

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



Extreme Data Analytics
as Enabler for
Science/SMEs/Industry

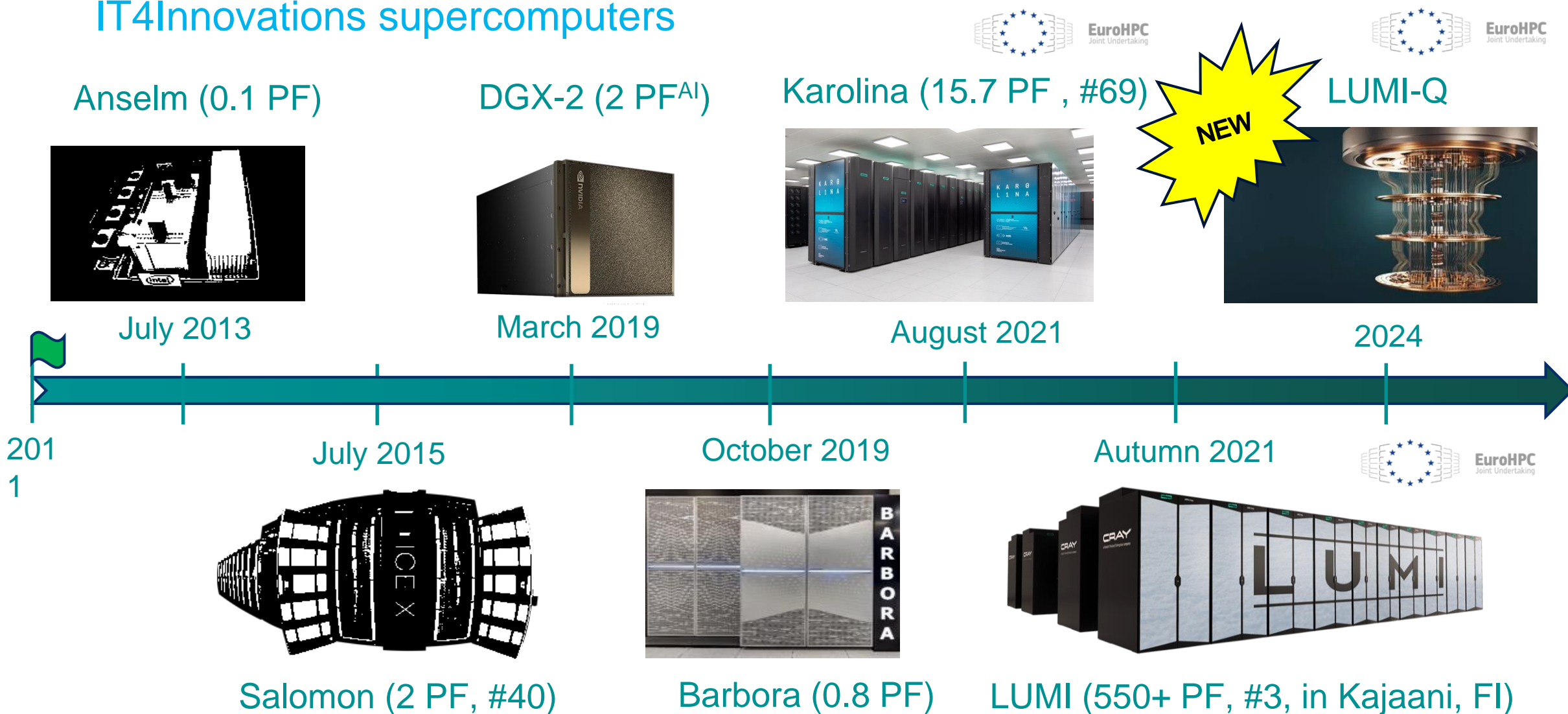
At Top-Level EU Supercopmting
Centres,
Cross-System
With Automatised Workflows

Connecting to
European Data Spaces,
EOSC and EUDAT



Data Management in HPC

IT4Innovations supercomputers

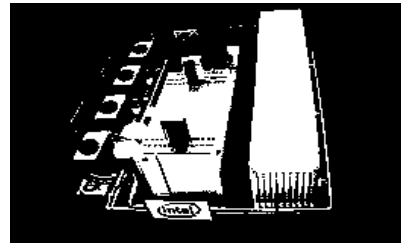


Data Management in HPC

IT4Innovations supercomputers & Storages



Anselm (0.1 PF)



July 2013

DGX-2 (2 PFAI)



March 2019

Karolina (15.7 PF , #69)



August 2021

LUMI-Q



2024

201
1

0,46
PB

July 2015

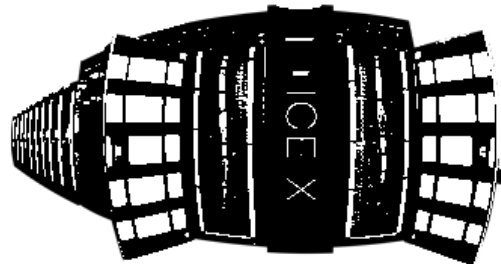
2,56
PB

October 2019

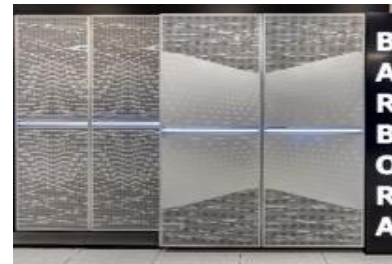
16 PB

Autumn 2021

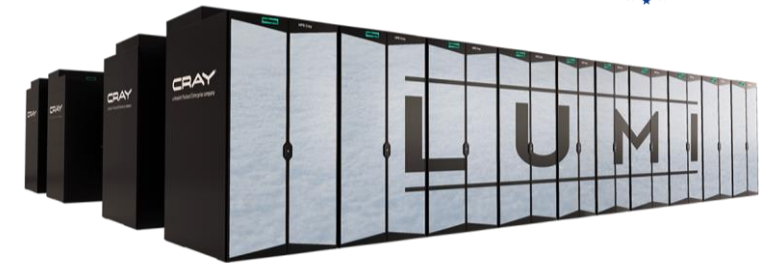
29 PB



Salomon (2 PF, #40)



Barbora (0.8 PF)

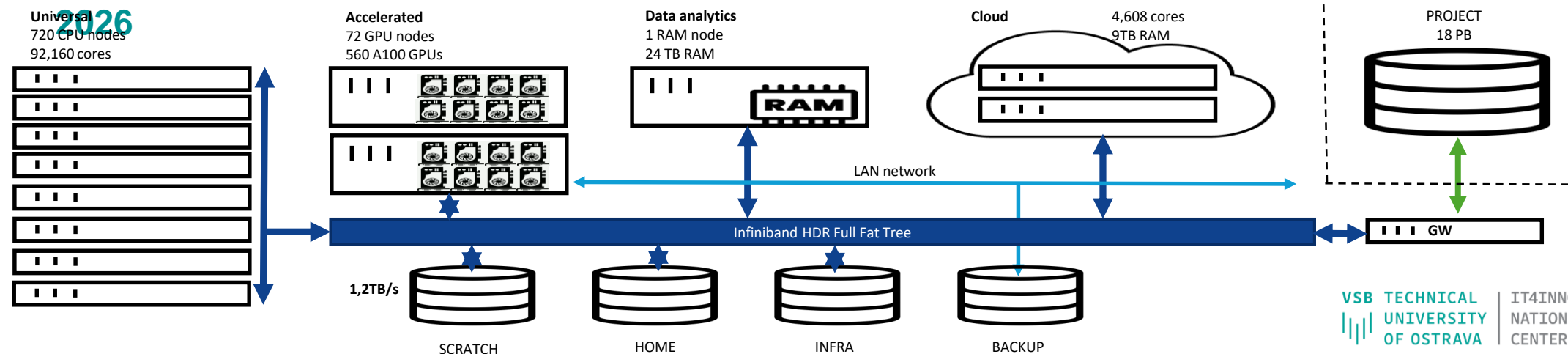
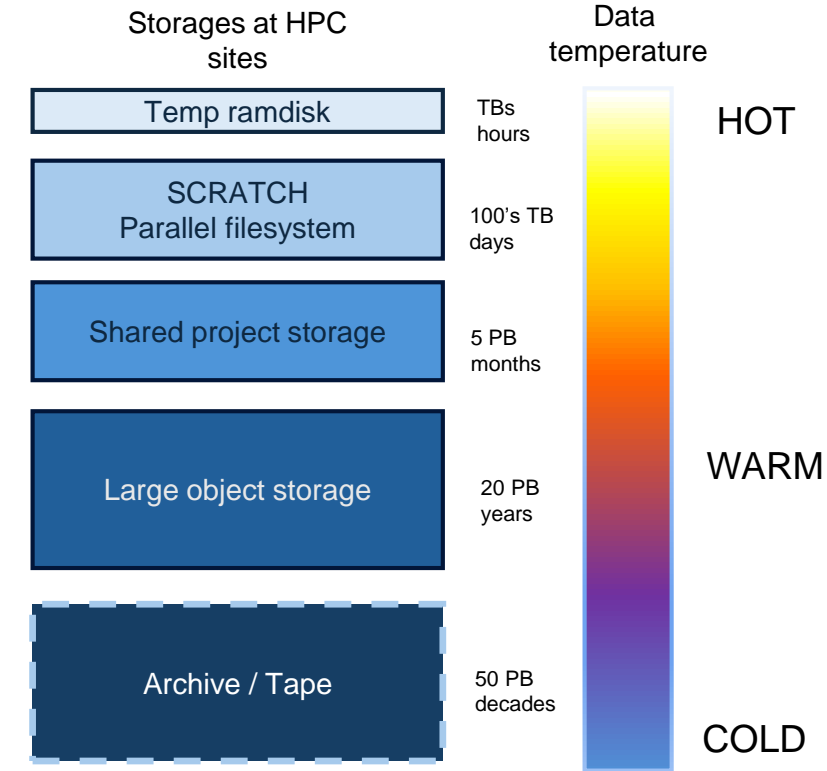


LUMI (550+ PF, #3, in Kajaani, FI)

