eosc | EuroScienceGateway

EuroScienceGateway

Leveraging the European compute infrastructures for data-intensive research guided by FAIR principles

Funded by the European Union
Grant agreement 101057388

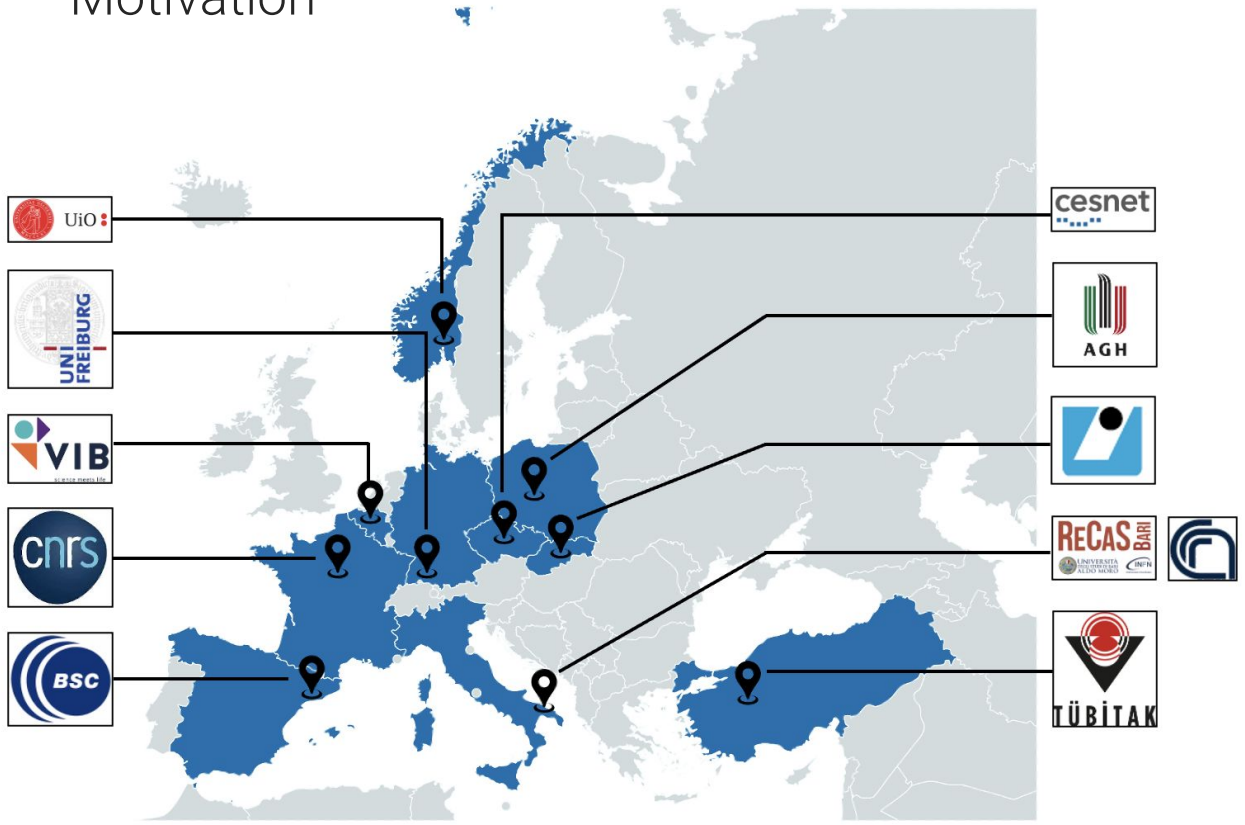
2024-10-02 by EuroScienceGateway

Overview

- The EuroScienceGateway project
- Broadening the login options for EuroScienceGateway
- BYOC: Bring Your Own Compute
- BYOS: Bring Your Own Storage
- Smart job scheduling across Europe

The EuroScienceGateway project

Motivation



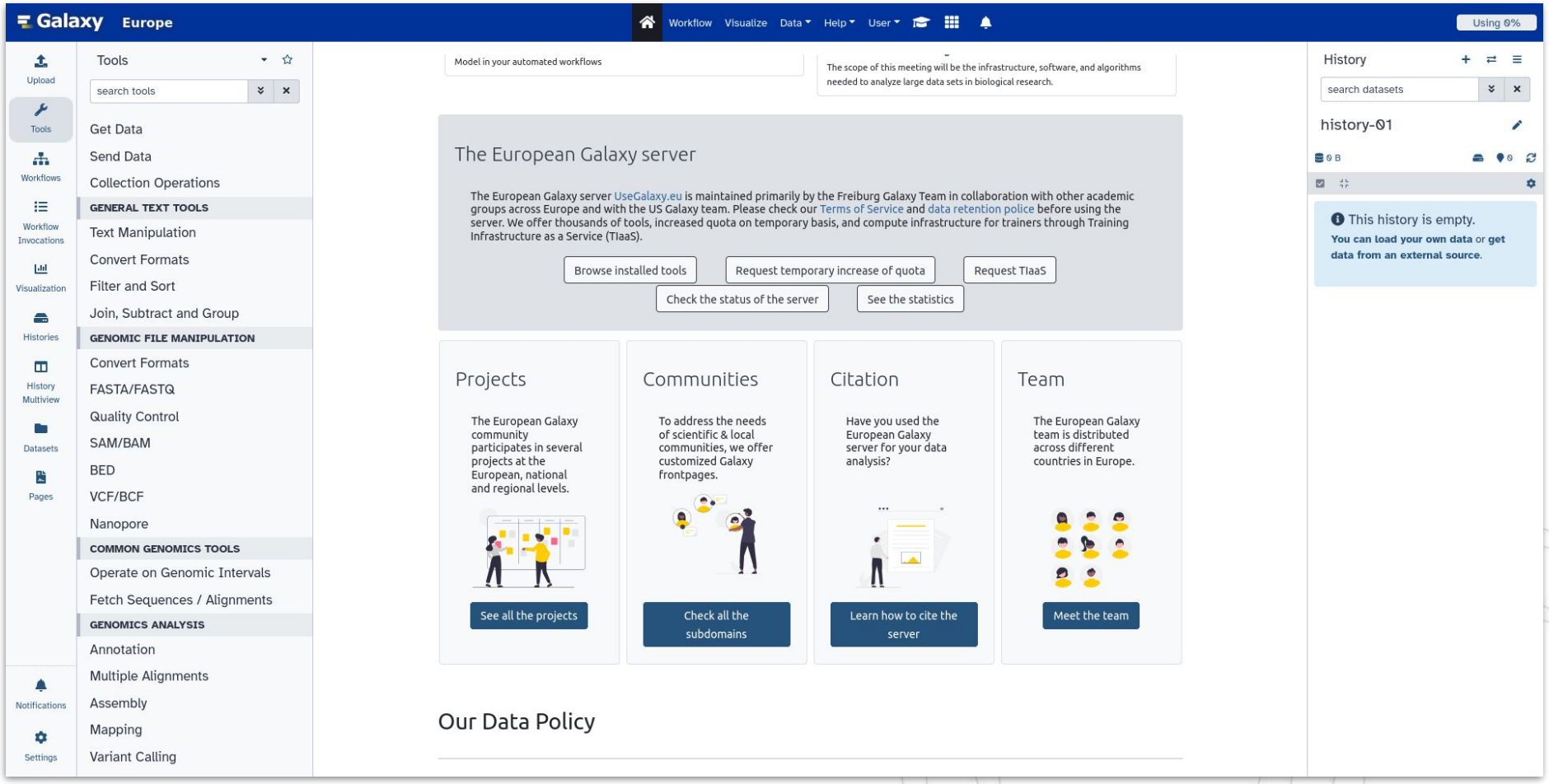
National Cloud and HPC infrastructures have been established, with differences in

- Hardware
- Configuration
- Software stack
- Authentication and Authorization
- Access typically targeted at local researchers

goal: provide efficient and structured access to data, tools and workflows supported by suitable IT infrastructures.

