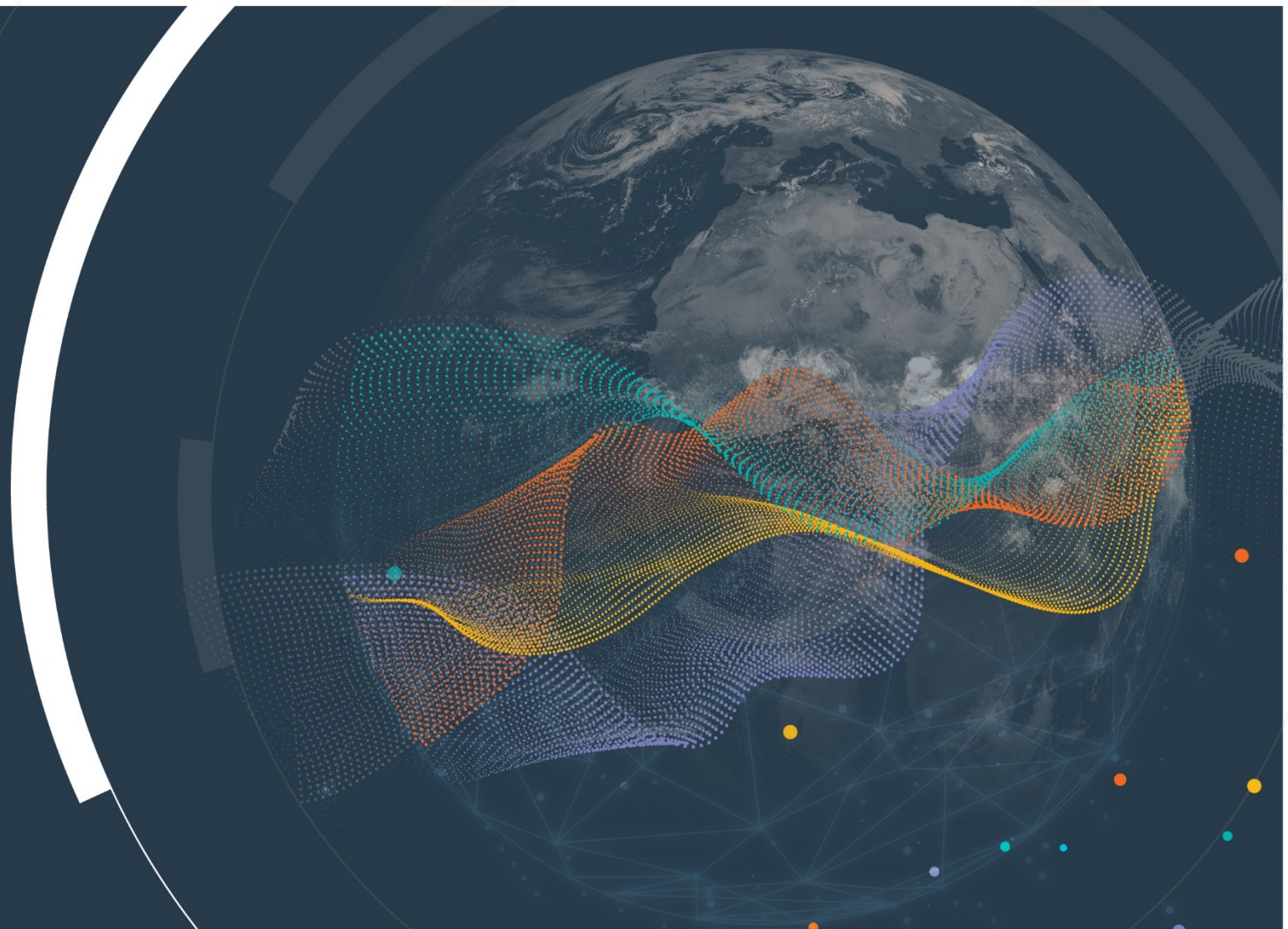
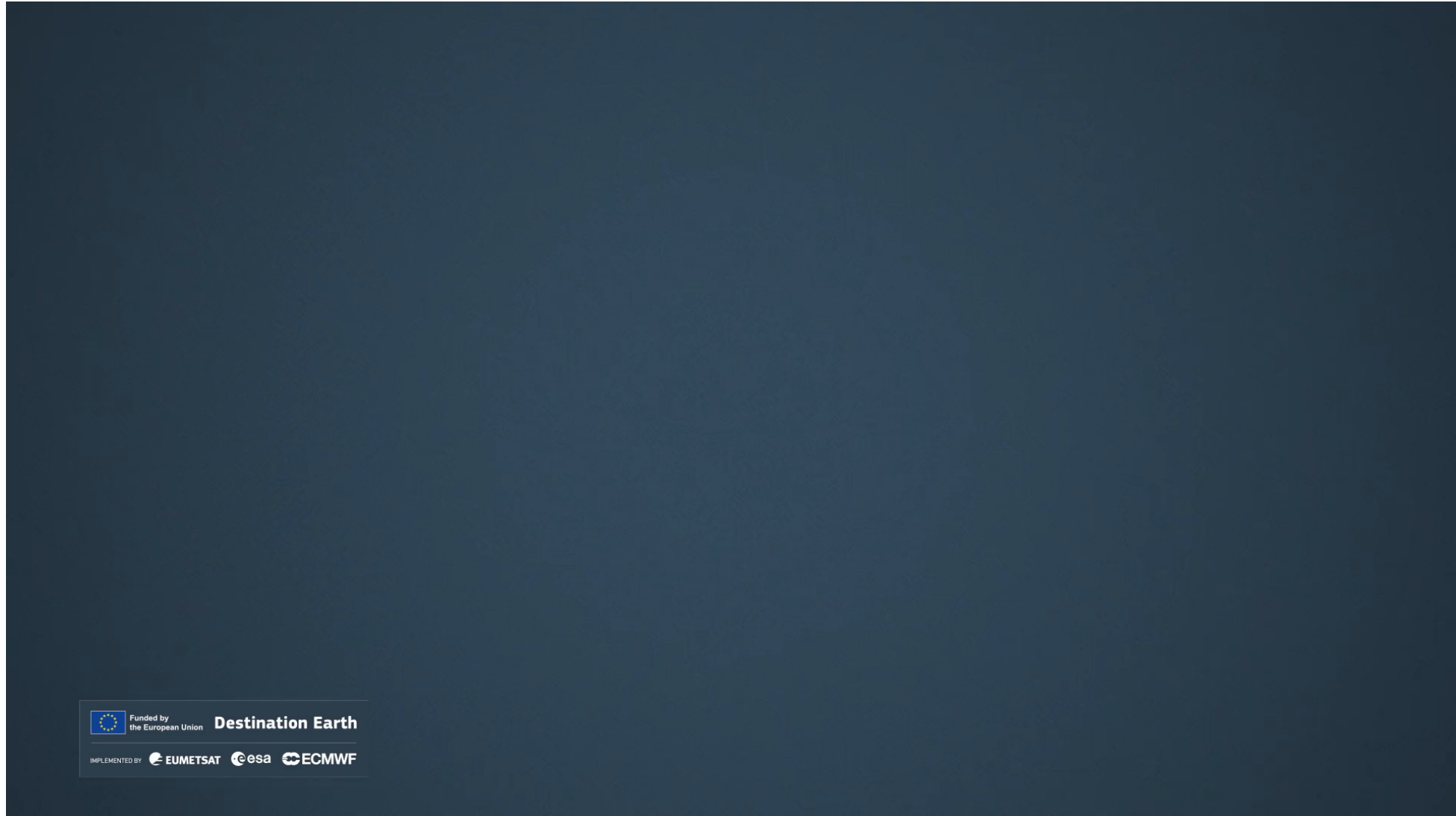