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



Data Management in HPC

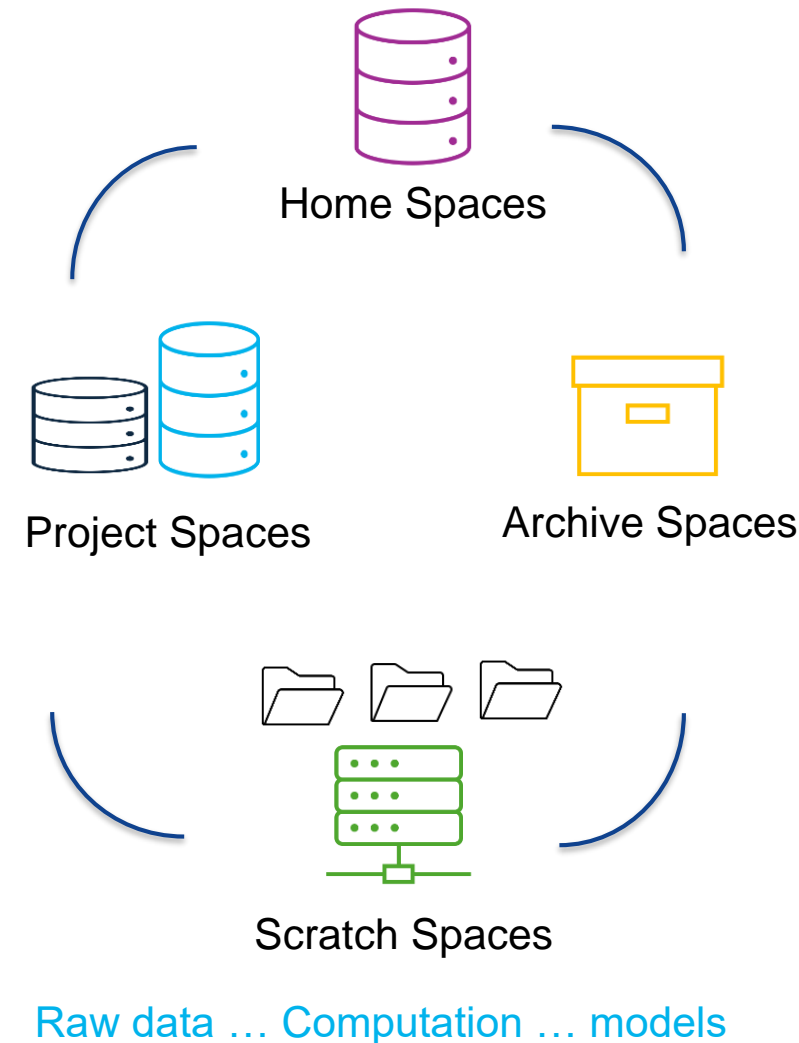
Support & User perspective

➤ 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.

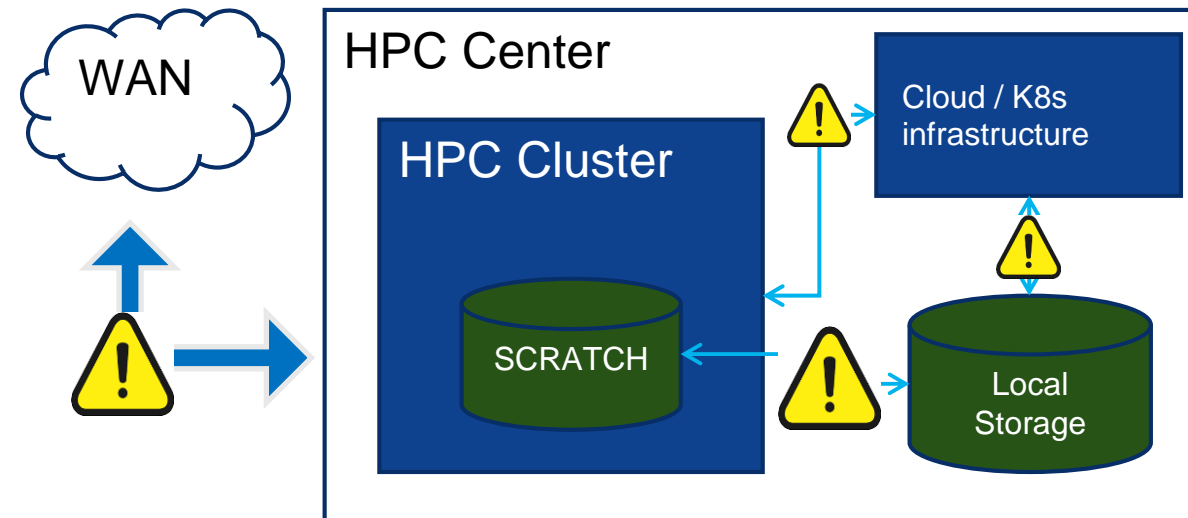


Data Management in HPC – selected techniques

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?
- 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
- Routing between European HPC centres is not always efficient
- **Tooling for staging/transfer** TBs of data does not exist

