DESTINATION EARTH

DATA LAKE

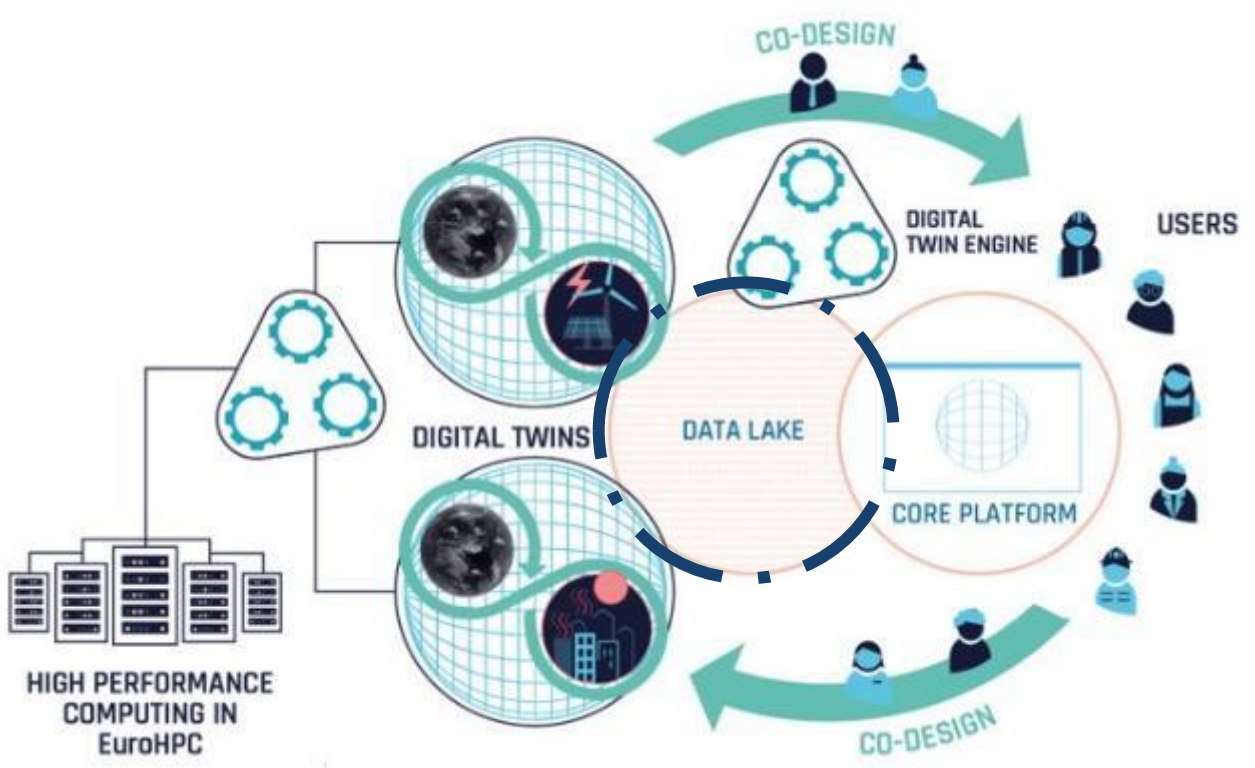
Michael Schick, Danaële Puechmaille

Destination Earth User eXchange – 13th November 2023

Funded by the European Union
Destination Earth implemented by ECMWF esa EUMETSAT

AGENDA

- DestinE Data Lake Pillars and Services
- DestinE Data Portfolio
- DestinE Data Lake Use cases



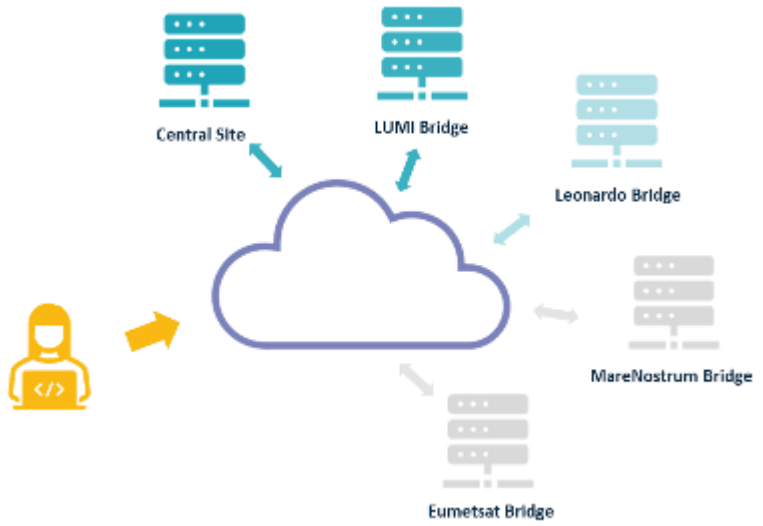
EUMETSAT EXPERTISE AND ROLE IN DESTINATION EARTH

