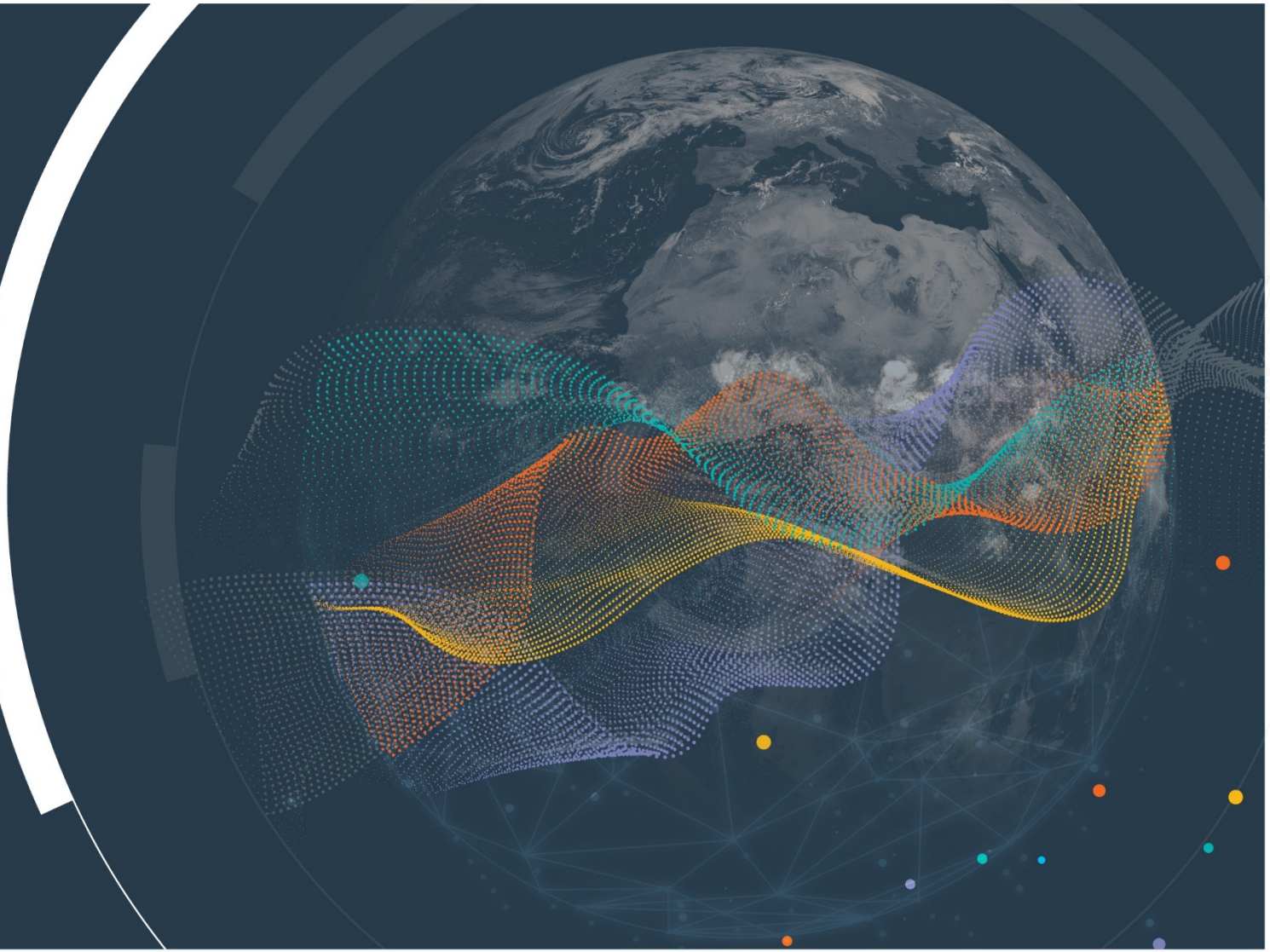
EUMETSAT

Funded by the European Union

Destination Earth

Destination Earth Data Lake Overview to EDGE Services

Borys Saulyak EUMETSAT

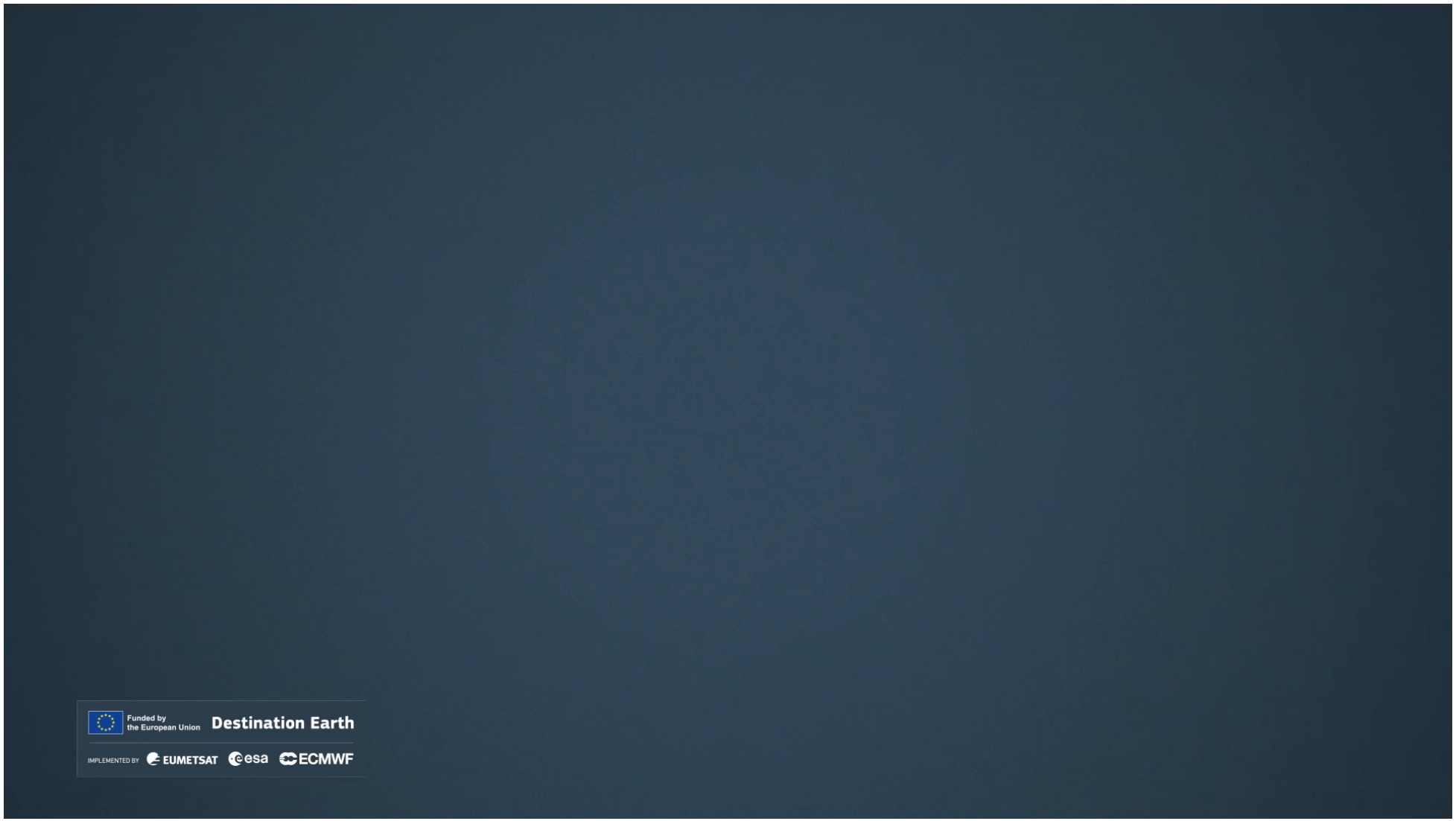


EUM/SCIR/TEM/21/1923005, v1A, 01 September 2024

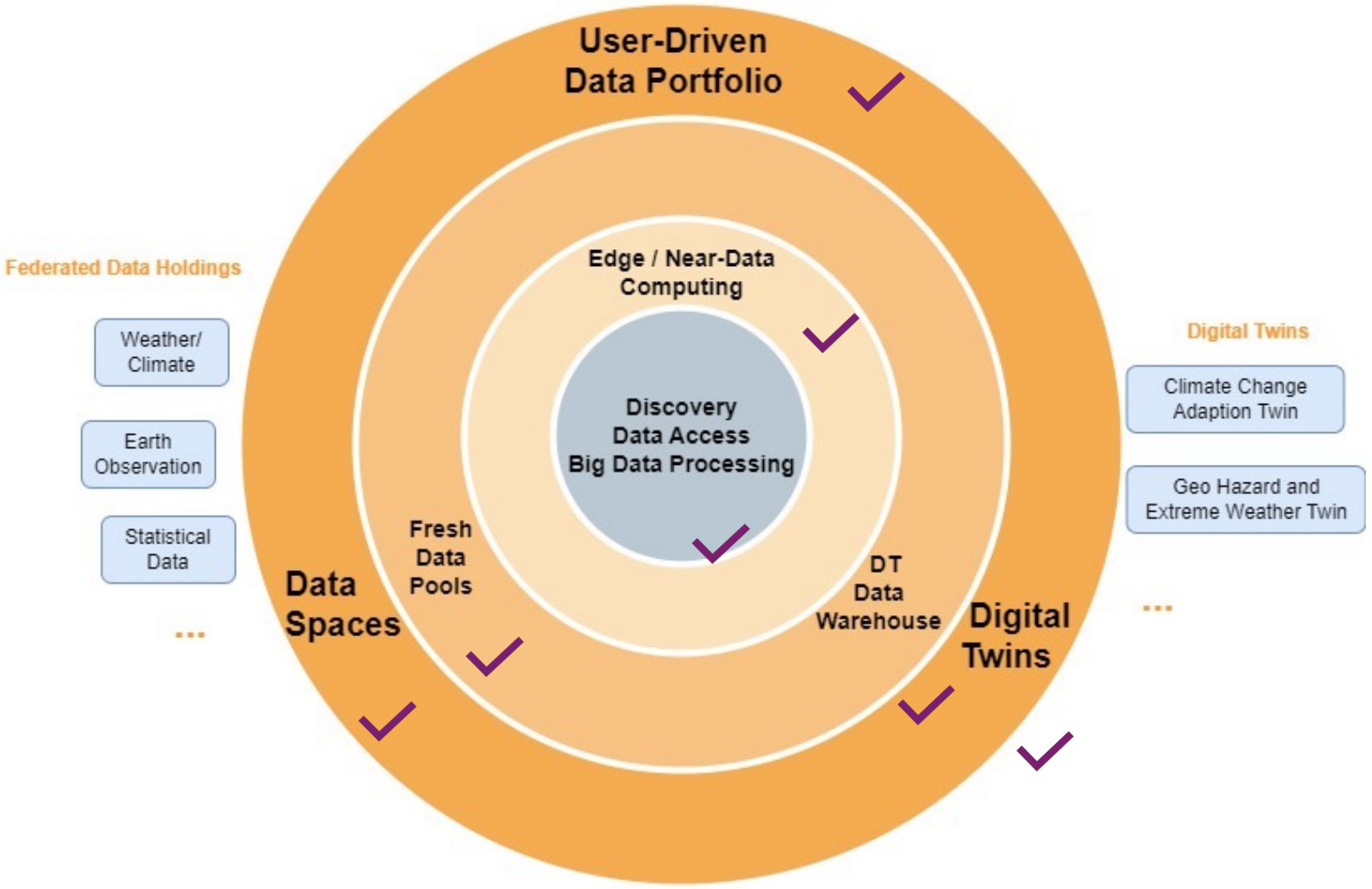
Classification

IMPLEMENTED BY ECMWF esa EUMETSAT

A possible journey using the data lake



DestinE Data Lake – in a nutshell



DEDL is self-standing component

- Built from geographically distributed physical elements
- Distributed services – seamless access

Discovery & Data Access

- Harmonisation of data access (HDA) to simplify data discovery & access
- External federated data spaces
- Digital Twin data (ECMWF):
  - Extreme Weather and Climate Change Adaptation
- DestinE User generated data

Big Data Processing