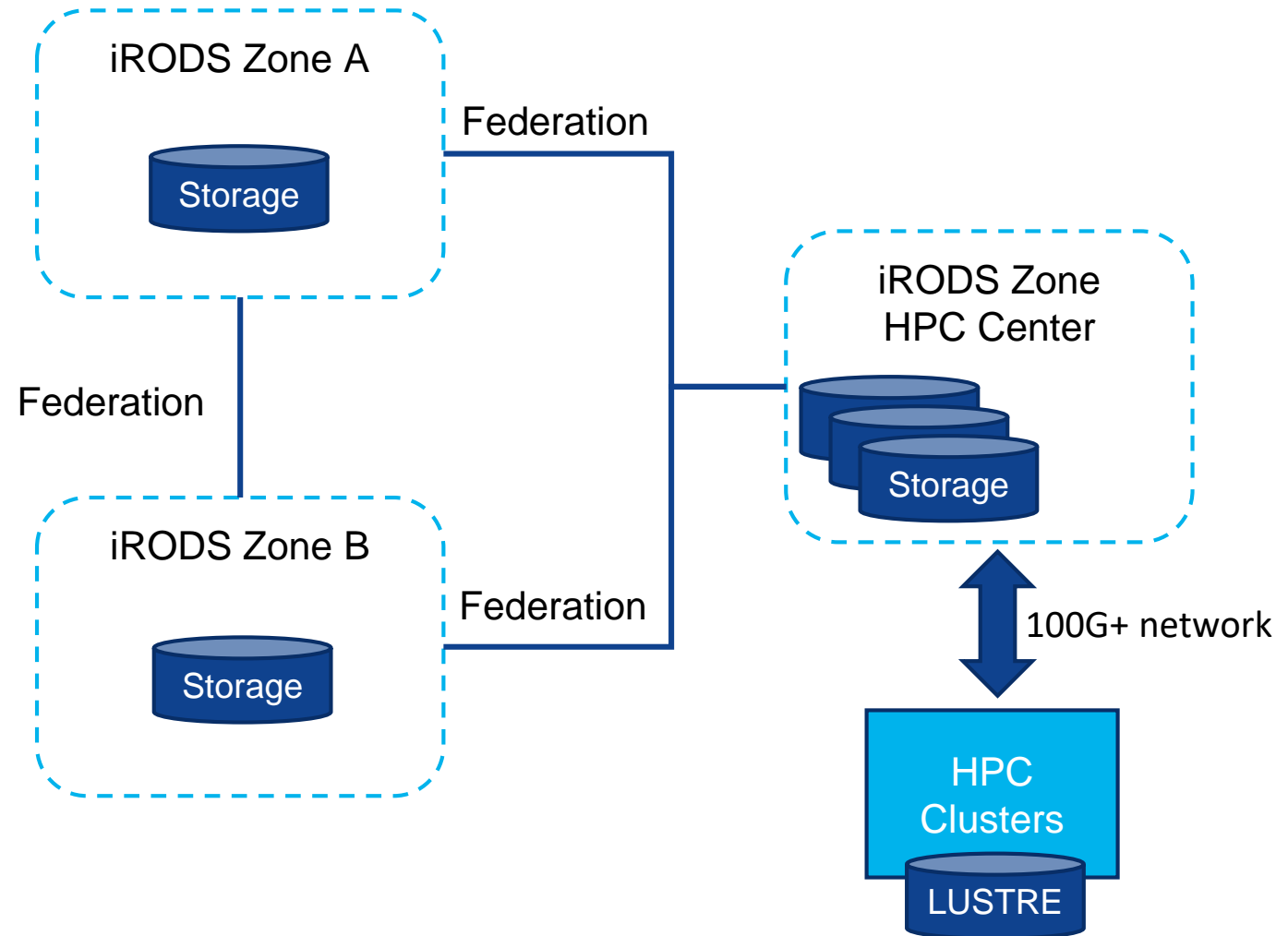
Data transfer challenges



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

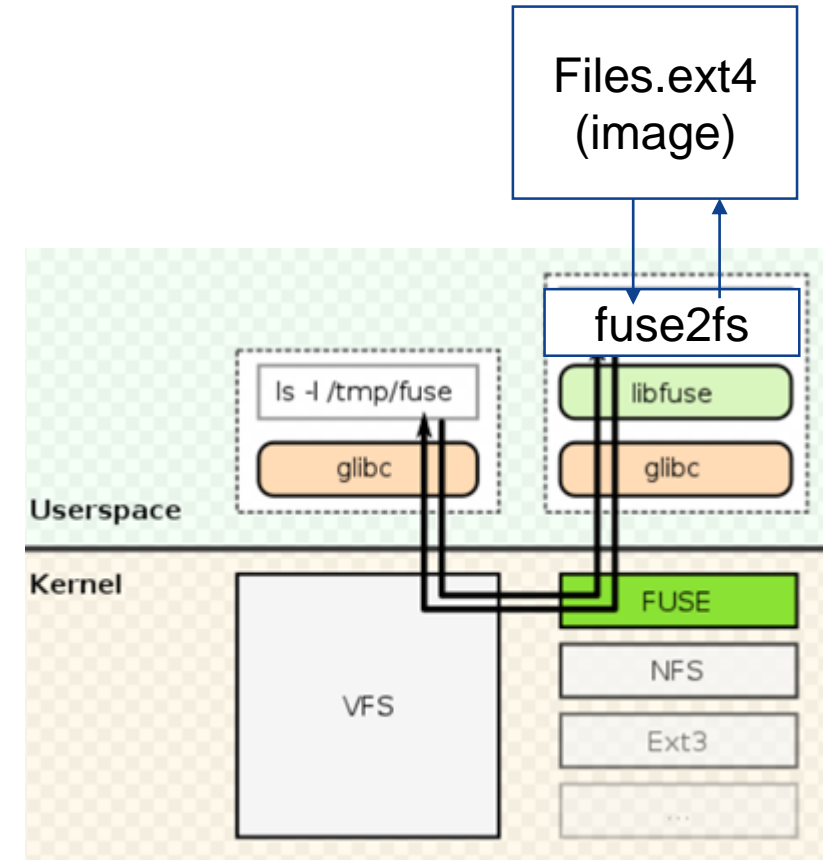


Data Management in HPC – selected techniques

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

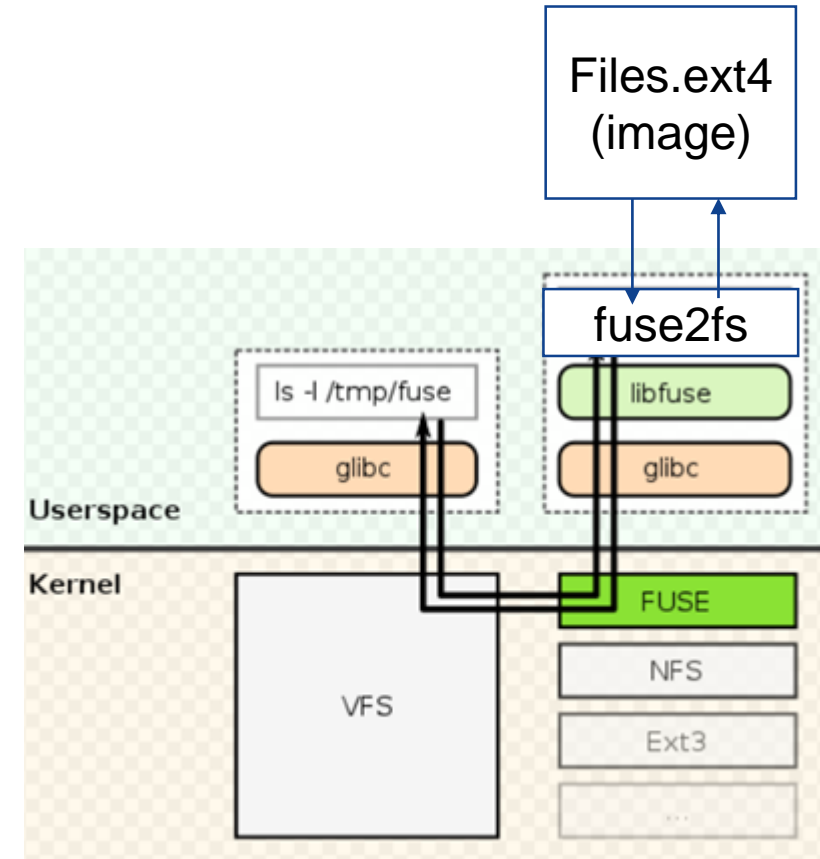
Handling large number of files on HPC clusters

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



HyperQueue



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

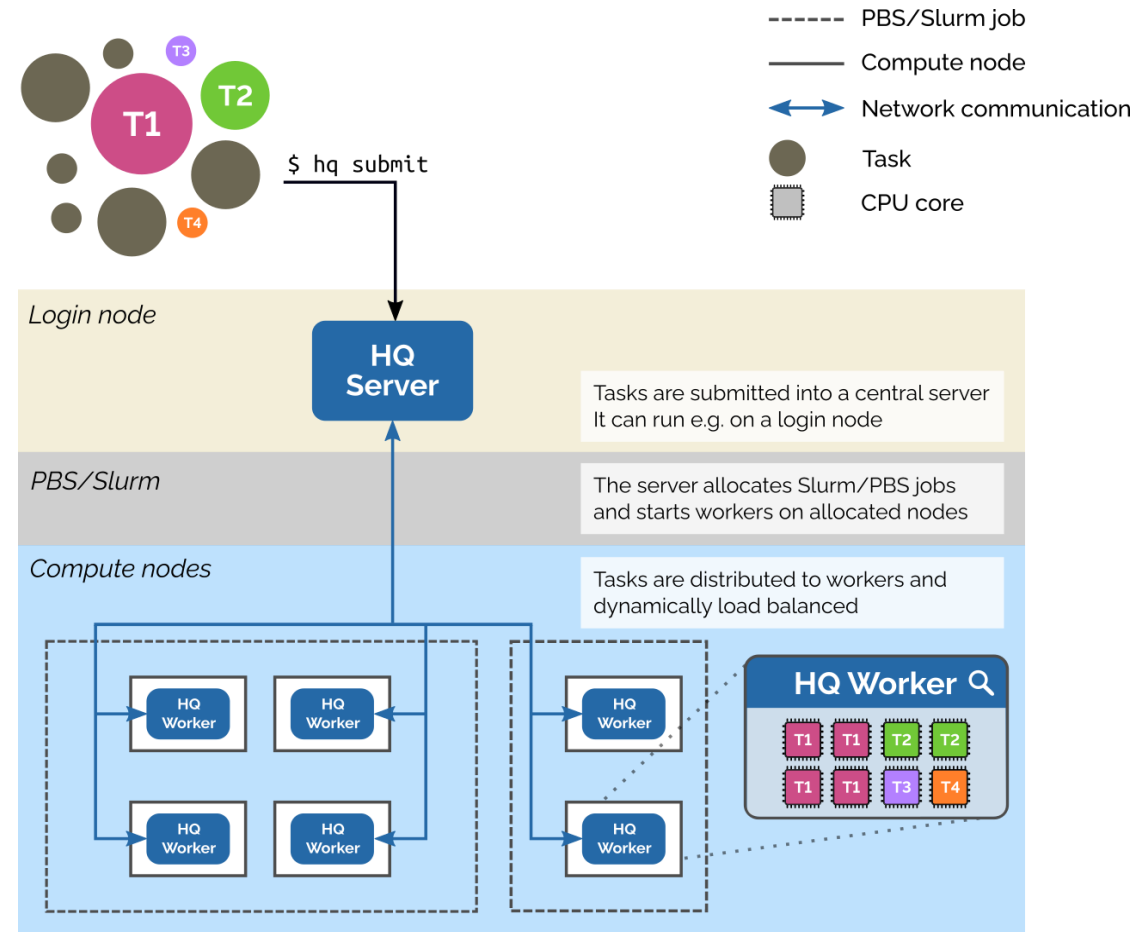
HyperQueue is an HPC-tailored distributed task runtime designed for simple and efficient execution of task graphs (workflows) on HPC clusters

```
#!/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/lt4innovations/hyperqueue>

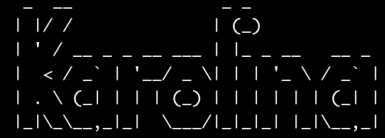
How to supercompute ...

- 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



```
-----  
GROMACS/4.5.5-gompi-2020c-ORCA-5.0.1  
GROMACS/2021.4-foss-cuda-2020b-PLUMED-2.7  
GROMACS/2022-foss-cuda-2020b  
GROMACS/2023-foss-2022a-CUDA-12.2.0-PLUMED-2.7  
GROMACS/2023-foss-2022a-CUDA-12.2.0-PLUMED-2.7  
GROMACS/2023.4-gompi-2020c-ORCA-5.0.1-don  
GROMACS/2024.1-foss-2023b  
-----  
OpenFOAM/v2312-foss-2023a OpenFOAM/9-  
-----  
ABINIT/9.10.3-intel-2022a  
ASE/3.22.1-gfbf-2023b  
Amber/22.0-foss-2022a-AmberTools-23.3-CUDA  
BEEF/0.1.1-intel-2020a  
CP2K/9.1-foss-2022a  
Critic2/1.1-stable-intel-2020b  
DFT-D4/3.6.0-intel-2022b-Python-3.10.8  
DFT-D4/3.6.0-intel-2022b-Python-3.10.8  
Libint/2.7  
Molden/7.3  
Molpro/mpp  
Molpro/mpp  
Molpro/mpp  
lines 1-27
```

```
Your public key has been saved in the  
The key fingerprint is:  
SHA256: fFp7L5cfGT0jmWRJ1TvVwHRJCzUR  
The key's randomart image is:  
+---[RSA 3072]----+  
|                 .  
|                .+.o+=00|  
|                 ..++*|  
|                 ..o|  
|                 ..o|  
+---+-----+  
|                 .  
|                .+.o+=00|  
|                 ..++*|  
|                 ..o|  
|                 ..o|  
+---+-----+
```



...running on Rocky 8.X

```
Public Service Announcement: Apptainer on the Karolina cluster  
Posted: (2024-05-10 10:23:47)  
  
Apptainer is now a part of the operating system, you do not need to load the  
module.  
  