Destination Earth Data Lake Overview to EDGE Services

Borys Saulyak EUMETSAT



A possible journey using the data lake



 Funded by
the European Union **Destination Earth**
IMPLEMENTED BY   

DestinE Data Lake – in a nutshell

DEDL is self-standing component

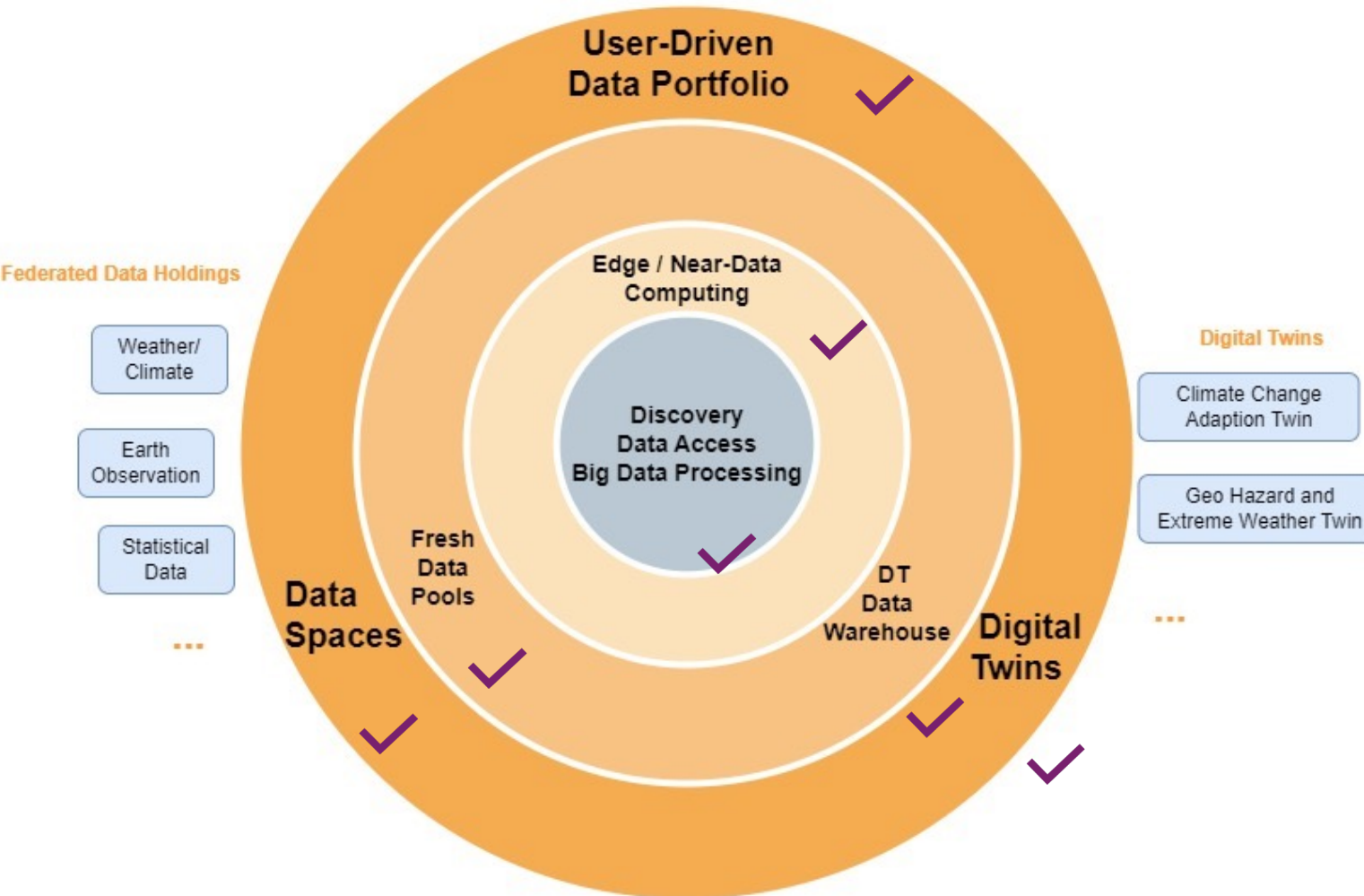
- Built from geographically distributed physical elements
- Distributed services – seamless access