- Processing near data including distributed computing & workflows
- Supports & enables AI/ML applications

DESTINE DATA LAKE – DISTRIBUTED INFRA

Eumetsat Bridge

GERMANY
DC in Darmstad

Leonardo Bridge

ITALY
CINECA DC in Bolonia

Mare Nostrum Bridge

SPAIN
DC in Barcelona

LUMI Bridge

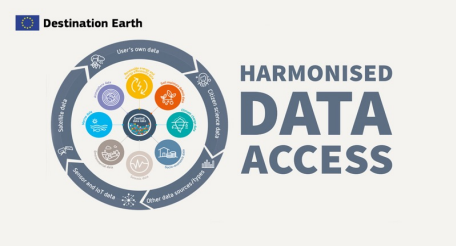
FINLAND
CSC DC in Kajaani

Central Site

POLAND
Jawczyce DC in Warsaw

DestinE Data lake services

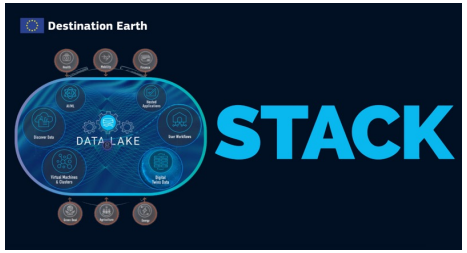
HDA



Harmonised Data Access: seamless access to
- DestinE Data – DT Outputs & User Generated Data for DestinE
- Federated Data

as per defined & evolving "DestinE Data Portfolio".
API => Spatio Temporal Asset Catalog (STAC)

Usage on request



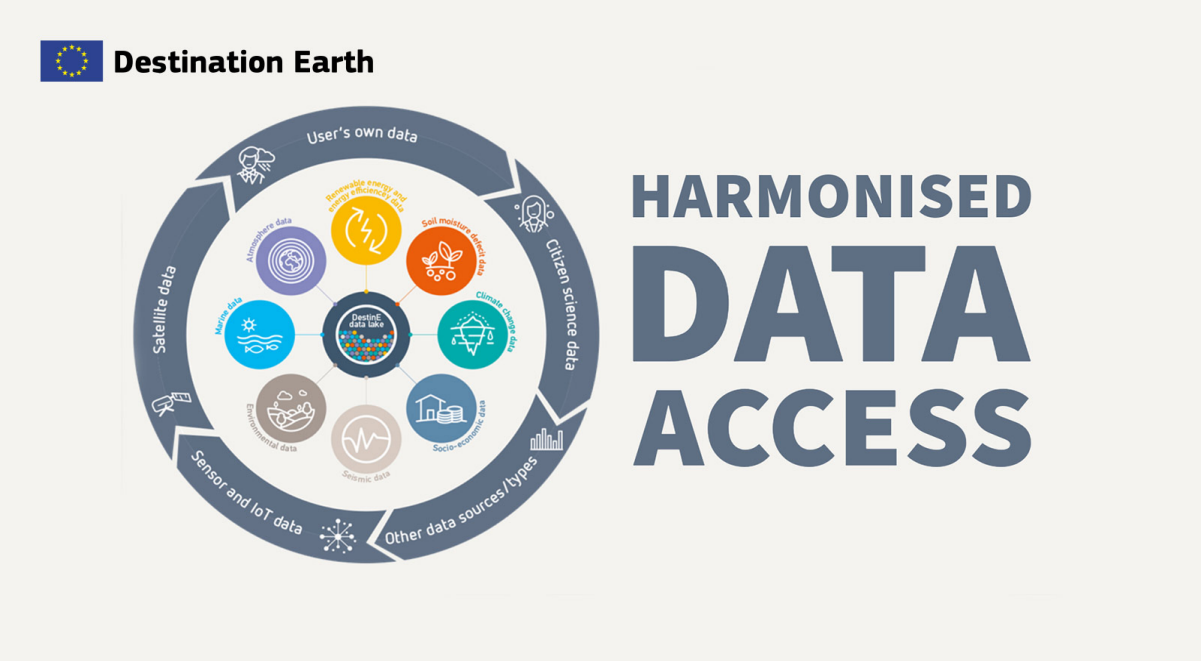
STACK: SaaS suite which enables near data processing
- JupyterHub, Dask/ Dask Gateway and Open Data Cube