$ apptainer --version  
apptainer version 1.3.1-1.el8  
  
Last login: Mon May 27 10:05:28 2024 from 89.24.247.60  
[mgolas@login2.karolina ~]$
```

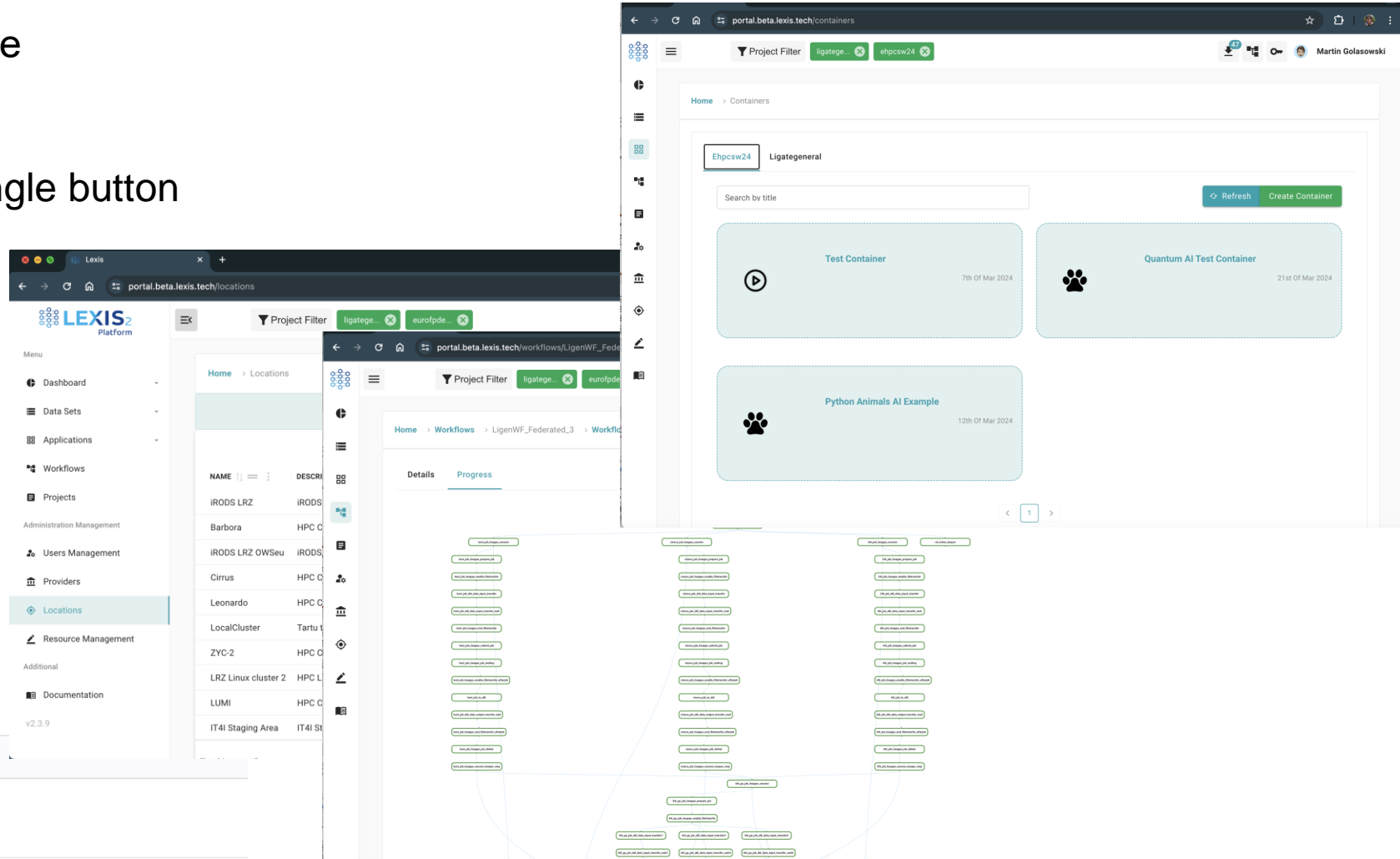
Easy and Safe access

- 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

The image displays several screenshots of the LEXIS2 Platform web interface. The top right screenshot shows the 'Containers' page with a search bar and buttons for 'Refresh' and 'Create Container'. Below the search bar are three container cards: 'Test Container' (7th Of Mar 2024), 'Quantum AI Test Container' (21st Of Mar 2024), and 'Python Animals AI Example' (12th Of Mar 2024). The middle screenshot shows the 'Locations' page with a table of locations:

NAME	DESCR
iRODS LRZ	iRODS
Barbora	HPC C
iRODS LRZ OWSeu	iRODS
Cirrus	HPC C
Leonardo	HPC C
LocalCluster	Tartu
ZYC-2	HPC C
LRZ Linux cluster 2	HPC L
LUMI	HPC C
IT4I Staging Area	IT4I St

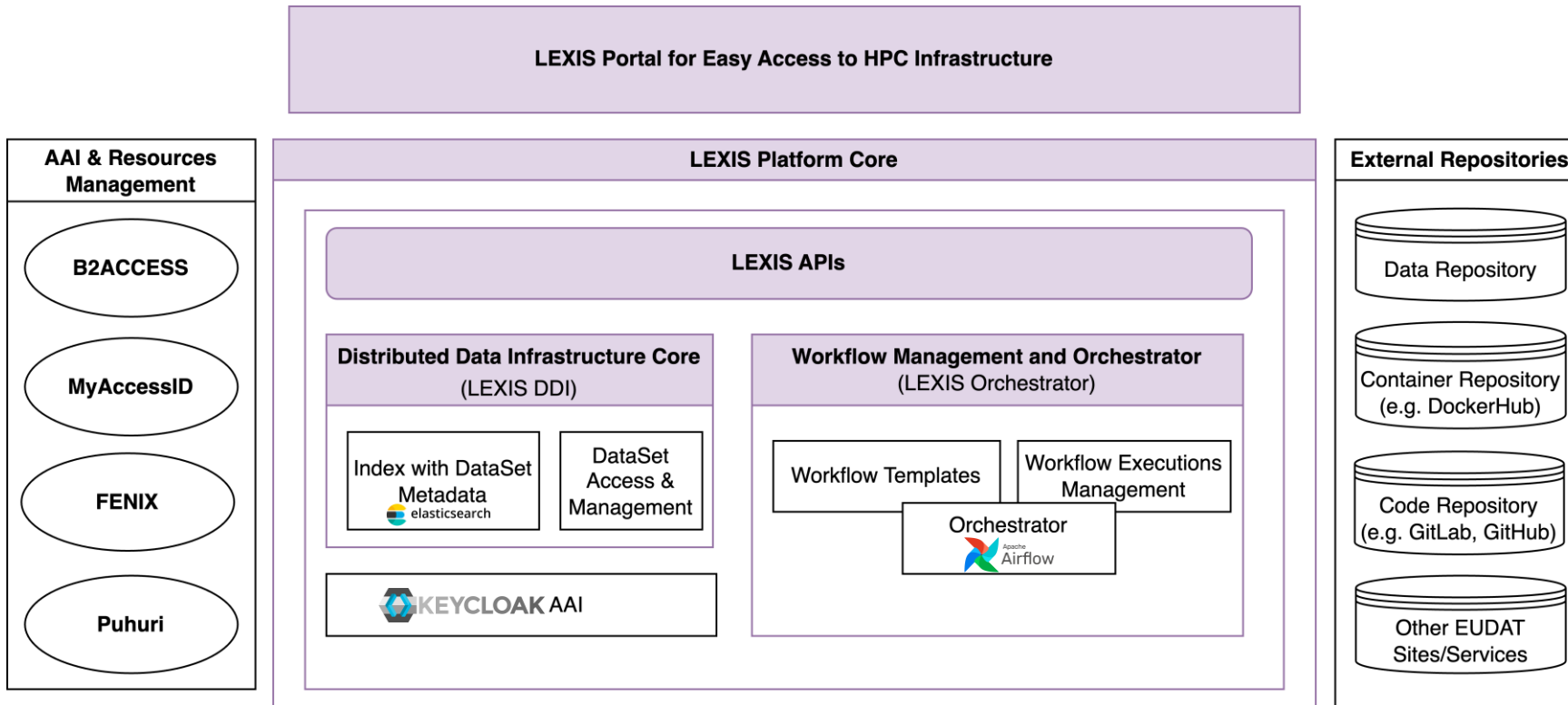
The bottom screenshot shows a detailed workflow diagram with multiple steps and nodes.

README.md

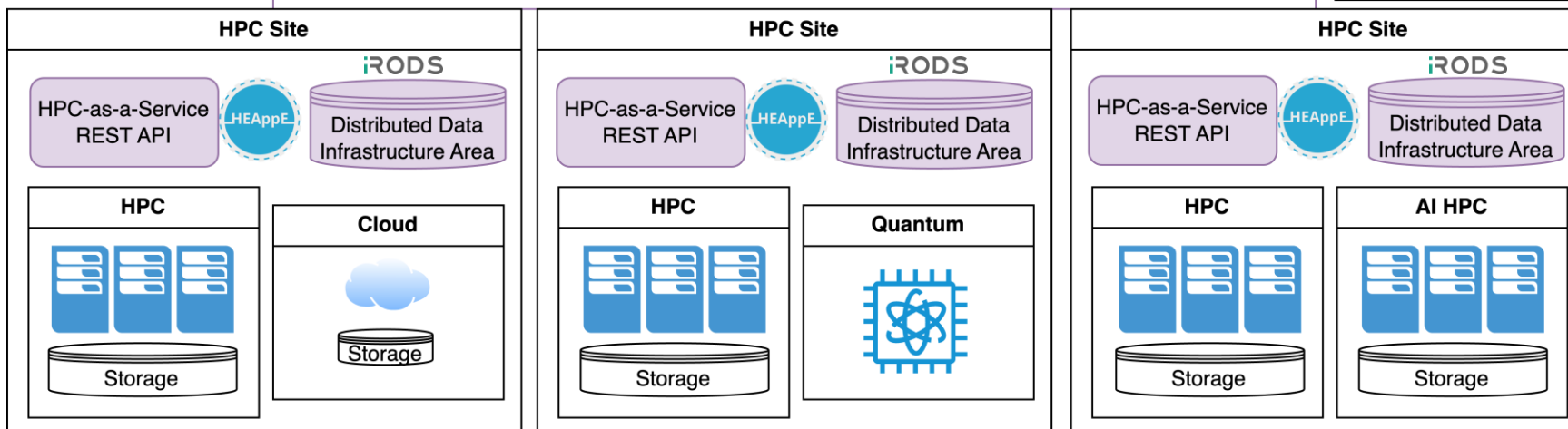
py4lexis

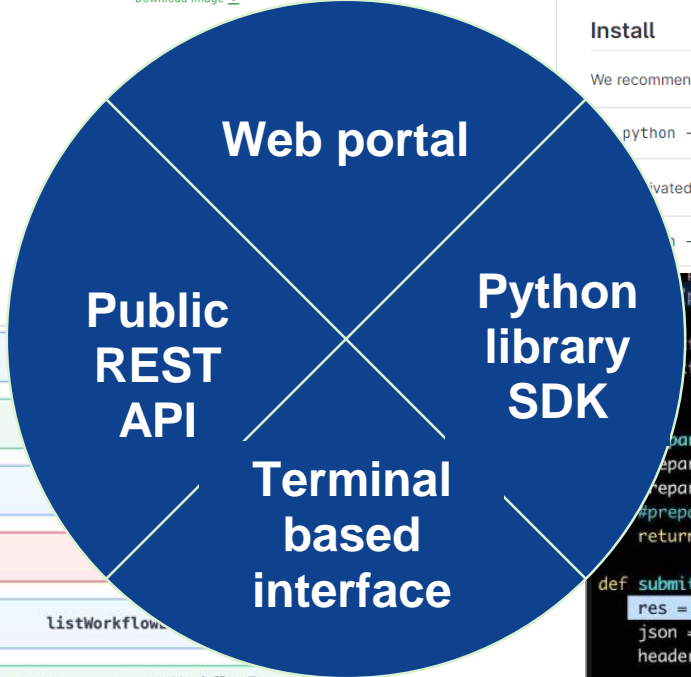
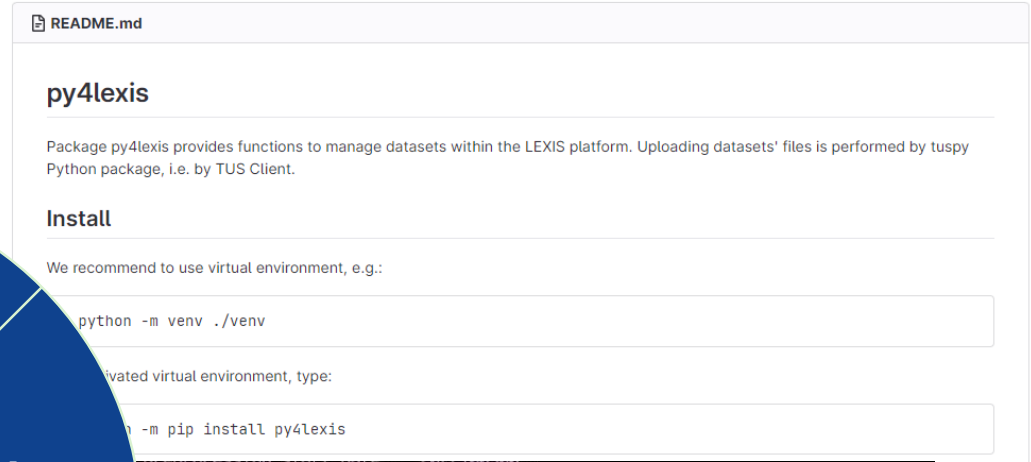
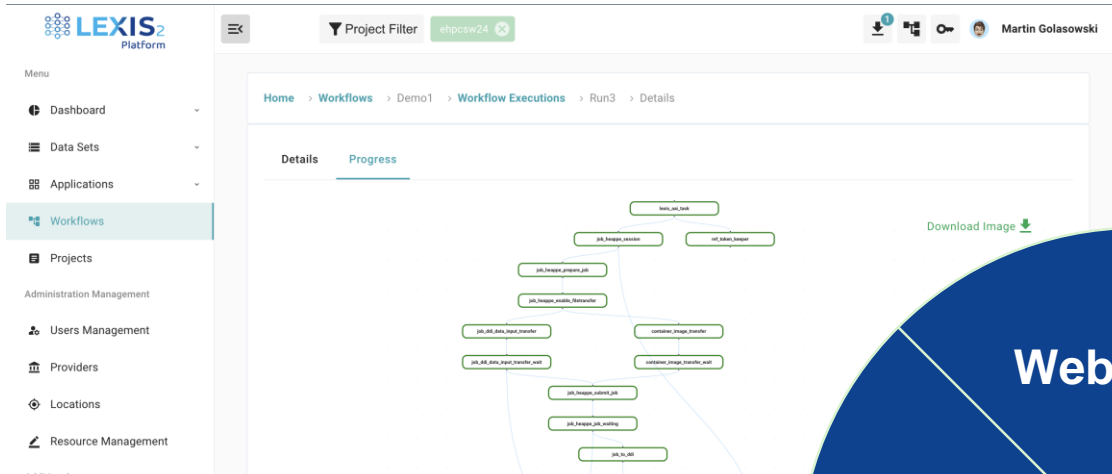
Package py4lexis provides functions to manage Python package, i.e. by TUS Client.

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)**





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.	listWorkflowExecutions
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