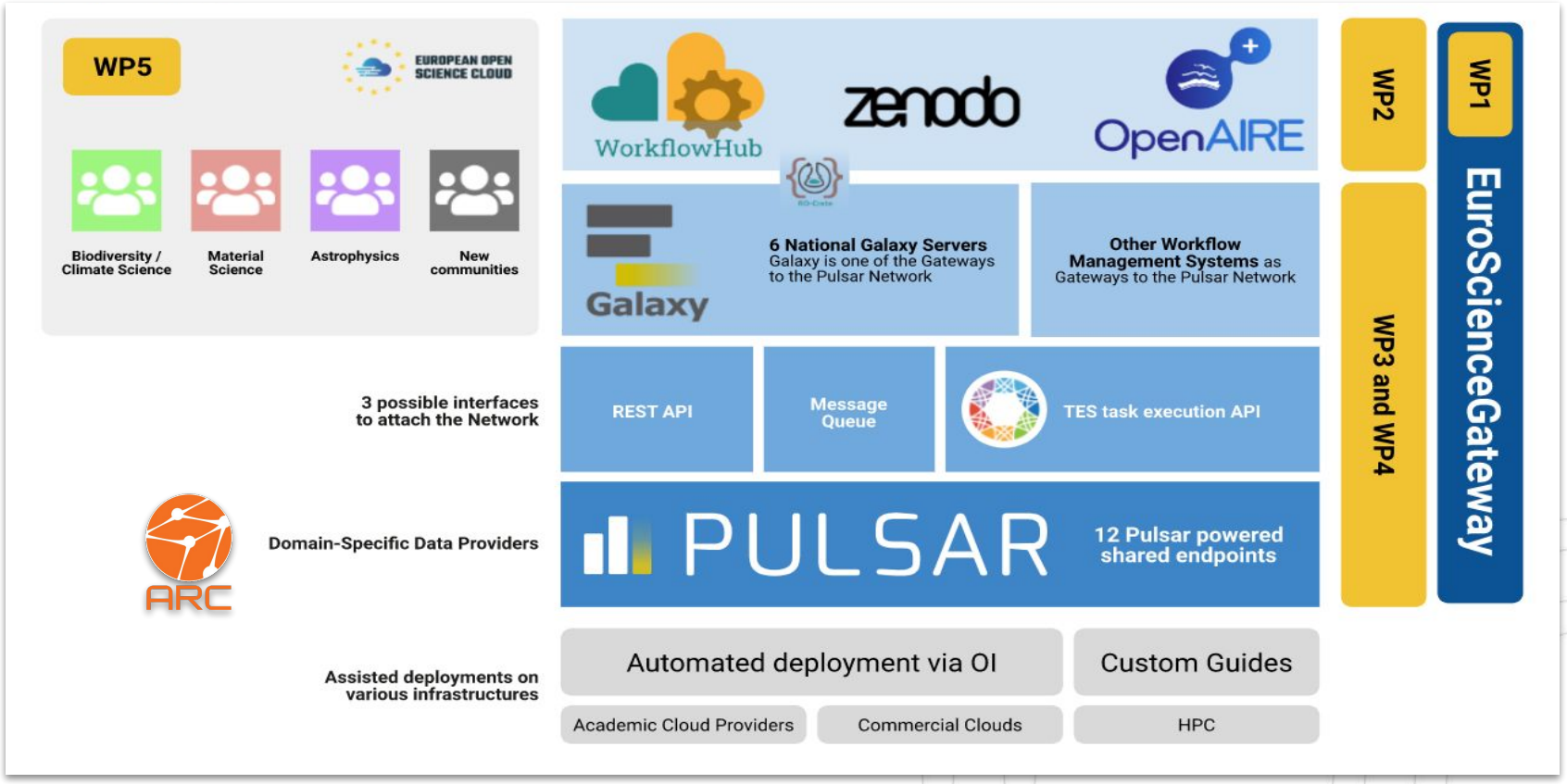
The EuroScienceGateway project

How?



The EuroScienceGateway project

How?



Partners


VIB
MANCHESTER 1824
The University of Manchester
Université de Paris
UNIVERSITÀ DEGLI STUDI FIRENZE
BIO
UKRI
Science and Technology Facilities Council
UNIVERSITAT DE BARCELONA
UNI FREIBURG
TÜBITAK
ULAKBİM
UNIVERSITAS OSLOENSIS
MDCCCXI
AGH
AGH University of Science and Technology
INFN
Istituto Nazionale di Fisica Nucleare
EPFL
cnrs
Consiglio Nazionale delle Ricerche
cesnet
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Institute of Informatics
Slovak Academy of Sciences
egi
CYFRONET

Overview

- The EuroScienceGateway project
- **Broadening the login options for EuroScienceGateway**
- BYOC: Bring Your Own Compute
- BYOS: Bring Your Own Storage
- Smart job scheduling across Europe

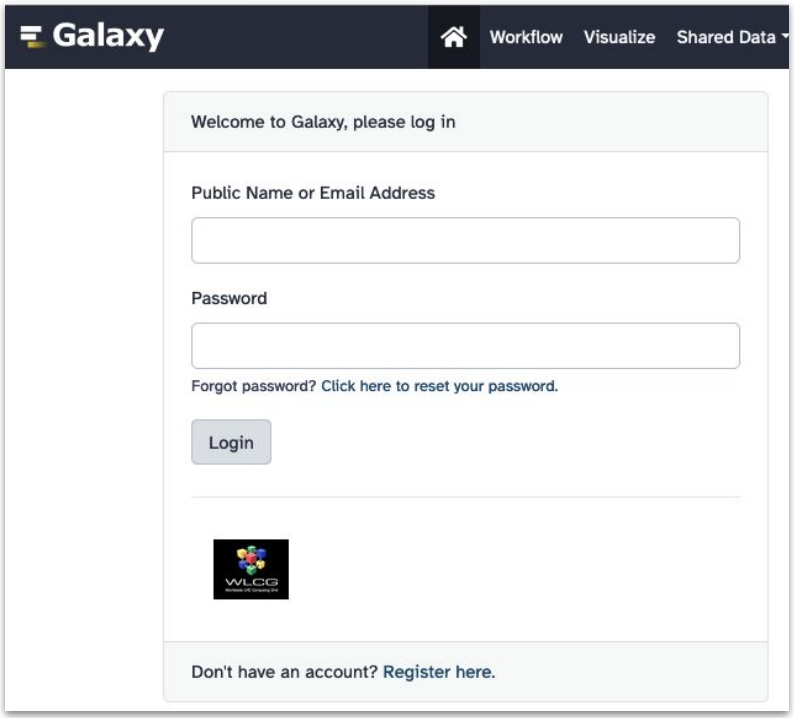
Broadening the login options for EuroScienceGateway

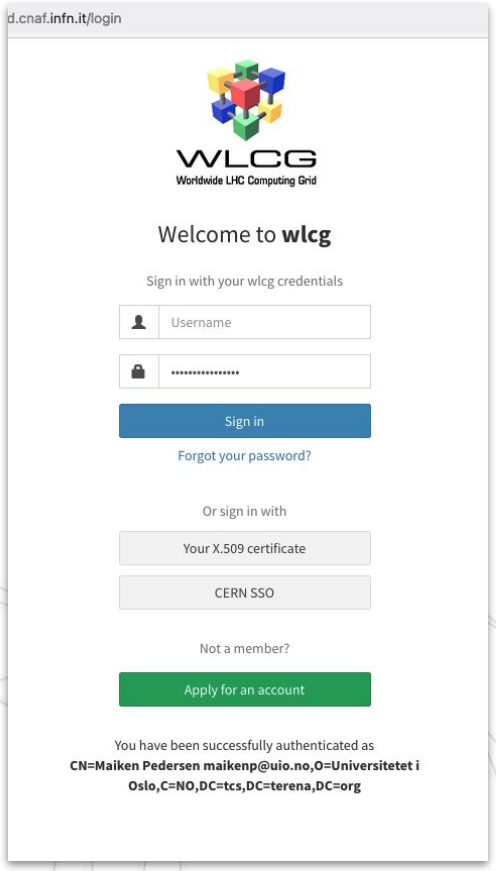
- Integration with WLCG IAM
- Integration with EGI Check-in

Broadening the login options for EuroScienceGateway

WLCG IAM: AAI solution for the Worldwide LHC Computing Grid

- WLCG IAM can be added to Galaxy thanks to [python-social-auth]
- Getting a token from WLCG IAM is one option in order to be allowed to submit jobs to ARC sites - IAM-services to trust are configurable in ARC

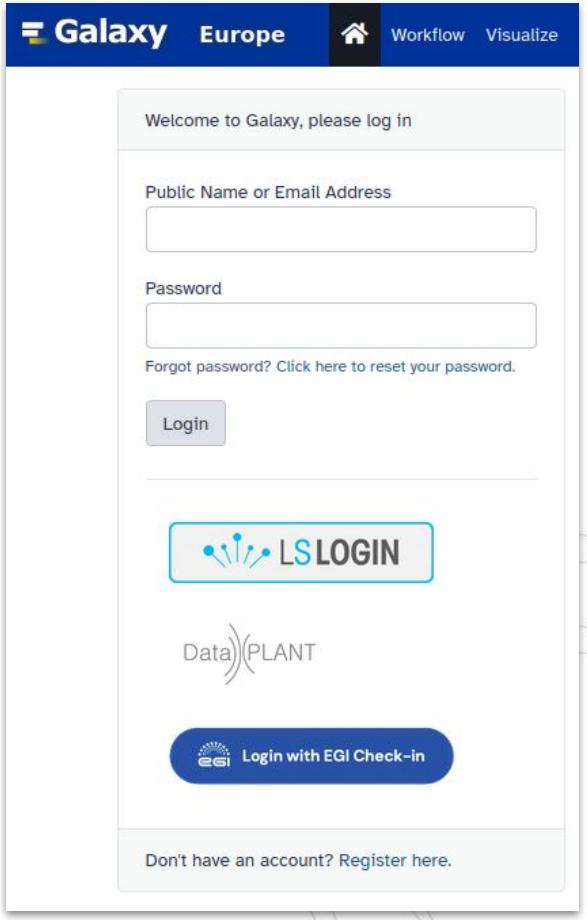


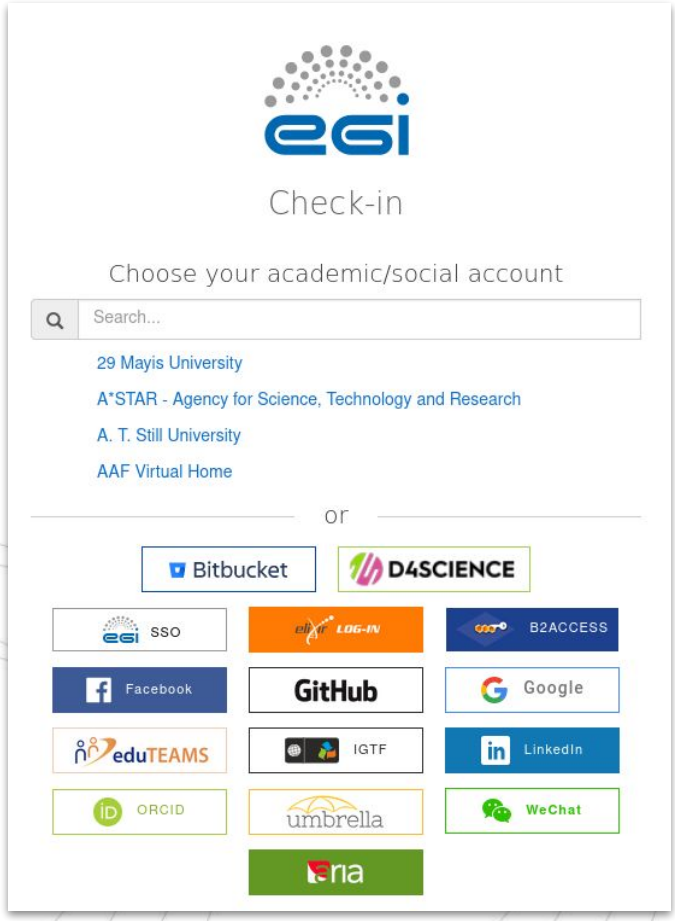


Broadening the login options for EuroScienceGateway

EGI Check-in: AAI solution for the EGI Federated Cloud

- Use your existing credentials to log into Galaxy
- With a token from EGI Check-in you can deploy compute and storage resources in the EGI Federated Cloud and connect them with Galaxy (more details later)
  - Can also be used to submit jobs to ARC