ISLET Compute: (IaaS/PaaS) enables near data processing by allowing users to manage and deploy virtualised workloads

ISLET Storage: S3 Object Storage to store user's data and processing results

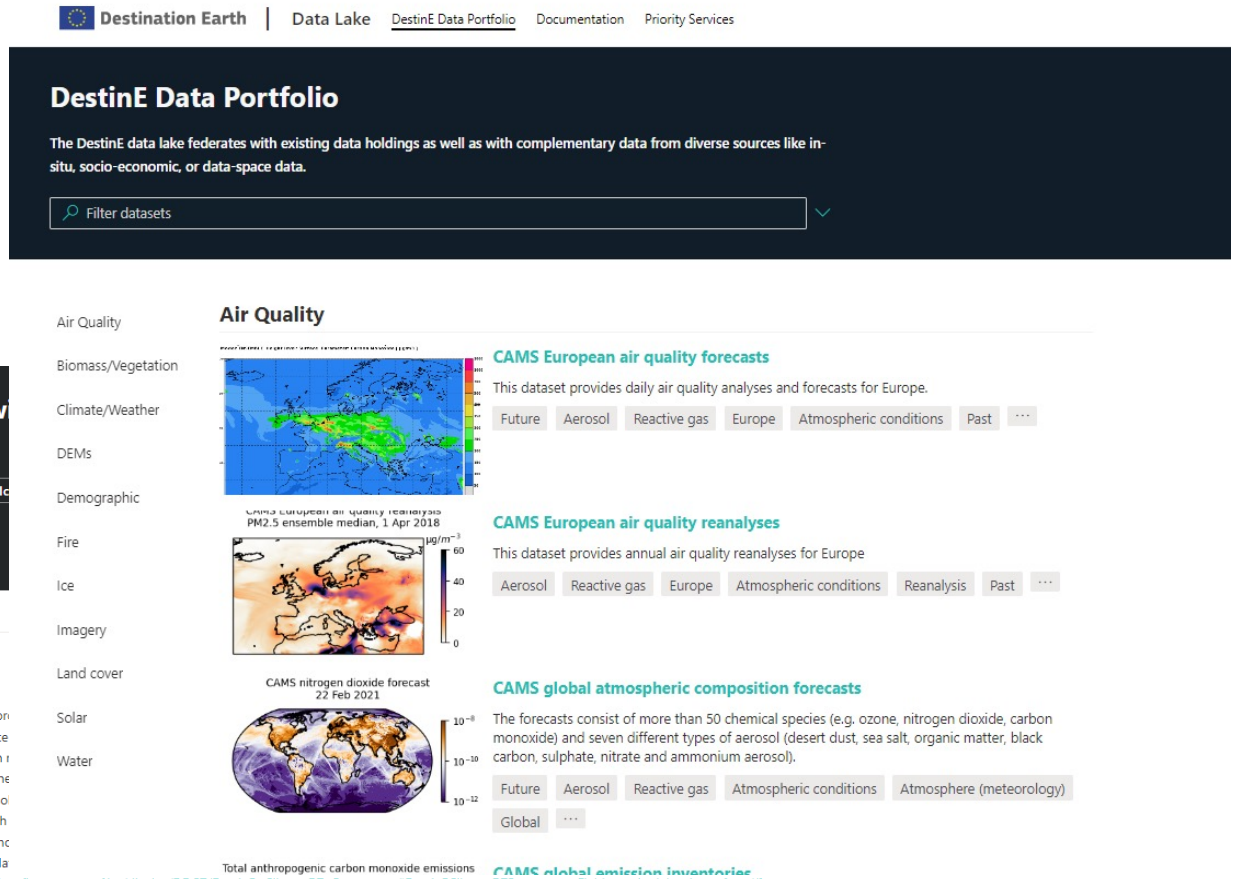
HOOK: allows to execute high level pre-defined or own workflows
- Data harvest: to harvest data from a federated data provider a priori of planned data processing or analysis task

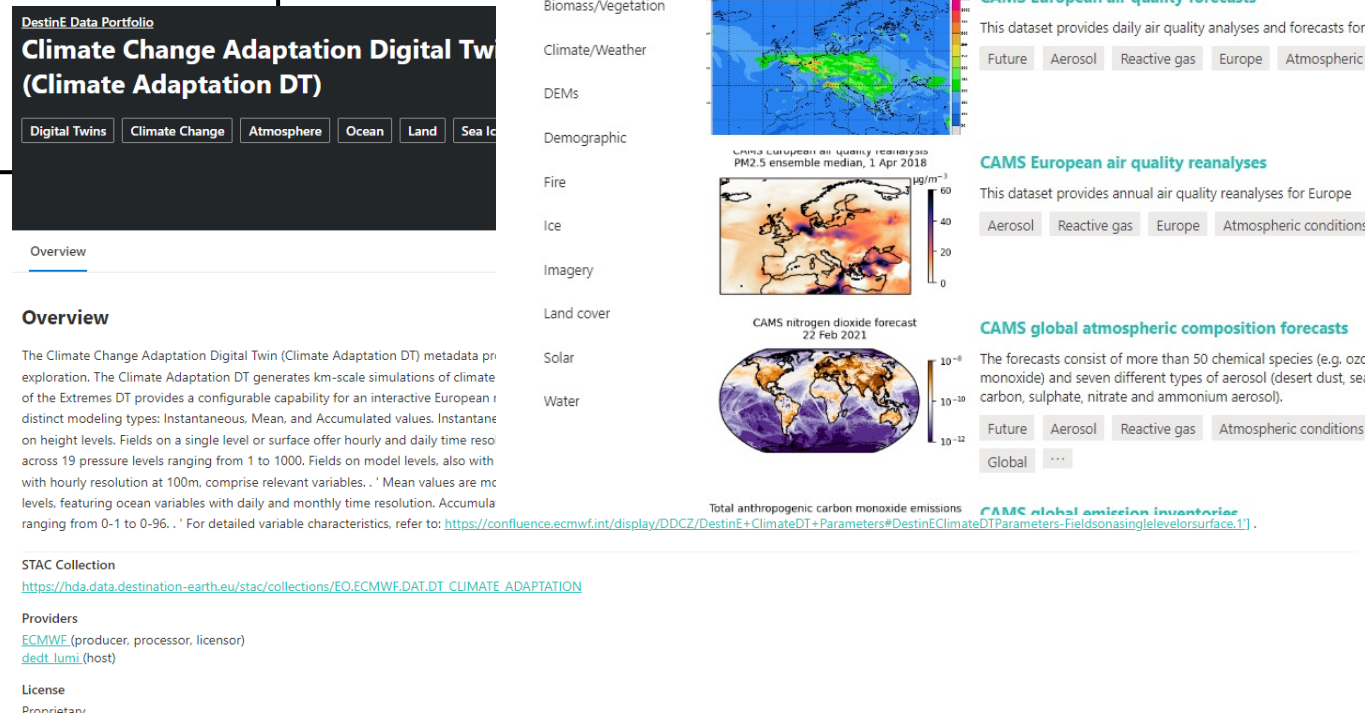
Data discovery and access



Harmonised Data Access

- **Discover** - DestinE Data Portfolio
  - Using Web Browser
  - Using HDA (API) - Notebooks
    - Platform (Insula)
    - Data Lake (STACK)