- EUMETSAT has a long-term proven expertise on ***Data Management*** and ***Distributed Cloud Solutions*** responding to operational user requirements/ needs
- ***Data Federation*** with existing European data providers is part of the EUMETSAT strategy on Big Data management
- Use Synergies to maximise benefits for users and Members States

DESTINE DATA LAKE – 3 PILLARS

EUMETSAT build Destination Earth Data Lake on three main pillars

- Data Federation
- DestinE Digital Twins data
- Facilitate distributed near data processing



DESTINE DATA LAKE

Self-standing component

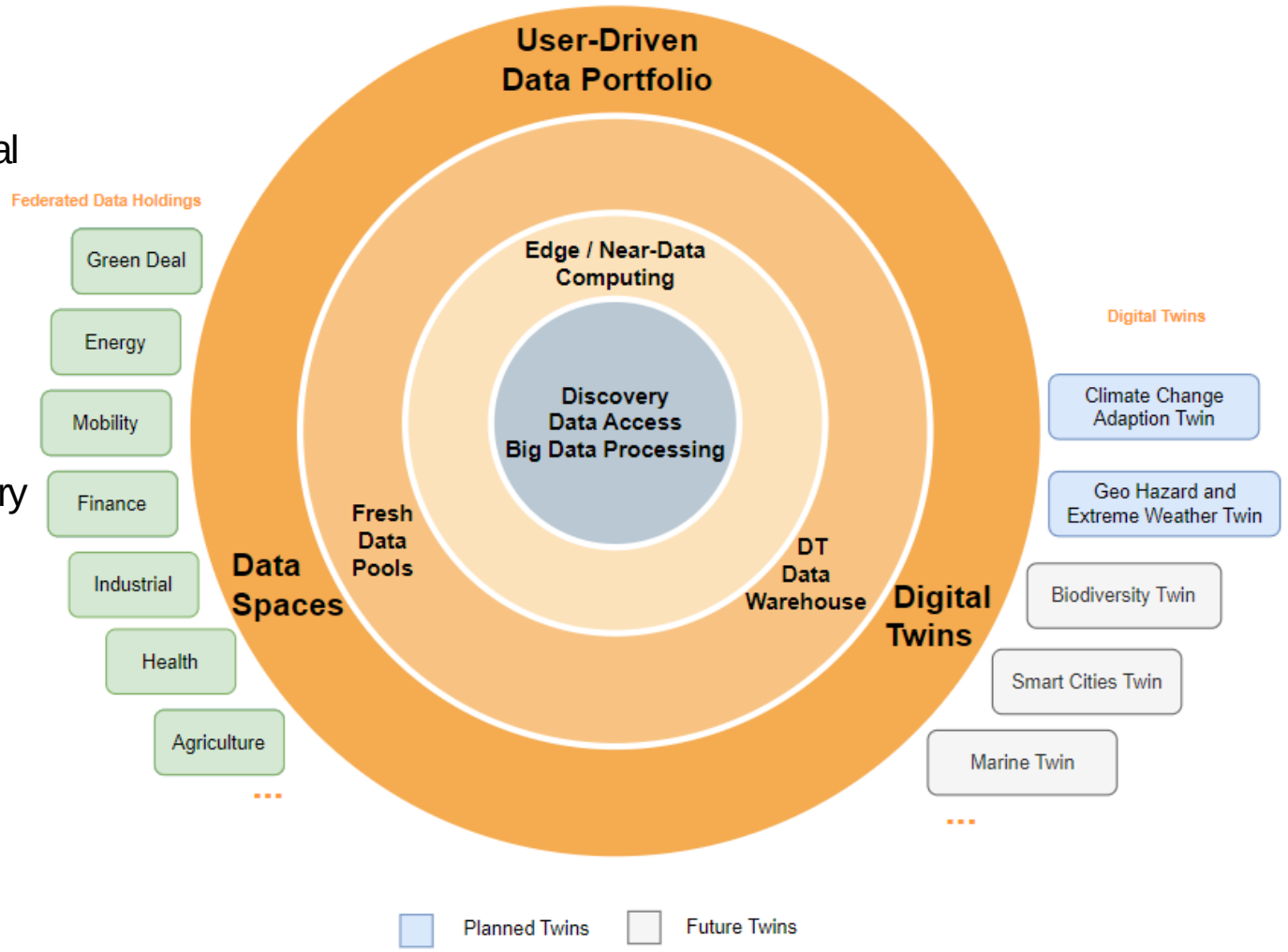
- Built from geographically distributed physical elements (central & edges)
- Distributed services – seamless access
- Implemented via European Industry