Overview

- The EuroScienceGateway project
- Broadening the login options for EuroScienceGateway
- **BYOC: Bring Your Own Compute**
- BYOS: Bring Your Own Storage
- Smart job scheduling across Europe

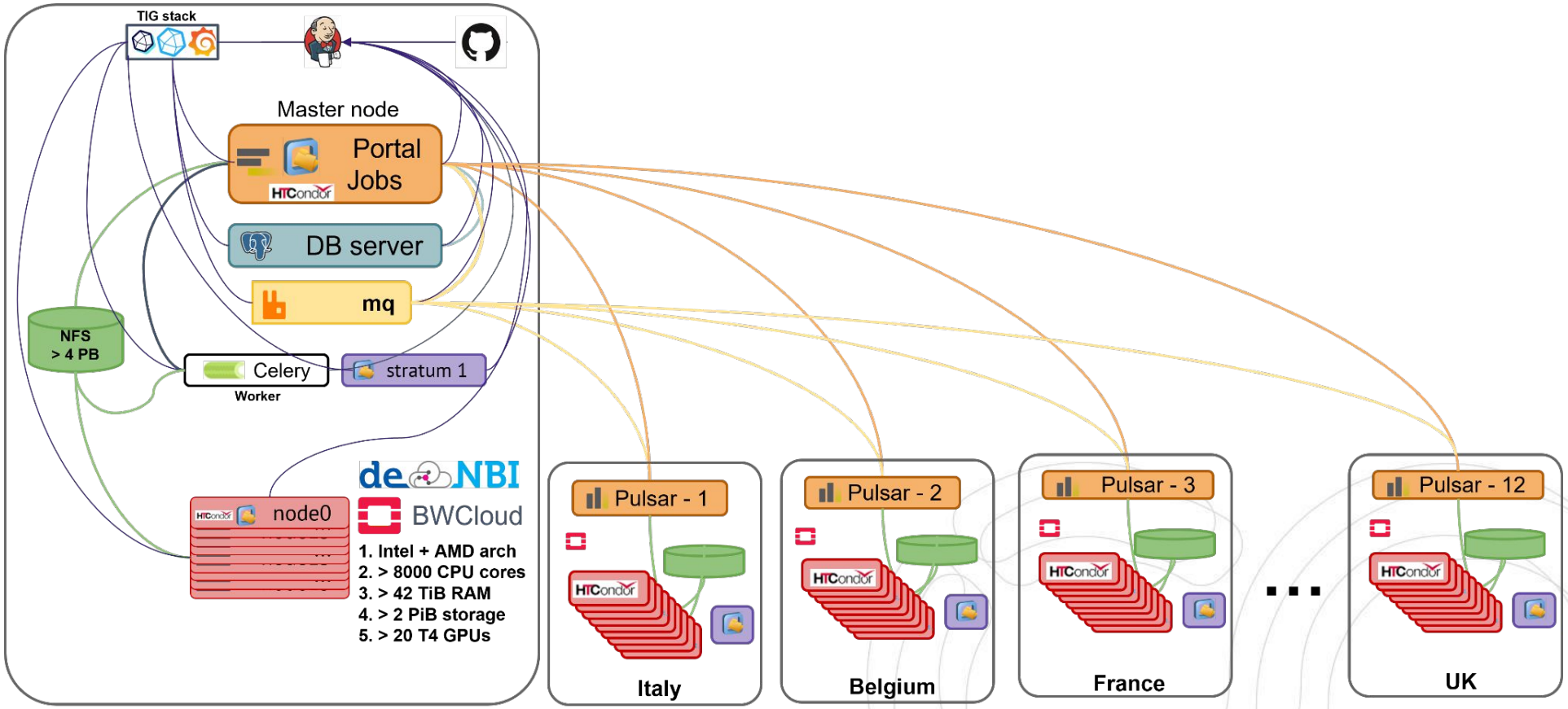
Bring Your Own Compute

Connecting Galaxy with Pulsar

- Overview of Galaxy and Pulsar deployments in EU

Bring Your Own Compute

Connecting Galaxy with Pulsar



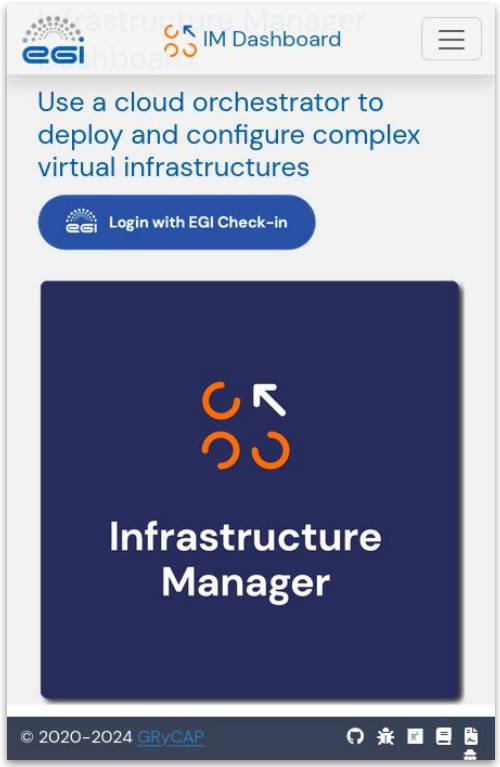
Bring Your Own Compute

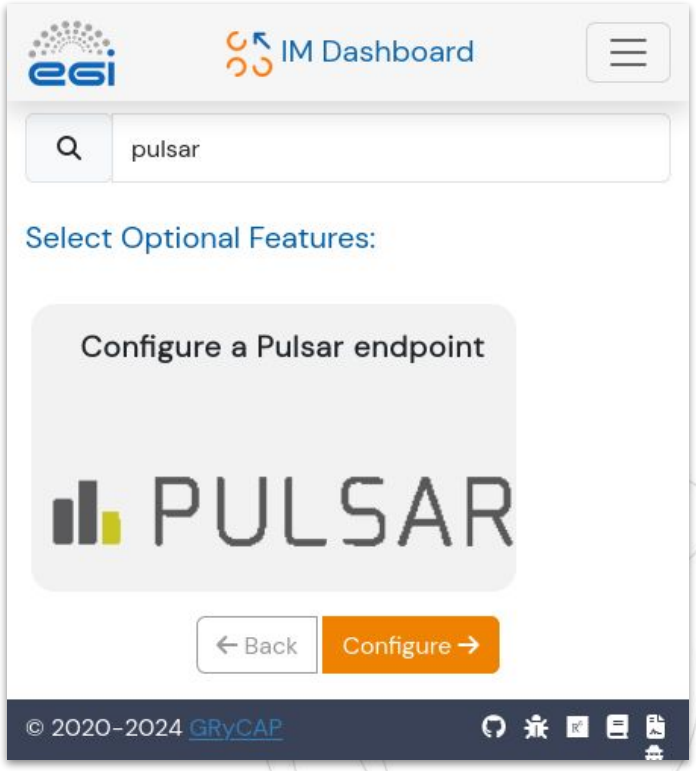
- Connect Galaxy with computing resources in the cloud
  - Connecting Galaxy with a managed ARC site
  - ARC deployment in the cloud with Infrastructure Manager
  - Pulsar deployment in the cloud with Infrastructure Manager
  - work by INFN to automate the connection of pulsar with Galaxy

Bring Your Own Compute

Pulsar deployment in the cloud with Infrastructure Manager

- Do have access to a cloud? Use Infrastructure Manager (https://im.egi.eu/) to deploy Pulsar and connect it to Galaxy

