National Czech Programme | eosc

Data management a LEXIS Platforma

Martin Golasowski, IT4I

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

MUNI ICS

Extreme Data Analytics in an Exascale Era with HPC-Cloud-AI Convergence

EXAHMIND

Extreme Data Analytics as Enabler for Science/SMEs/Industry

+

At Top-Level EU Supercopmuting Centres, Cross-System With Automatised Workflows

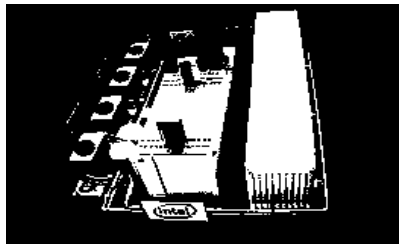
+

Connecting to European Data Spaces, EOSC and EUDAT

Data Management in HPC

IT4Innovations supercomputers

Anselm (0.1 PF)



July 2013

DGX-2 (2 PF^{AI})

March 2019

Karolina (15.7 PF , #69)

August 2021

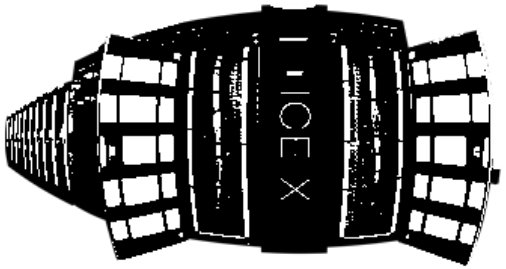
NEW

LUMI-Q

2024

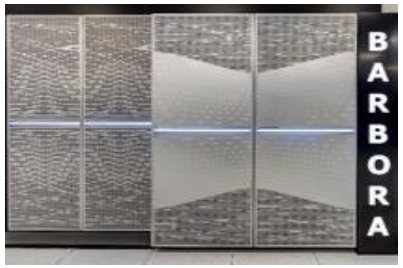
2011

July 2015



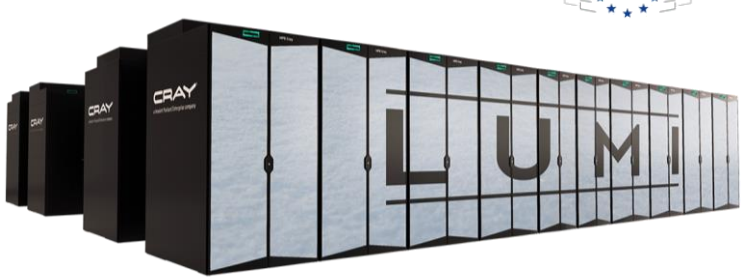
Salomon (2 PF, #40)

October 2019



Barbora (0.8 PF)

Autumn 2021

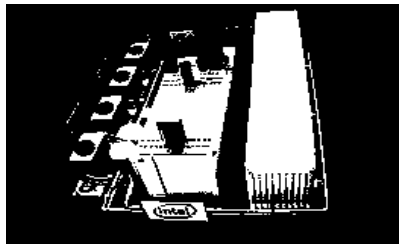


LUMI (550+ PF, #3, in Kajaani, FI)

Data Management in HPC

IT4Innovations supercomputers & Storages

Anselm (0.1 PF)



DGX-2 (2 PF^{AI})

Karolina (15.7 PF , #69)

NEW

LUMI-Q

2011

July 2013

0,46 PB

July 2015

March 2019

2,56 PB

October 2019

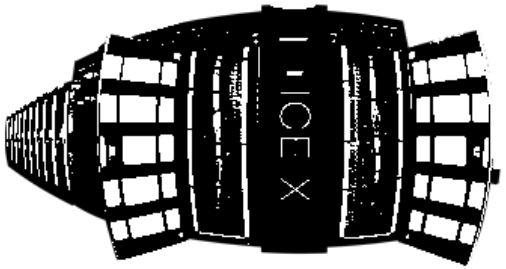
August 2021

16 PB

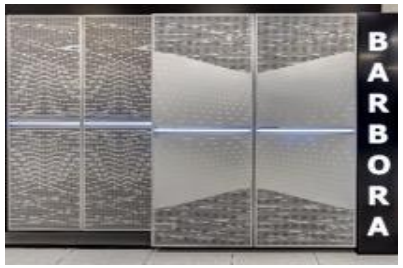
Autumn 2021

2024

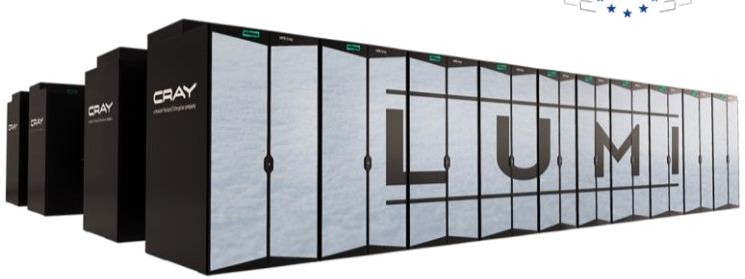
29 PB



Salomon (2 PF, #40)



Barbora (0.8 PF)

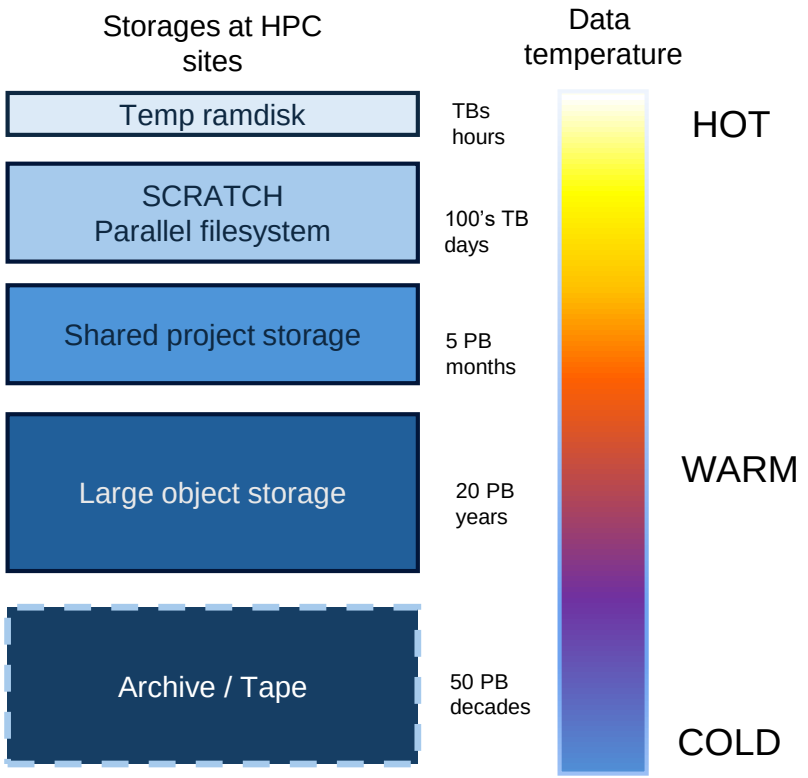


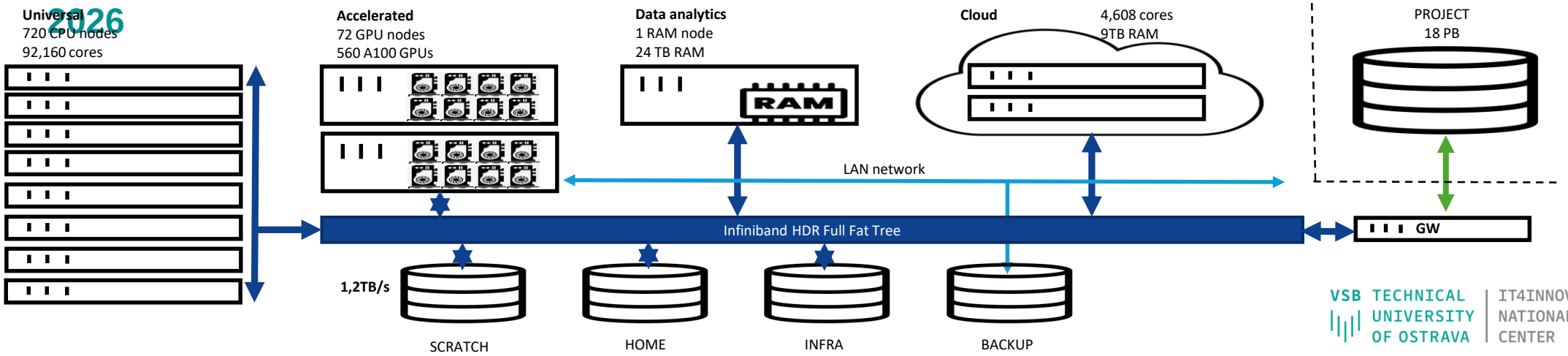
LUMI (550+ PF, #3, in Kajaani, FI)

Data Management in HPC

Karolina – heterogeneous infrastructure

- In operation from **2021**
- EuroHPC supercomputer
  - 65% Czech funding
  - 35% EU funding (for EuroHPC users)
- Total investment approx. **EUR 15 million**
- Total theoretical performance **15.7 PFlop/s**
  - #69 Top500 in June 2021 (Karolina GPU)
  - #8 Green500 in November 2021 (Karolina GPU)
- Expected end of operation for EuroHPC users **2026**



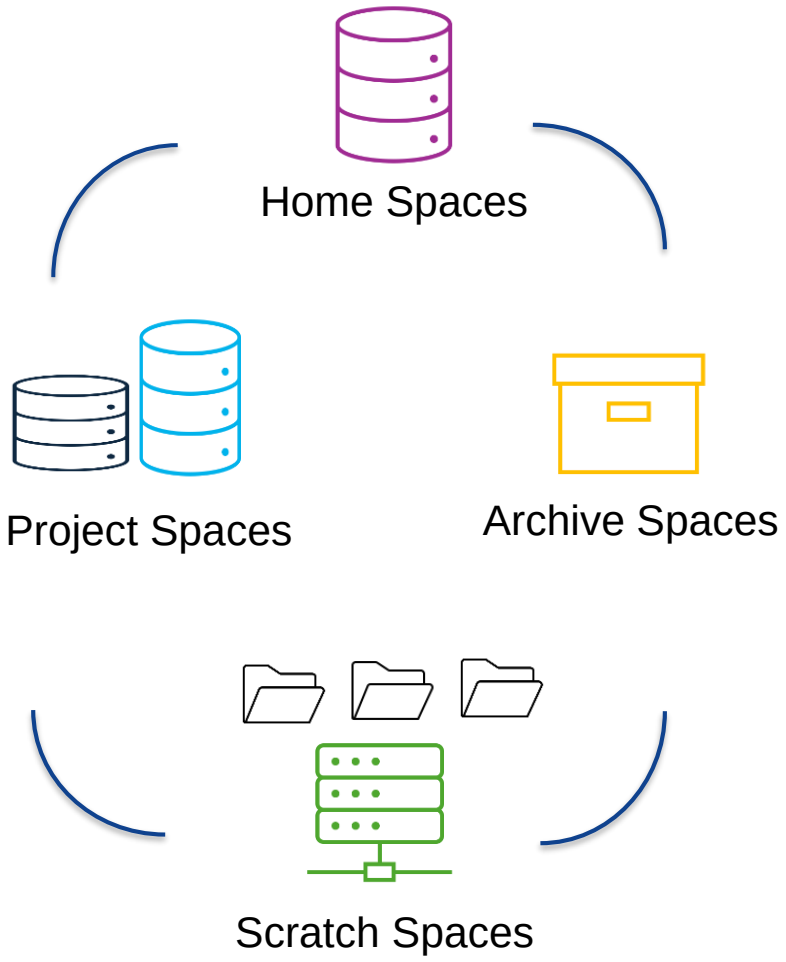


VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

Data Management in HPC

Support & User perspective

- **User perspective**
  - Easy
    - Data transfer + computation execution
    - Provenance collection (monitoring, environment..)
  - They projects usually have limited life time (e.g. Open Access Competition rules)
  - Data usually have to survive beyond computational project life time (FAIR)
  - New era - AI training & inference
- **HPC operations support perspective**
  - HPC clusters are designed as general purpose systems
  - Applications with generative AI and LLM open new type of challenges for HPC centres
  - Data policies and life cycle monitoring
  - etc.