```
preprocessing_start_date": "20221008",
preprocessing_dataset_path_geographical_data_path": "project/proj934a1763631a81e
flowExecutionName": "Run3",
flowExecutionConsent": True

prepare_data():
prepared = WRF_EXEC_PARAMS
prepared["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: | ██████████-----| 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 through the LEXIS Portal
- LLM/AI applications **without access** to data

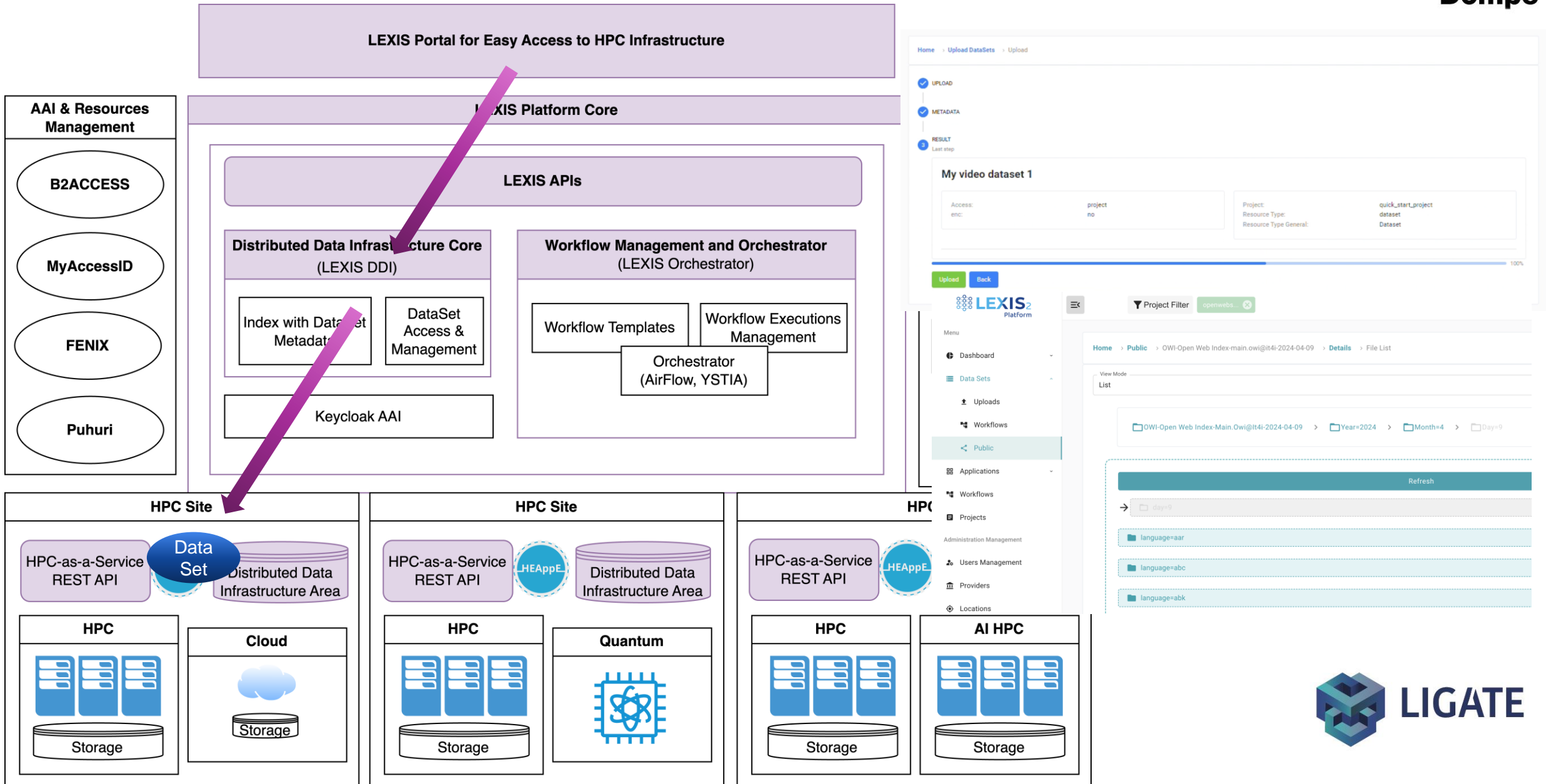


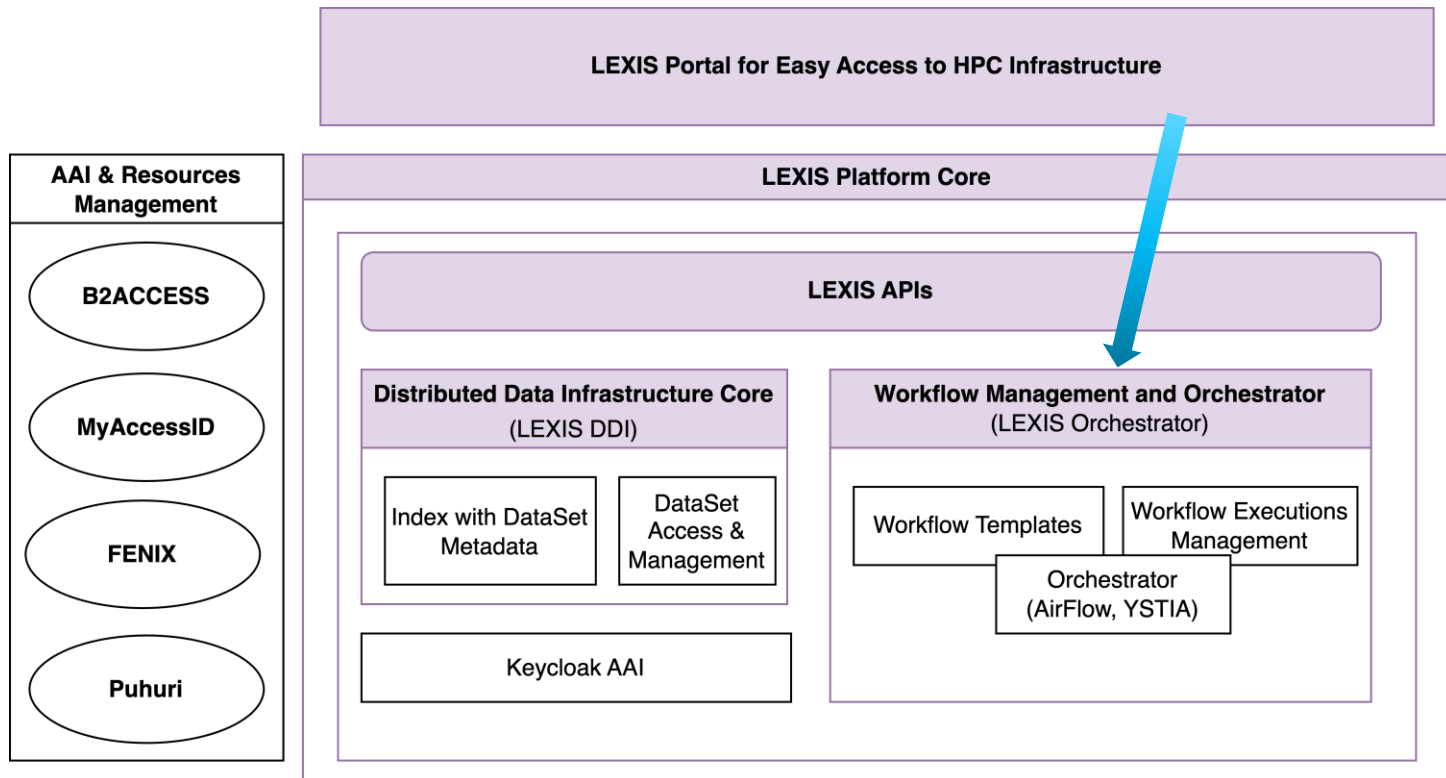
LEXIS Platform

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





× Create workflow execution

Name
Test Run 1

Input path
Animals AI input ✕ ▼

■ de19e3b2-8177-11ef-9de3-0242c0a8700a

application_executable_args
animals_dir annots.csv

Output dataset metadata :