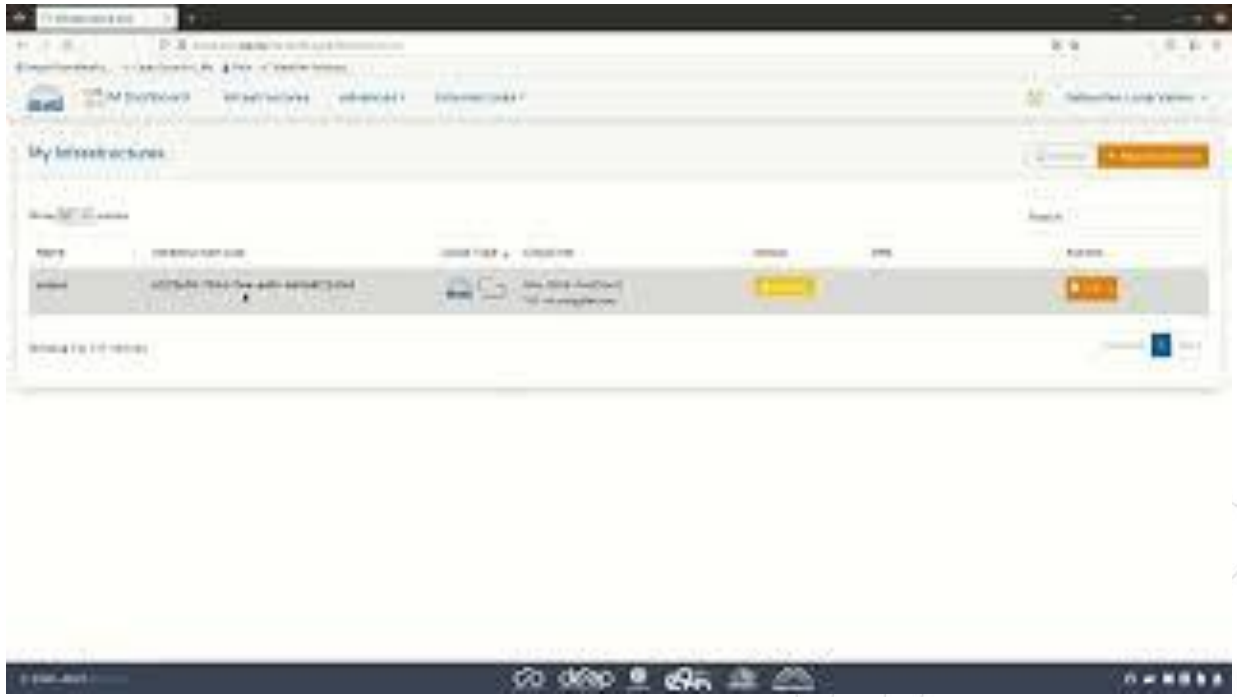




Bring Your Own Compute

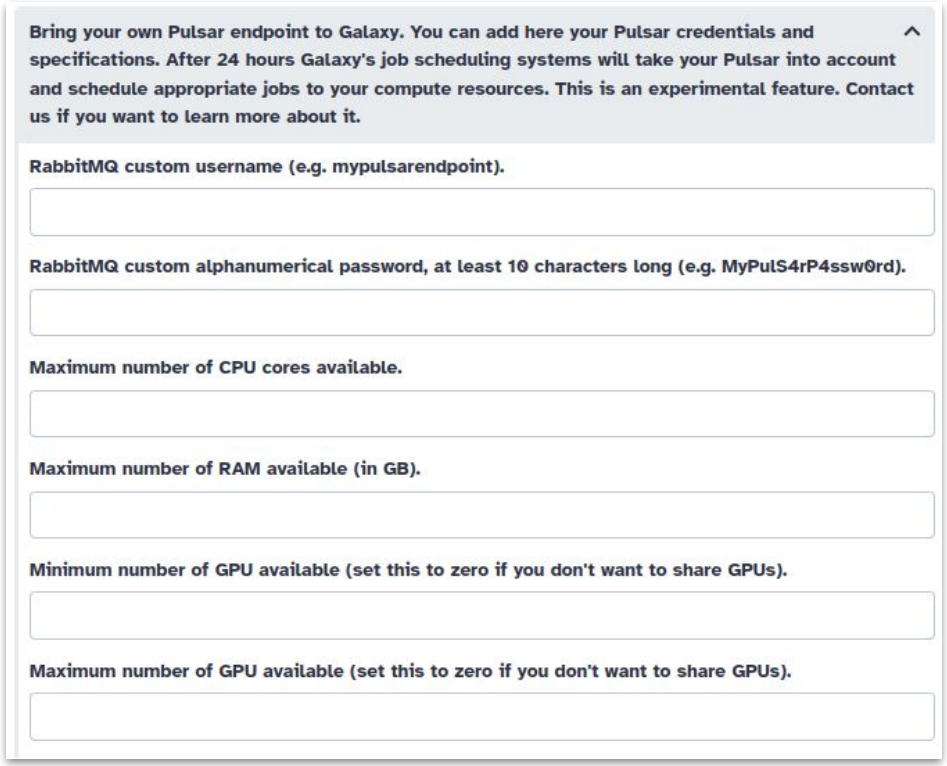
Pulsar deployment in the cloud with Infrastructure Manager

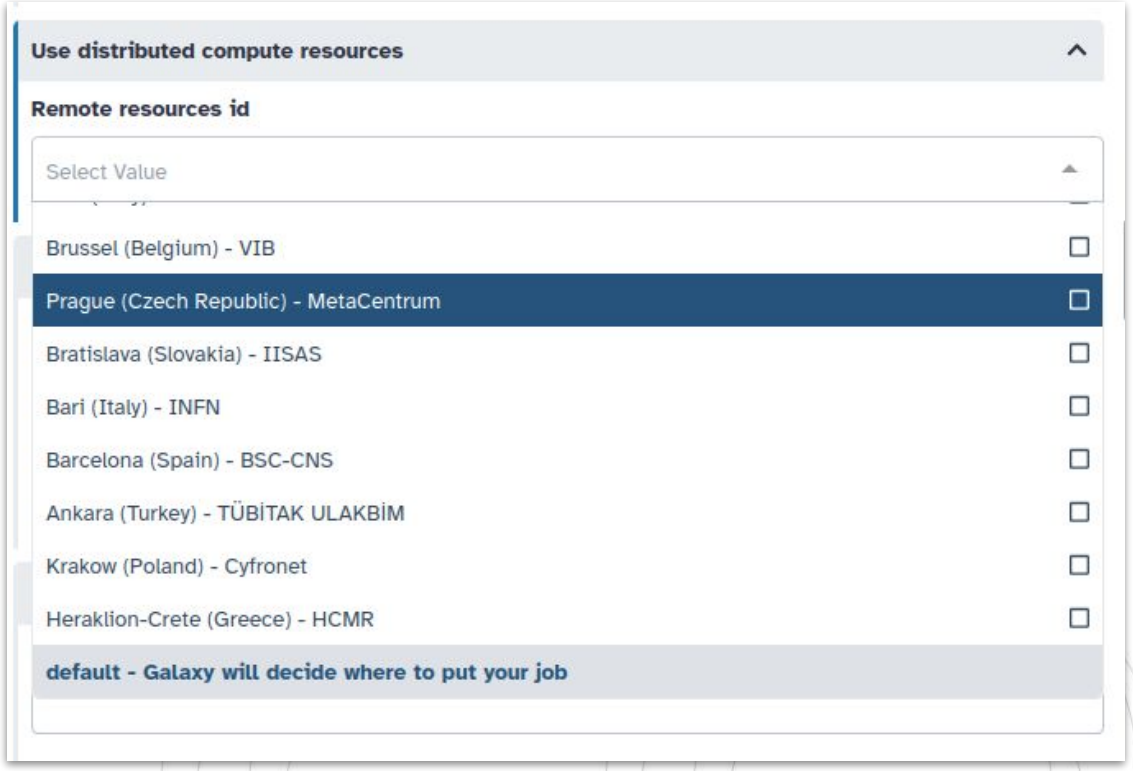
- See tutorial in https://galaxyproject.org/news/2023-10-31-esg-byoc-im/ and come to the booth to try it live!



Bring Your Own Compute

Connecting Galaxy with Pulsar



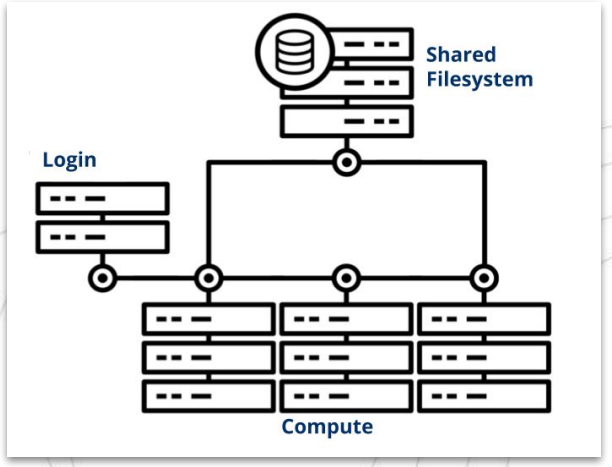


Bring Your Own Compute

Connecting Galaxy with a managed ARC site - testing only, not yet in production

- Do you have access to a managed cluster with ARC? Log into Galaxy with WLCG IAM or EGI Check-in and configure the ARC endpoint in Galaxy

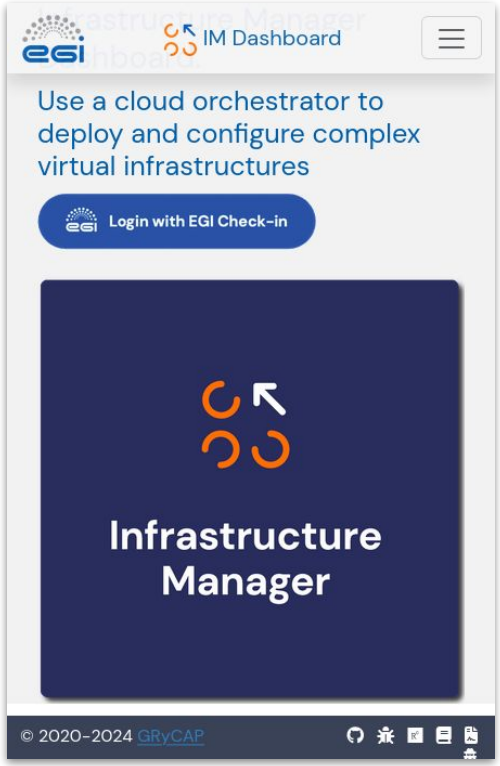
- Galaxy can execute jobs in ARC sites thanks to the prototype job runner [developed within the project]