Raw data … Computation … models

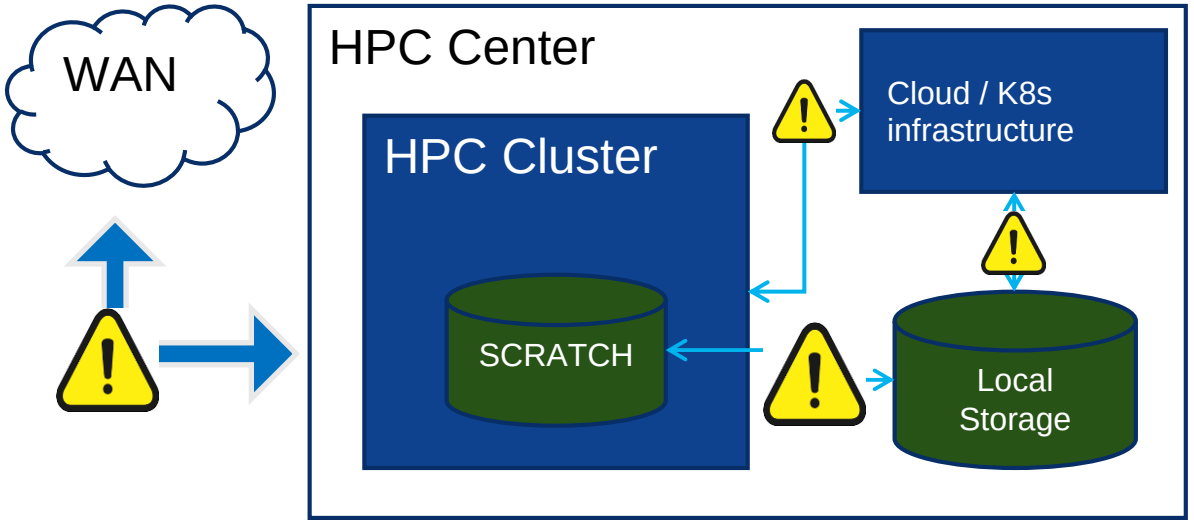
What happens between ??

Data Management in HPC – selected techniques

The data must be transferred efficiently

- **Moving of TBs** of data to HPC centres is complicated
  - Network restrictions of outgoing traffic
  - HPC centres are not primary data source provider
  - Allocations per project – longer term storage?
- Routing between European HPC centres is not always efficient
- **Tooling for staging/transfer** TBs of data does not exist
- HPC Centres rely on **legacy SSH protocols, not suitable for large data transfer**
- Integration of vastly different storage concepts (DBs, Object, POSIX) to common usable platform

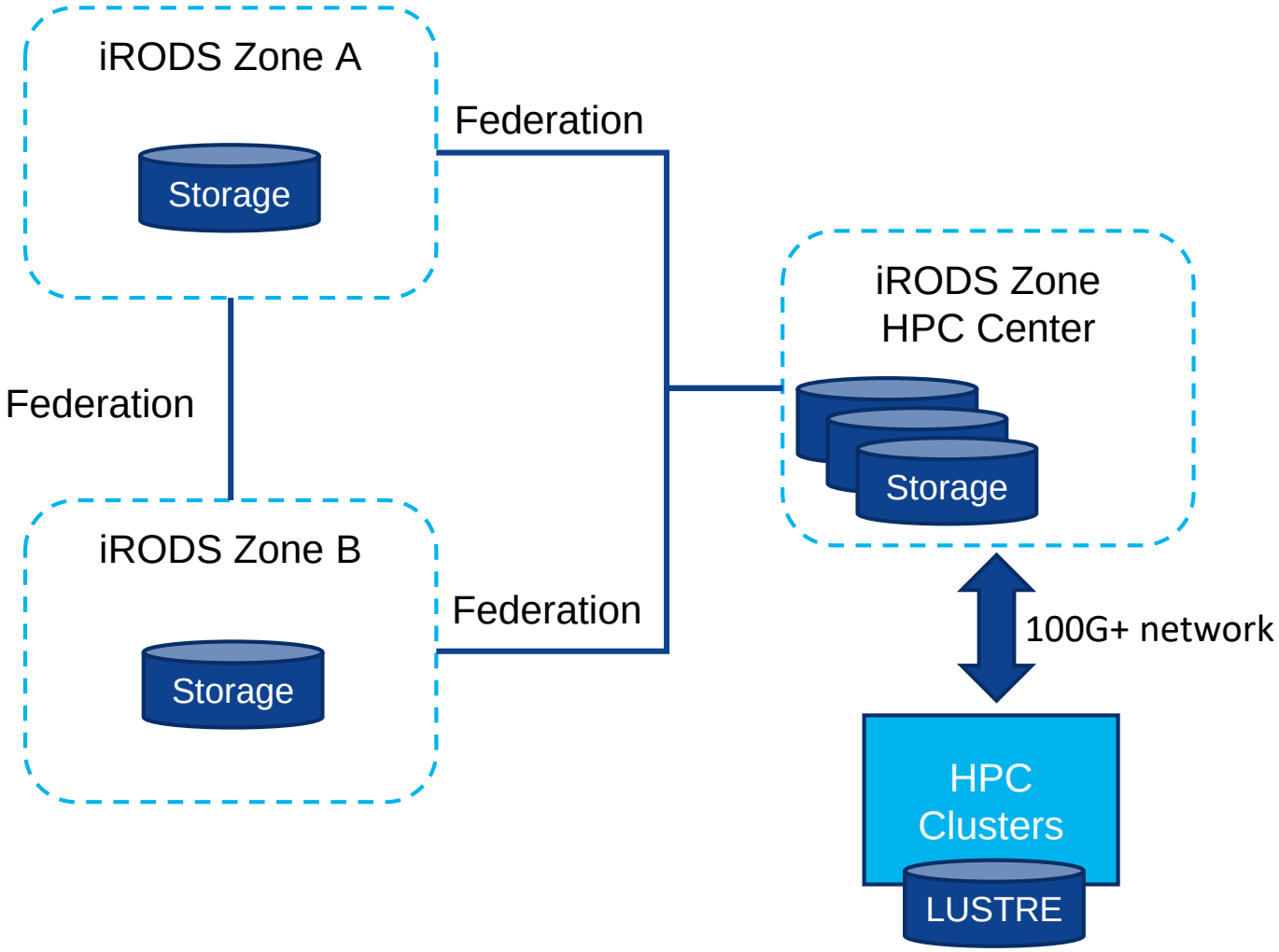
Data transfer challenges



Data Management in HPC – selected techniques

The data must be transferred

- Same transfer protocols across data centers
- Data movement close to compute
- Autonomy of each center must be maintained
- All data centers can connect to common AAI and use fast WAN networks

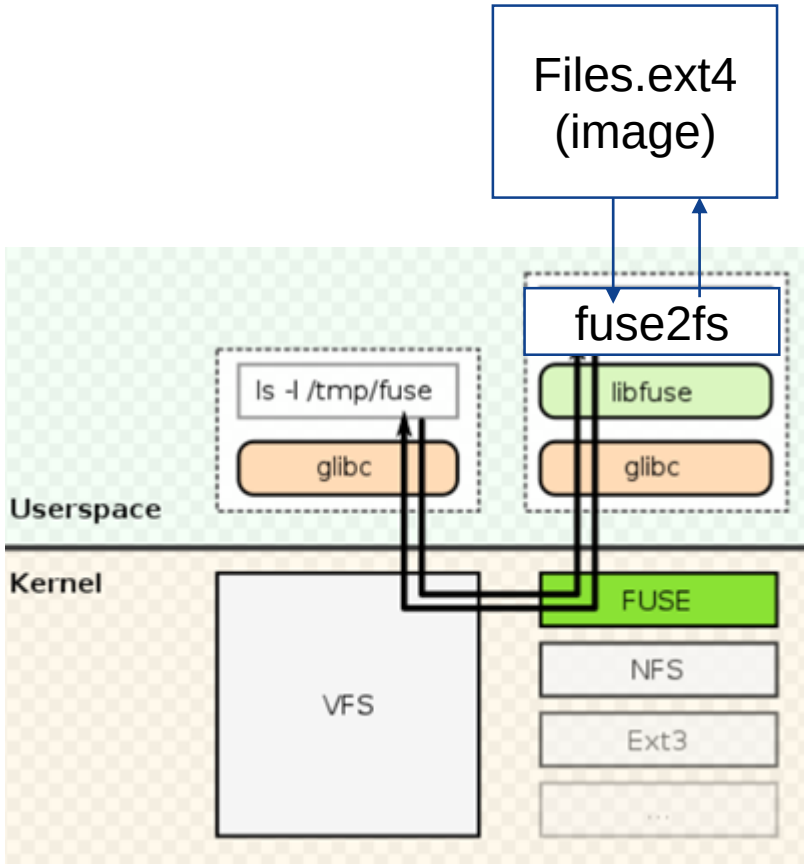


Data Management in HPC – selected techniques

Handling large number of files on HPC clusters

The problem with shared filesystems

- Inode (file count) numbers are limited
- HPC centers set limits on number of files
- Traversing large trees, stat() may be (very) time consuming
- Metadata servers may get overloaded on open(), close(),
  stat(), seek() and other I/O operations

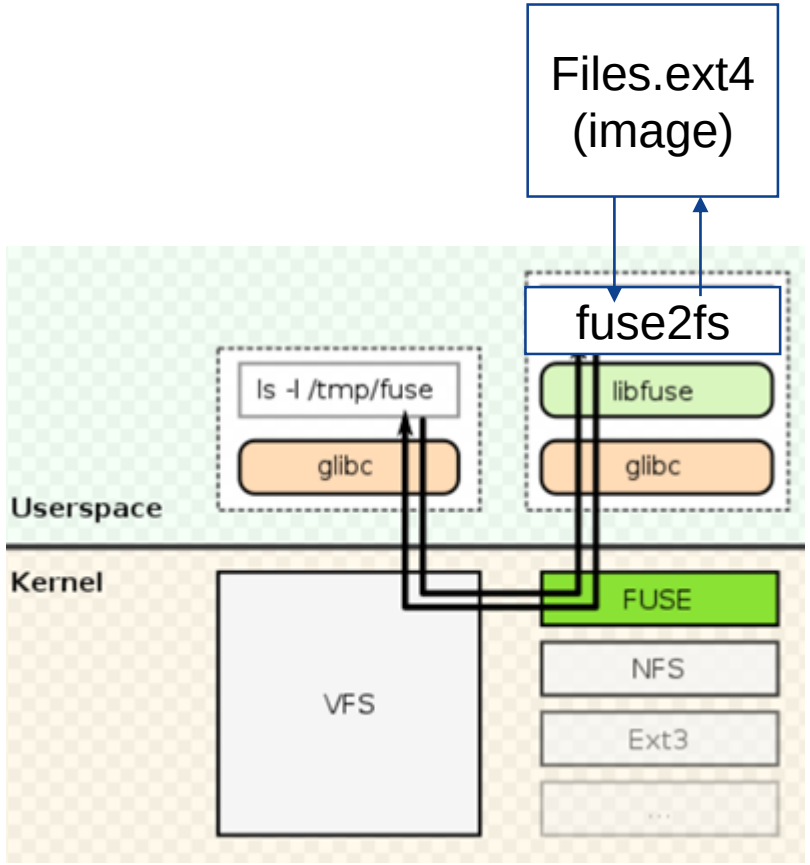


Data Management in HPC – selected techniques

Handling large number of files on HPC clusters

Mitigation

- Filesystem in User Space
- Single file (image) on shared filesystem can host millions of files internally.
  Access via userspace process
- **fuse2fs** (mounts ext2 family images)
- **archivemount** (mounts tar, zip, rar, etc... archives)
- Other FUSE clients, can mount other images, remote repos, object storages,
  even e-mail IMAP archives)
- Read only mount can and should be done in parallel, across multiple nodes



Solution

- Use adequate object storage (S3, iRODS, ADIOS, etc.) plus staging mechanism to parallel file system if needed