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



LUMI Bridge



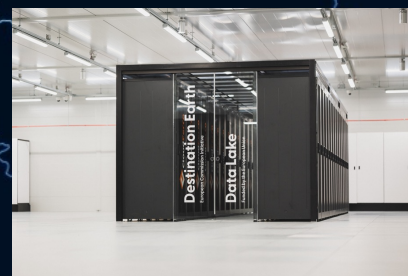
Leonardo Bridge



Central Site

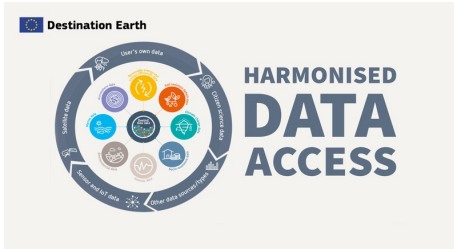


Mare Nostrum Bridge



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



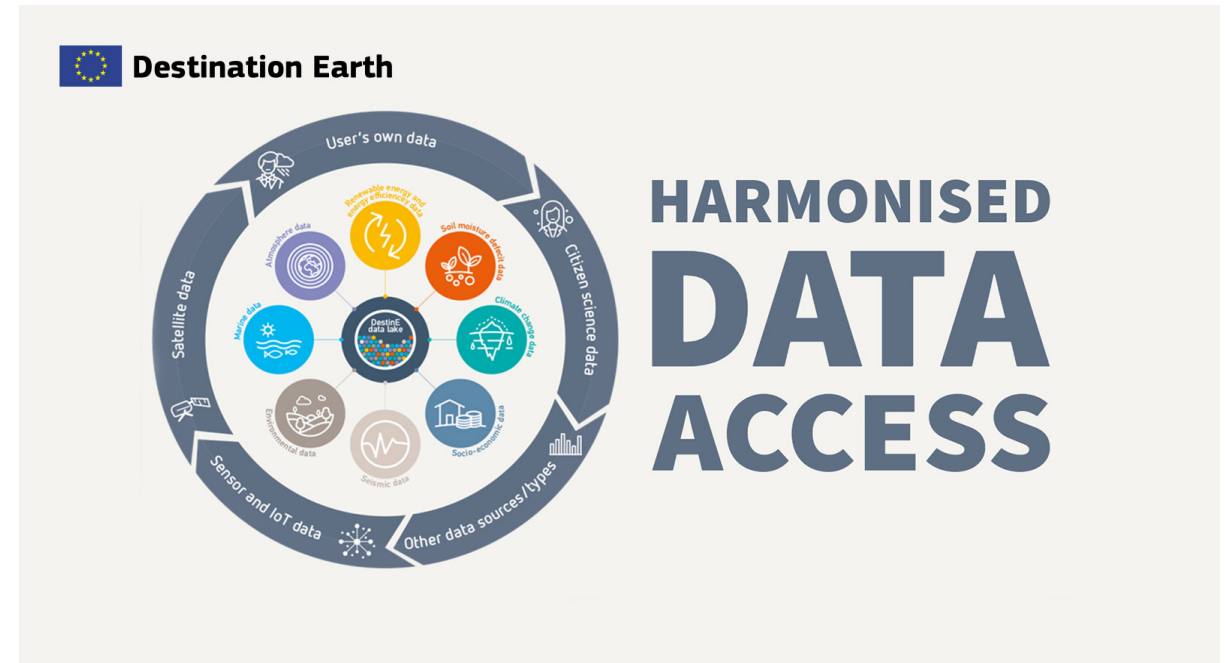
ISLET Storage: S3 Object Storage to store user’s data and processing results



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

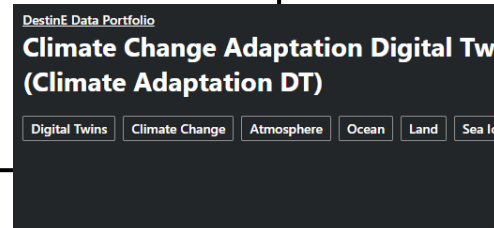
Data discovery and access



Harmonised Data Access

- **Discover** - DestinE Data Portfolio

- Using Web Browser
- Using HDA (API) - Notebooks
 - Platform (Insula)
 - Data Lake (STACK)



Overview

Overview

The Climate Change Adaptation Digital Twin (Climate Adaptation DT) metadata provides exploration. The Climate Adaptation DT generates km-scale simulations of climate of the Extremes DT provides a configurable capability for an interactive European distinct modeling types: Instantaneous, Mean, and Accumulated values. Instantaneous on height levels. Fields on a single level or surface offer hourly and daily time resolution across 19 pressure levels ranging from 1 to 1000. Fields on model levels, also with hourly resolution at 100m, comprise relevant variables. Mean values are model levels, featuring ocean variables with daily and monthly time resolution. Accumulated ranging from 0-1 to 0-96. For detailed variable characteristics, refer to: <https://confluence.ecmwf.int/display/DDCZ/DestinE+ClimateDT+Parameters#DestinEClimateDTParameters-Fieldsonasinglelevelorsurface.1>

STAC Collection

https://hda.data.destination-earth.eu/stac/collections/EO.ECMWF.DAT.DT.CLIMATE_ADAPTATION

Providers

[ECMWF](#) (producer, processor, licensor)

[dedt_lumi](#) (host)

License

Proprietary

DestinE Data Portfolio

The DestinE data lake federates with existing data holdings as well as with complementary data from diverse sources like in-situ, socio-economic, or data-space data.

Air Quality

Air Quality

Biomass/Vegetation

Climate/Weather

DEMs

Demographic

Fire

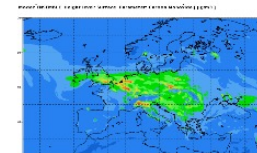
Ice

Imagery

Land cover

Solar

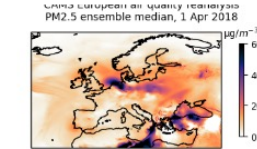
Water



CAMS European air quality forecasts

This dataset provides daily air quality analyses and forecasts for Europe.

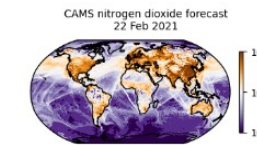
[Future](#) [Aerosol](#) [Reactive gas](#) [Europe](#) [Atmospheric conditions](#) [Past](#) ...



CAMS European air quality reanalyses

This dataset provides annual air quality reanalyses for Europe

[Aerosol](#) [Reactive gas](#) [Europe](#) [Atmospheric conditions](#) [Reanalysis](#) [Past](#) ...



CAMS global atmospheric composition forecasts

The forecasts consist of more than 50 chemical species (e.g. ozone, nitrogen dioxide, carbon monoxide) and seven different types of aerosol (desert dust, sea salt, organic matter, black carbon, sulphate, nitrate and ammonium aerosol).

