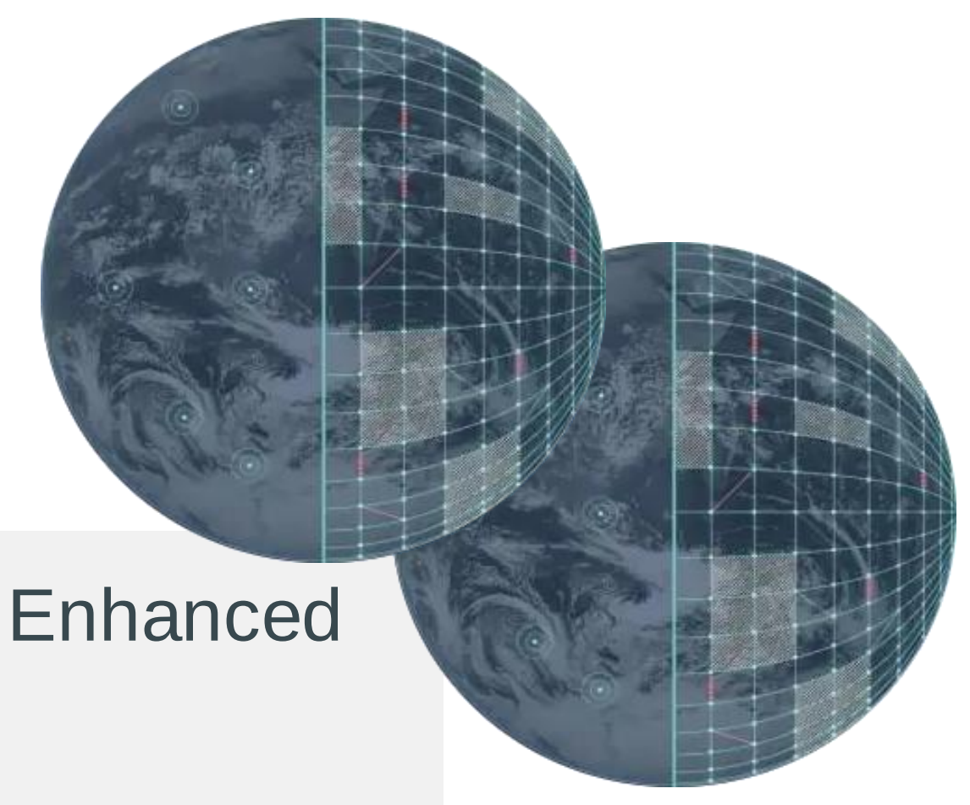
Towards Destination Earth Data Lake (DEDL) Service

M. Ziółkowski[1], M. Drupka[1] J. Musiał[1],
J. Bojanowski[1], P. Mujta[1], M. Bylicki[1], A. Lambare[2]
A. Le Carvennec[2], T. Hilton[2], C. Reimer[3], M. Schick[4]

CloudFerro   a Sopra Steria company   eodc   EUMETSAT

Introduction to DestinE Initiative

The objective of the **Destination Earth initiative (DestinE)** is to develop a very high precision **digital replica of the Earth** that models enviromental processes governing activities on the planet. This will enable the monitoring and **simulation of natural processes** as well as the impact of human activity. It also facilitates the creation and testing of sustainable development scenarios, aligning with the European Union's Green Deal and the EC Digital Strategy priorities.



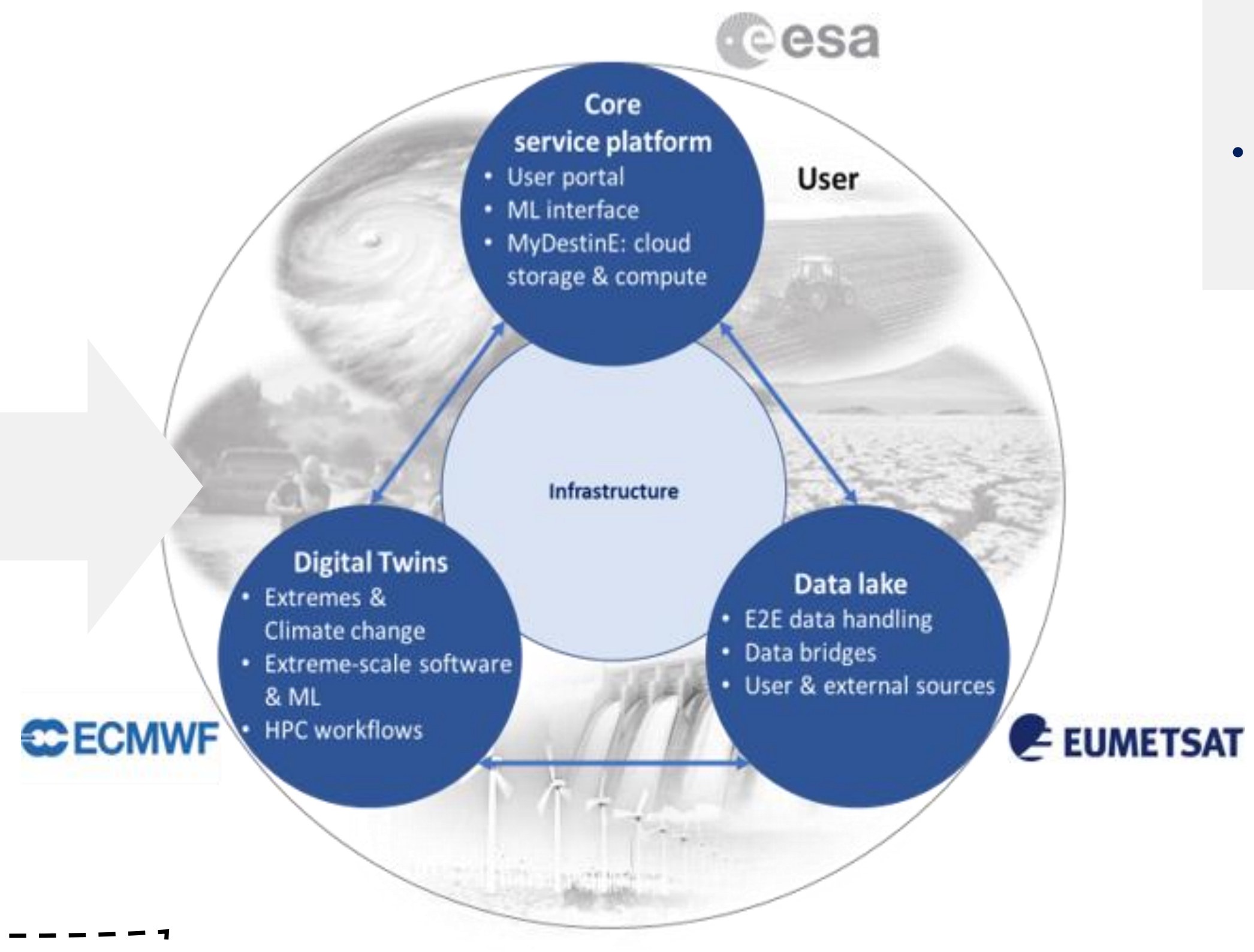
Implementing Thematic Digital Twins For Enhanced Enviromental Resilience