Data Management in HPC – selected techniques

Artificial intelligence is not just training …

Training

- Finite amount of data - **batch job**
- Need to transfer data closest to the GPUs - efficient staging
- Store the results back

Inference

- Batch or **request based**
- Request based unsuitable for batch scheduling - resource wasting
- Need to load the models efficiently

HPC Cluster

slurm workload manager

HEAppE

HyperQueue

EXA4MIND

CPU Partition

GPU Partition

Meta scheduling for better utilisation of AI applications

HyperQueue

Motivation

- It is challenging to run complex task workflows on HPC clusters
  - Workflows are heterogeneous, have dependencies
  - Clusters are heterogeneous
  - Allocation managers are not accustomed to enormous amounts of smaller, less resource intensive tasks

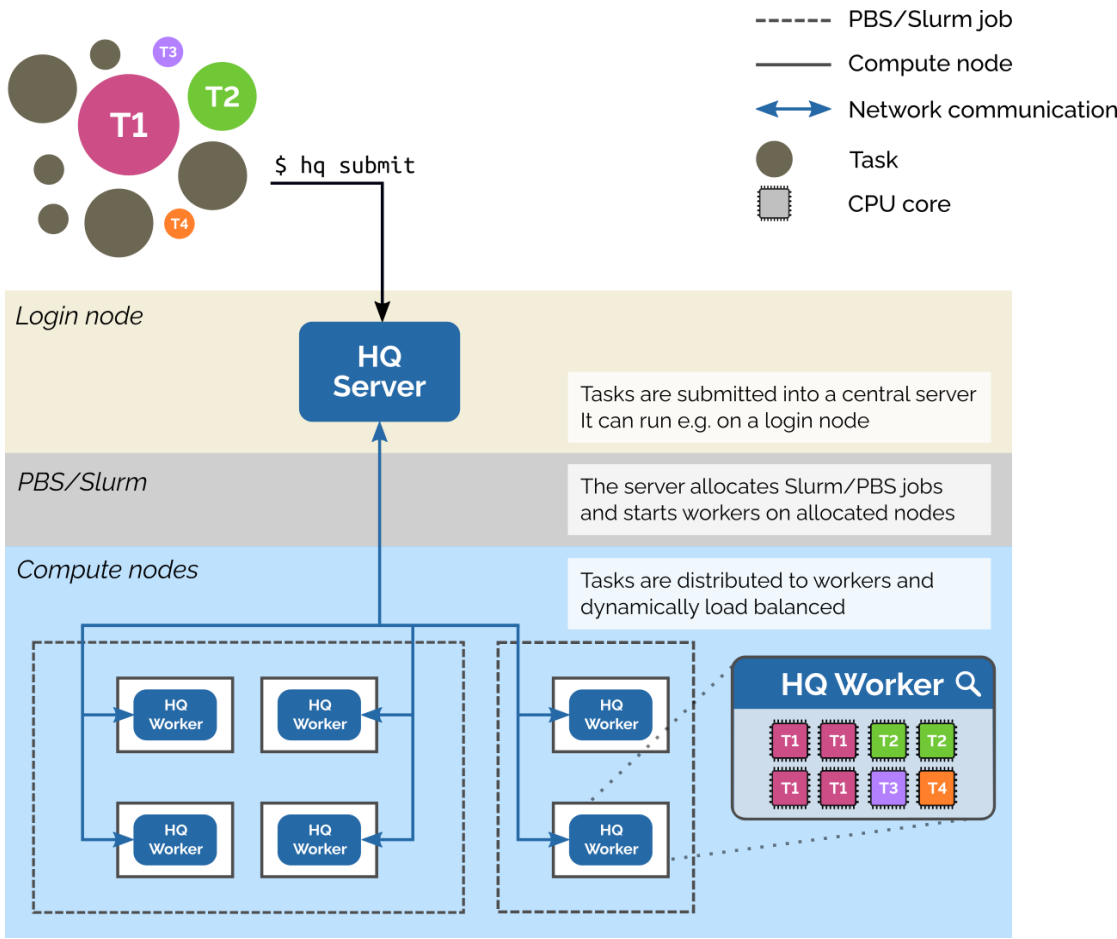
HyperQueue is an HPC-tailored distributed task runtime designed for simple and efficient execution of task graphs (workflows) on HPC clusters

```bash
#!/bin/bash

while :
do
  # Submit a job and wait for it to complete
  ./hq submit --wait ./compute.sh

  # Read the output of the job
  output=$(./hq job cat last stdout)

  # Decide if we should end or continue
  if [ "${output}" -eq 0 ]; then
      break
  fi
done
```



Open-source (MIT-licensed) available at GitHub:
https://github.com/It4innovations/hyperqueue

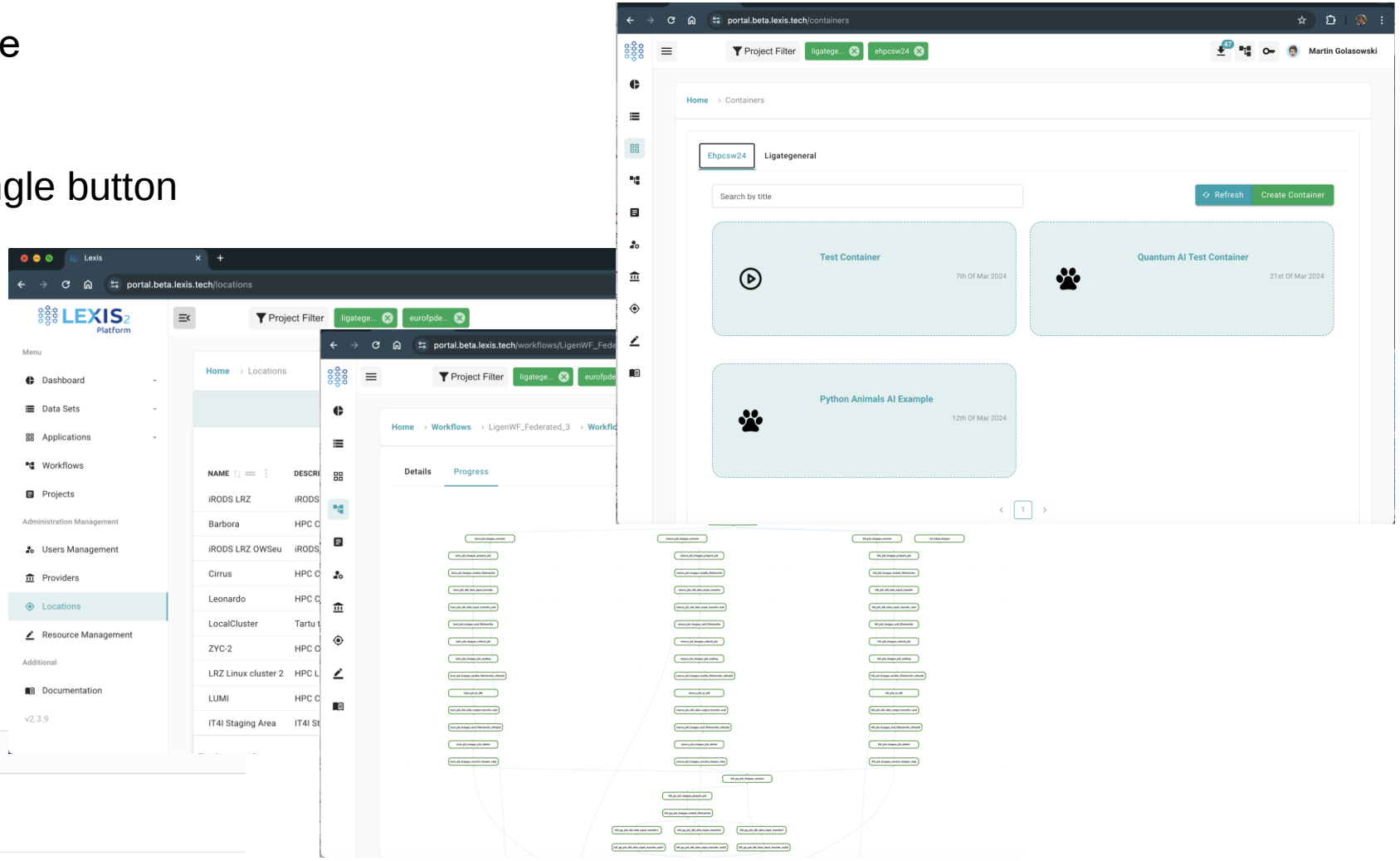
How to supercompute …

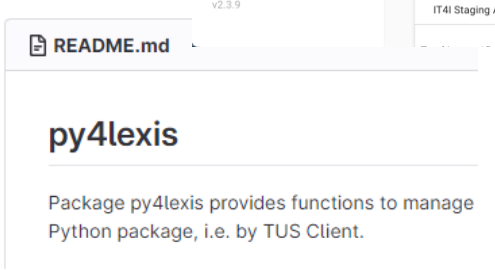
- Pick a supercomputer
- Fill out a request for allocation
- Get it approved
- Set up an account
- Set up a SSH key
- Know how to use Linux terminal
- Login in to the Supercomputer
- Learn how to launch jobs in SLURM
- Learn about modules
- Learn about storages and data transfer
- … finally compute something

Easy and Safe access

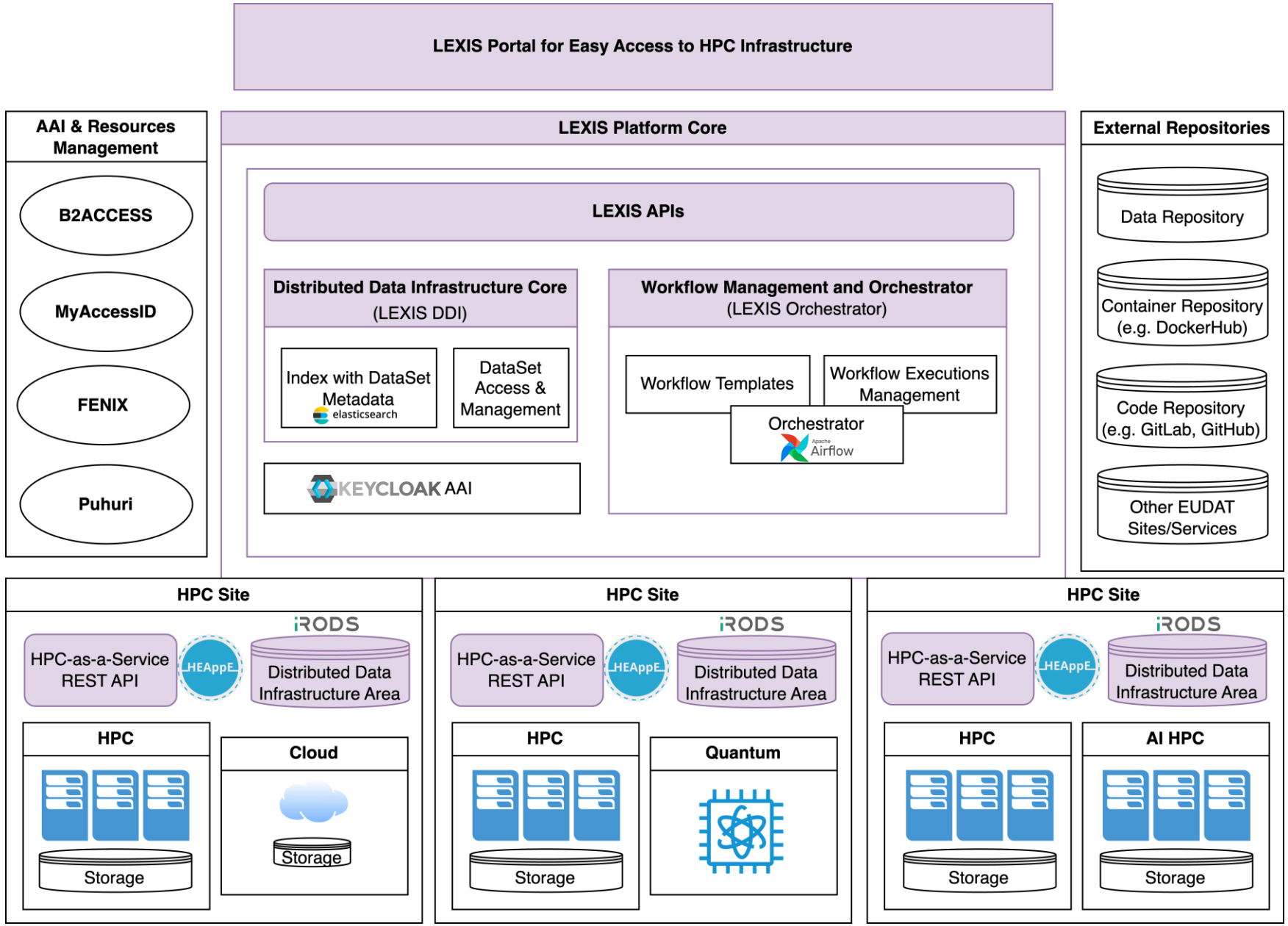
- Let’s have a nice web interface
- Allocations at one place
- Across multiple clusters
- Launch applications with a single button
- Get logs
- Manage data in iRODS
- Use common web login
- … and many more