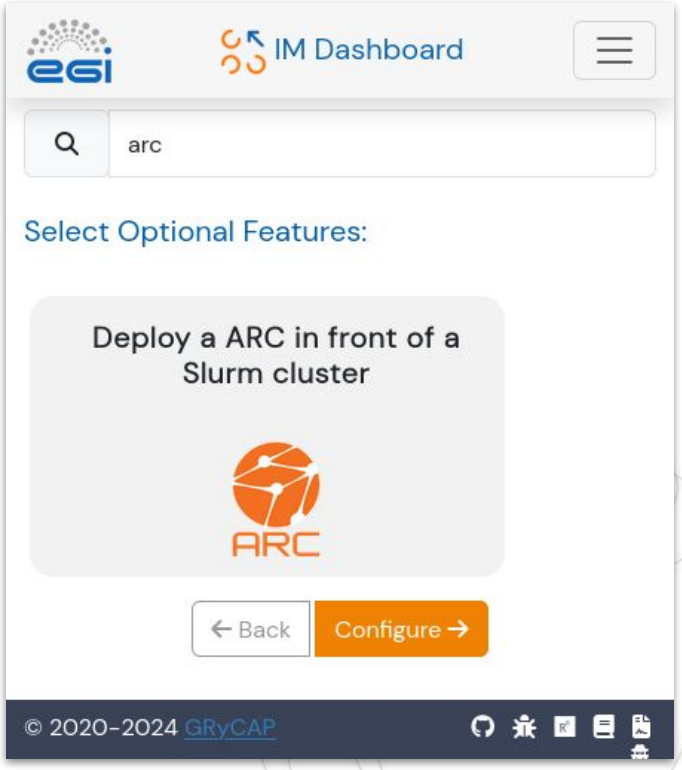


Bring Your Own Compute

ARC deployment in the cloud with Infrastructure Manager

- Do have access to a cloud? Use Infrastructure Manager (https://im.egi.eu/) to deploy ARC and connect it to Galaxy





Overview

- The EuroScienceGateway project
- Broadening the login options for EuroScienceGateway
- BYOC: Bring Your Own Compute
- **BYOS: Bring Your Own Storage**
- Smart job scheduling across Europe

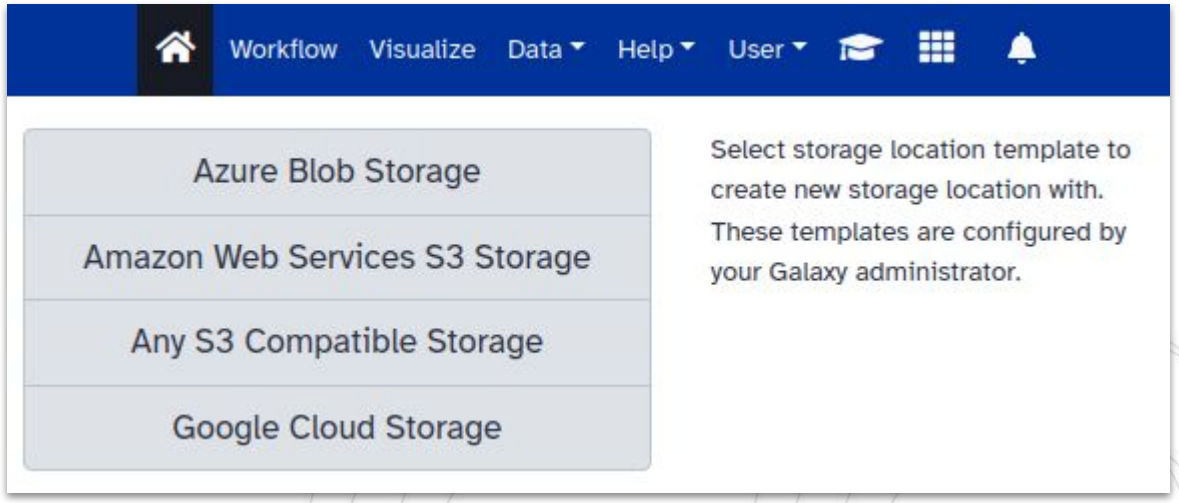
Bring Your Own Storage

Connect Galaxy with storage resources in the cloud

- Do you have access to object storage? Here is how to connect it to Galaxy
  - Tutorial: https://galaxyproject.org/news/2024-09-20-esg-byos-im/

- Select object storage type and provide
  - Name of the bucket / Container
  - Access (ID) key / Storage account
  - Secret key / Account key

- “Any S3 Compatible Storage” will also require
  - URL to the API endpoint



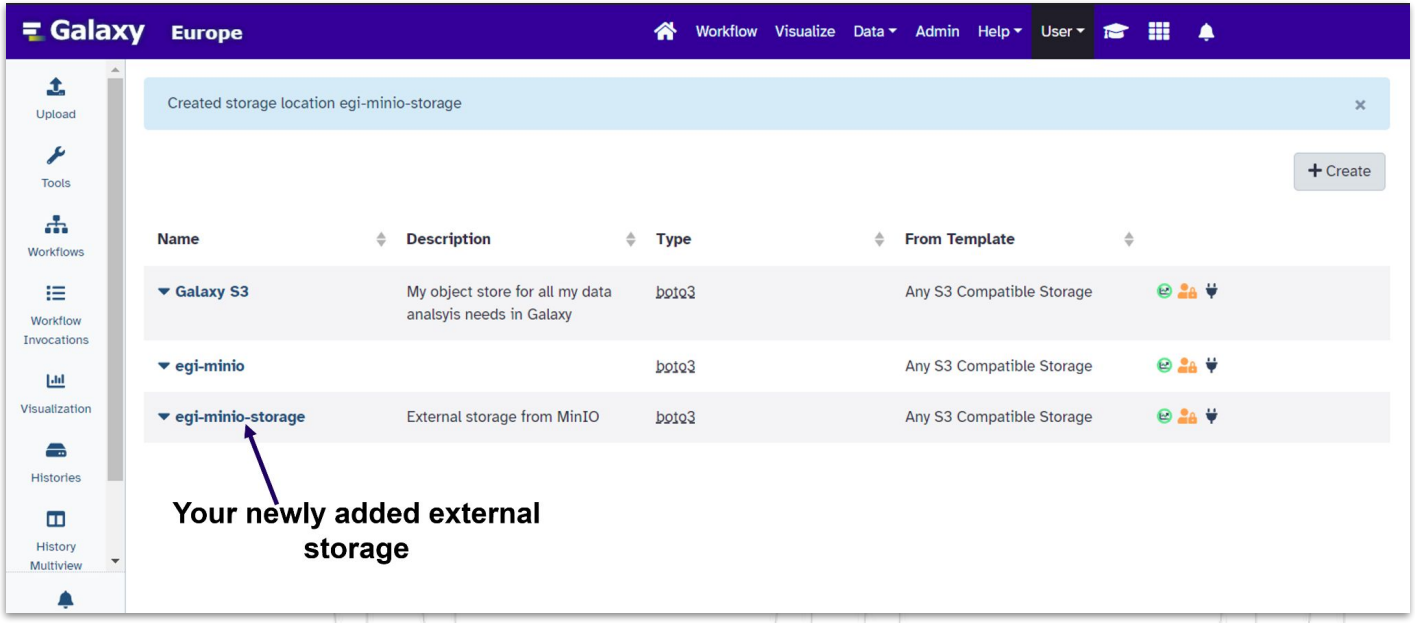
Bring Your Own Storage

Connect Galaxy with storage resources in the cloud

- Do you have access to object storage? Here is how to connect it to Galaxy
  - Tutorial: https://galaxyproject.org/news/2024-09-20-esg-byos-im/

- Select object storage type and provide
  - Name of the bucket / Container
  - Access (ID) key / Storage account
  - Secret key / Account key

- “Any S3 Compatible Storage” will also require
  - URL to the API endpoint

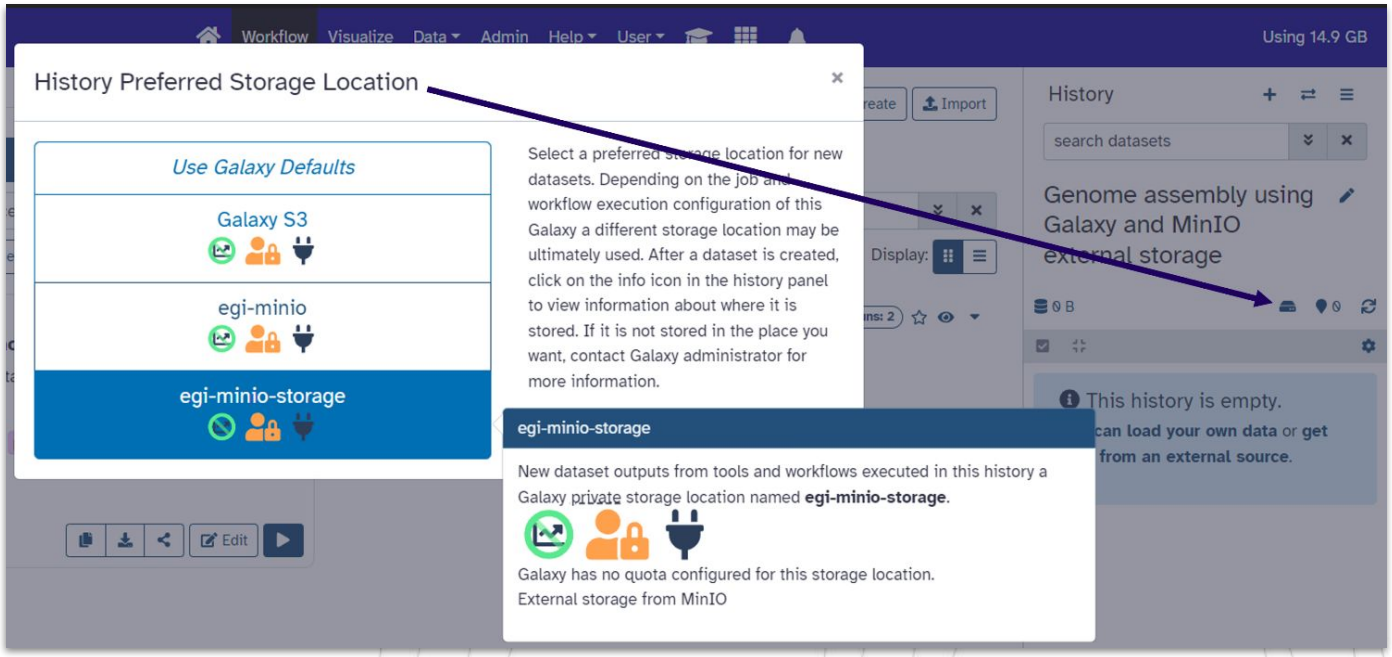


Bring Your Own Storage

Connect Galaxy with storage resources in the cloud

- Do you have access to object storage? Here is how to connect it to Galaxy and a tutorial is available here.