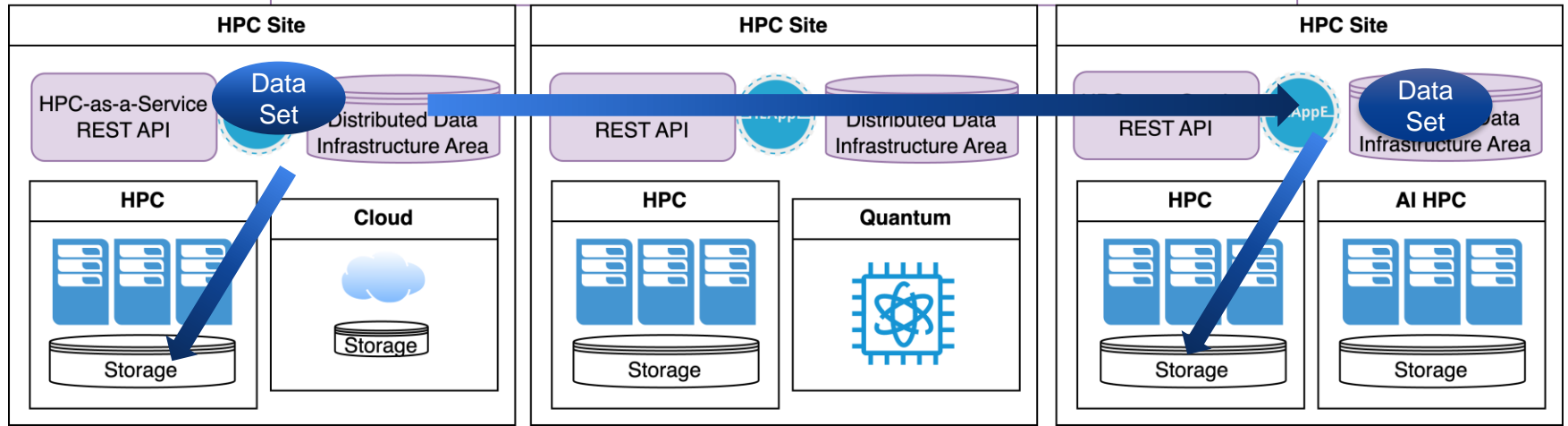
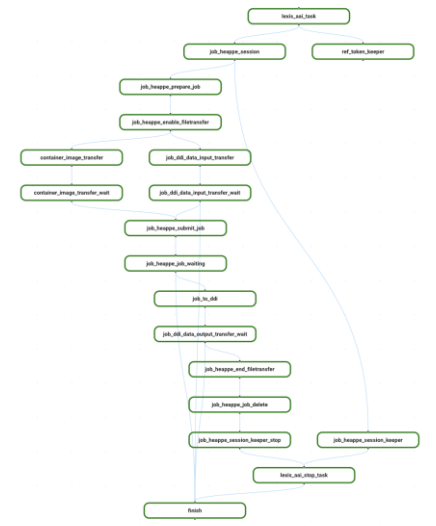
title :

value
Animals AI Video Example 3 output dataset

Other EUDAT Sites/Services

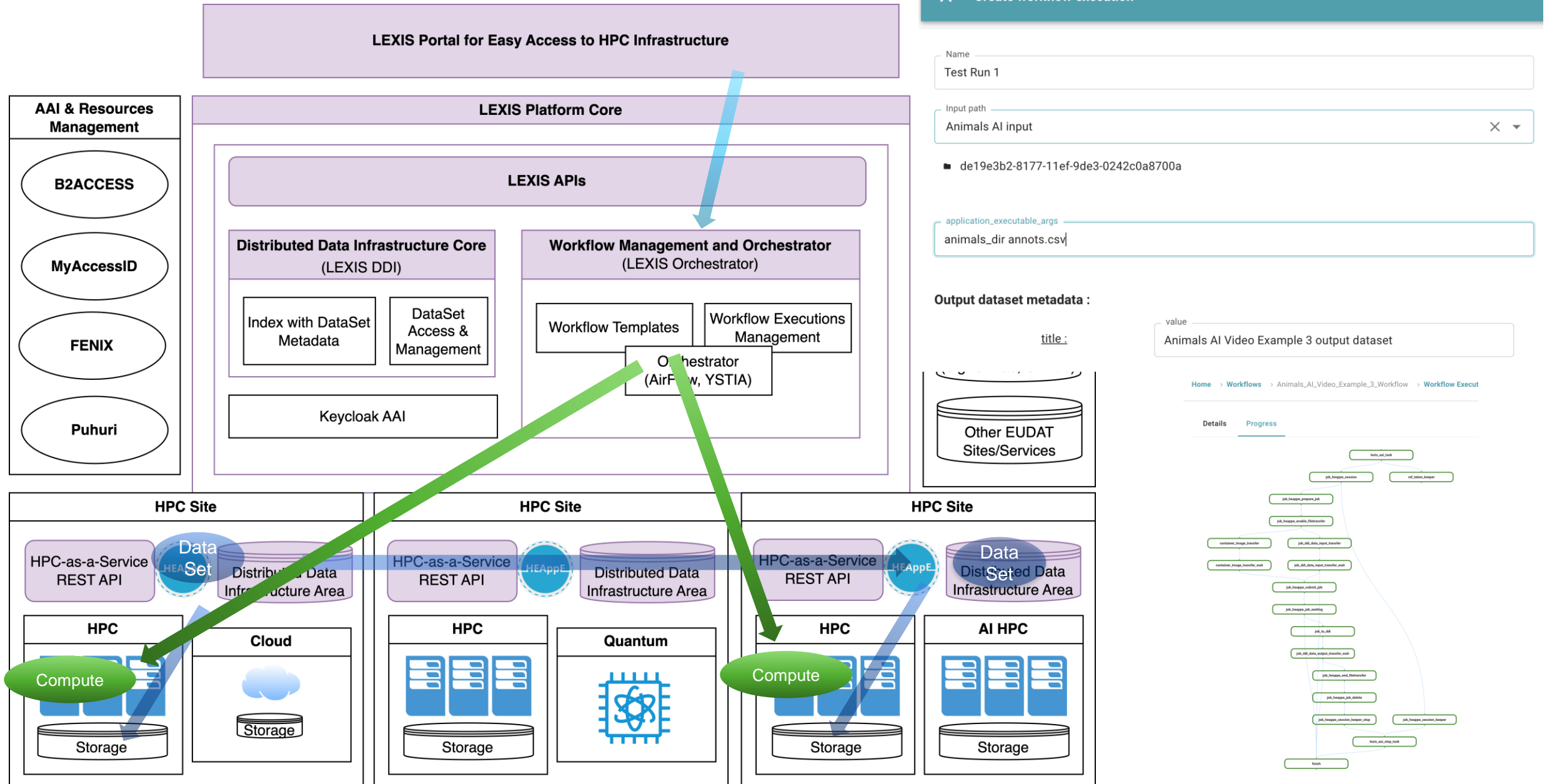
Home > Workflows > Animals_AI_Video_Example_3_Workflow > Workflow Execut