Visit for more:
docs.lexis.tech
opencode.it4i.eu/lexis-platform





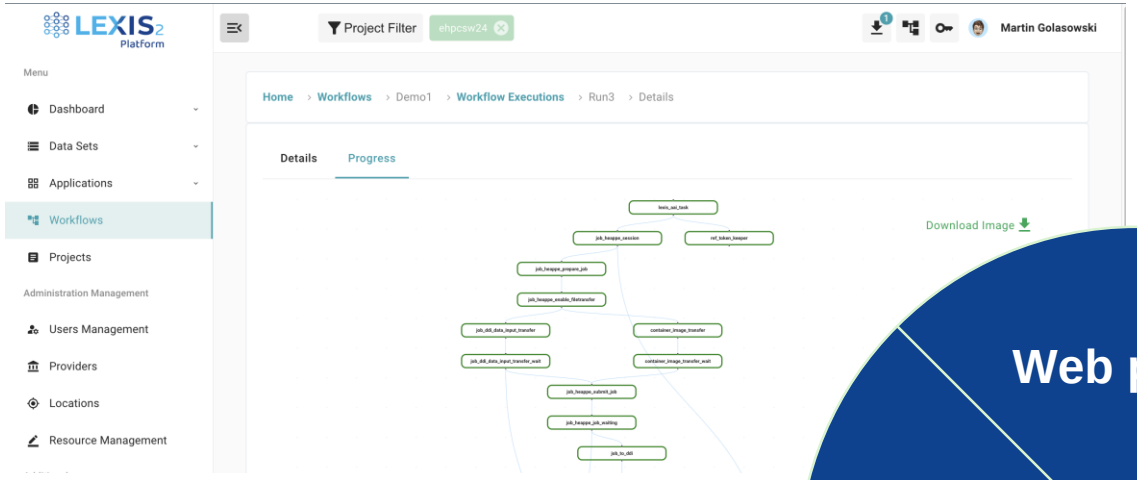
LEXIS Platform



- Federation of **European computing centres**
- **Hiding** of technical and **operational differences** across organizations
- **HPC & Cloud** service providers, Data providers
- Unified & distributed **data management**
- Orchestration
- **Federated Authentication & Authorization** Infrastructure (AAI)

LEXIS Platform

LEXIS2 Platform



workflowManagement Actions relating to management of Workflows and Workflow Executions

GET	/workflow	Return list of available LEXIS Workflows	
POST	/workflow	Create a new LEXIS Workflow on the system	
GET	/workflow/{workflowId}	Return detailed info on LEXIS Workflow for given Workflow ID	
DELETE	/workflow/{workflowId}	Delete LEXIS Workflow on the system	
GET	/workflow/{workflowId}/execution	List the current available LEXIS Workflow Executions.	listWorkflow
POST	/workflow/{workflowId}/execution	TODO: Needs implemented with TOSCA 1.3 Capabilitise. Create a new LEXIS Workflow Execution by providing remaining inputs	createWorkflowExecution
GET	/workflow/{workflowId}/execution/{workflowExecutionId}	Returns LEXIS Workflow Execution detail.	getWorkflowExecutionDetail
DELETE	/workflow/{workflowId}/execution/{workflowExecutionId}	Cancel a LEXIS Workflow Execution.	cancelWorkflowExecution

Web portal

Public REST API

Python library SDK

Terminal based interface

README.md

py4lexis

Package py4lexis provides functions to manage datasets within the LEXIS platform. Uploading datasets' files is performed by tuspy Python package, i.e. by TUS Client.

Install

We recommend to use virtual environment, e.g.:

```
python -m venv ./venv
```

ivated virtual environment, type:

```
-m pip install py4lexis
```

```
preprocessing_dataset_path_geographical_data_path": "project/proj934a1763631a81e

flowExecutionName": "Run3",
flowExecutionConsent": True

are_data():
epared = WRF_EXEC_PARAMS
repared["inputParameters"]["preprocessing_start_date"] = datetime.datetime.now().str
#prepared["inputParameters"]["preprocessing_start_date"] = '2022120400'
return prepared

def submit_wf_execution(token):
    res = requests.post(LEXIS_API + WORKFLOW_ID + '/execution',
    json = prepare_data(),
    headers = {'Authorization': 'Bearer ' + token})

    if res.status_code != 201:
        logger.error(res.content)
        return

    return res.json()
```

```
(venv) mgolas@Martins-MacBook-Pro ~/I/L/L/Py4Lexis (develop)> python examples/upload.py
Progress:  |[redacted]----------------------------------| 30.0% Uploaded
```

LEXIS Platform

Selected use-cases

LIGATE Project

- Application for molecular docking simulation – private IP by DOMPÉ
- LEXIS provides access to workflows with this application running on HPC
- *Without direct access* to the binary or source code

OpenWebSearch.eu

- European open web index processed through LEXIS on several HPC locations (LRZ, IT4I, CSC, DLR)
- Public indices made available trhough the LEXIS Portal
- LLM/AI applications **without access** to data

LEXIS Platform

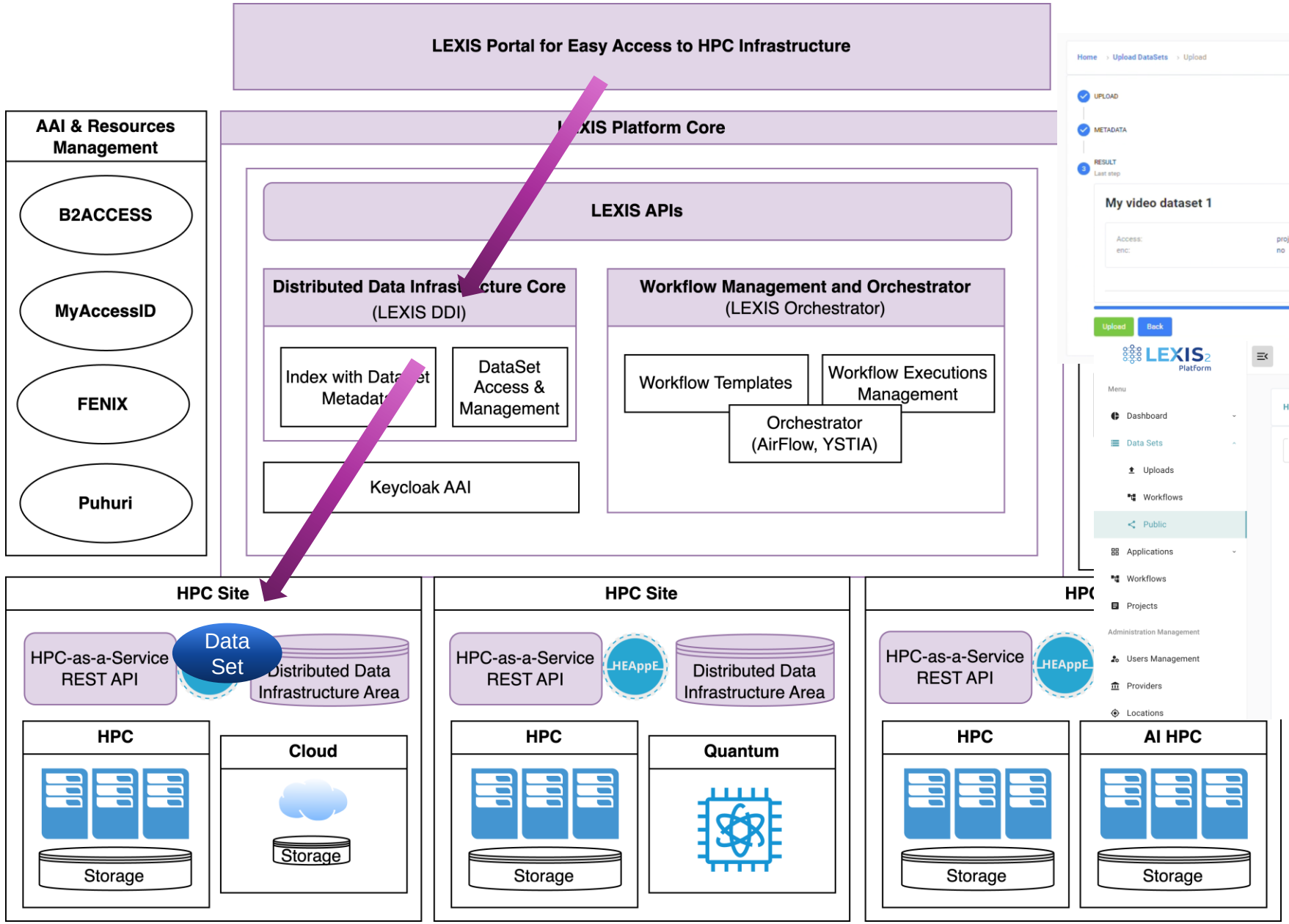
User story - SME DOMPE & LiGen SW

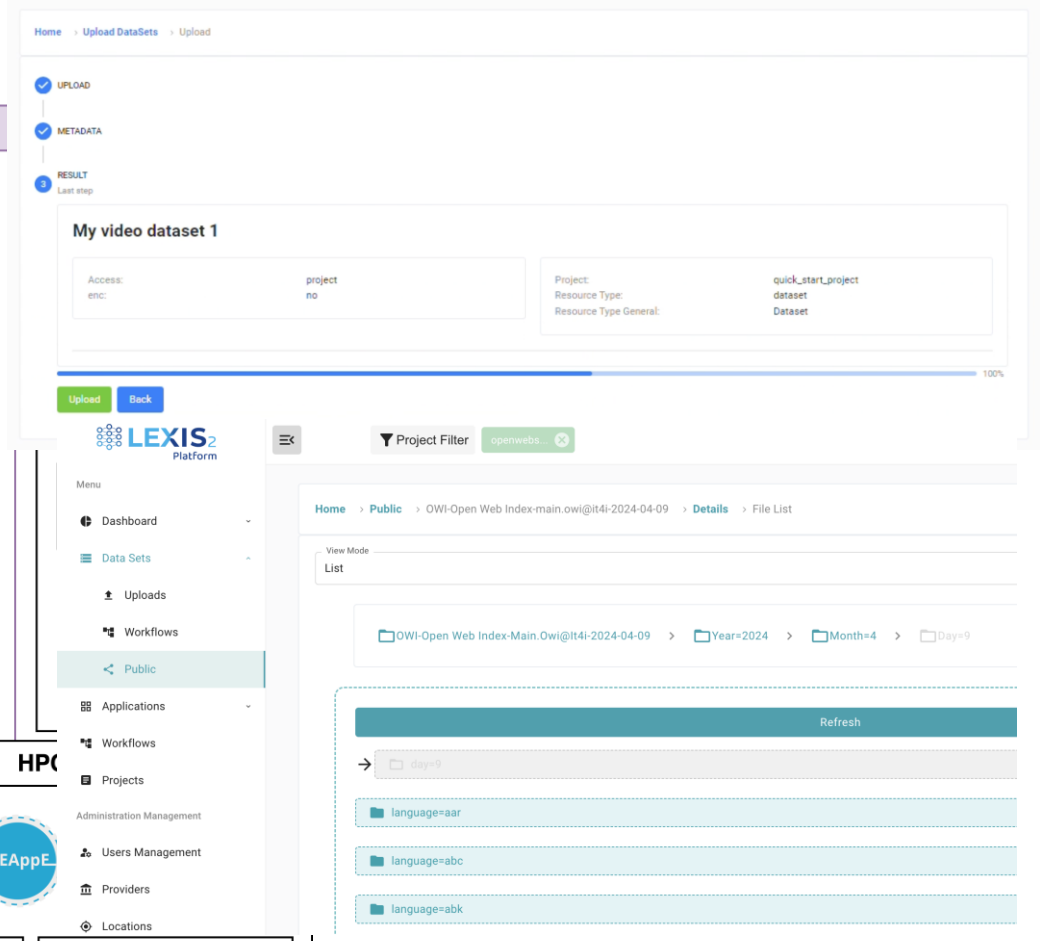
- DOMPE has its own code co-developed with POLIMI & CINECA
- DOMPE LiGen code IP does not allow to share LiGen SW (even binary)
- DOMPE would like to extend its customer base thanks to LEXIS Platform and EuroHPC infrastructure without losing IP
- DOMPE would like to make available LiGen to non-profit academic research and public institutions as a Platform-as-a-Service solution for drug discovery