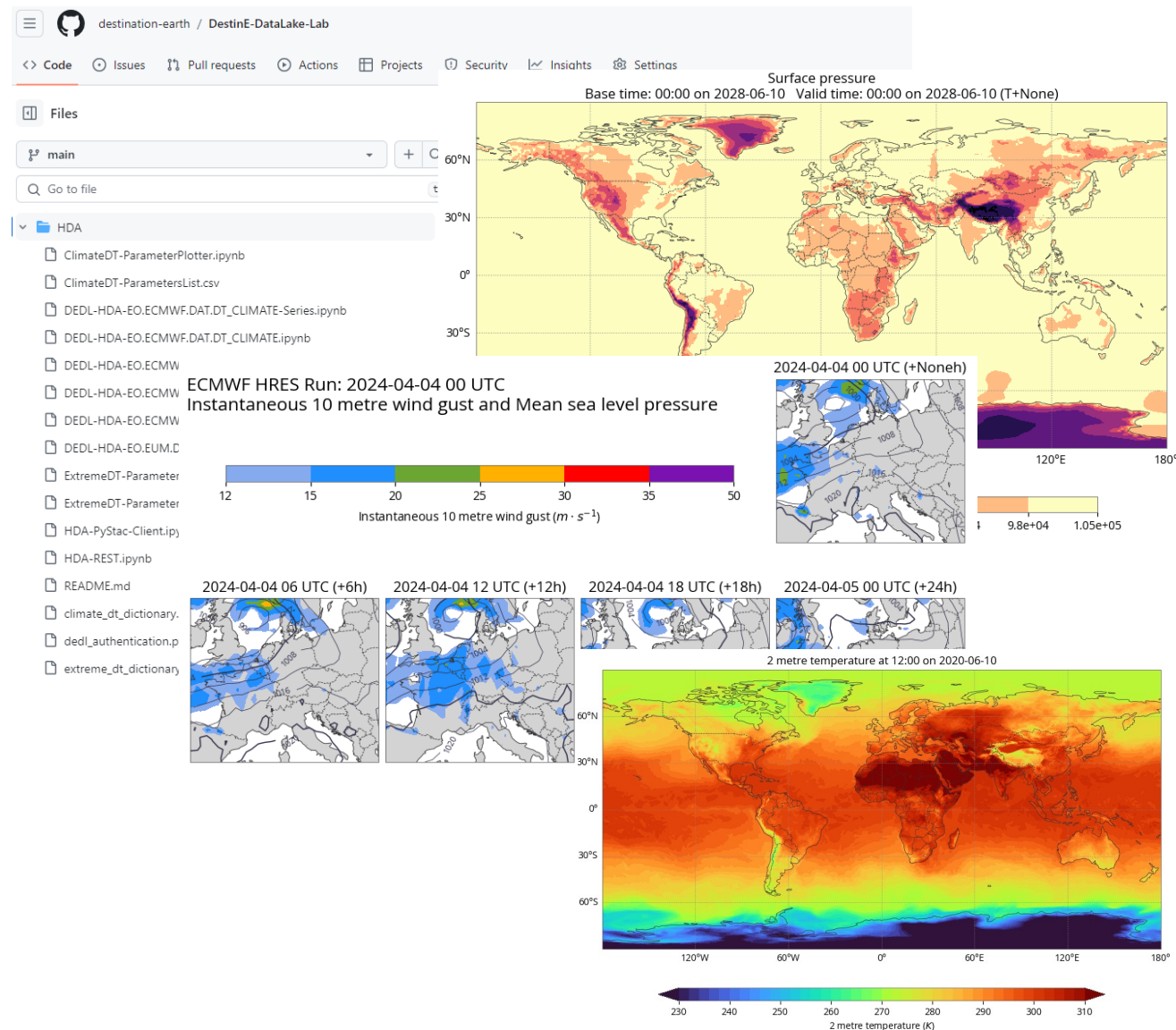
[Future](#) [Aerosol](#) [Reactive gas](#) [Atmospheric conditions](#) [Atmosphere \(meteorology\)](#) [Global](#) ...