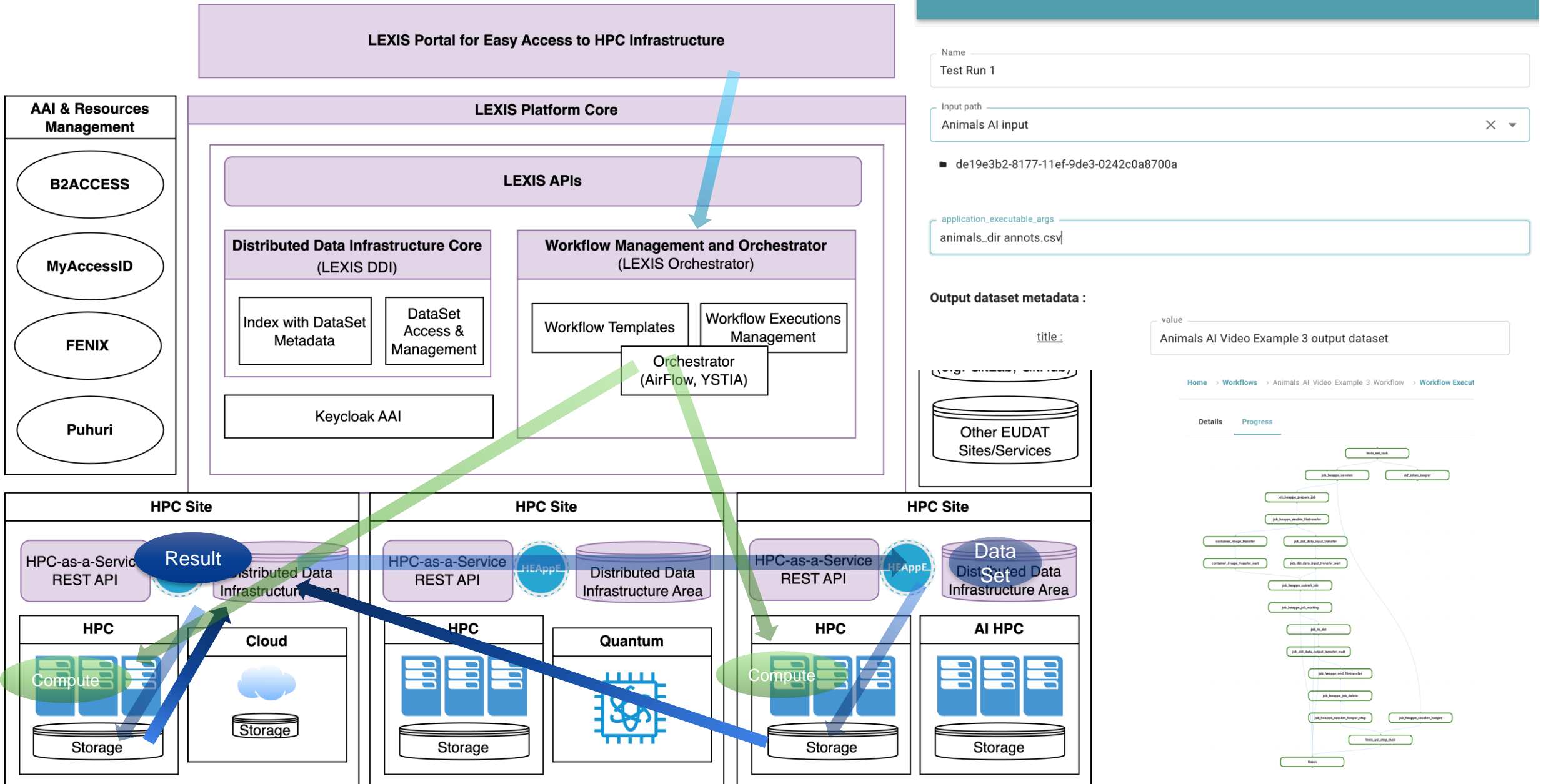
Details Progress

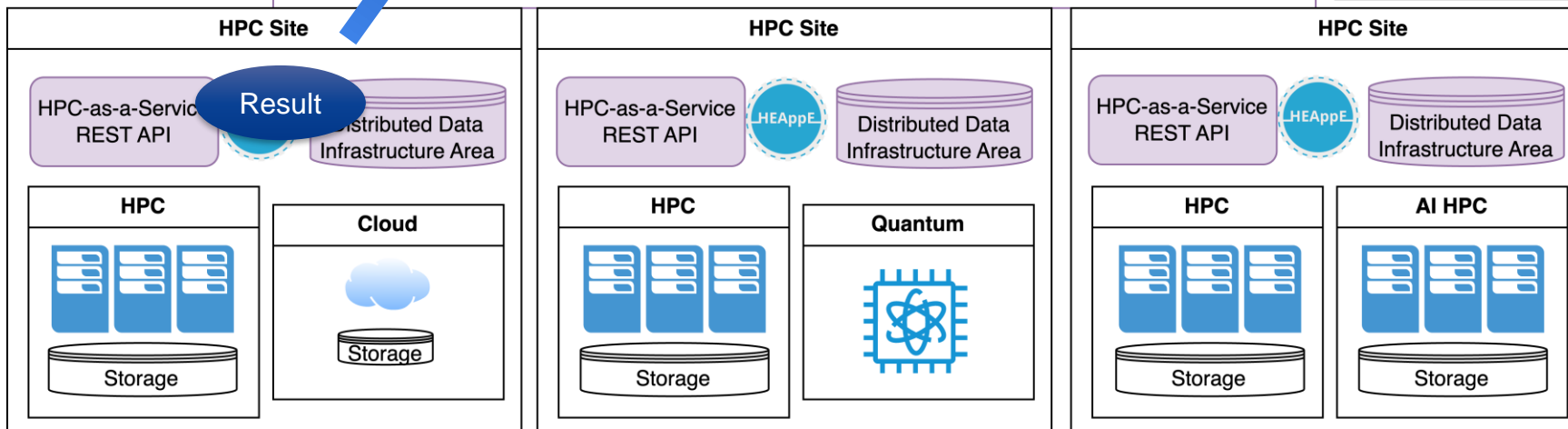
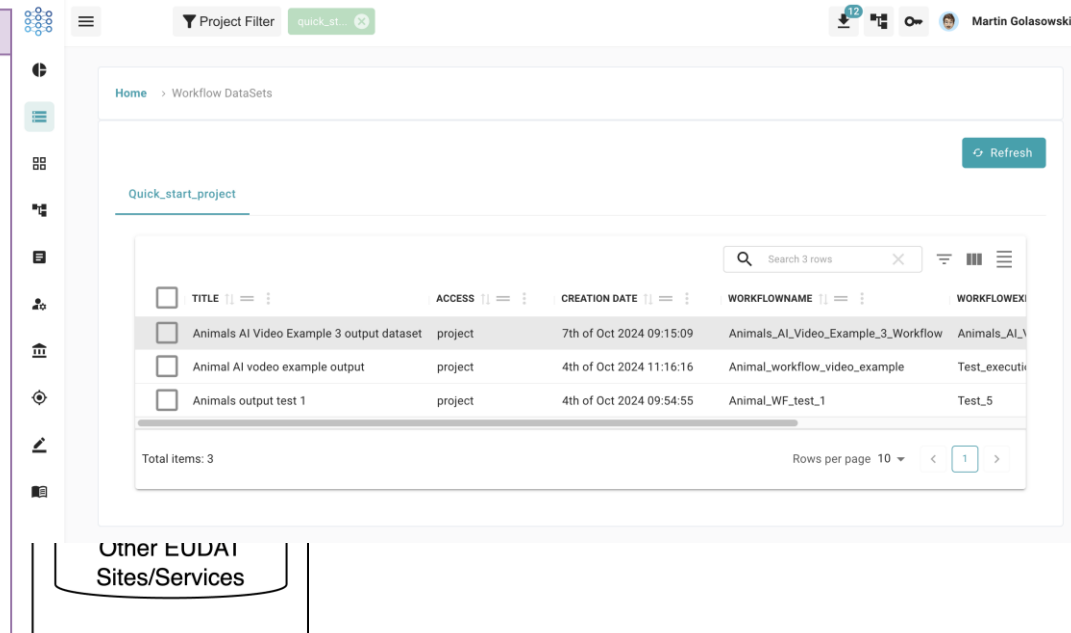
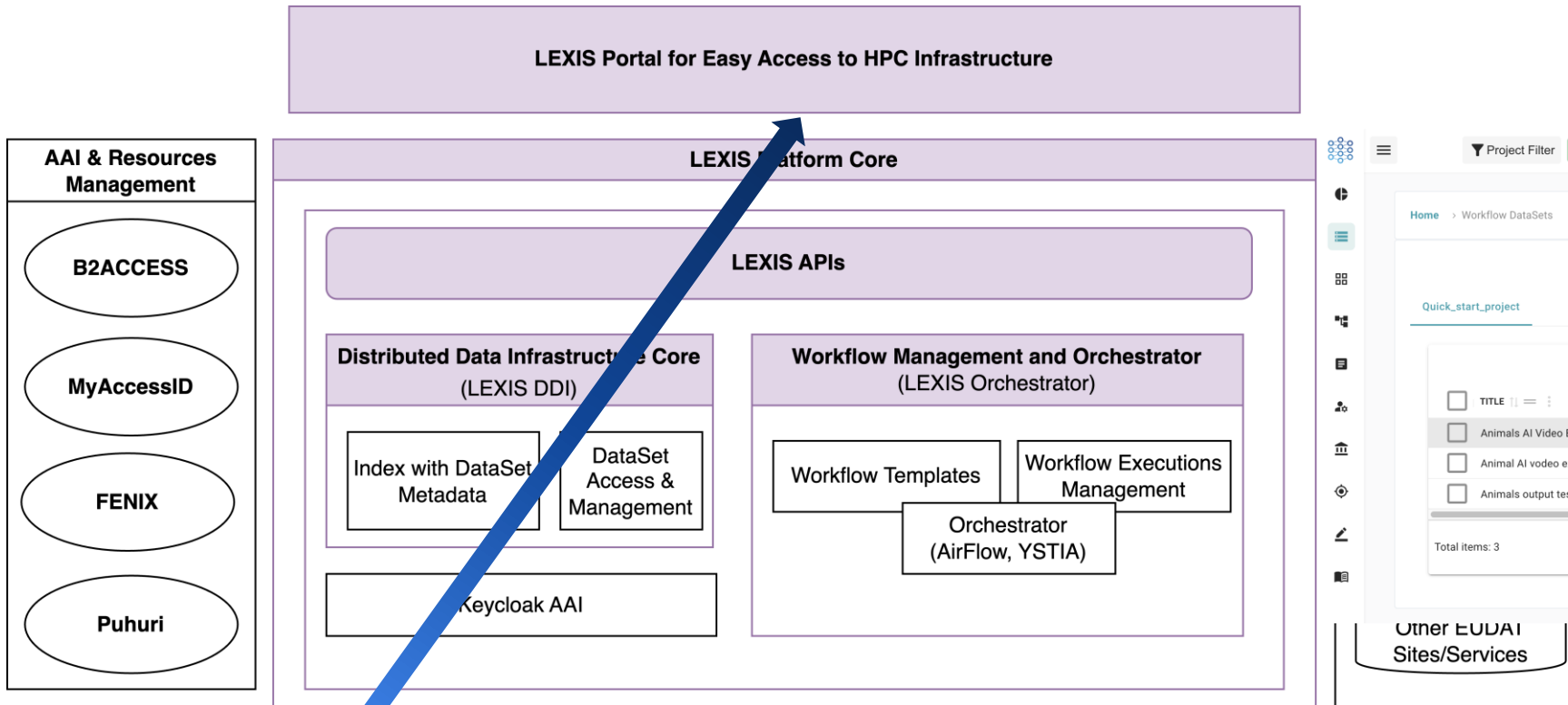