LEXIS Platform

User story - Data update

Dompé

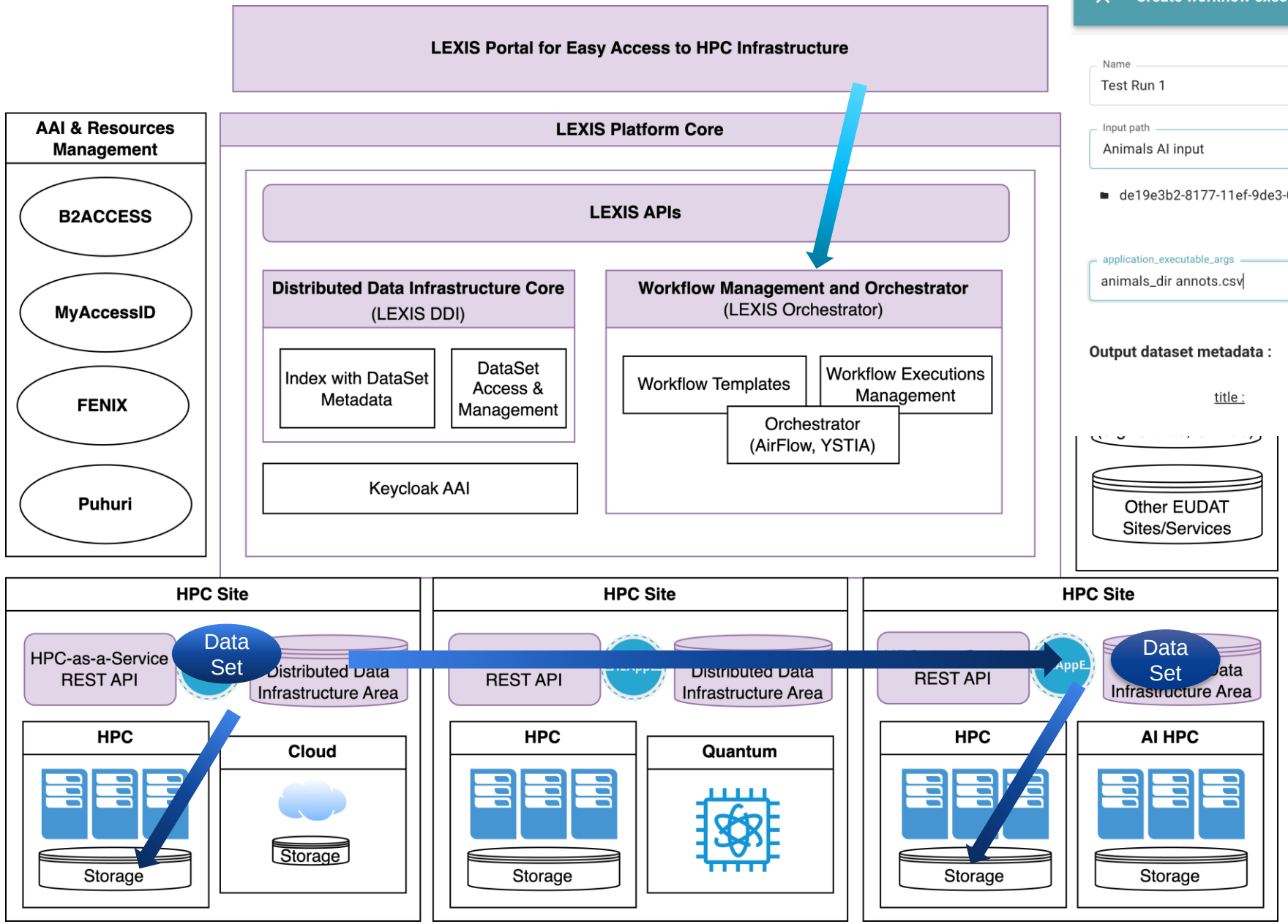


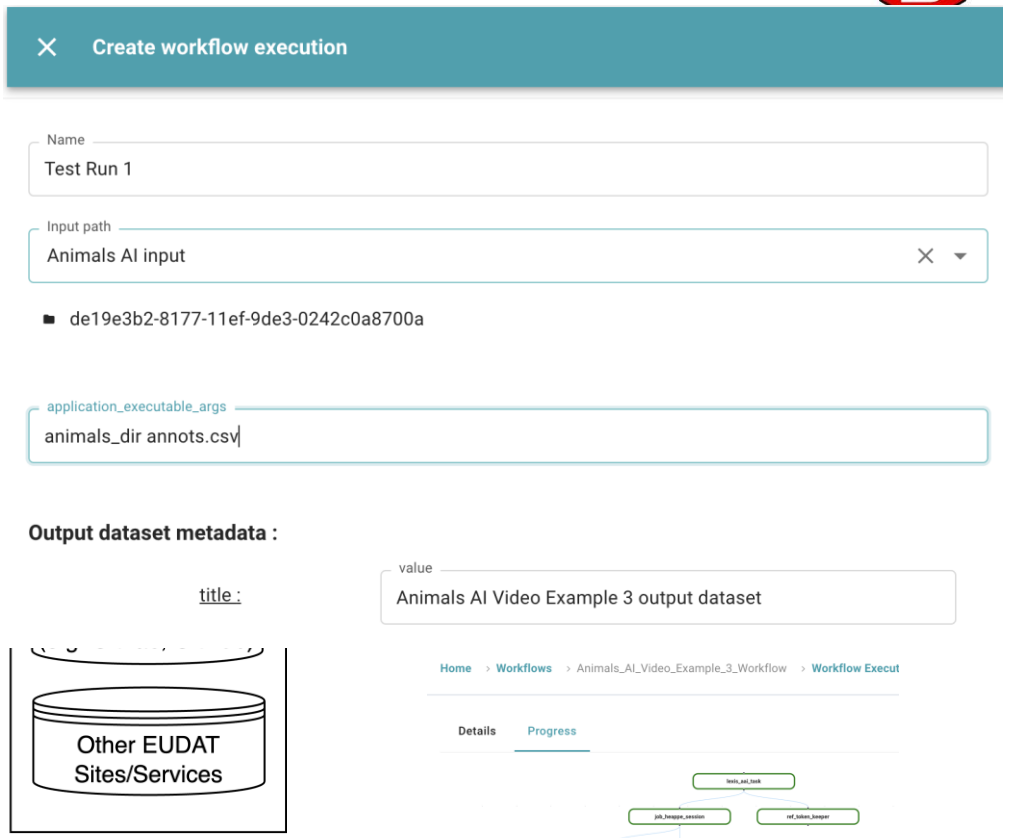


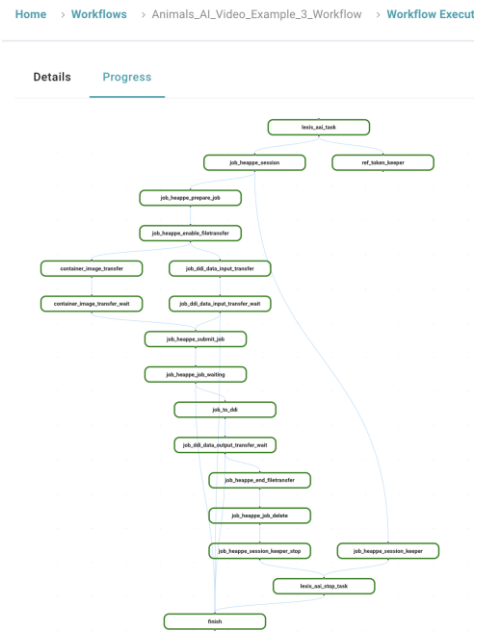
LIGATE

LEXIS Platform

User story - Workflow execution

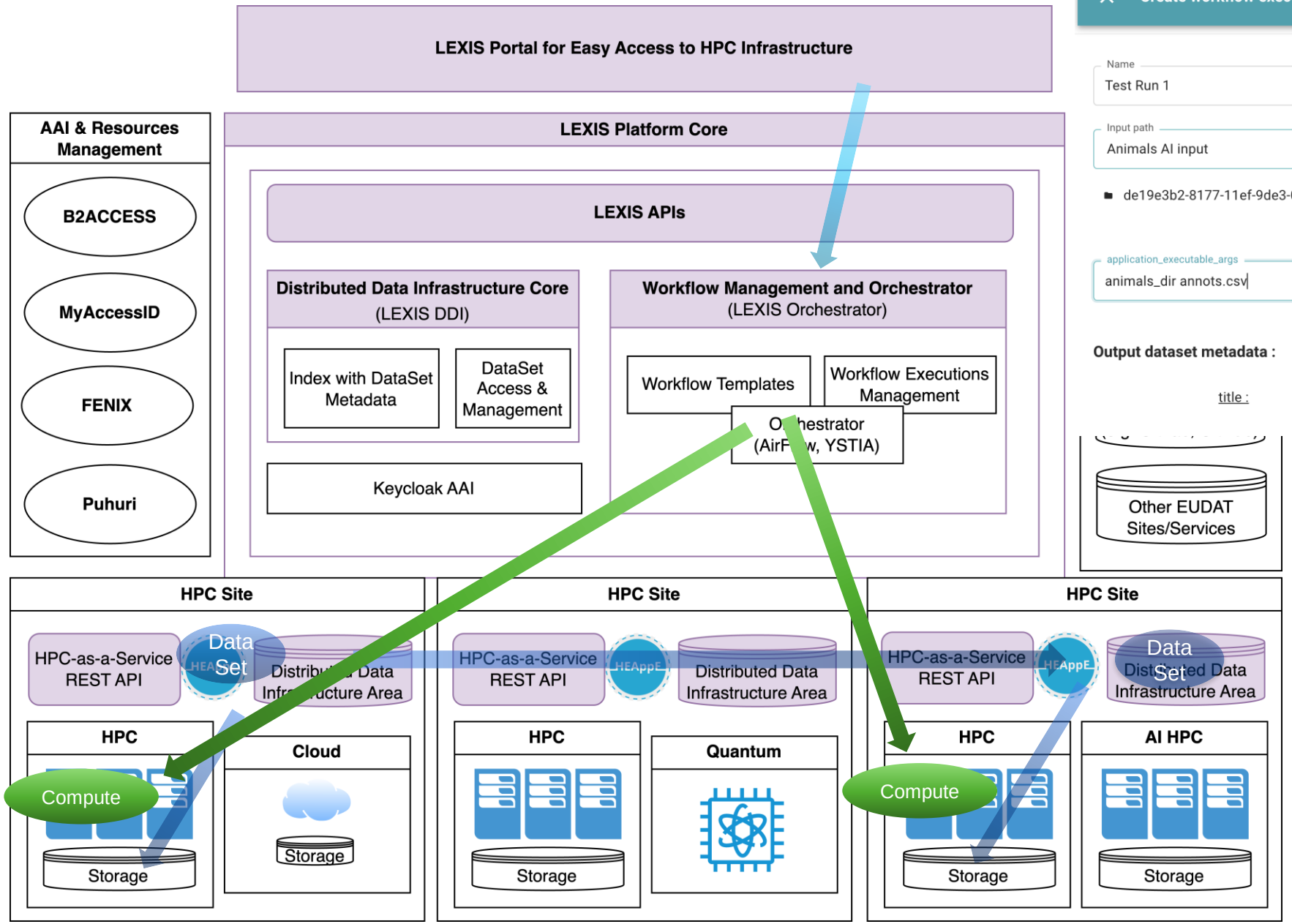


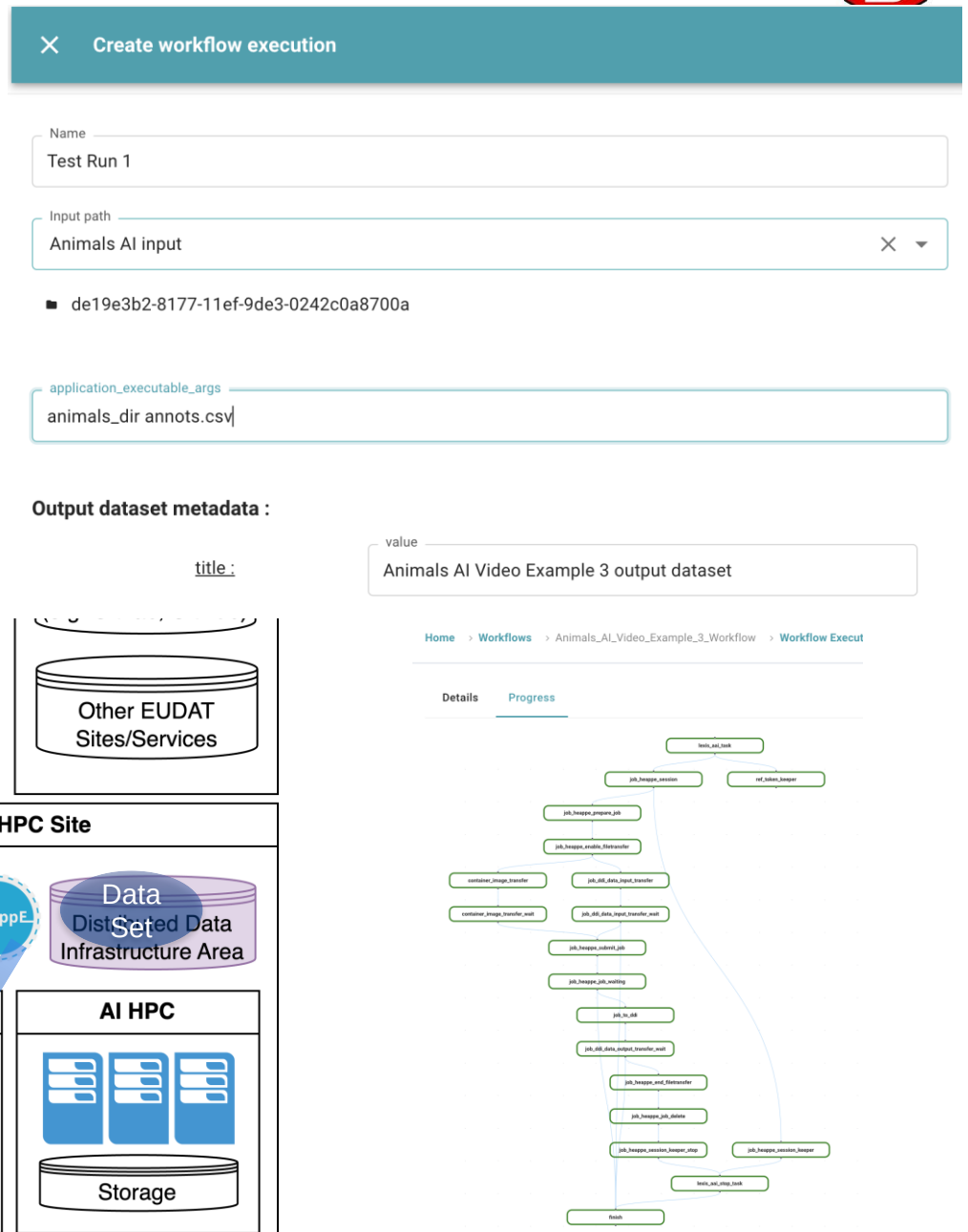




LEXIS Platform

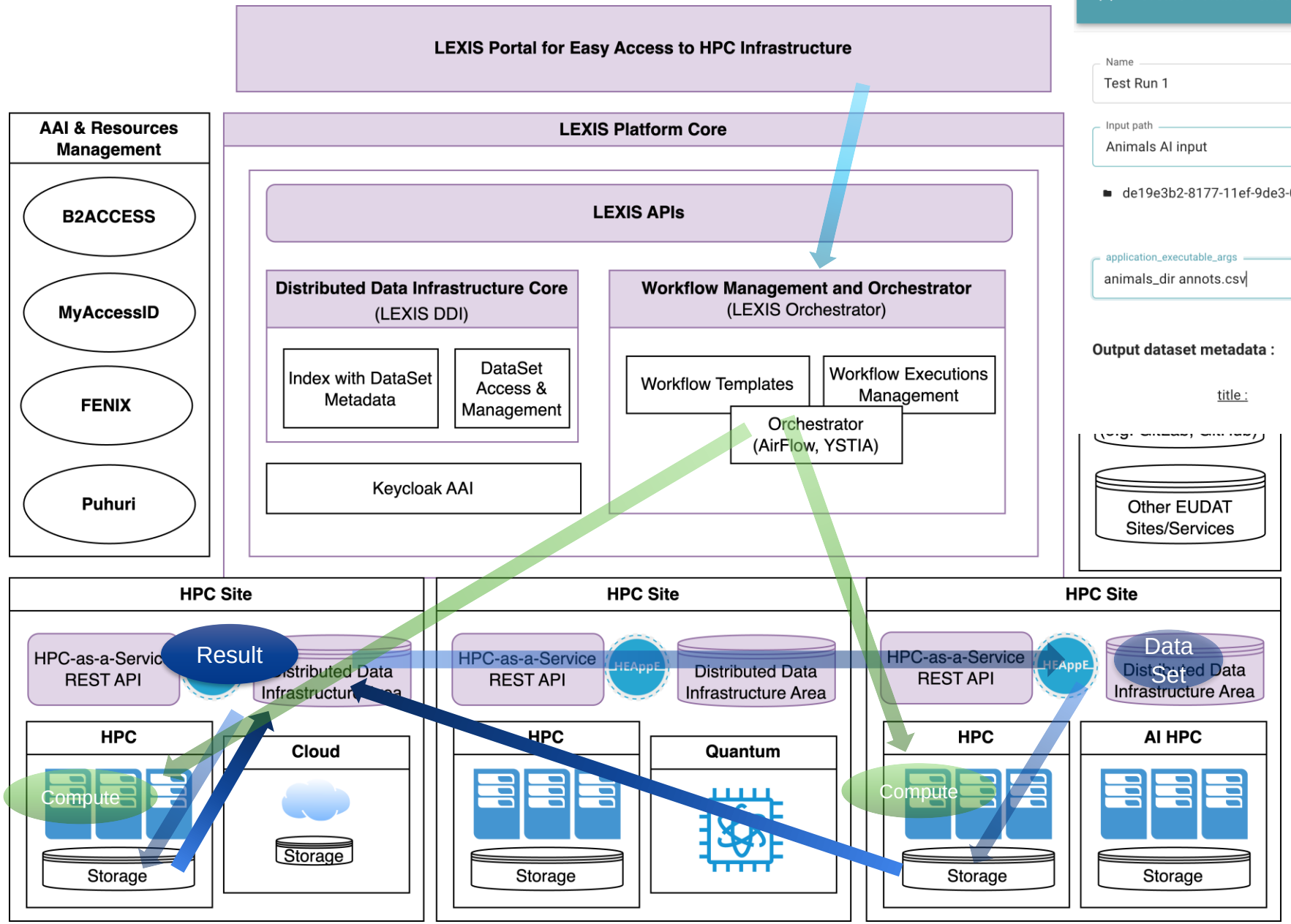
User story - Workflow execution

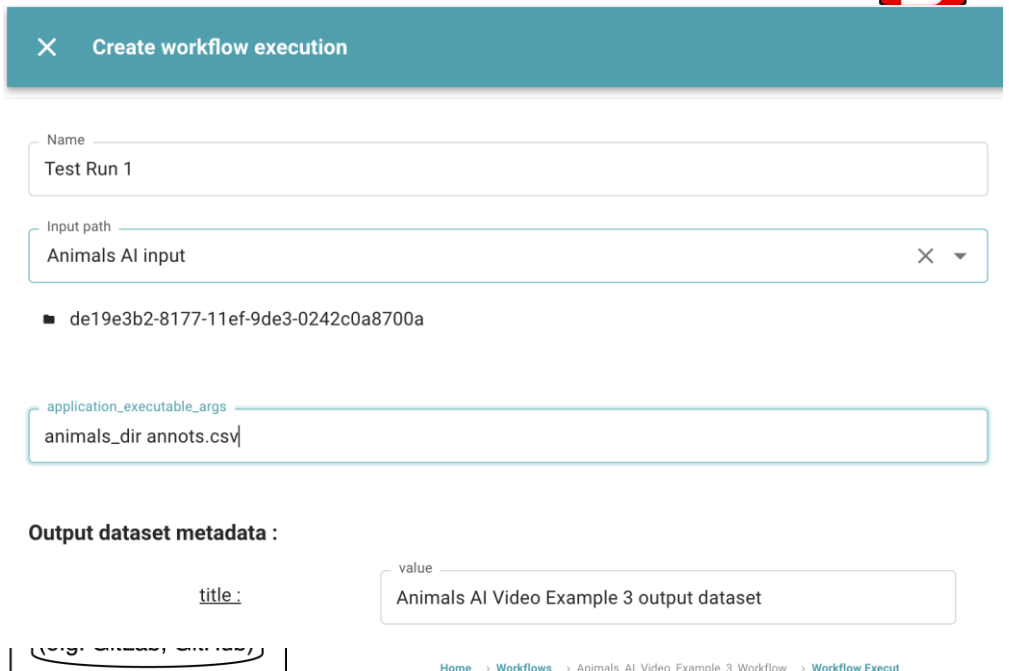


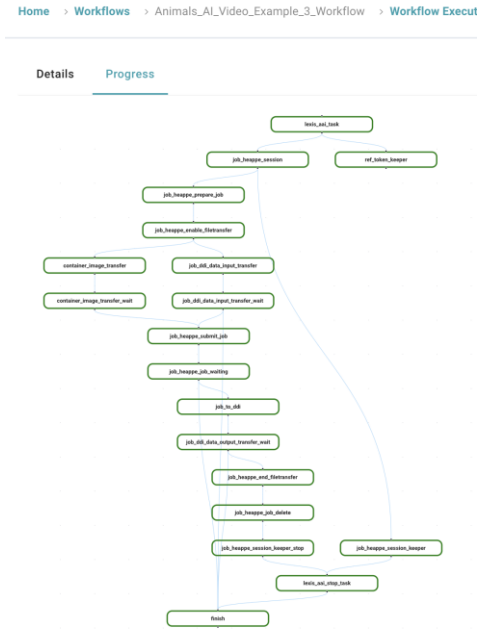