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

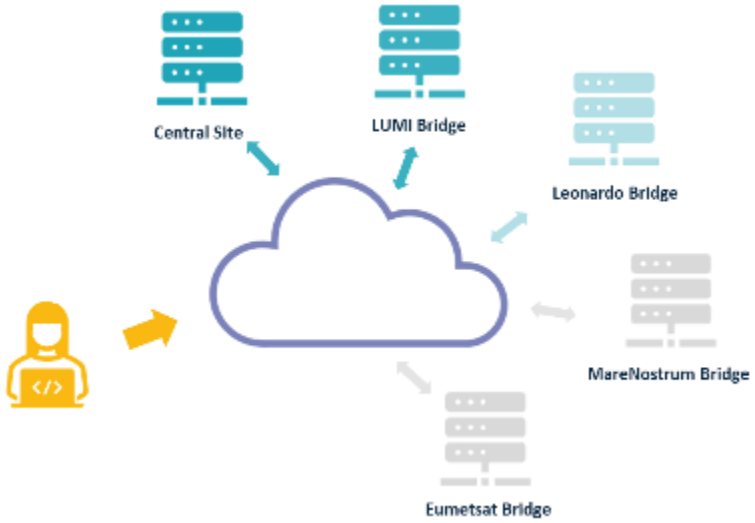
Big Data Processing

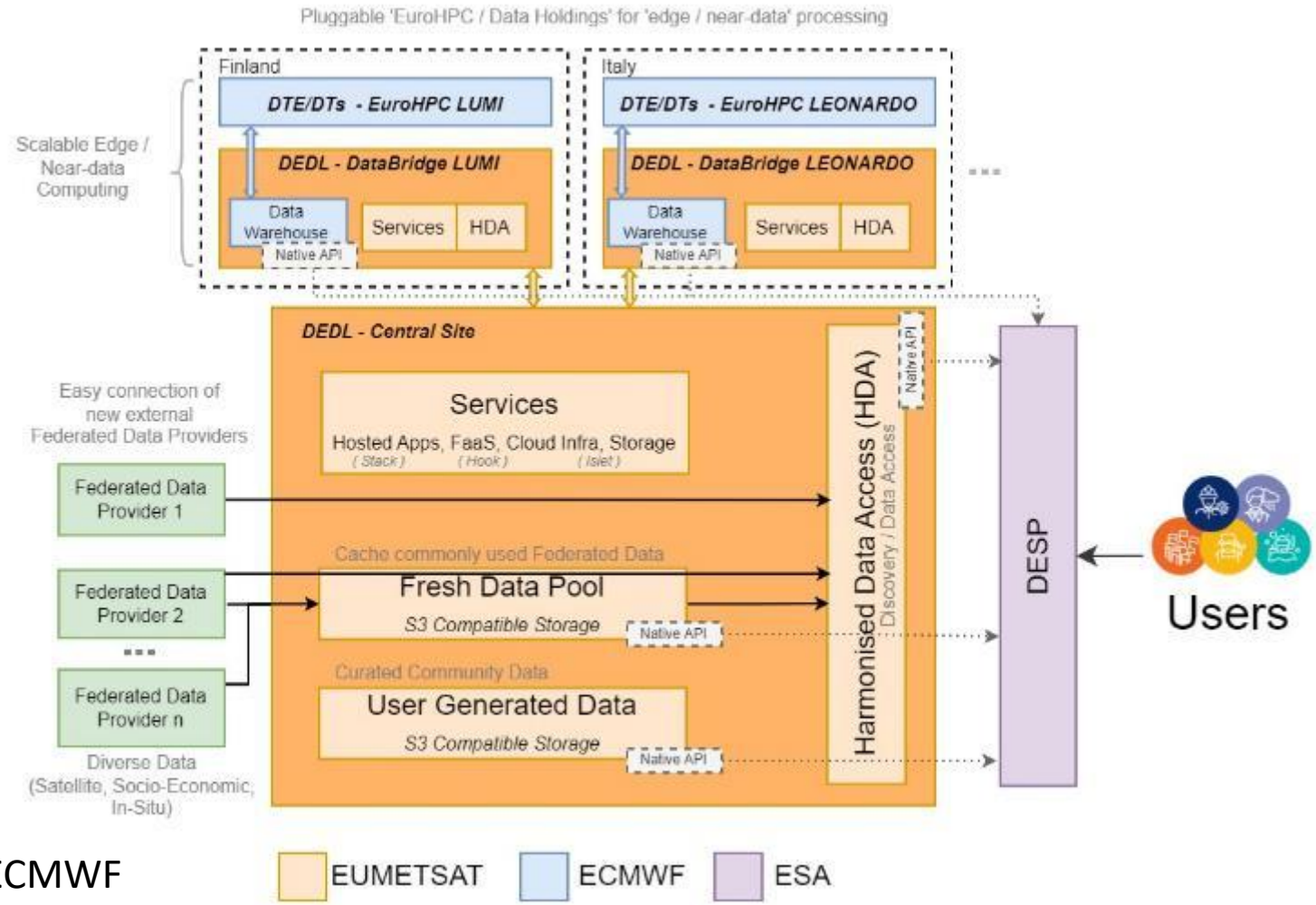
- Processing near data including distributed computing & workflows
- Supports & enables AI/ML applications