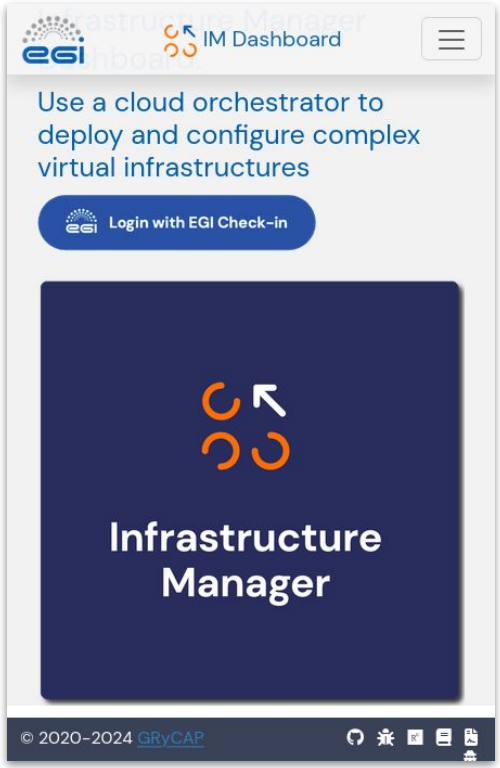
- The new storage can be used
  - per `History`
  - per `Tool`
  - per `Workflow`
  - or `Default` (for everything)

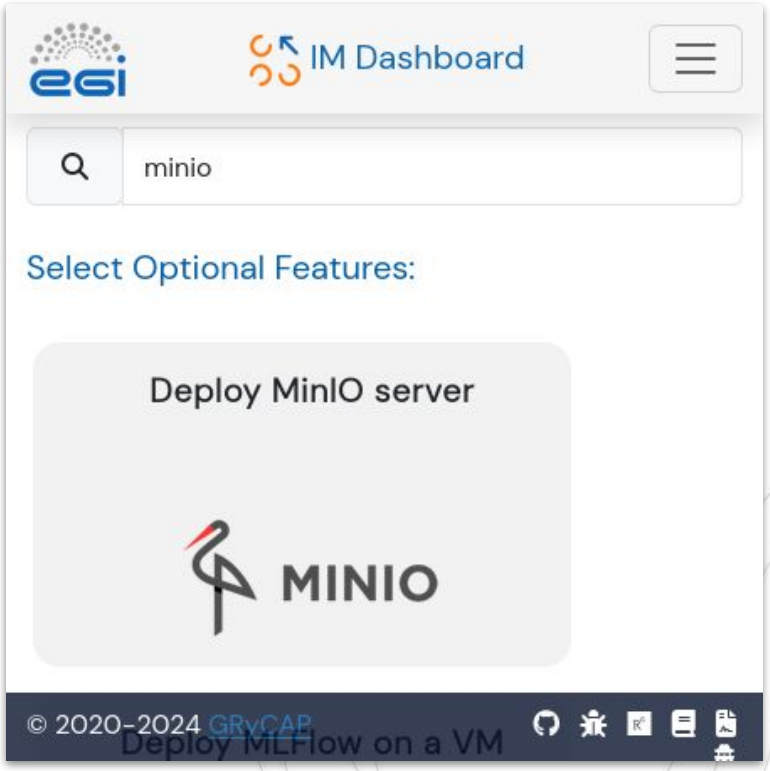


Bring Your Own Storage

Connect Galaxy with storage resources in the EGI Federated Cloud

- Use Infrastructure Manager (https://im.egi.eu/) to deploy MinIO and connect it to Galaxy (see tutorial)