LEXIS Platform

User story - Workflow execution



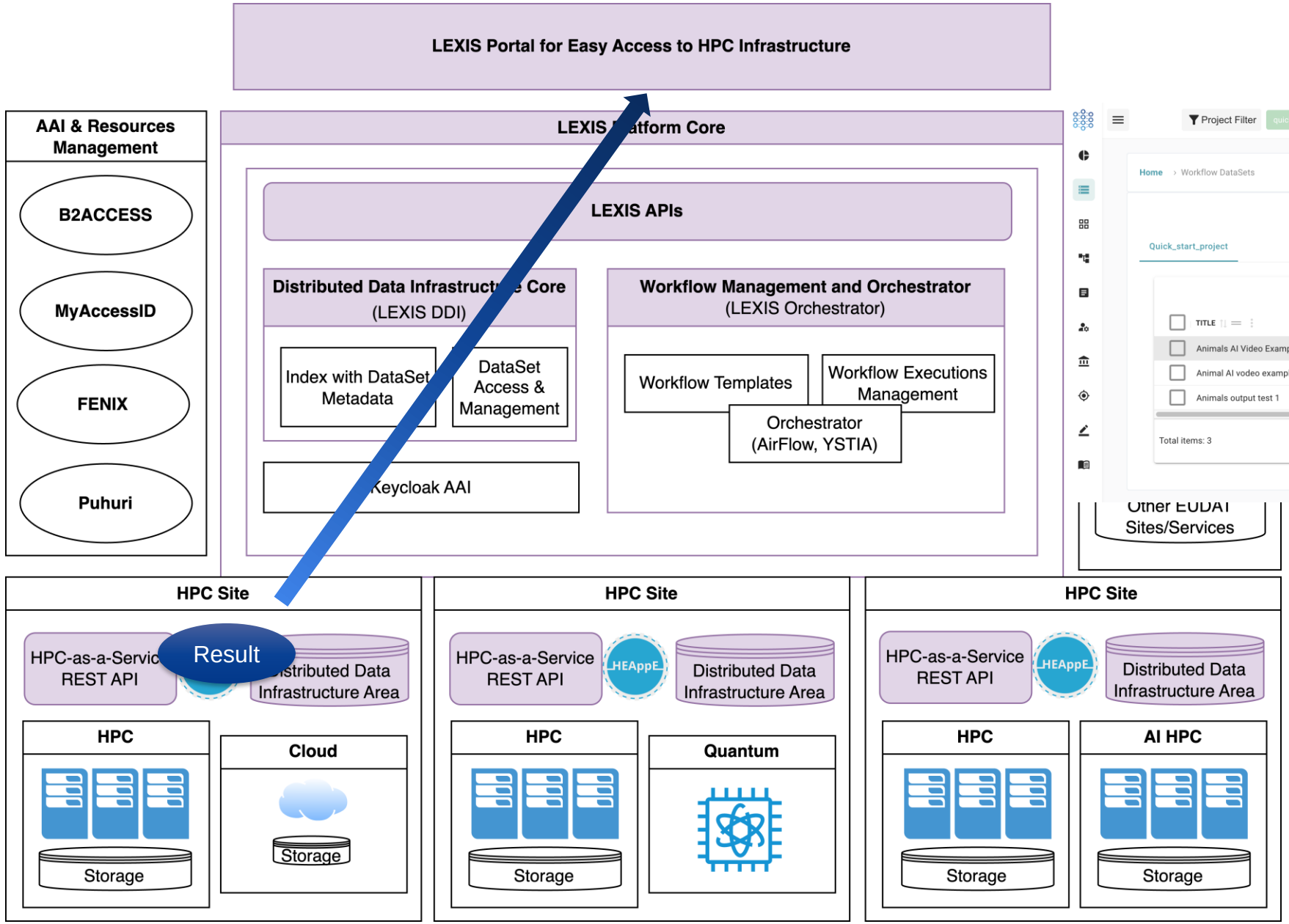


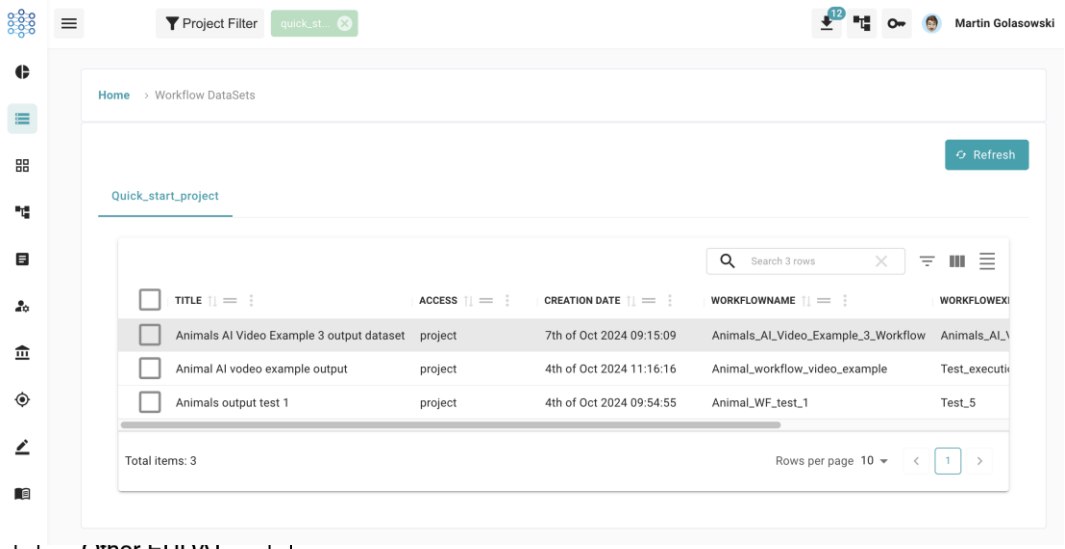


LEXIS Platform

User story - Result download

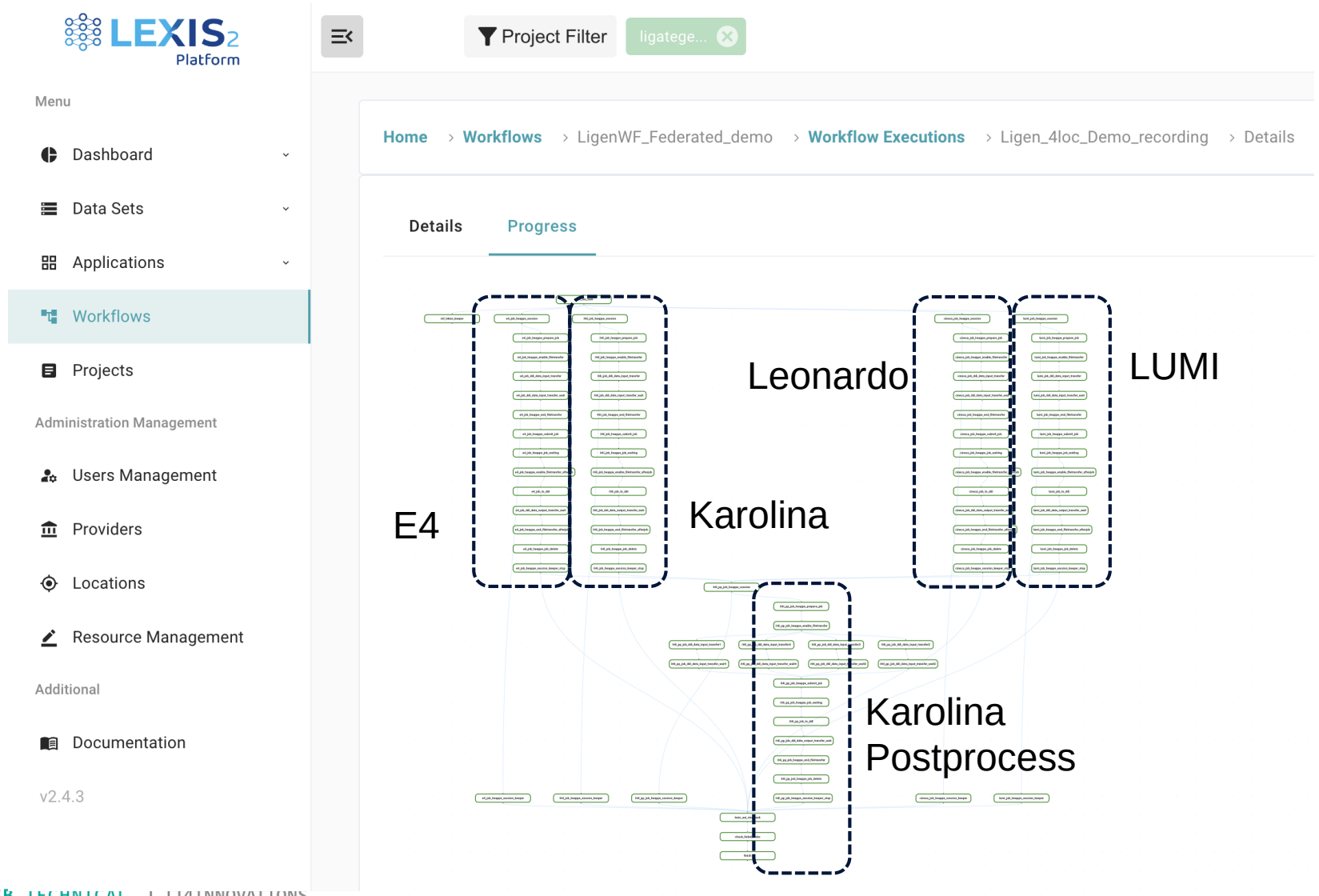
Dompé





LIGATE

Federated execution on 4 HPC clusters at once



LEXIS2 Platform

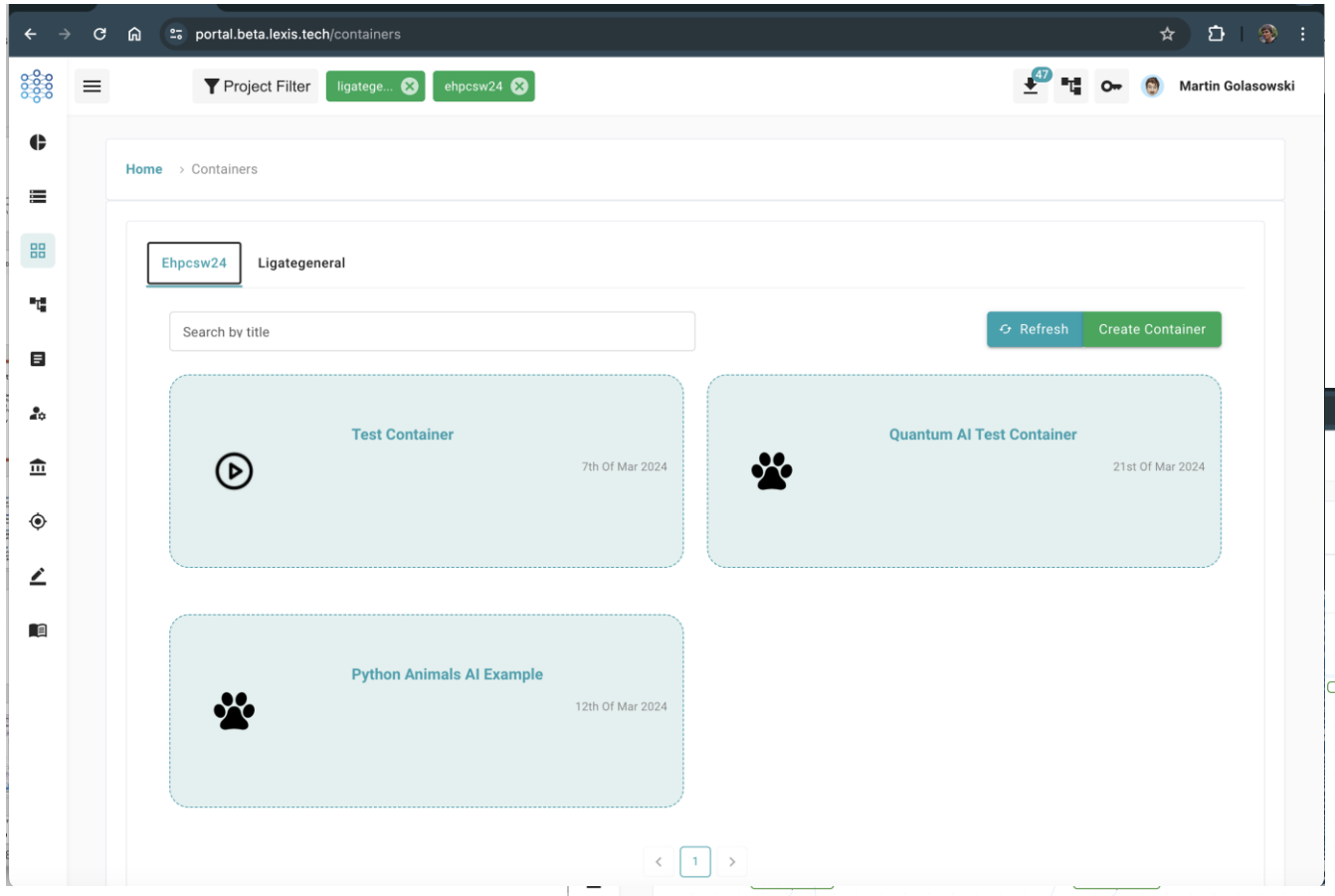
Executed on 4 locations

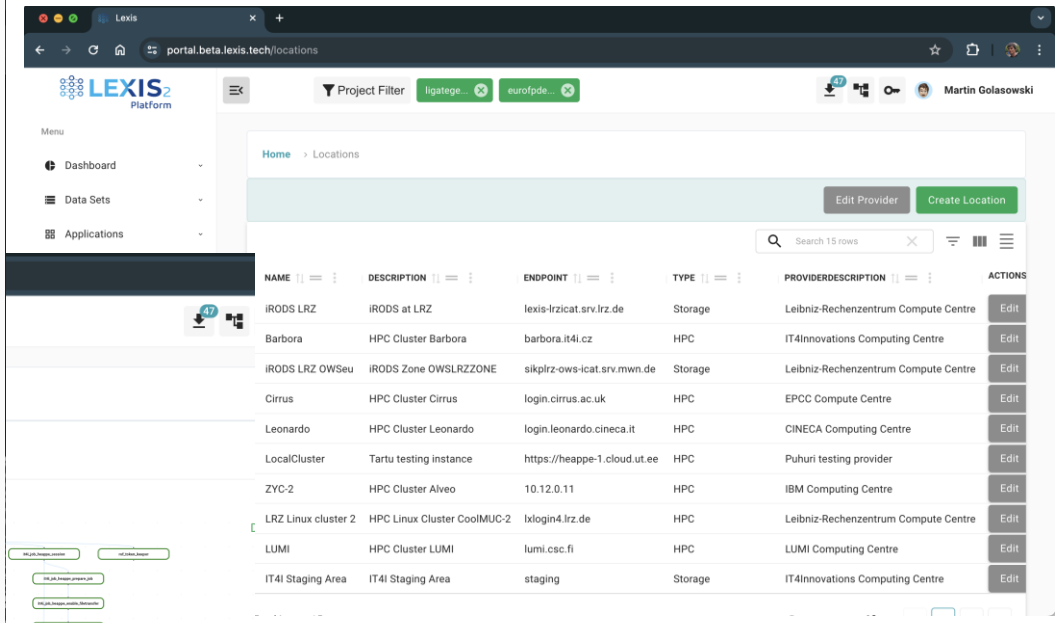
- LUMI
- Leonardo (CINECA)
- Karolina (IT4I)
- E4 AMD Cluster