DESTINE DATA LAKE ARCHITECTURE

- Access to Digital Twins Data
- Data Federation
- Distributed Infrastructure
- DTE-DEDL Reference Architecture Interface*





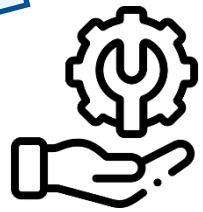
*Reference Architecture – Jointly developed with ECMWF

DESTINE DATA LAKE SERVICES

Discovery Services

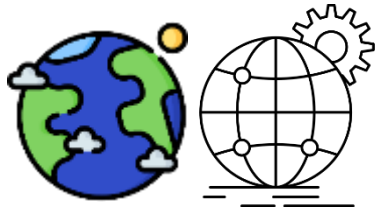
DE Data Portfolio
Datasets

Discovery

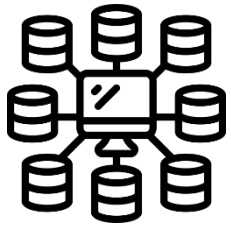


DEDL Services

Data Access Services

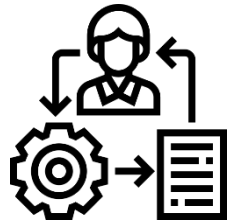


Digital
Twin Outputs



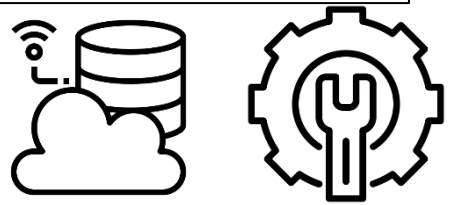
Federated
Datasets

Fresh
Data Pool



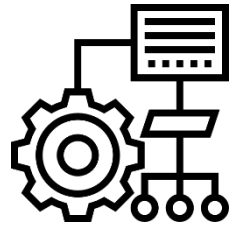
User-
Generated Data

Big Data Processing Services



Infrastructure& tools
(Islet Service)

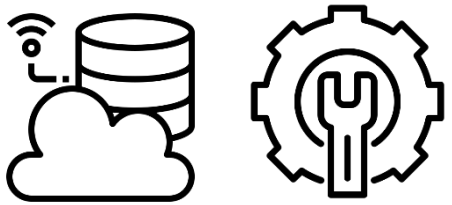
Hosted Applications
(Stack Service)



Functions
(Hook Service)

DESTINE DATA LAKE BIG DATA PROCESSING SERVICES

Infrastructure & Tools



Islet Service

- VMs, GPUs, Object Storage, k8s clusters
- blueprints (VMs, libraries & tools for data science and AI/ML)

For Users who

- **set up and manage their own development environment**
- **deploy already existing processing chains**

Hosted Applications

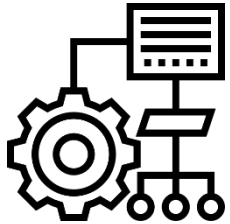
Stack Service

DEDL-provided off-the-shelf working environments and applications
(JupyterHub ecosystem, DASK Gateway)

For Users who

- **want ready-to-use applications and environments**

Functions



Hook Service

Predefined processing workflows/ functions
User-defined workflows
System or User-defined data cubes

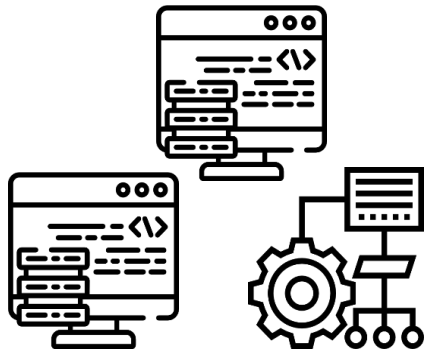
For Users who

- **want ready-to-use building blocks for their applications**
- **want advanced processing services**