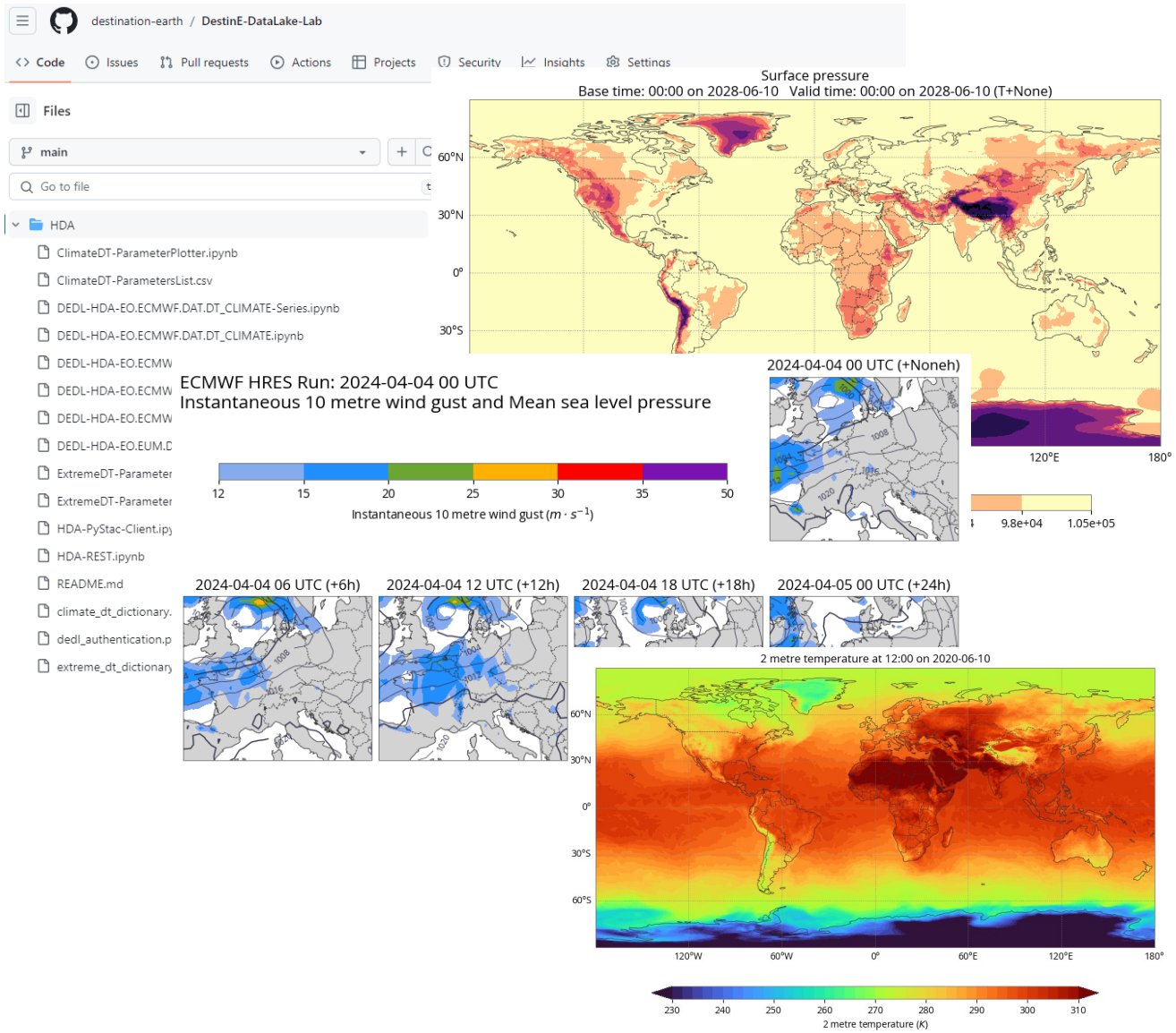




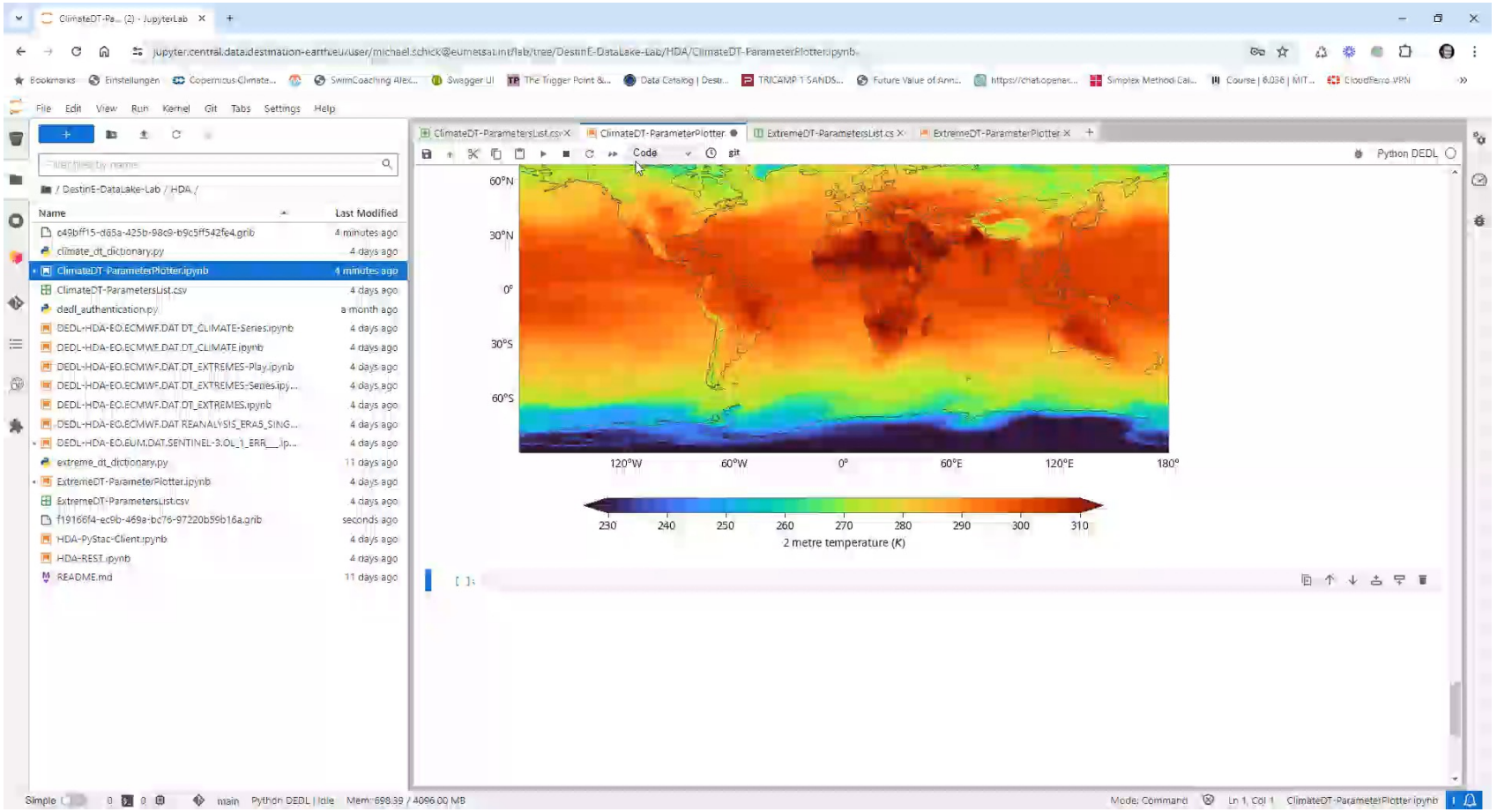
Harmonised Data Acces

- Harmonised Data Access (STAC)
  - REST EAPI client
  - PyStac-Client
  - ODAG (simple python API)
- Access DestinE Data Portfolio
- Datasets from multiple provider
- Protected datasets (DT Data)
- Quotas Group based

- Notebooks DestinE-DataLakeLab
  - From Insula, STACK
    - Climate DT plotter
    - Extreme DT plotter
    - HDA-PySTAC-Client
    - ....