Total anthropogenic carbon monoxide emissions

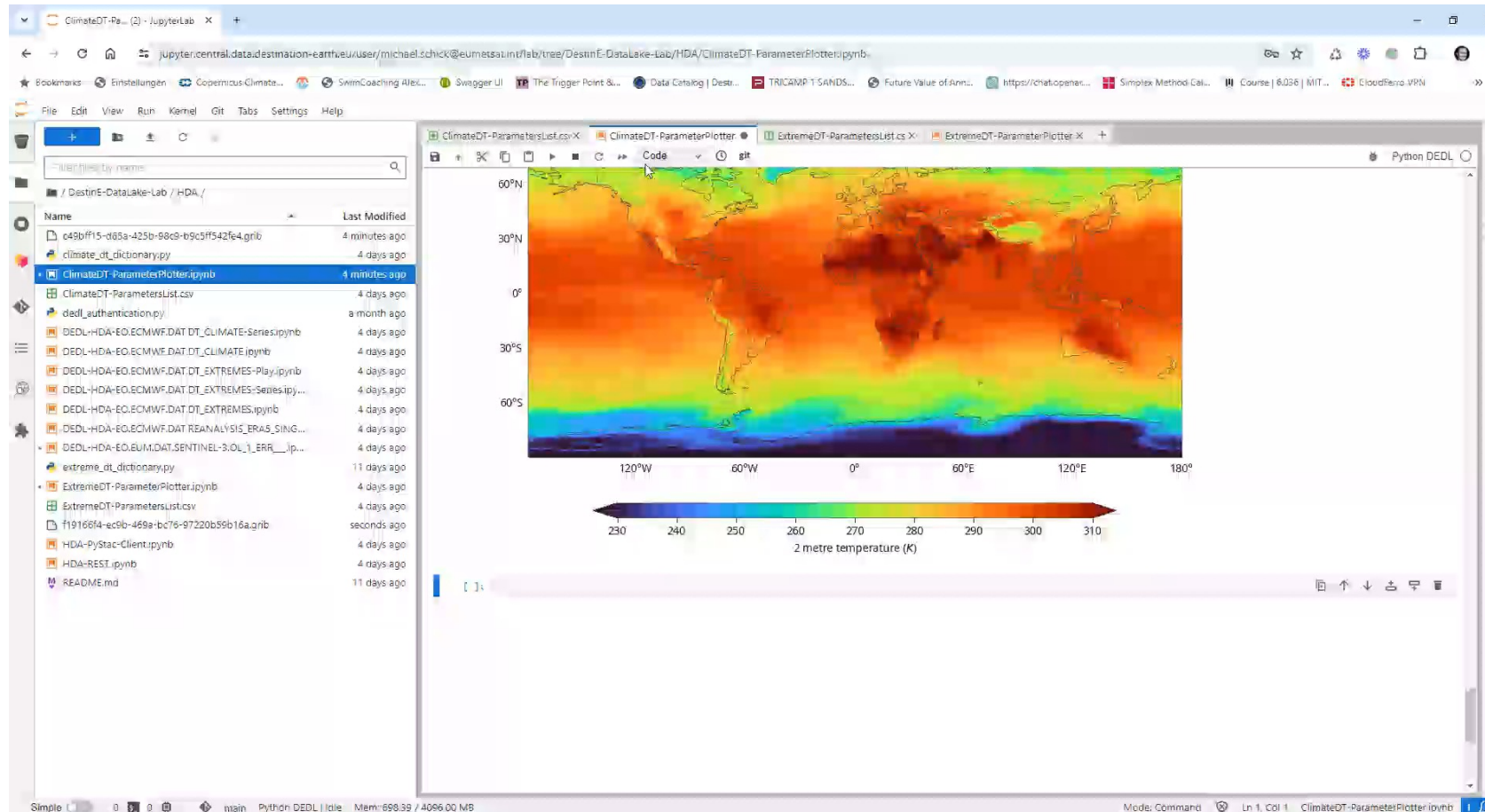
CAMS global emission inventories

Harmonised Data Access

- 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



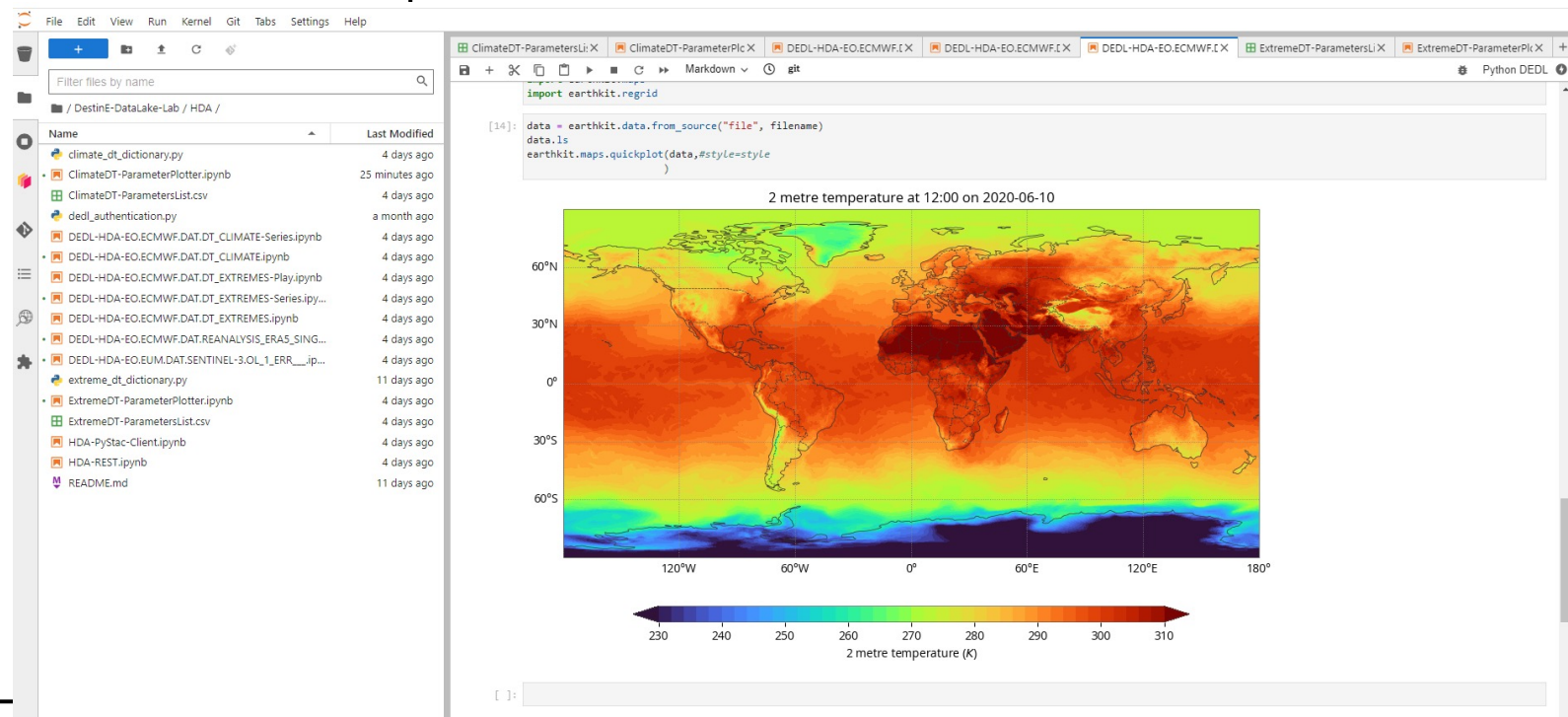
STACK service

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

Funded by
the European Union

Destination Earth

IMPLEMENTED BY



STACK service - Dask 101

Overview

Content

- Dask API introduction
- dask.distributed
- DestinE DataLake Dask Cluster

Duration: 20 min.

What is Dask?

Dask is a Python library for parallel and distributed computing.

Dask addresses the challenge of scaling Python code from a single machine to large clusters of machines. In the world of data science and scientific computing, Python is a popular language due to its ease of use and extensive libraries like NumPy, Pandas, and scikit-learn. However, these libraries often struggle to handle large datasets that exceed the memory capacity of a single machine or require parallel processing for efficient computation.