DEDL – BIG DATA PROCESSING SERVICES

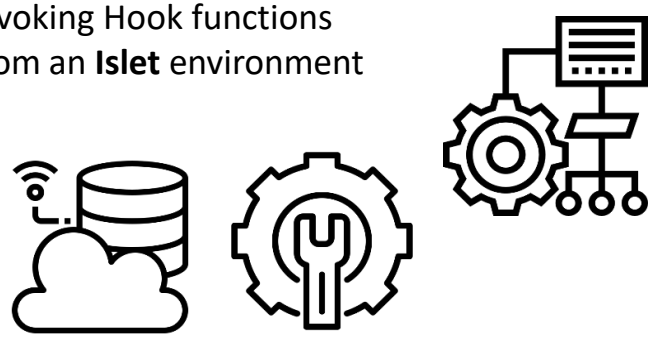
Users can pick and mix big data processing service offerings

Stack (JupyterHub) + **Islet-Storage** (Uploading own data & Storing results)

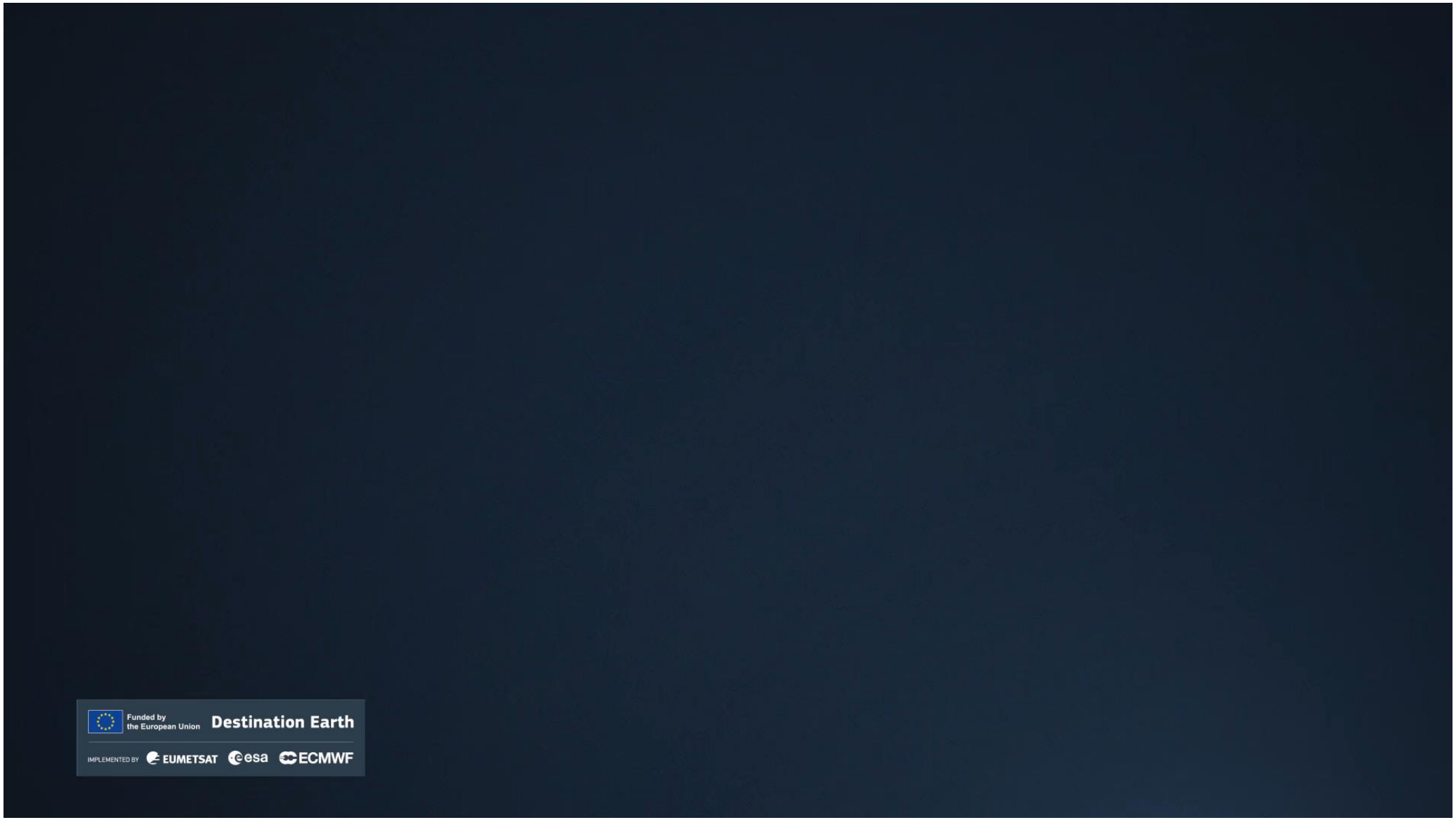


Using **Stack** application (e.g. DASK Gateway) + **Hook** functions in a **Stack** environment (JupyterHub)