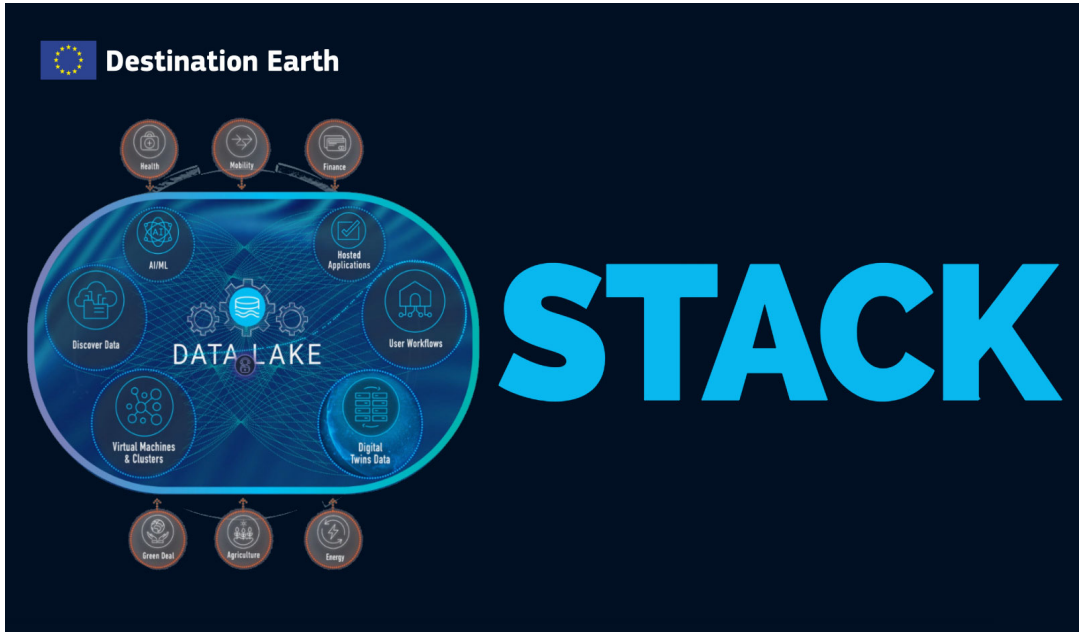


ClimateDT-plotter



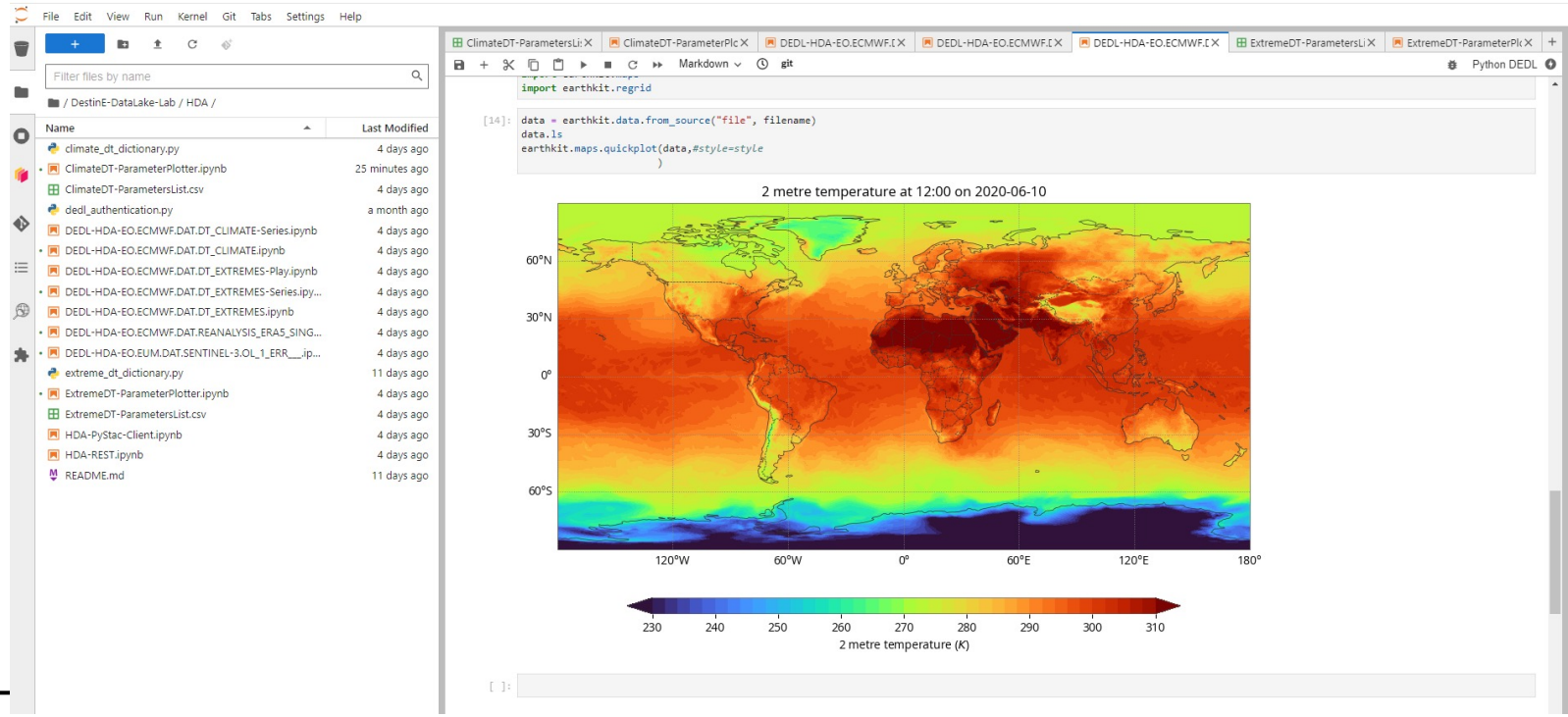
STACK service

Distributed processing



STACK: JupyterHub + JEG

- Jupyter Hub
  - Development Environment
- Jupyter Enterprise Gateways (JEG)
  - @edges – Lumi, MN, Leonardo
  - execute notebooks on edges
- Object Storage Access to FDP
- Plug-Ins
  - Gitlab
  - Object Storage (S3)
  - EODAG
- Quota Group based
- DestinE-DataLakeLab examples



STACK: DASK

What is DASK ?