Dask provides parallel computing capabilities and allows Python developers to work with larger-

STACK: Data cube

Two options:

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

The screenshot shows a Jupyter Notebook with the following content:

- File Explorer:** Lists notebooks like `odc_app_bracciano_lake.ipynb`, `odc_app_jrc_training.ipynb`, and `odc_app_test.ipynb`.
- Code Cell 1:**

```

import datacube
%matplotlib inline
import matplotlib.pyplot as plt
import numpy as np

from IPython.display import Image
from matplotlib.colors import ListedColormap
from matplotlib.patches import Patch

print(datacube.__version__)
            
```
- Code Cell 2:**

```

bc.Datacube(app = 'water_extents_bracciano_lake', env = 'datacube')
products(with_pandas=True, dataset_count=False)
            
```
- Table:**

	name	description	license	default_crs	default_resolution
time					
prec	climatedt_2t_prec	digital twin forecast output, 2 metre temperat...	None	None	None
j2a	sentinel2_l2a	Sentinel-2a and Sentinel-2b imagery, processed...	None	None	None
- Code Cell 3:**

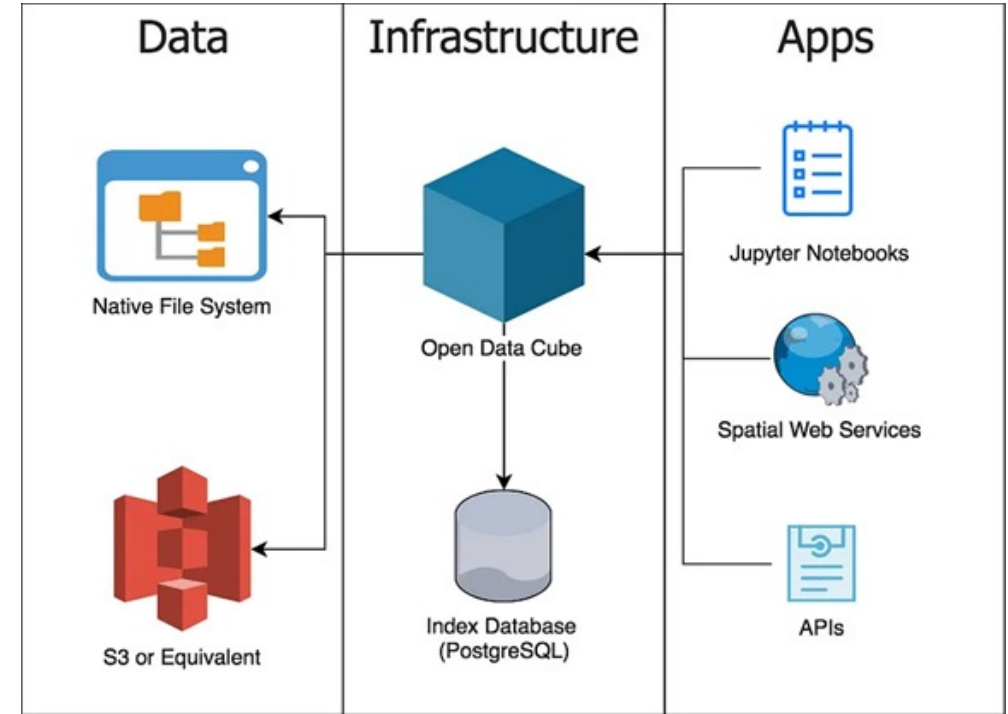
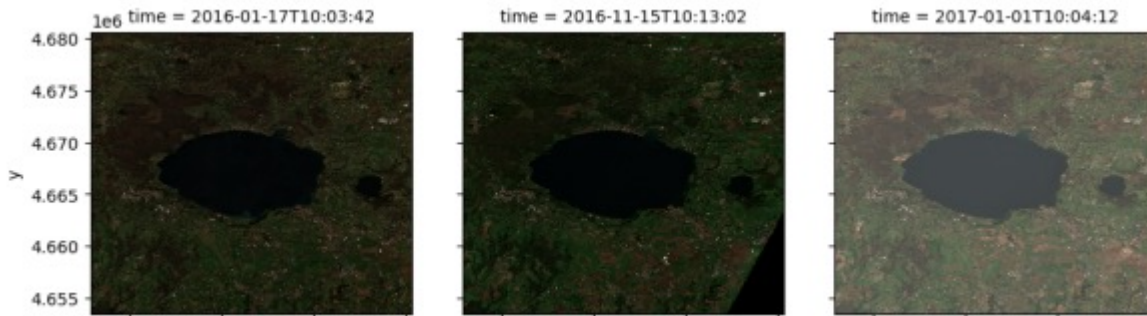
```

img = bc["prec"].mean(["latitude", "longitude"]).plot(figsize=(16, 4))
plt.title("Air temperature at 2m")
            
```
- Figure:** A 2x2 grid of heatmaps showing air temperature at 2m for different times: 2034-05-01T12:00:00, 2034-07-01T12:00:00, 2034-09-01T12:00:00, and 2034-11-01T12:00:00. The y-axis is latitude (41.8 to 42.4). Below the grid is a line plot of the temperature time series.

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.

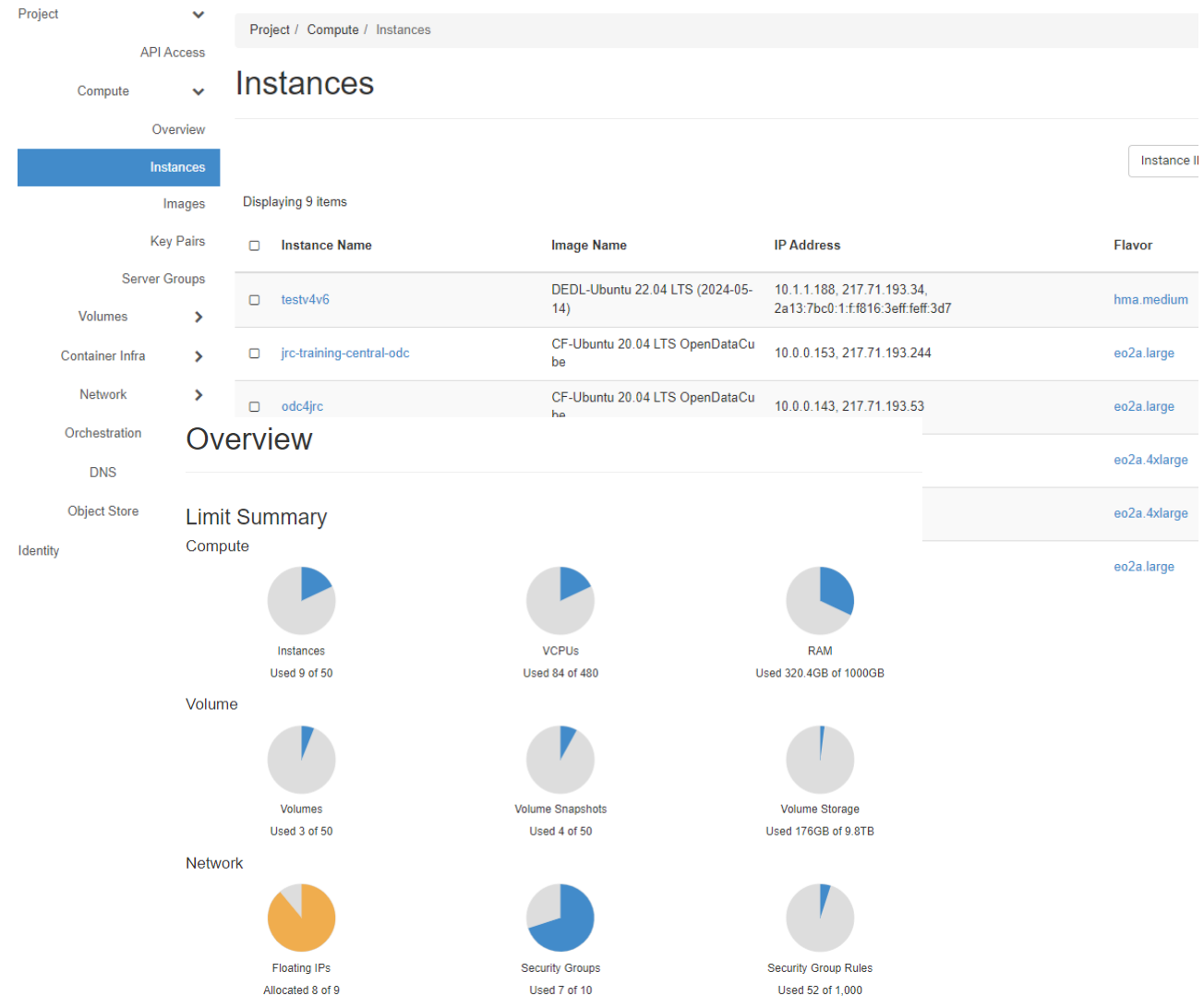
At its core, the ODC is a set of Python libraries and PostgreSQL database that helps you work with geospatial raster data.