Invoking Hook functions from an **Islet** environment

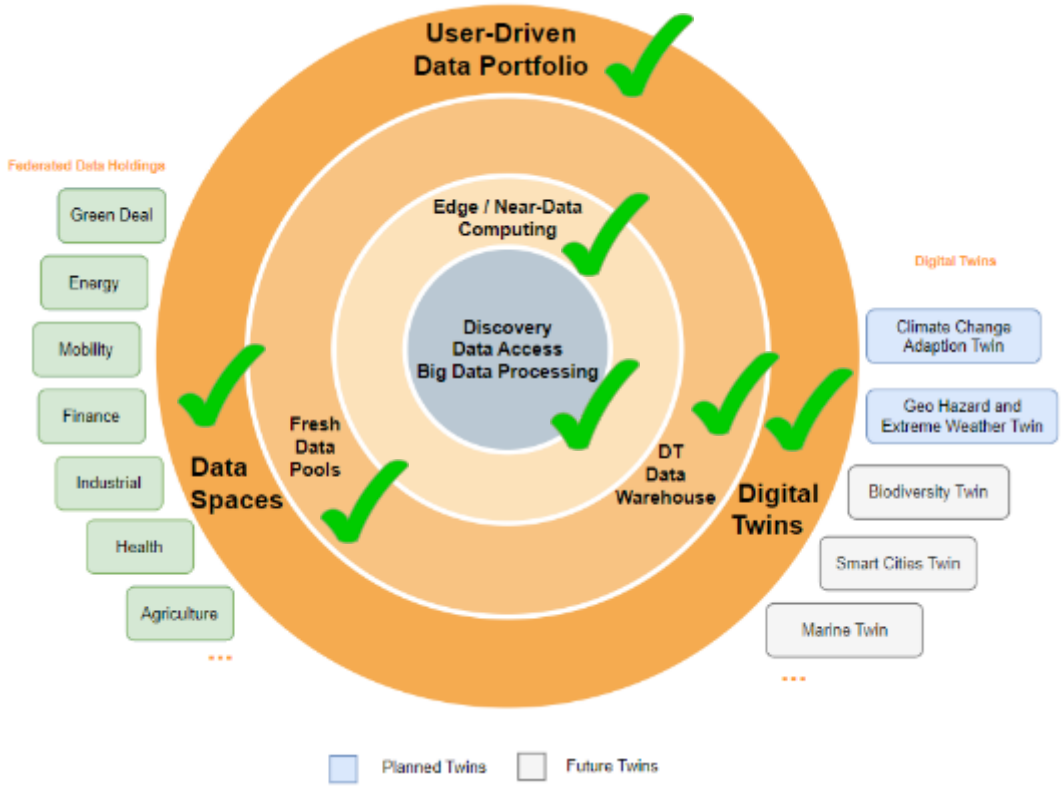


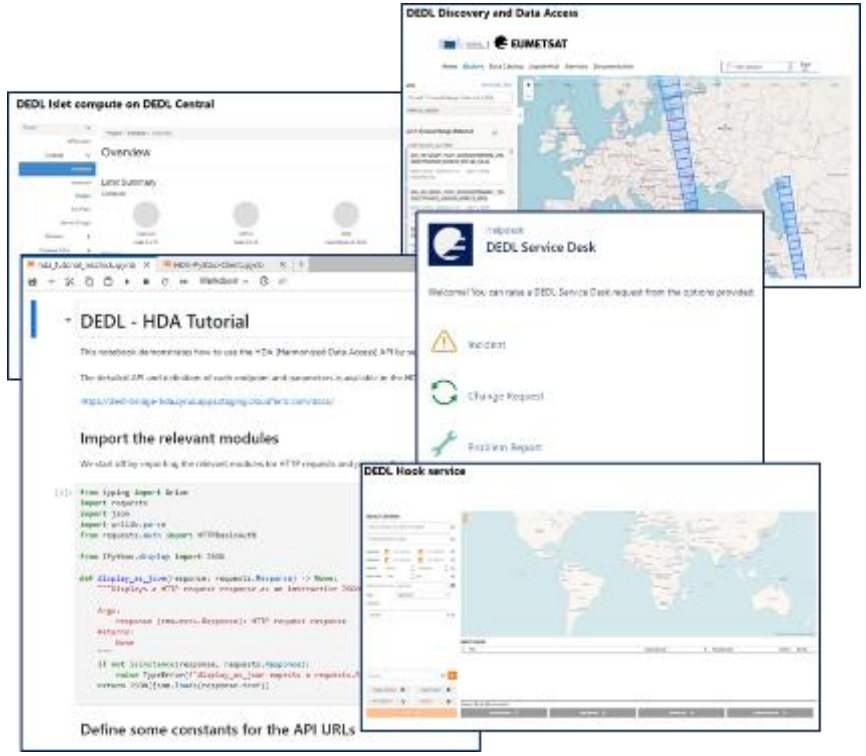
Invoking **Stack** applications (e.g. DASK Gateway) in an **Islet** environment