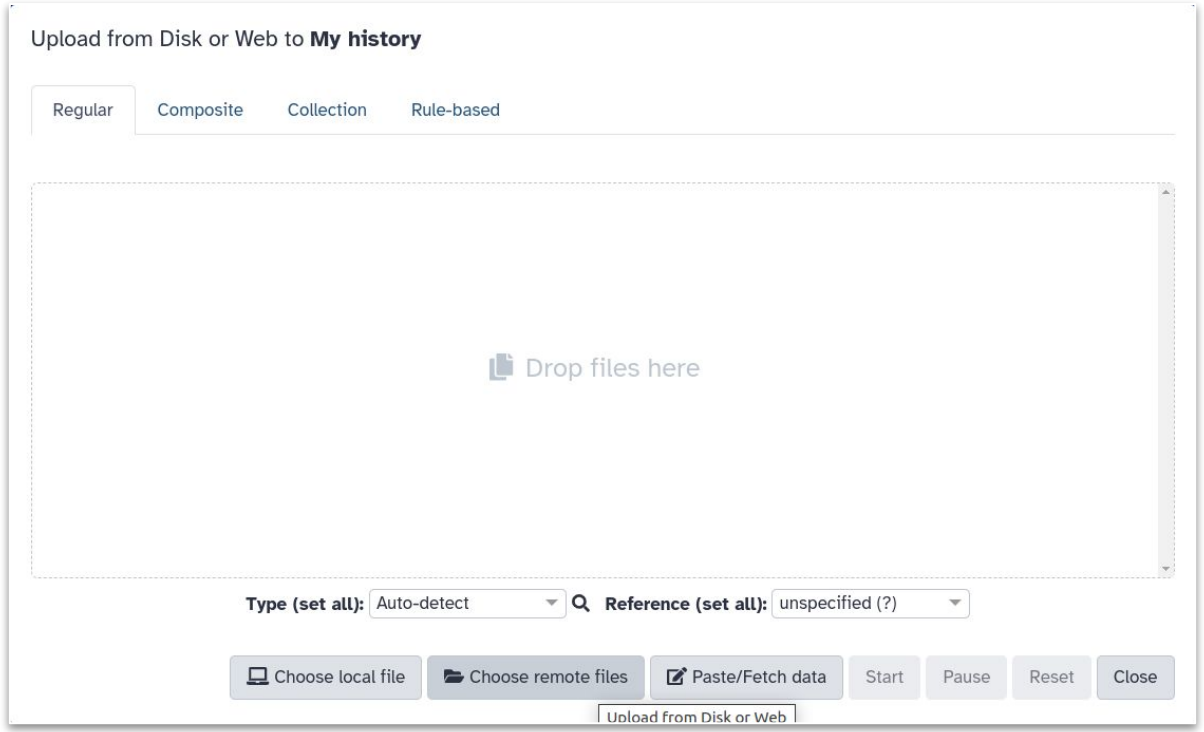


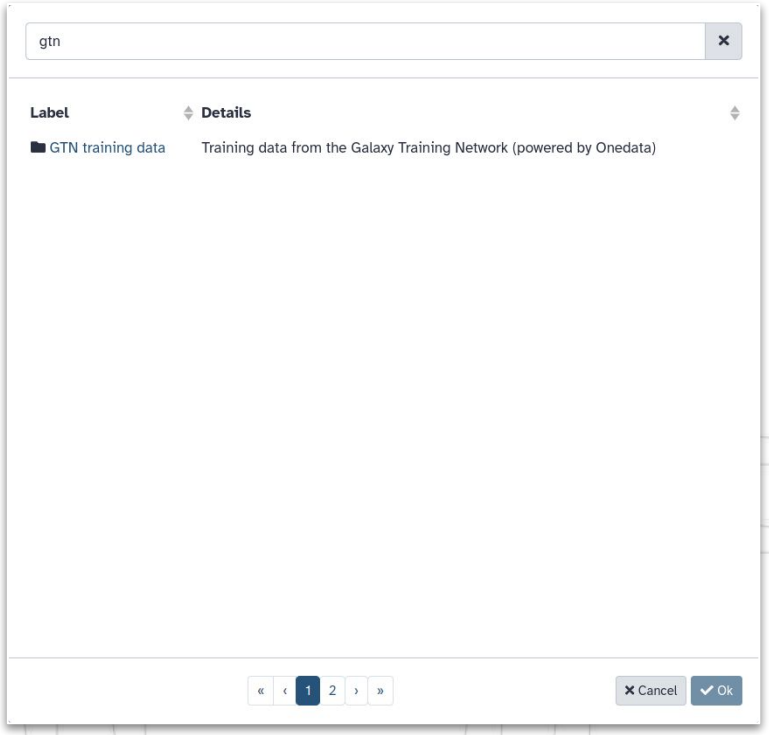


Bring Your Own Storage

Connect Galaxy with storage resources in EGI DataHub

1. GTN training data is now hosted in EGI DataHub and available on *usegalaxy.eu* for anyone to import.

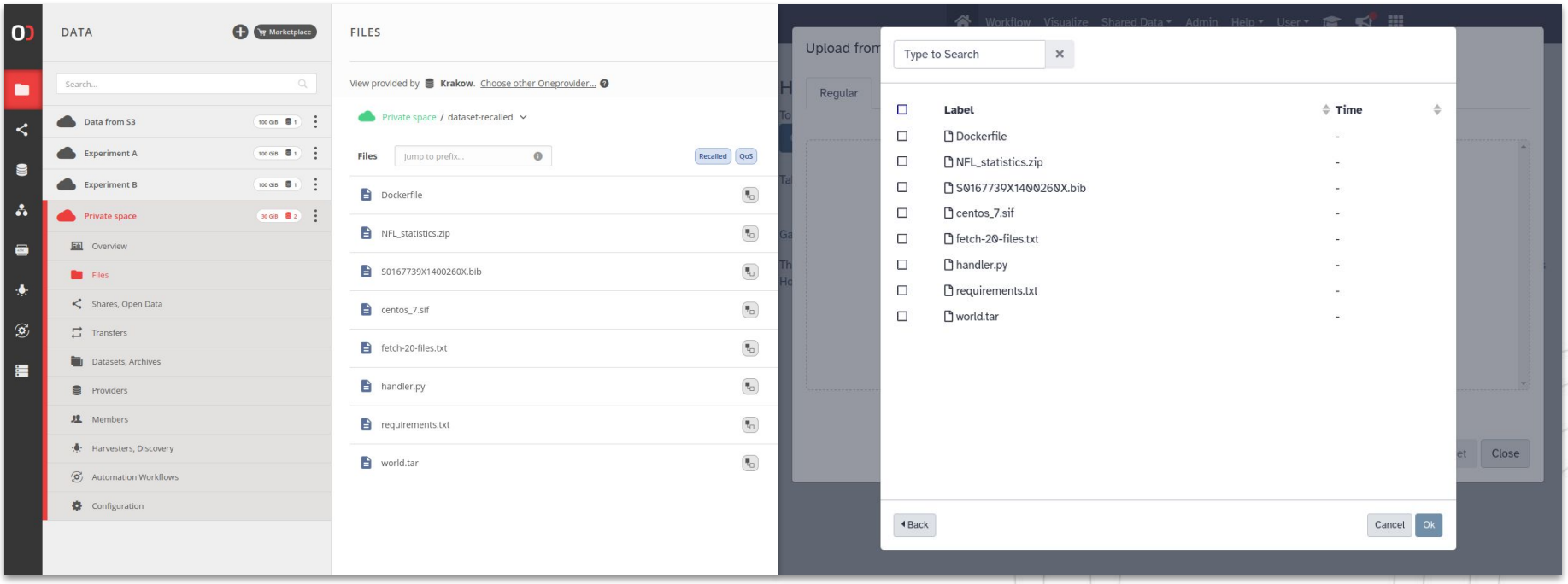




Bring Your Own Storage

Connect Galaxy with storage resources in EGI DataHub *or any Onedata environment*

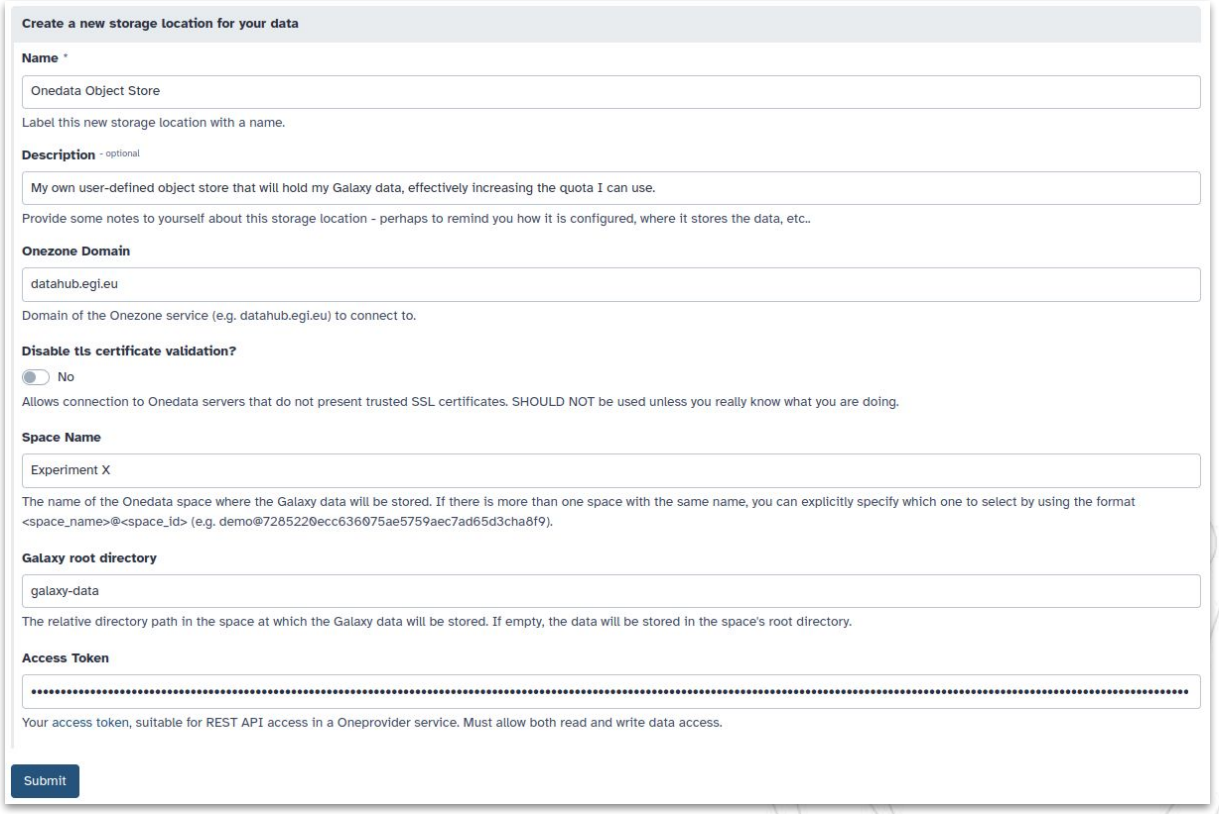
2. Configure personal Onedata files source credentials to import or export your datasets.



Bring Your Own Storage

Connect Galaxy with storage resources in EGI DataHub *or any Onedata environment*

3. Configure a personal Onedata storage location for Galaxy user data (Object Store) — available in an upcoming release.



Overview

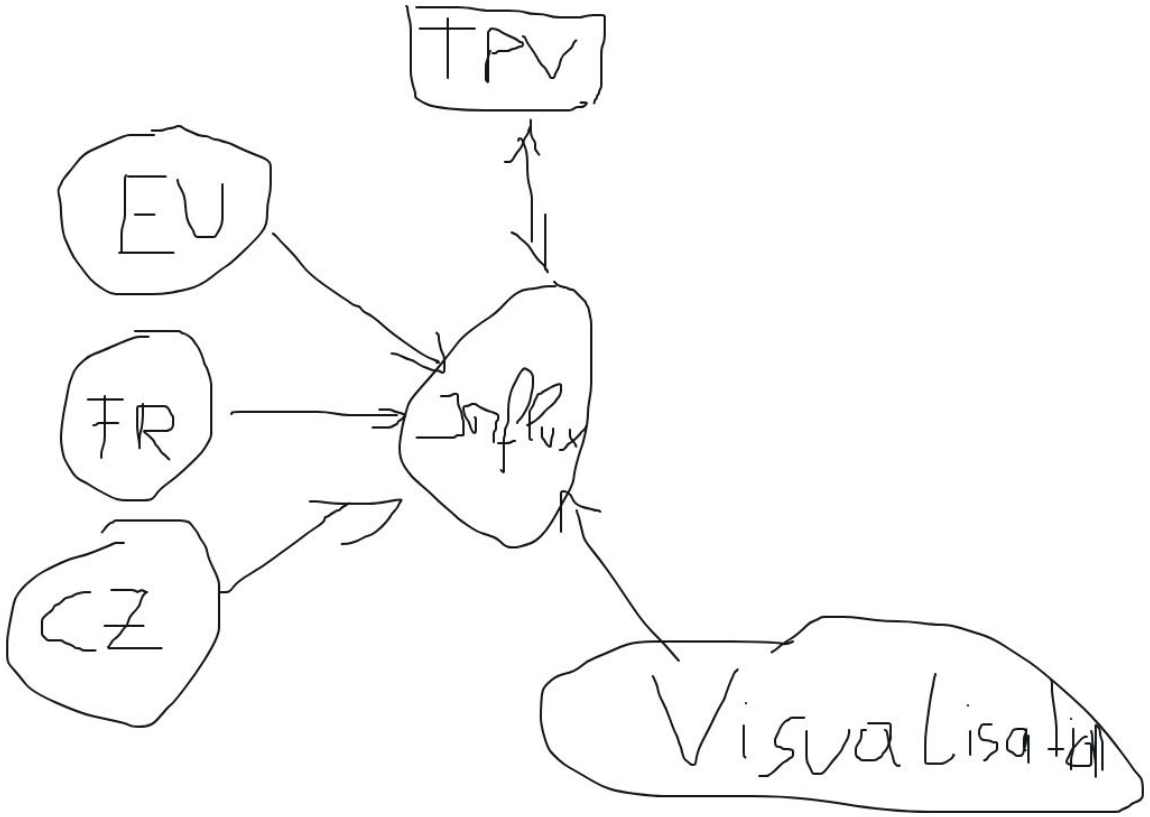
- The EuroScienceGateway project
- Broadening the login options for EuroScienceGateway
- BYOC: Bring Your Own Compute
- BYOS: Bring Your Own Storage
- **Smart job scheduling across Europe**

Smart scheduling

- How do we efficiently schedule jobs?
- Gather usage statistics from pulsar endpoints
- TPV, metascheduling algorithm, and TPV metascheduler API
- Visualization of jobs across Europe