LEXIS Platform

Containers & Scripts

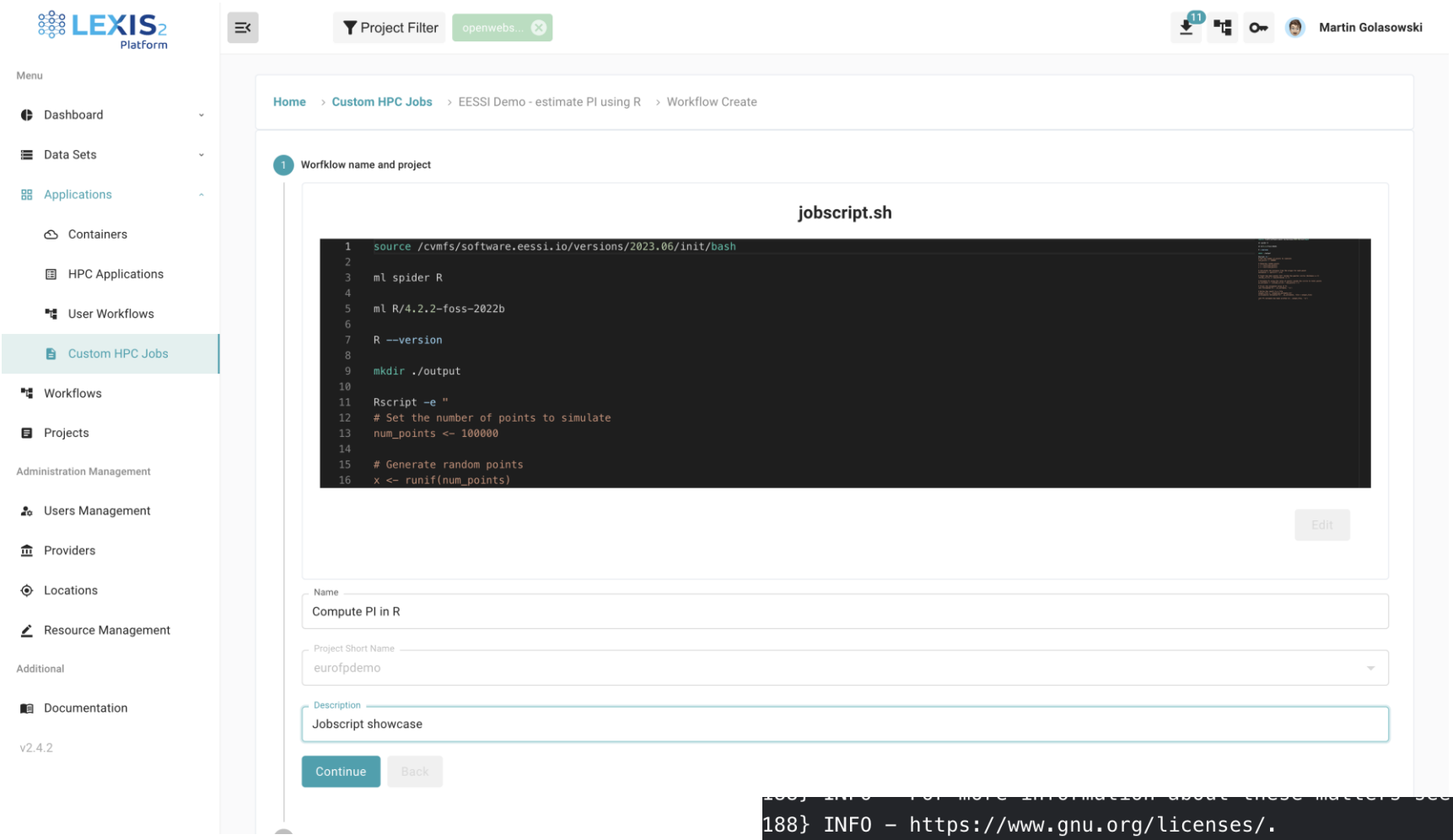
LEXIS2 Platform



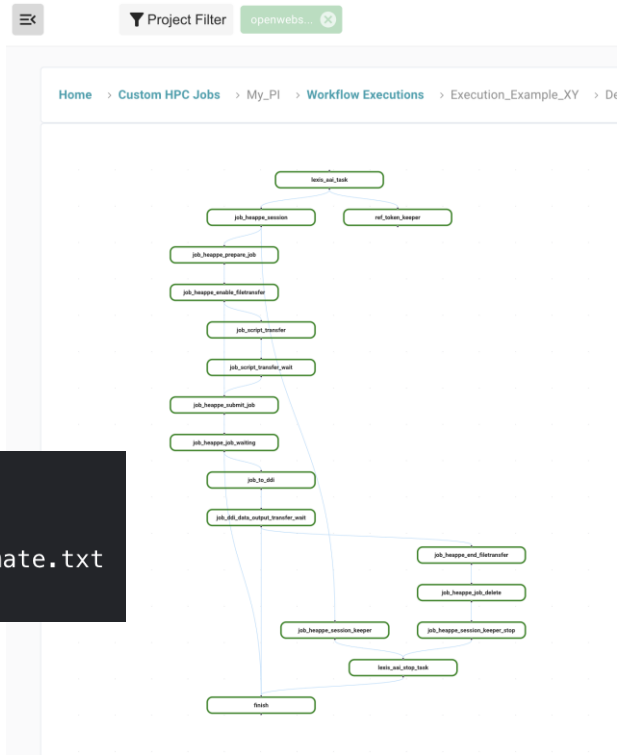


LEXIS Platform

Containers & Scripts

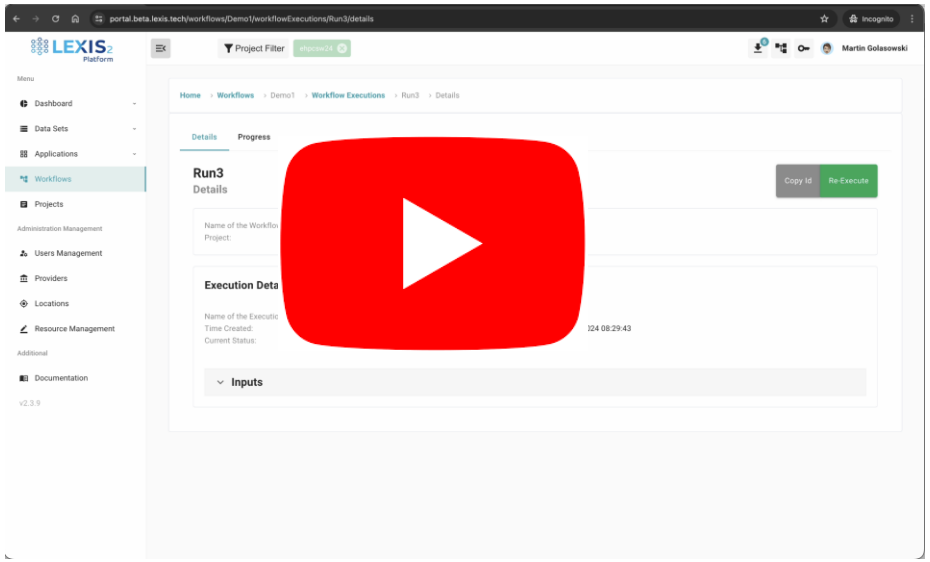


- Run custom R script
- Use common EESSI HPC module
- Workflow & output in GUI




```
188} INFO - https://www.gnu.org/licenses/.
188} INFO - Estimated Pi: 3.1408
188} INFO - Pi estimate has been written to output/pi_estimate.txt
 - Success criteria met. Exiting.