Funded by the European Union
Destination Earth
IMPLEMENTED BY EUMETSAT esa ECMWF

DESTINE DATA LAKE - STATUS

- Increment#1 (Minimum Viable Service) accepted in October 2023
- Ready for integration with the Digitial Twins and the Core Platform Service





DESTINE DATA LAKE IN MORE DEPTH

- Poster - Data Lake Services (CloudFerro)
  - Architecture & Services
  - Q&A

- Poster - Harmonised Data Access (HDA)(CS France)
  - Access to Federated Data & DT Outputs using HDA
  - STAC based API
  - Python wrapper EODAG

- Session A4
  - Multi-cloud processing with DASK: demonstrating the capabilities of the DestinE (EODC)
  - DestinE Data Lake Harmonized Data Access (CS Group)

Poster session
Mon. 13th Nov 16:45 – 19:00
Session A4 : Data Lake and Data Spaces
Tue. 14th Nov 15:45 – 17:15

DESTINE DATA LAKE SERVICE PROVIDER

CloudFerro

eodc

ES GROUP

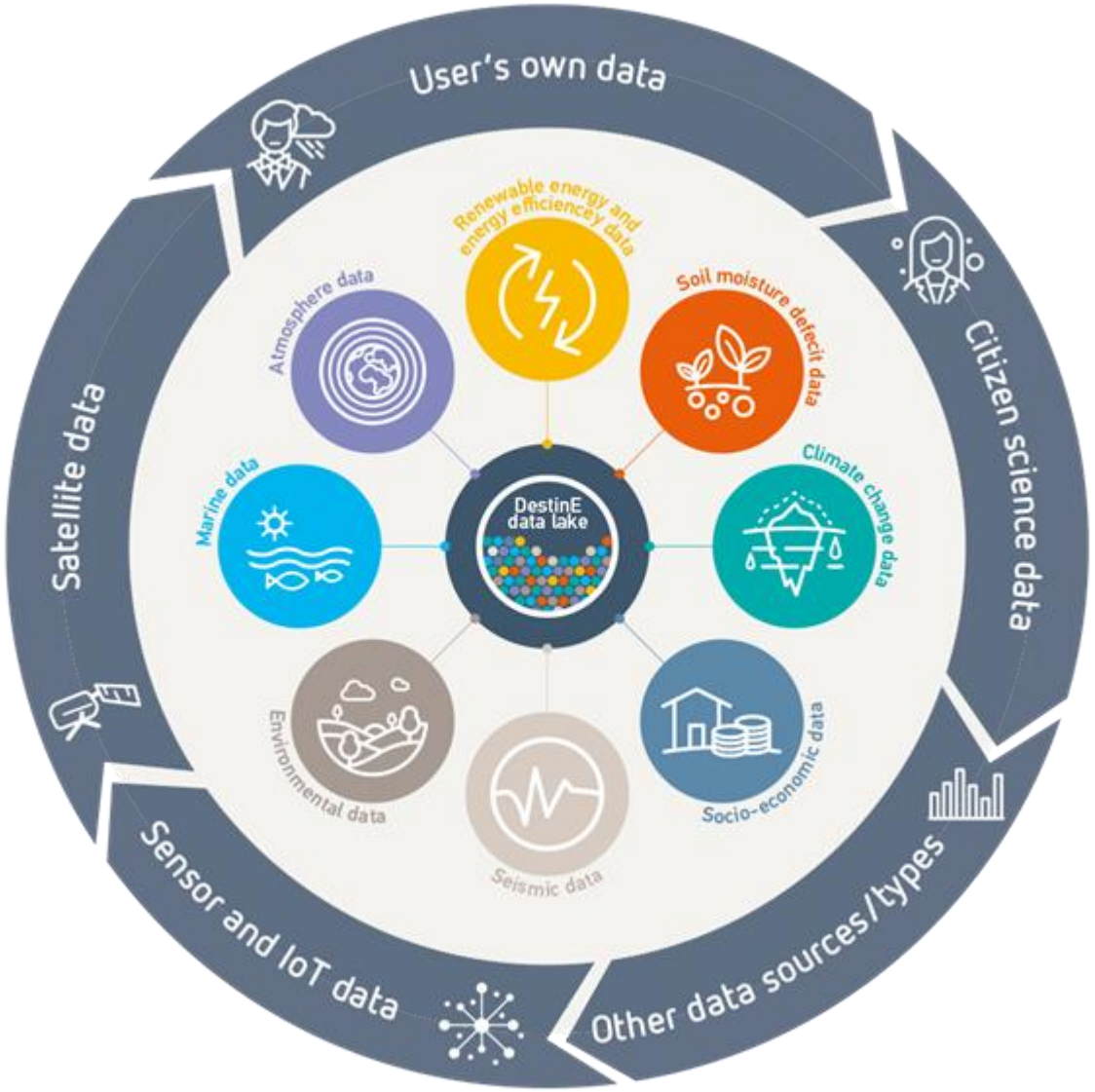
DestinE Data Bridge Infrastructure in LUMI

