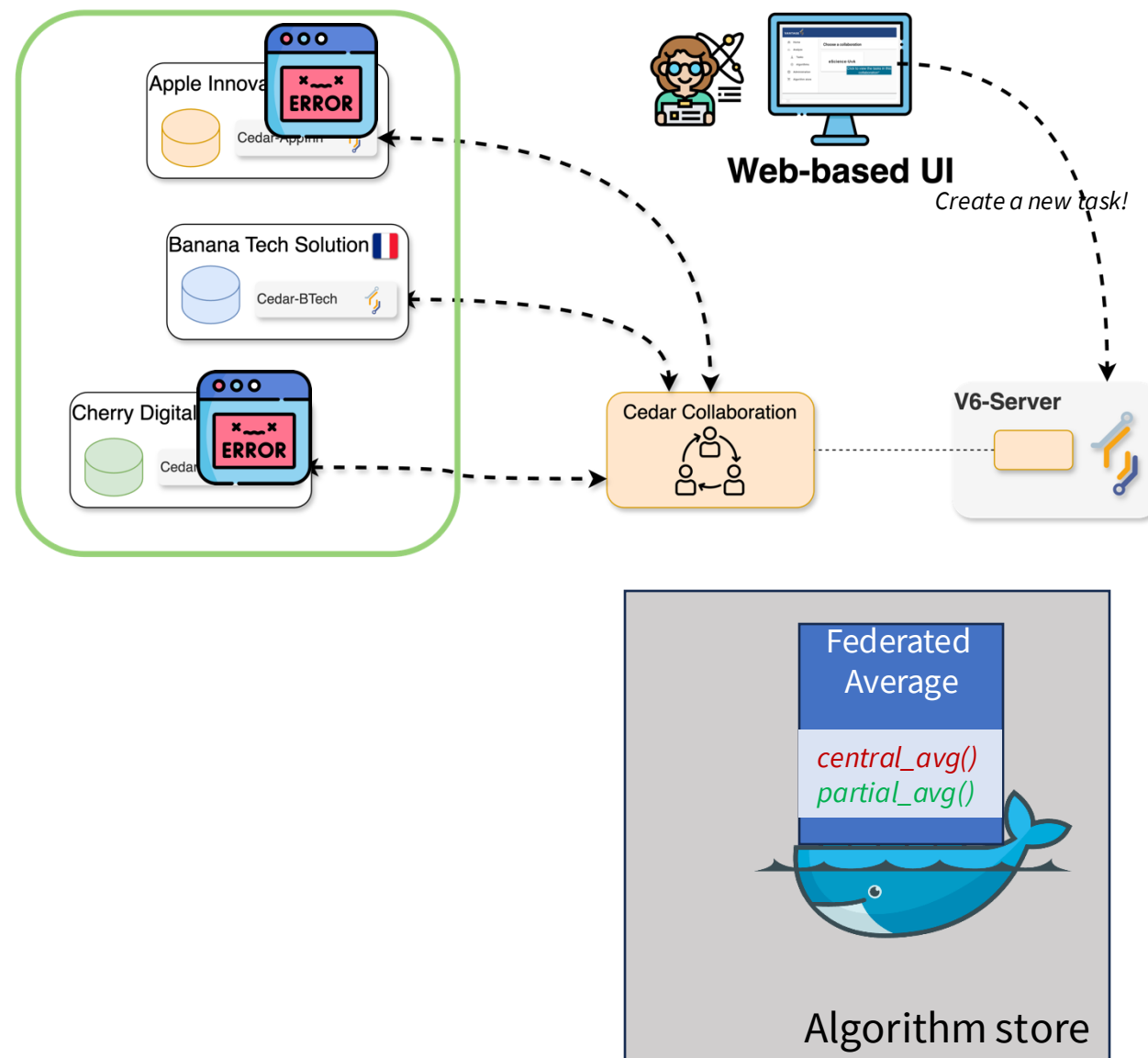


Running a PET: federated analysis



Problems at the code level?

- What if...
 - You run the **central_average** function (**Federated Average algorithm**), and later on a node gets a runtime error? (e.g., a code bug, unexpected input, etc.)



Running a federated algorithm

⚡ CHALLENGE 6

Handling problems through the UI!

- Use the study that has **no** offline nodes.
- Choose the **Average algorithm**.
- Select the **central_average** function and one of the organizations within the study.
- Choose the default database.
- Choose the **ageGroup** variable as the input.

Look at the error logs. What was the cause of the failure of the **partial_average**? What about the **central_average**?

gender	age	height	weight	isOverweight	ageGroup
M	39	152	108	False	30 - 40
M	8	118	106	False	0 - 10
M	16	161	110	True	10 - 20
M	94	110	115	True	90 - 100
M	47	117	152	True	40 - 50
F	29	127	110	True	20 - 30
M	5	95	65	False	0 - 10
M	39	142	196	False	30 - 40
F	20	189	112	False	20 - 30
F	84	145	116	False	80 - 90

⚡ HINT

The error logs contain the error messages generated by the Python interpreter while running the algorithm functions. Look for messages that hint the root cause of the problem.

Running a federated algorithm

⚡ CHALLENGE 6

FUNCTION **partial_average**(dataset, column_name)

```
1 INFO("Extracting column", column_name)
2 column_values = dataset[column_name]
3
4 INFO("Computing partials")
5 local_sum = SUM(column_values)
6 local_count = SIZE(column_values)
7
8 RETURN {"sum": local_sum, "count": local_count}
```

END FUNCTION

Look at the error logs. What was the cause of the failure of the **partial_average**? What about the **central_average**?

⚡ HINT

FUNCTION **central_average**(client, column_name)

```
1 study_orgs = client.study.list()
2
3
4 INFO("Requesting partial computation")
5 task = client.task.create(
6     function = "partial_average"
7     column = column_name
8     organizations = study_orgs
9 )
10
11 INFO("Waiting for results")
12 results = client.wait_for_results(task)
13
14 global_sum = 0
15 global_count = 0
16
17 FOR EACH output IN results
18     global_sum = global_sum + output.sum
19     global_count = global_count + output.count
20
21 average = global_sum/global_count
22
23 RETURN {"average": average}
```

END FUNCTION

⚡ HINT

Getting familiar with the UI

⚡ SOLUTION

1. This time, the **partial_average** function on all the data nodes crashes, which lead to the failure of the Main process (**central_average**).
2. The **partial_average** crashed while trying to parse the input as a number.
The **central_average** crashed due to a division-by-zero.

Task

Main process

!

Name: asds
Node: gui-test-collaboration - Cherry Digital
Error during execution

Show logs

Child processes

!

Name: subtask
Node: gui-test-collaboration - Cherry Digital
Error during execution

Show logs

!

Name: subtask
Node: gui-test-collaboration - Date Palm Technologies
Error during execution

Show logs

!

Name: subtask
Node: gui-test-collaboration - Banana Tech Solutions
Error during execution

Show logs

!

Name: subtask
Node: gui-test-collaboration - Apple Innovations
Error during execution

Show logs

Play around with other algorithms

⚡ CHALLENGE 7

Try to run other (more advanced) algorithms!

2 Select algorithm

Select which algorithm you want to run

Algorithm* ▼

- Contingency table
- Kaplan-Meier
- Average
- Description

age	months	censor	days
56	31	1	926
58	56	0	1690
62	79	0	2374
44	52	0	1569
47	55	0	1646
62	34	1	1010
63	88	0	2646
68	70	0	2105
35	103	0	3075
64	94	0	2832

⚡ **HINT**
This time, choose the *kaplan_meier*
(csv) database!