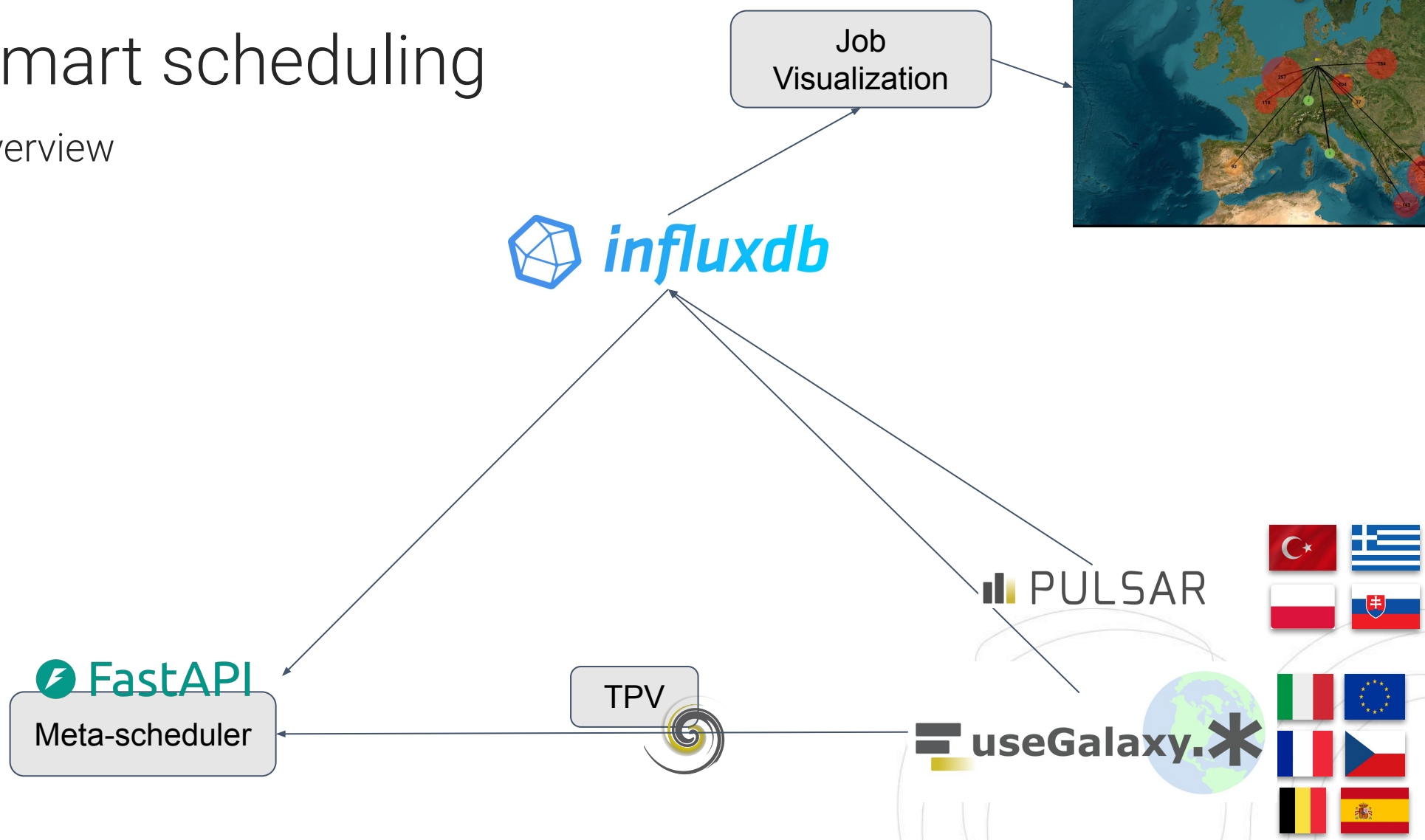
Smart scheduling

Brainstorming over Zoom :)



Smart scheduling

Overview



Smart scheduling

Central API endpoint

- Implements **matchmaking** logic,
  ranking available pulsar destinations based on:
  - current - and historically collected **(tool-destination) metrics**
  - **data locality** (geolocation) of the objectstore/destination

- Currently, simple weighting of the different metrics

- Working on more advanced
  Fuzzy/Adaptive-based matchmaking algorithms

```
tools:
  default:
    ...
    rank: |
      ...

      # currently object store info is stored in a yaml
      request_data["static_objectstores_info"] = objectstore_info

      # dataset info
      request_data["static_dataset_info"] = helpers.get_dataset_attributes(job.input_datasets)

      # static job info
      request_data["static_job_info"] = {"tool_id": tool.id, "mem": mem, "cores": cores, "gpus": gpus,}

      # current destination info
      dest_info = []
      for dest in candidate_destinations:
        dest_dict = {"id": dest.id}
        dest_dict.update(dest.context)

        dest_dict["queued_job_count"] = model.Query(model.Job).filter(
            model.Job.state == "queued",
            model.Job.destination_id == dest.dest_name).count()
        dest_dict["running_job_count"] = model.Query(model.Job).filter(
            model.Job.state == "running",
            model.Job.destination_id == dest.dest_name).count()

        dest_info.append(dest_dict)

      request_data["current_dest_info"] = dest_info

      # Send a POST request to the API endpoint
      response = requests.post(api_url, json=request_data)

      sorted_candidate_destinations = sorted(candidate_destinations,
                                             key=lambda x: sorted_destination_ids.index(x.id))
      sorted_candidate_destinations
```

Smart scheduling

Central API endpoint

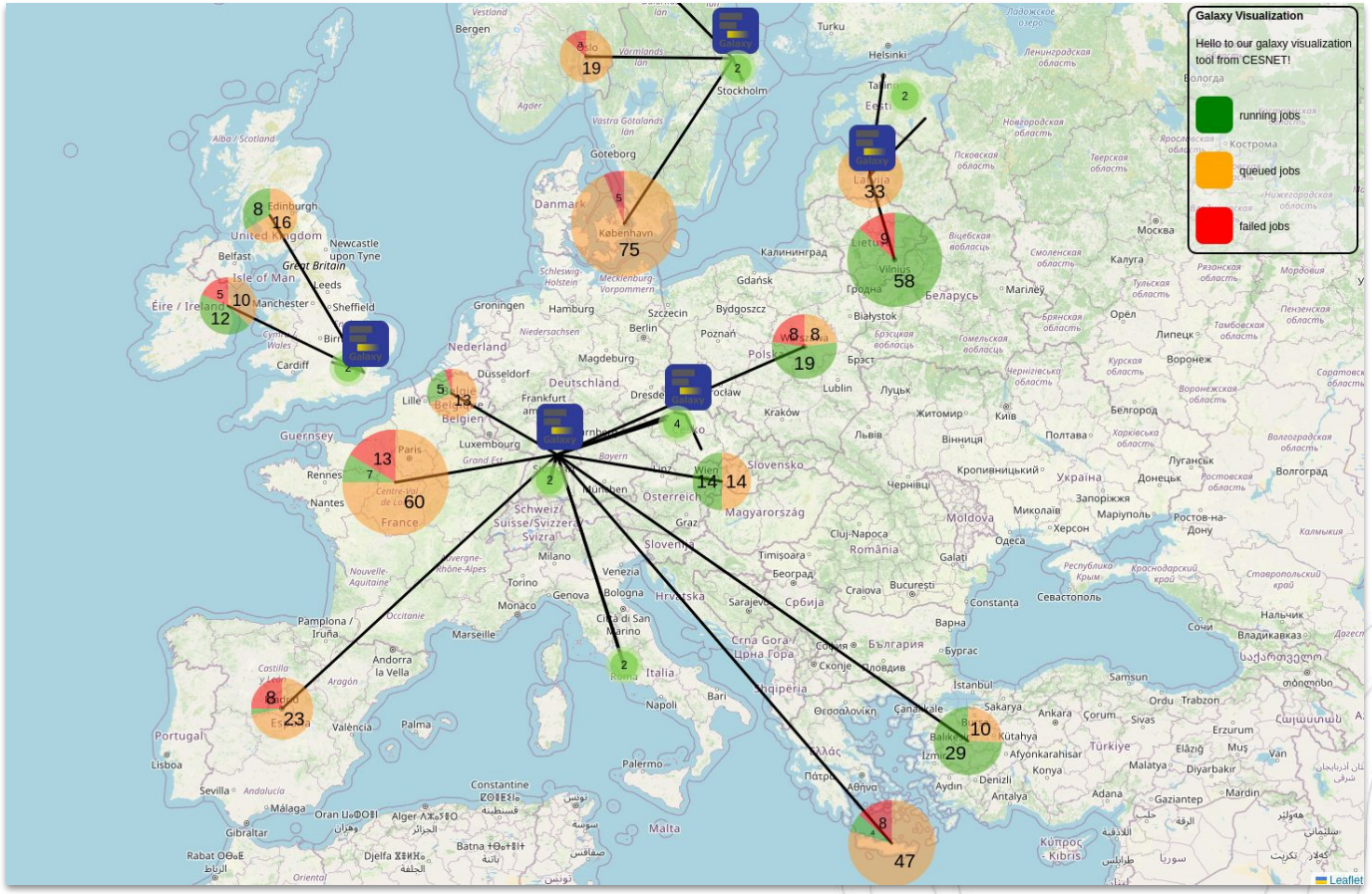
- Implements **matchmaking** logic,
  ranking available pulsar destinations based on:
  - current - and historically collected **(tool-destination) metrics**
  - **data locality** (geolocation) of the objectstore/destination

- Currently, simple weighting of the different metrics

- Working on more advanced
  Fuzzy/Adaptive-based matchmaking algorithms

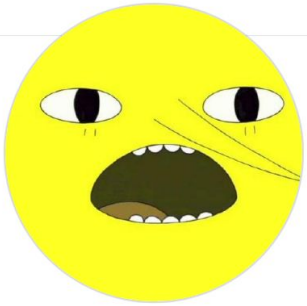
Smart scheduling

Visualization of jobs across Europe



Thanks!

All ESG members, specially:



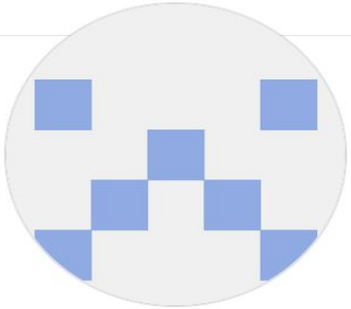
Paul De Geest

Sanjay Kumar Srikakulam

Abdulrahman Azab

Björn Grüning

Enol Fernández



Maiken Pedersen

Łukasz Opioła

Martin Demko

Tomáš Vondrák