DESTINE DATA PORTFOLIO

- **Evolving Data Portfolio**
- User **influenced and agreed with EC**
  - Managed and Controlled by DestinE JCCB

- Digital Twins Data
  - Climate Change Adaptation
  - Extreme Weather and Geo hazards

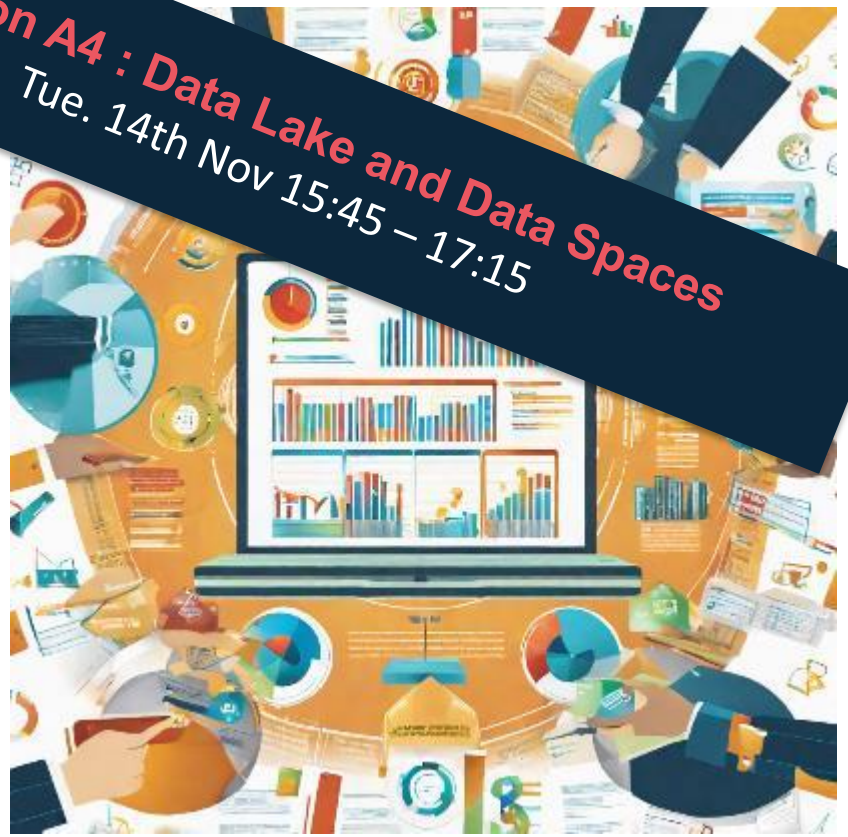
- Federated datasets
  - Contributing missions (EUMETSAT, ESA, ECMWF)
  - Copernicus Satellites & Services data
  - Eurostat
  - ISIMIP
  - IAGOS
- Harmonized APIs: STAC compliant



DATA GOVERNANCE

- Data Classification
- Data Policy & Data Stewardship
- Data Access Control
- Data Lifetime and Retention Management
- Data Attribution
- Data Provenance
- Data Quality
- DestinE Data Governance

Session A4 : Data Lake and Data Spaces
Tue. 14th Nov 15:45 – 17:15



DESTINE DATA LAKE USE CASES

- EUMETSAT organised two workshops to identify use case candidates suitable for validation of Destination Earth Data Lake component. Those use cases will be procured in Q4 2023 with target to be demonstrated by the end of Phase I.

- Those use cases will be in the areas of
  - Agriculture and Biodiversity
  - Health
  - Energy
  - and potentially also in the area of support to developing AI/ML applications

Five out of nine use cases selected for procurement

DESTINE DATA LAKE USE CASES

- Others use cases are developed by DestinE Service provider to demonstrate direct usage of developed services
  - Drought & Flooding (EODC) and Hydrology (CS France)

Pakistan Flooding - 2023

Danube Delta

THANK YOU VERY MUCH