```

AI workflow demo with Python in Apptainer

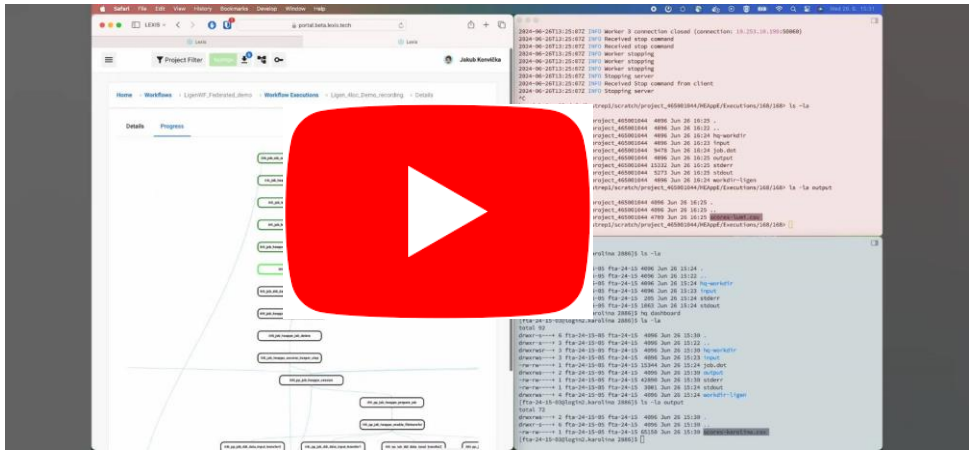


Check it out!

LEXIS2 Platform

Documentation
https://docs.lexis.tech

Urgent computing federated workflow execution



Check it out!

National Czech Programme | eosc

Thank you!

Questions?

martin.golasowski@vsb.cz

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

MUNI ICS

cesnet

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | IT4INNOVATIONS NÁRODNÍ SUPERPOČÍTAČOVÉ CENTRUM