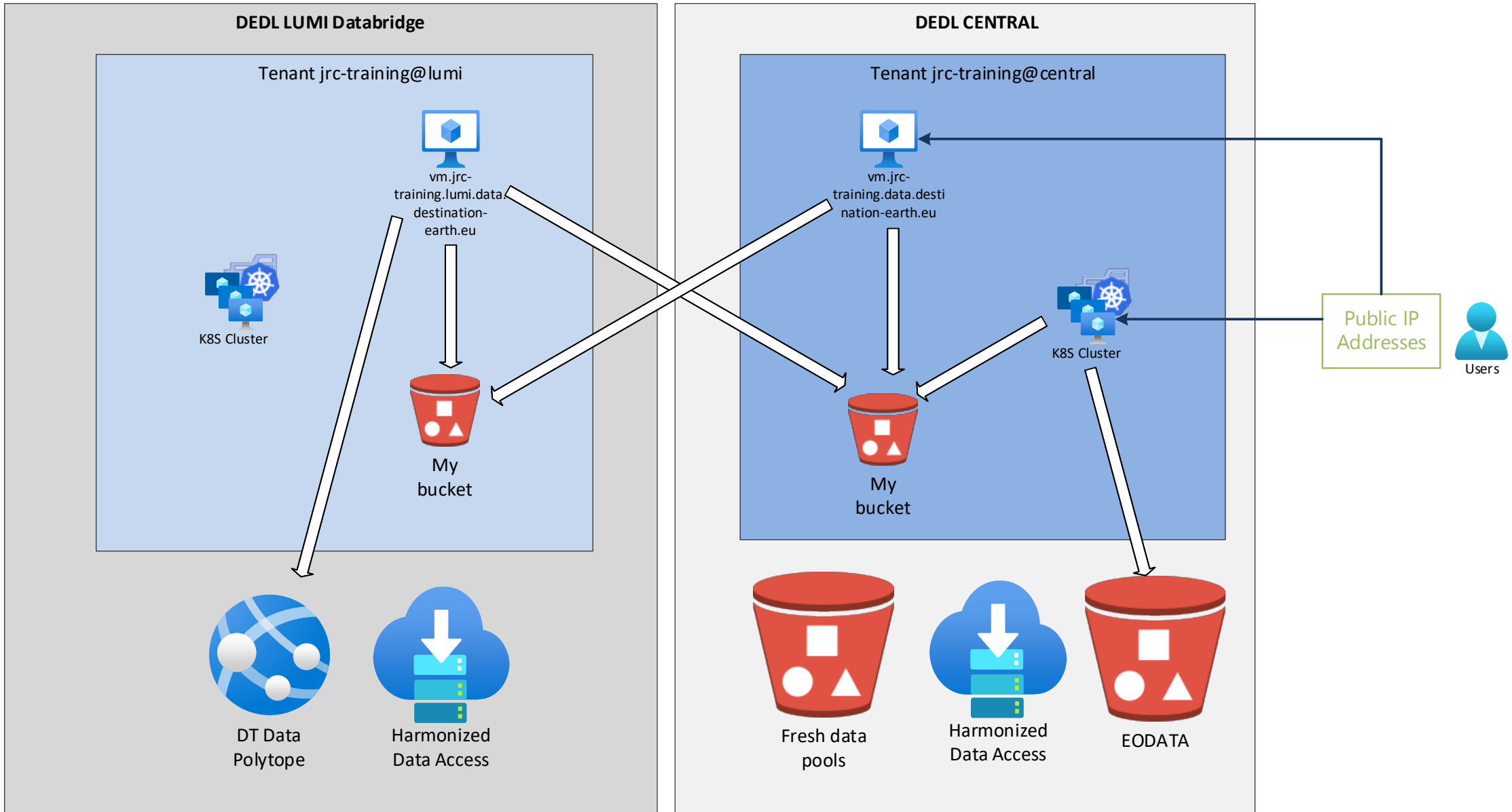
ISLET service IaaS/PaaS



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





HOOK service

User workflows



Pre-Defined processors

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

- Build your own workflow

Ordering Panel

- [Processors](#)
- [Orders](#)
- [Processing Site](#)
- [Sync Config](#)

Processors

-- Select Brand --

Display Name	Version	Name
Copernicus DEM Mosaic	1.0	copdem
DEDL Hello World	1.1.0	dedl_hello_world
Data harvest	0.0.1	data-harvest
DataCube Loader	0.1.0	datacube_loader
ODP Test	1.0.0	odp-test
Sentinel-1 Coherence/Interferometry	1.0.0	card_cohinf
Sentinel-1: Terrain-corrected backscatter (Private)	3.6.2	card_bs_private
Sentinel-1: Terrain-corrected backscatter	3.6.2	card_bs
Sentinel-2: C2RCC	1.1.1	c2rcc