LEXIS Platform

User story - Workflow execution



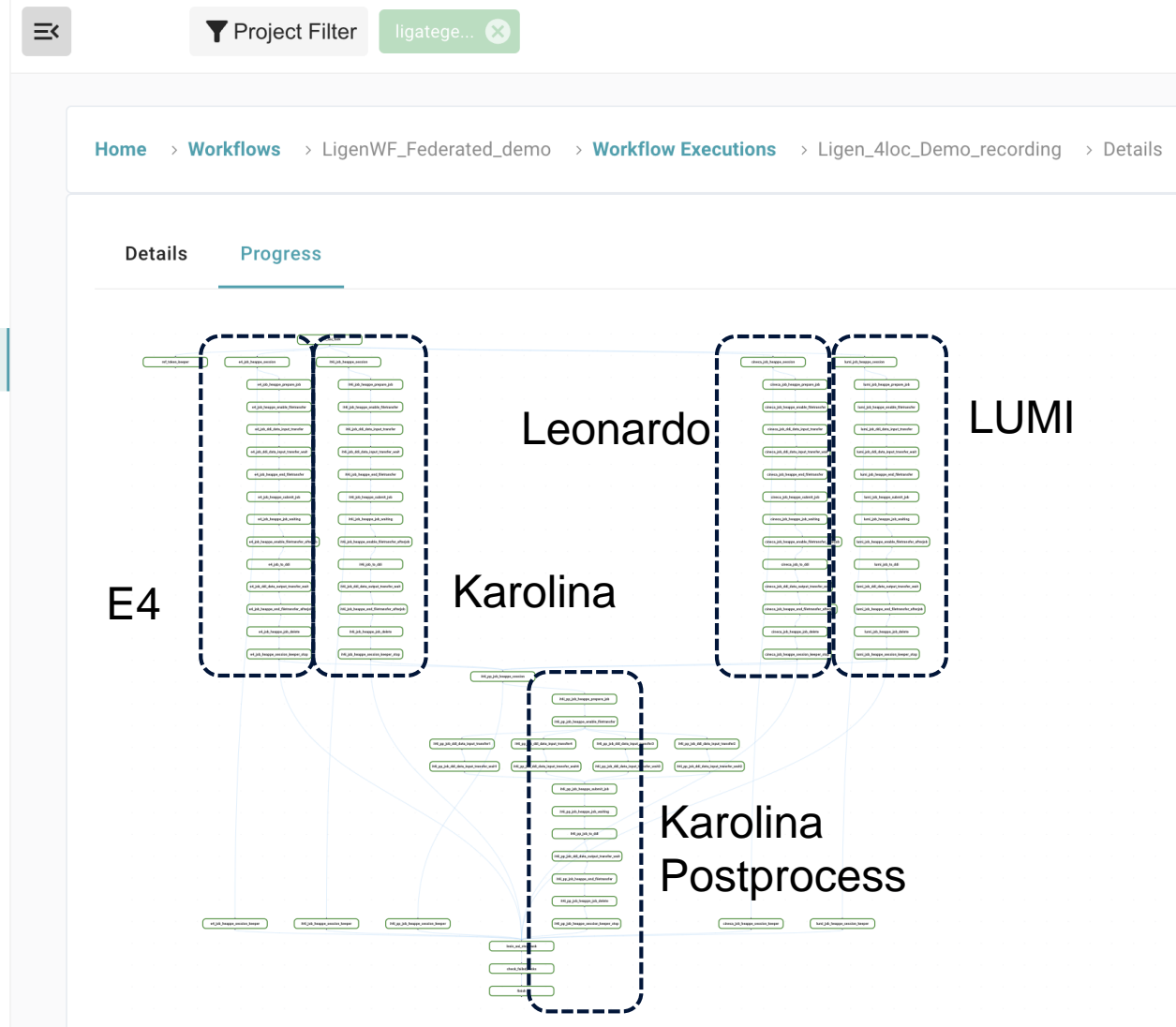




Federated execution on 4 HPC clusters at once



- Menu
- Dashboard
- Data Sets
- Applications
- Workflows**
- Projects
- Administration Management
- Users Management
- Providers
- Locations
- Resource Management
- Additional
- Documentation
- v2.4.3



Executed on 4 locations

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

LEXIS Platform

Containers & Scripts



portal.beta.lexis.tech/containers

Project Filter: ligatege... ehpcsw24

Home > Containers

Search by title

Refresh Create Container

Test Container 7th Of Mar 2024

Quantum AI Test Container 21st Of Mar 2024

Python Animals AI Example 12th Of Mar 2024

portal.beta.lexis.tech/locations

Project Filter: ligatege... eurofpde...

Home > Locations

Edit Provider Create Location

NAME	DESCRIPTION	ENDPOINT	TYPE	PROVIDERDESCRIPTION	ACTIONS
IRODS LRZ	iRODS at LRZ	lexis-lrzicat.srv.lrz.de	Storage	Leibniz-Rechenzentrum Compute Centre	Edit
Barbora	HPC Cluster Barbora	barbora.it4i.cz	HPC	IT4Innovations Computing Centre	Edit
IRODS LRZ OWSeu	iRODS Zone OWSLRZZONE	sikprlz-ows-icat.srv.mwn.de	Storage	Leibniz-Rechenzentrum Compute Centre	Edit
Cirrus	HPC Cluster Cirrus	login.cirrus.ac.uk	HPC	EPCC Compute Centre	Edit
Leonardo	HPC Cluster Leonardo	login.leonardo.cineca.it	HPC	CINECA Computing Centre	Edit
LocalCluster	Tartu testing instance	https://heappe-1.cloud.ut.ee	HPC	Puhuri testing provider	Edit
ZYC-2	HPC Cluster Alveo	10.12.0.11	HPC	IBM Computing Centre	Edit
LRZ Linux cluster 2	HPC Linux Cluster CoolMUC-2	lxlogin4.lrz.de	HPC	Leibniz-Rechenzentrum Compute Centre	Edit
LUMI	HPC Cluster LUMI	lumi.csc.fi	HPC	LUMI Computing Centre	Edit
IT4I Staging Area	IT4I Staging Area	staging	Storage	IT4Innovations Computing Centre	Edit



Project Filter openwebs...

Home > Custom HPC Jobs > EESSI Demo - estimate PI using R > Workflow Create

1 Workflow name and project

```
jobscrip.sh
```

```
1 source /cvmfs/software.eessi.io/versions/2023.06/init/bash
2
3 ml spider R
4
5 ml R/4.2.2-foss-2022b
6
7 R --version
8
9 mkdir ./output
10
11 Rscript -e "
12 # Set the number of points to simulate
13 num_points <- 100000
14
15 # Generate random points
16 x <- runif(num_points)
```

Name
Compute PI in R

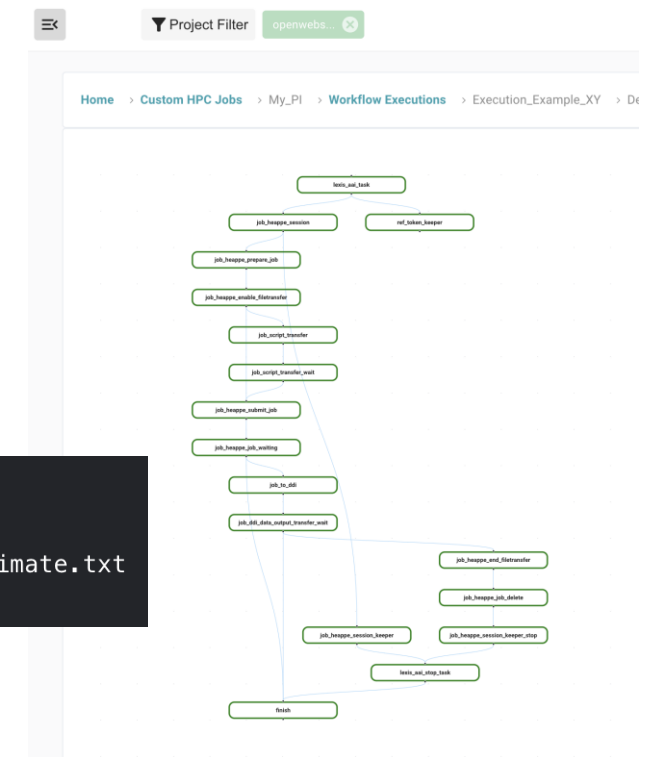
Project Short Name
eurofpdemo

Description
Jobscrip showcase

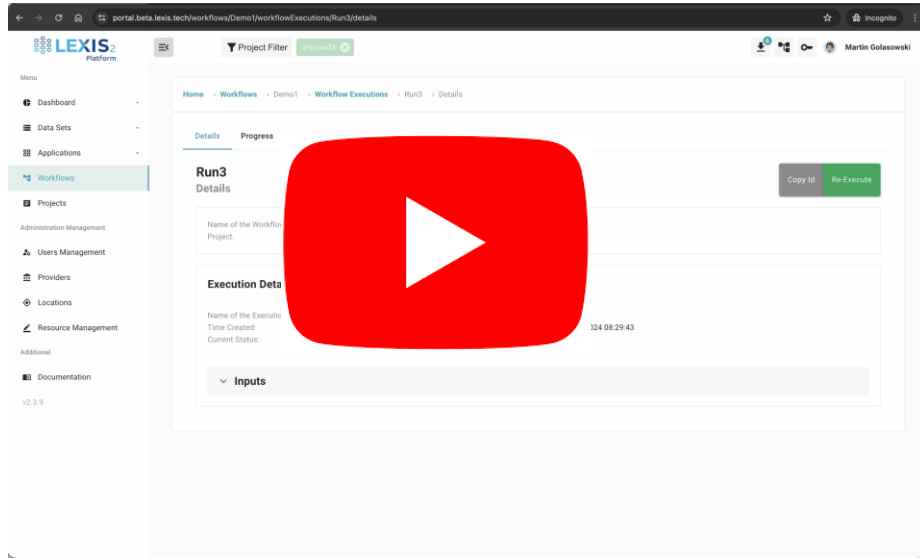
Continue Back

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.

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



AI workflow demo with Python in Apptainer

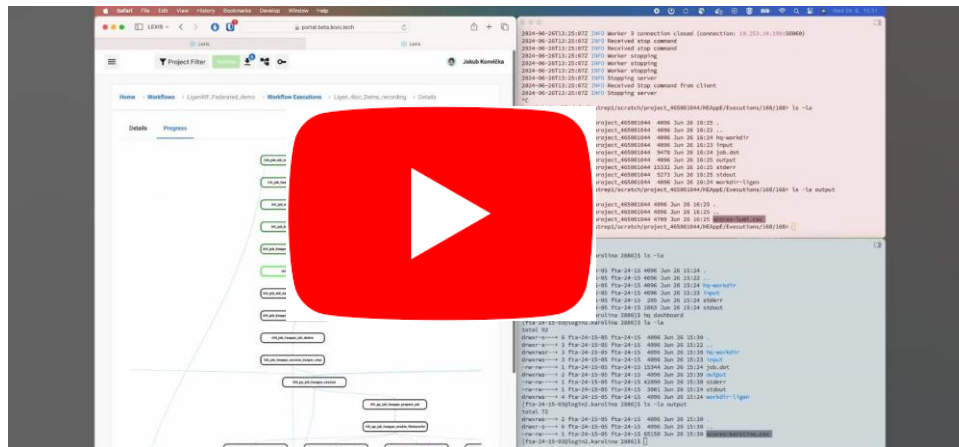


[Check it out!](#)



Documentation
<https://docs.lexis.tech>

Urgent computing federated workflow execution



[Check it out!](#)

Thank you!

Questions?

`martin.golasowski@vsb.cz`



Spolufinancováno
Evropskou unií



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

MUNI
ICS



VŠB TECHNICKÁ
UNIVERZITA
OSTRAVA

IT4INNOVATIONS
NÁRODNÍ SUPERPOČÍTAČOVÉ
CENTRUM