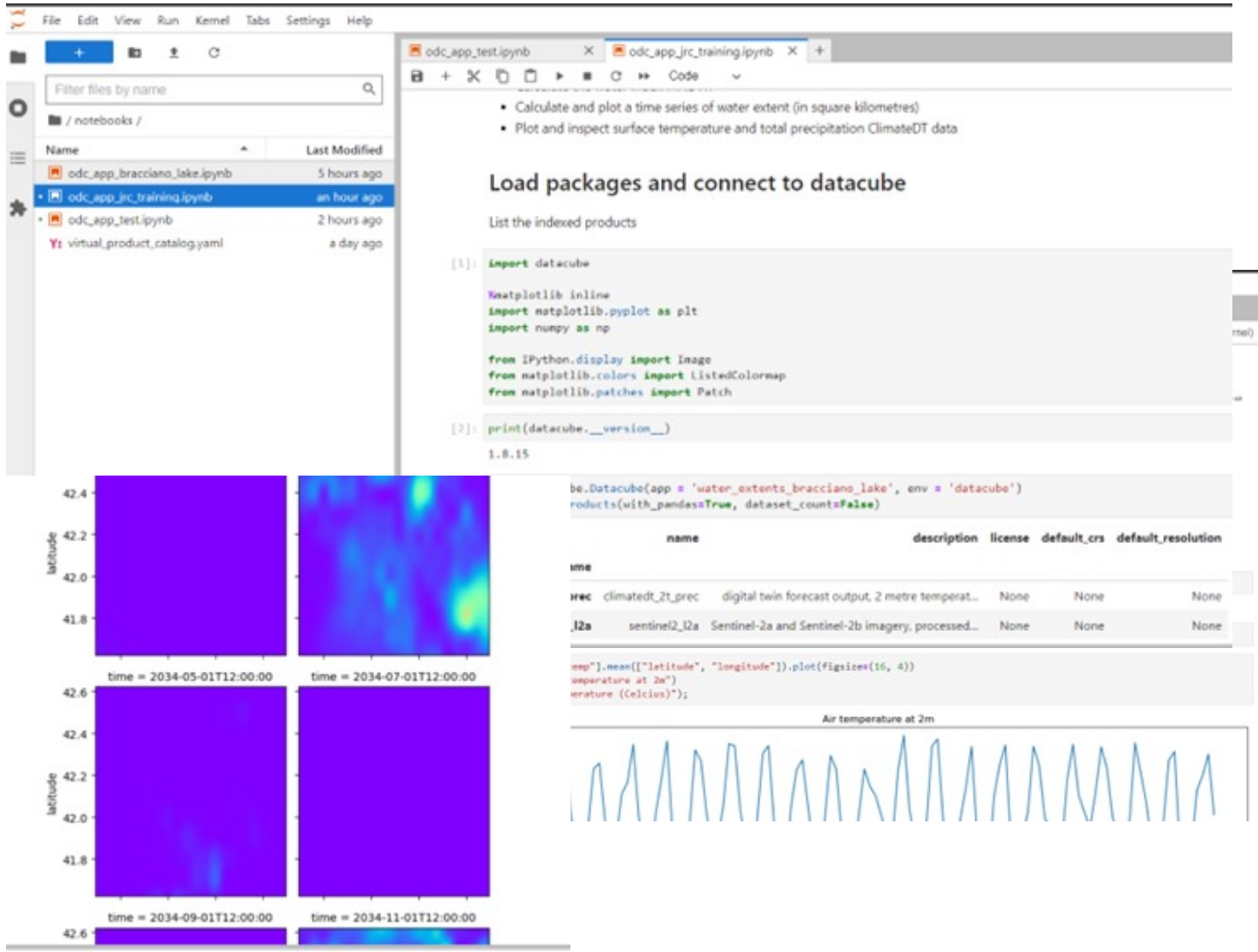
- Python Library for parallel and distributed computing
  - API
  - Client
  - Gateway
- Gateways on Central & Edges

- DASK usage example with a notebook
- Flood use case

STACK: Data cube

Two options:

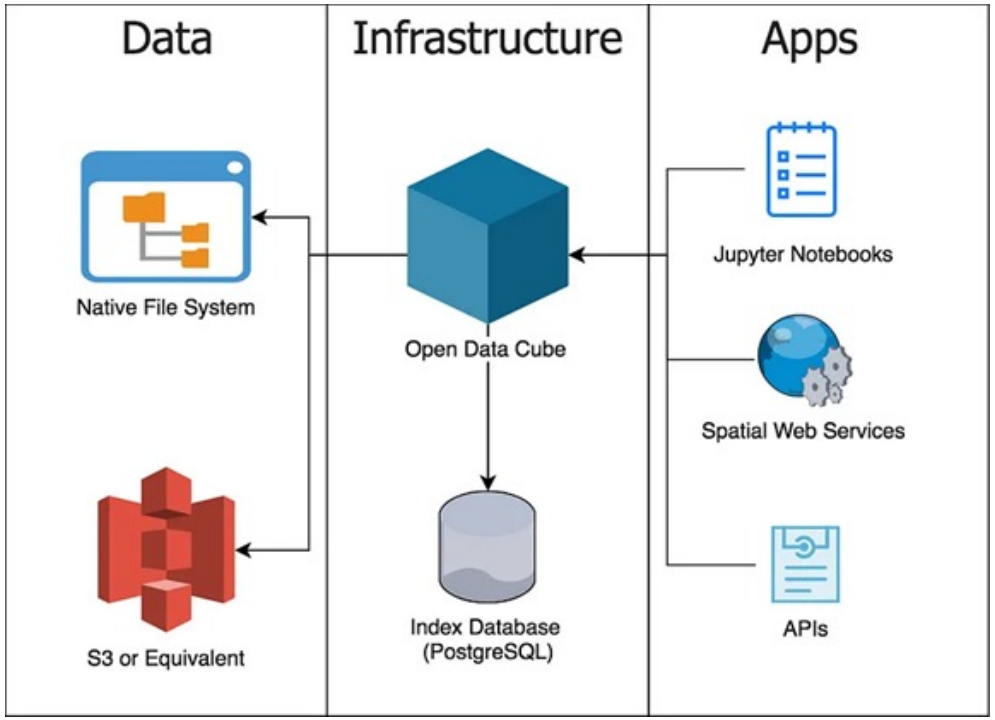
- Build your own cube:
  - ISLET – compute/storage
  - STACK - to access & interact
- DEDL provided data cube
  - Extreme DT data (protected)



STACK: Build your own cube

Use Open Data Cube image
Islet service

- Instantiate a VM with ODC
- Run the ODC environment
- Index products and datasets
- Run an ODC application



The Open Data Cube (ODC) is an Open-Source Geospatial Data Management and Analysis Software.

At its core, the ODC is a set of Python libraries and PostgreSQL database that helps you work with geospatial raster data.