Ordering Panel

Search

-- Select Brand -- -- Select Processor -- -- Select Status -- add_date Ascending

25/04/2024 14:37 to 26/05/2024 14:37

Ordering Panel

- [Processors](#)
- [Orders](#)
- [Processing Site](#)
- [Sync Config](#)

Orders

Id	Status	Order Name	Add Date	Processor
15686	done	availability_check	25-05-2024 14:30:03	odp-test (1.0.0)
15685	done	availability_check	25-05-2024 14:20:04	odp-test (1.0.0)
15684	done	availability_check	25-05-2024 14:10:04	odp-test (1.0.0)
15683	done	availability_check	25-05-2024 14:00:03	odp-test (1.0.0)
15682	done	availability_check	25-05-2024 13:50:04	odp-test (1.0.0)
15681	done	availability_check	25-05-2024 13:40:04	odp-test (1.0.0)
15680	done	availability_check	25-05-2024 13:30:04	odp-test (1.0.0)
15679	done	availability_check	25-05-2024 13:20:04	odp-test (1.0.0)
15678	done	availability_check	25-05-2024 13:10:03	odp-test (1.0.0)

HOOK Service

Interact with workflows using

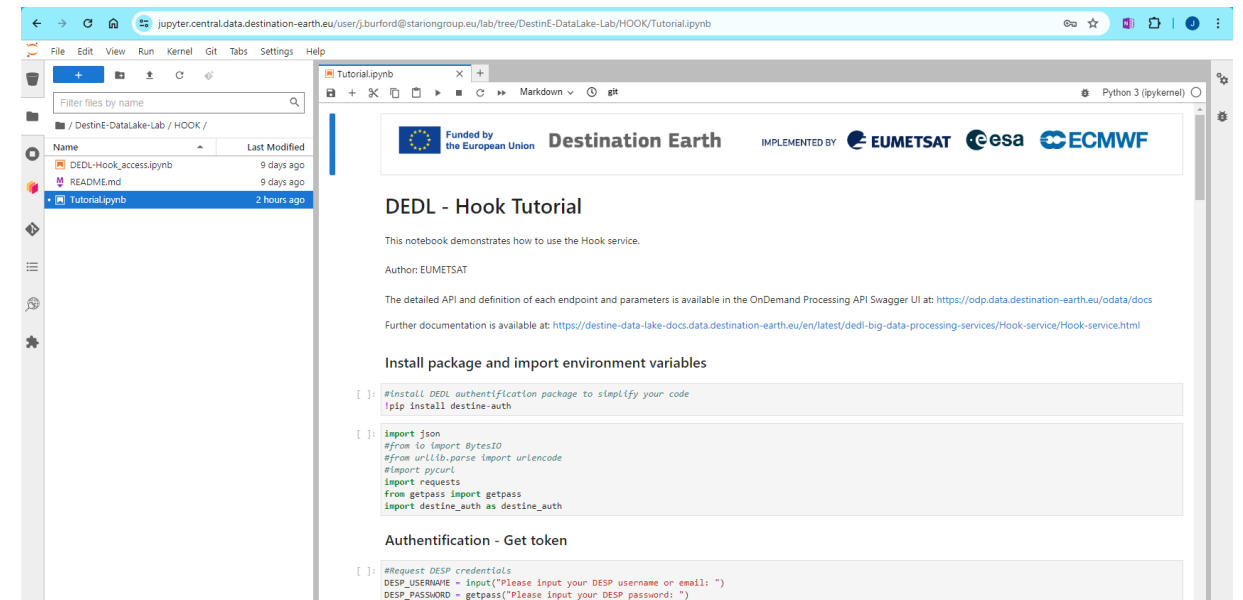
- REST API
- openEO API

User Interface:

- JupyterLab (STACK)
- openEO (UI)

Notebook HOOK/Tutorials

- Discover processors/ workflows
- Execute Data-Harvest workflow



The screenshot shows a JupyterLab interface in a browser. The left sidebar displays a file explorer for the directory '/DestinE-DataLake-Lab / HOOK /', listing files: 'DEDL-Hook_access.ipynb' (9 days ago), 'README.md' (9 days ago), and 'Tutorial.ipynb' (2 hours ago). The main area shows the 'Tutorial.ipynb' notebook with the following content:

DEDL - Hook Tutorial

This notebook demonstrates how to use the Hook service.

Author: EUMETSAT

The detailed API and definition of each endpoint and parameters is available in the OnDemand Processing API Swagger UI at: <https://odp.data.destination-earth.eu/odata/docs>

Further documentation is available at: <https://destine-data-lake-docs.data.destination-earth.eu/en/latest/dedi-big-data-processing-services/Hook-service/Hook-service.html>

Install package and import environment variables

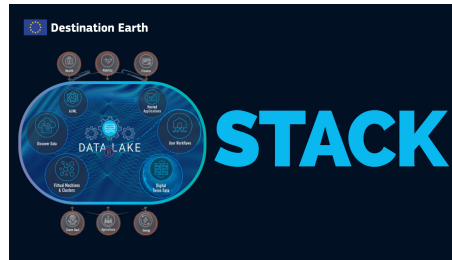
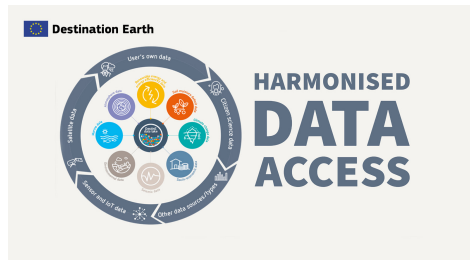
```
[ ]: #Install DEDL authentication package to simplify your code
pip install destine-auth
```

```
[ ]: #import json
#from io import BytesIO
#from urllib.parse import urlencode
#import pycurl
import requests
from getpass import getpass
import destine_auth as destine_auth
```

Authentication - Get token

```
[ ]: #Request DESP credentials
DESP_USERNAME = input("Please input your DESP username or email: ")
DESP_PASSWORD = getpass("Please input your DESP password: ")
```

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)

Pakistan Flood 2022 (Cloud Ferro, EODC)