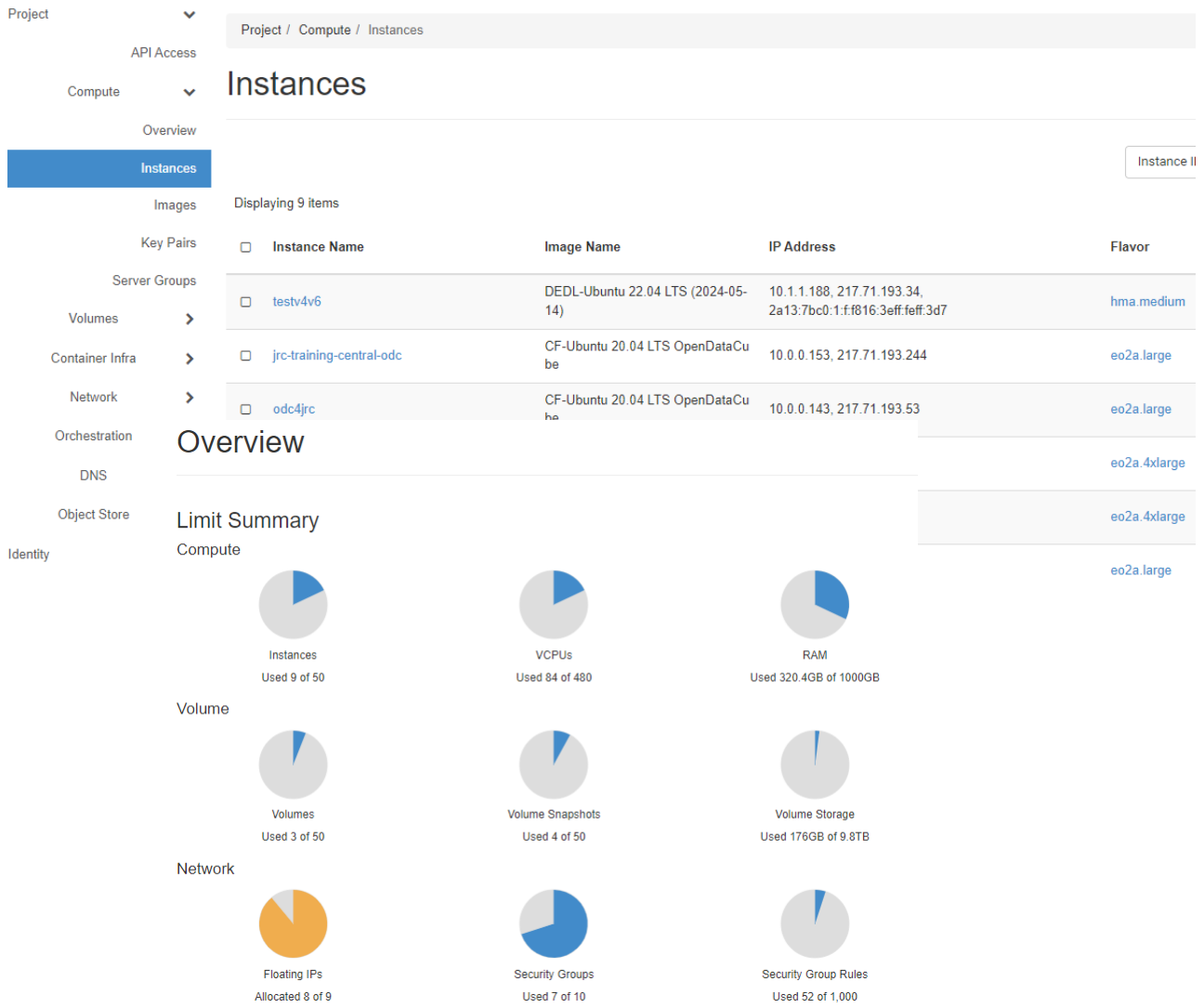
ISLET service
IaaS/PaaS

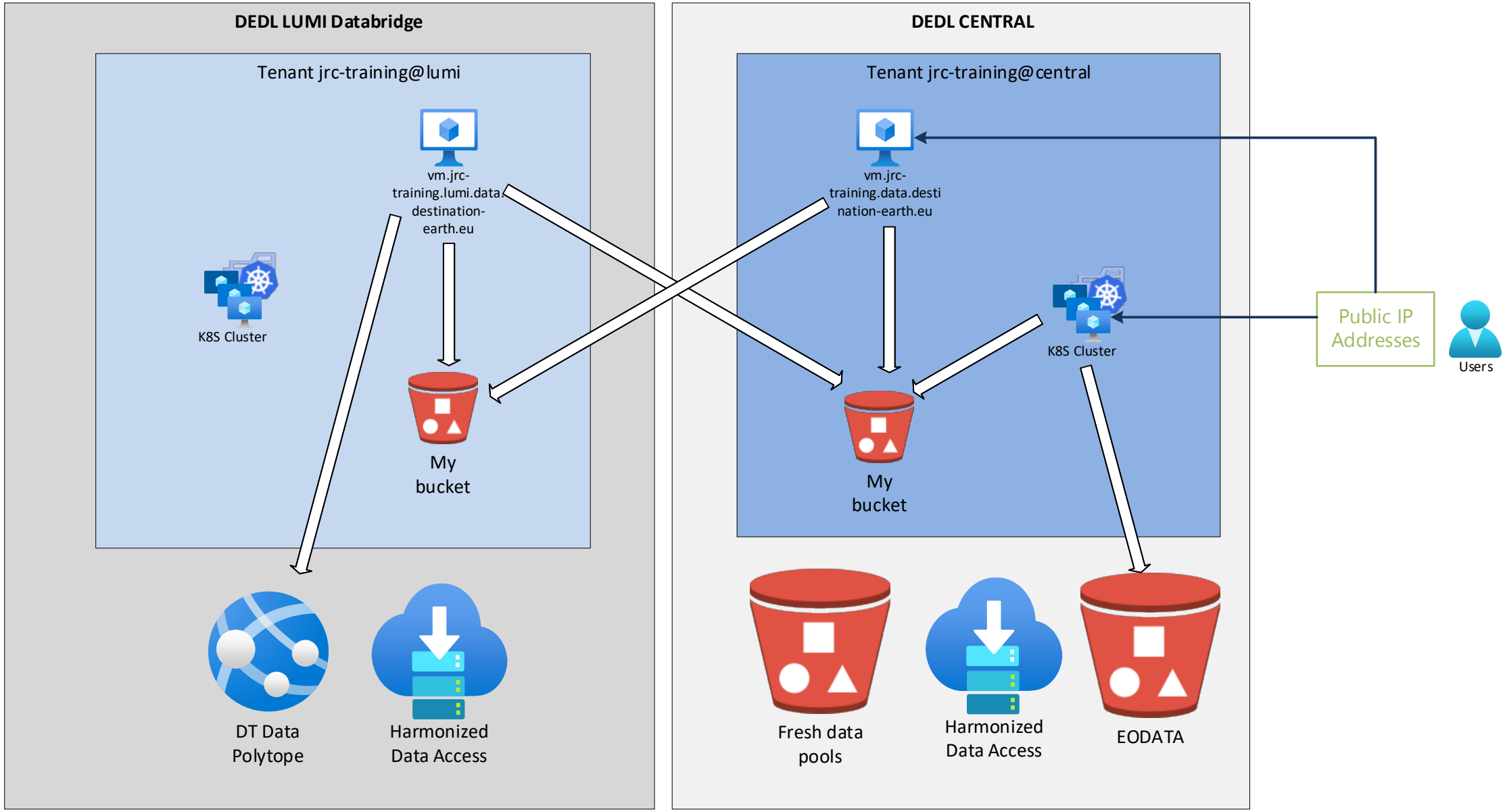
ISLET service

Funded by the European Union
Destination Earth

Running custom workload adjacent to the data, in particular:

- Provision of virtual infrastructure (VMs, storage, network, load-balancers, security)
- Provision of Kubernetes clusters
- Provision of S3 buckets, manage access to S3 buckets
- Access to supported OS images with tools pre-installed
- Manage data backups
- Deploy applications in VMs and K8S
- Make your own services and make them available for the users
- Exchange data with other DestinE users




DEDL LUMI Databridge
Tenant jrc-training@lumi
vm.jrc-training.lumi.data.destination-earth.eu
K8S Cluster
My bucket
DT Data Polytope
Harmonized Data Access
DEDL CENTRAL
Tenant jrc-training@central
vm.jrc-training.data.destination-earth.eu
K8S Cluster
My bucket
Fresh data pools
Harmonized Data Access
EODATA
Public IP Addresses
Users

HOOK service

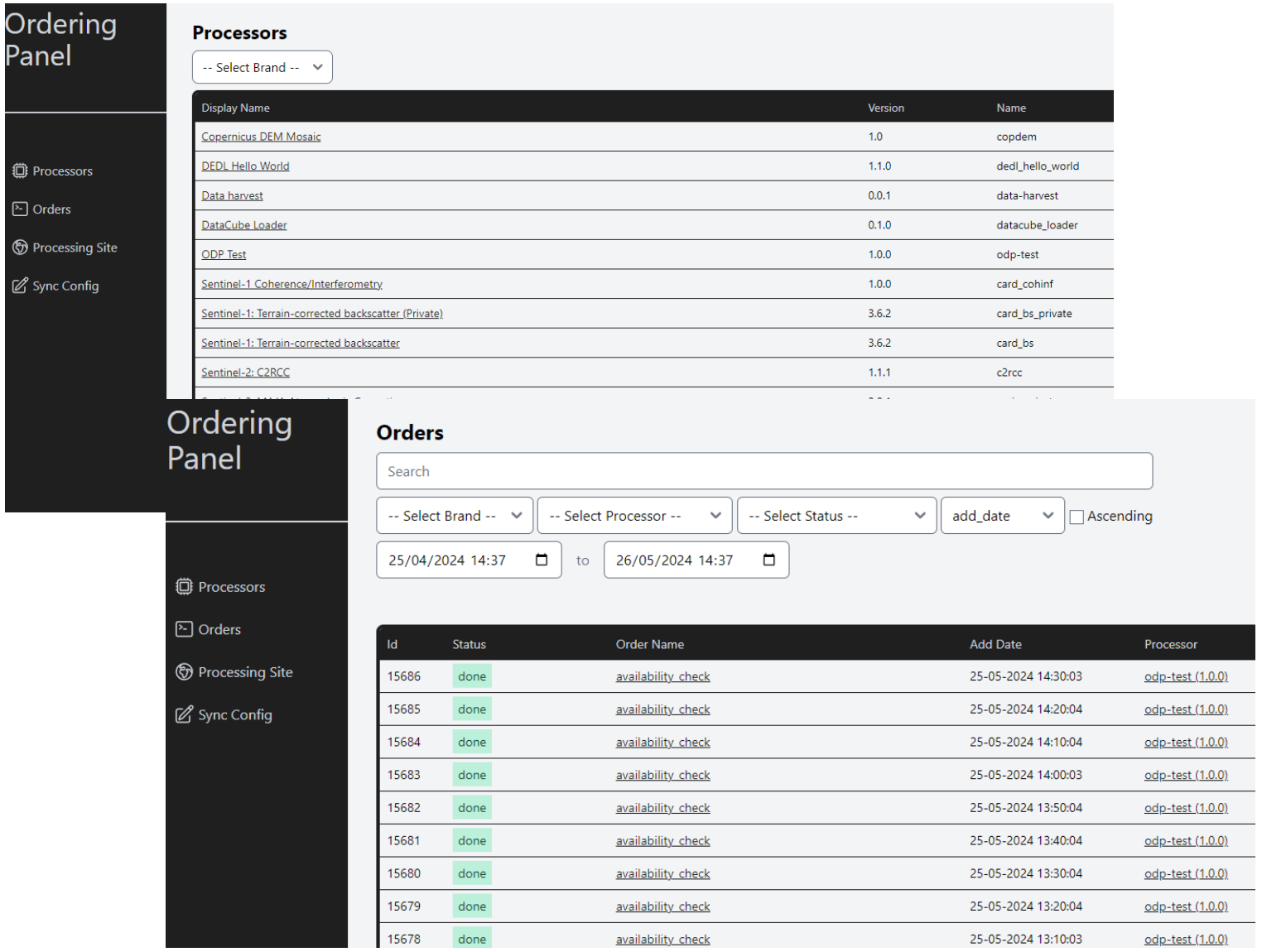
User workflows

HOOK

Pre-Defined processors

- Data Harvest
- Results:
  - S3 private bucket (ISLET)
  - Temporary bucket

- Build your own workflow

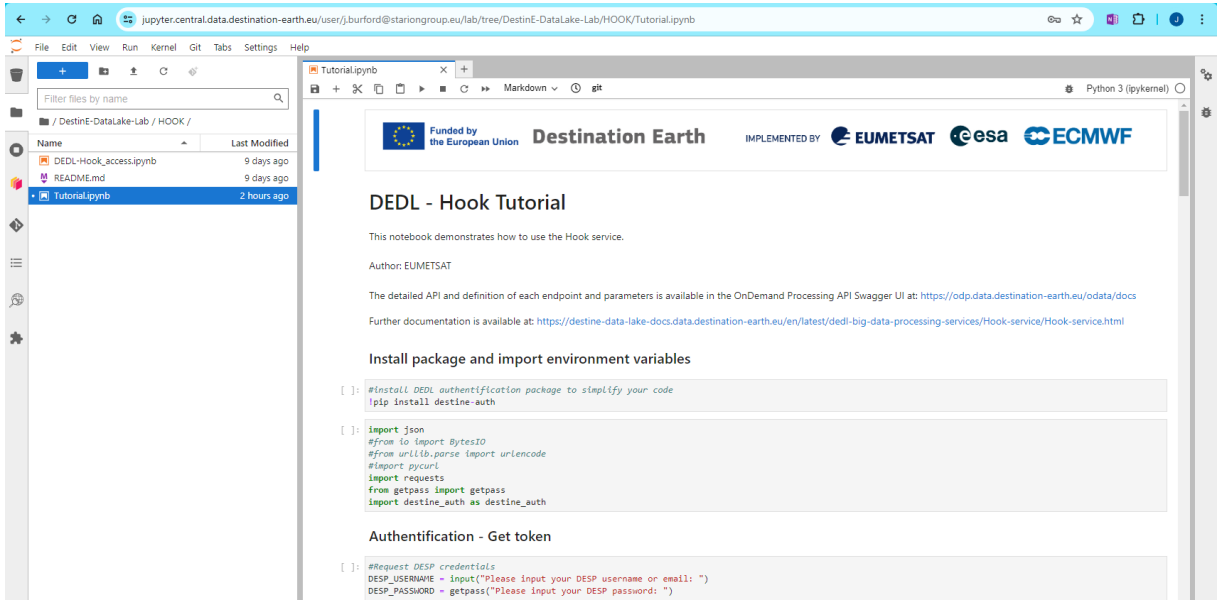


HOOK Service

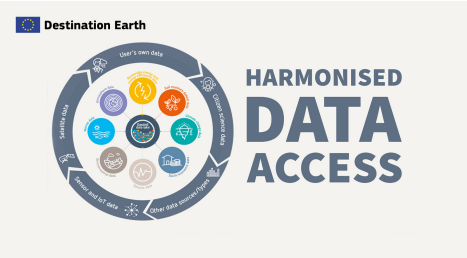
Interact with workflows using
- REST API
- openEO API

User Interface:
- JupyterLab (STACK)
- openEO (UI)

Notebook HOOK/Tutorials
- Discover processors/ workflows
- Execute Data-Harvest workflow



DestinE services on the Lake



- Documentation
- Public github with notebooks examples DestinE-DataLake-Lab
- DestinE Data Portfolio

DestinE use cases in DEDL

Use cases completed & delivered during Phase I

Description/ Details available on DestinE community website

Danube Delta Water Reservoir Monitoring (Cloud Ferro, CS Group)

Italy Drought 2022 (Cloud Ferro, EODC)

Pakinstan Flood 2022 (Cloud Ferro, EODC)