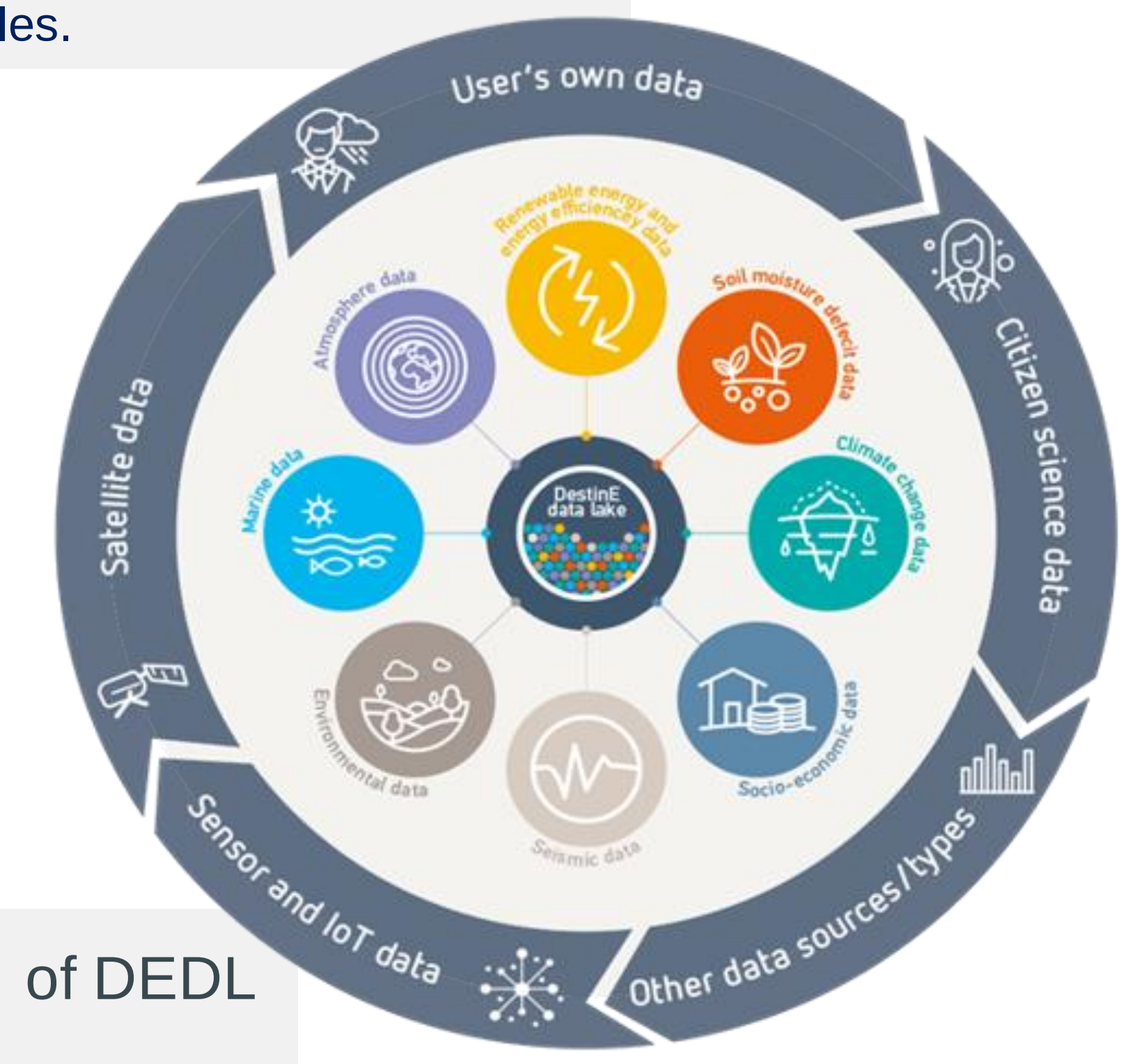
Currently two thematic DT are being implemented:

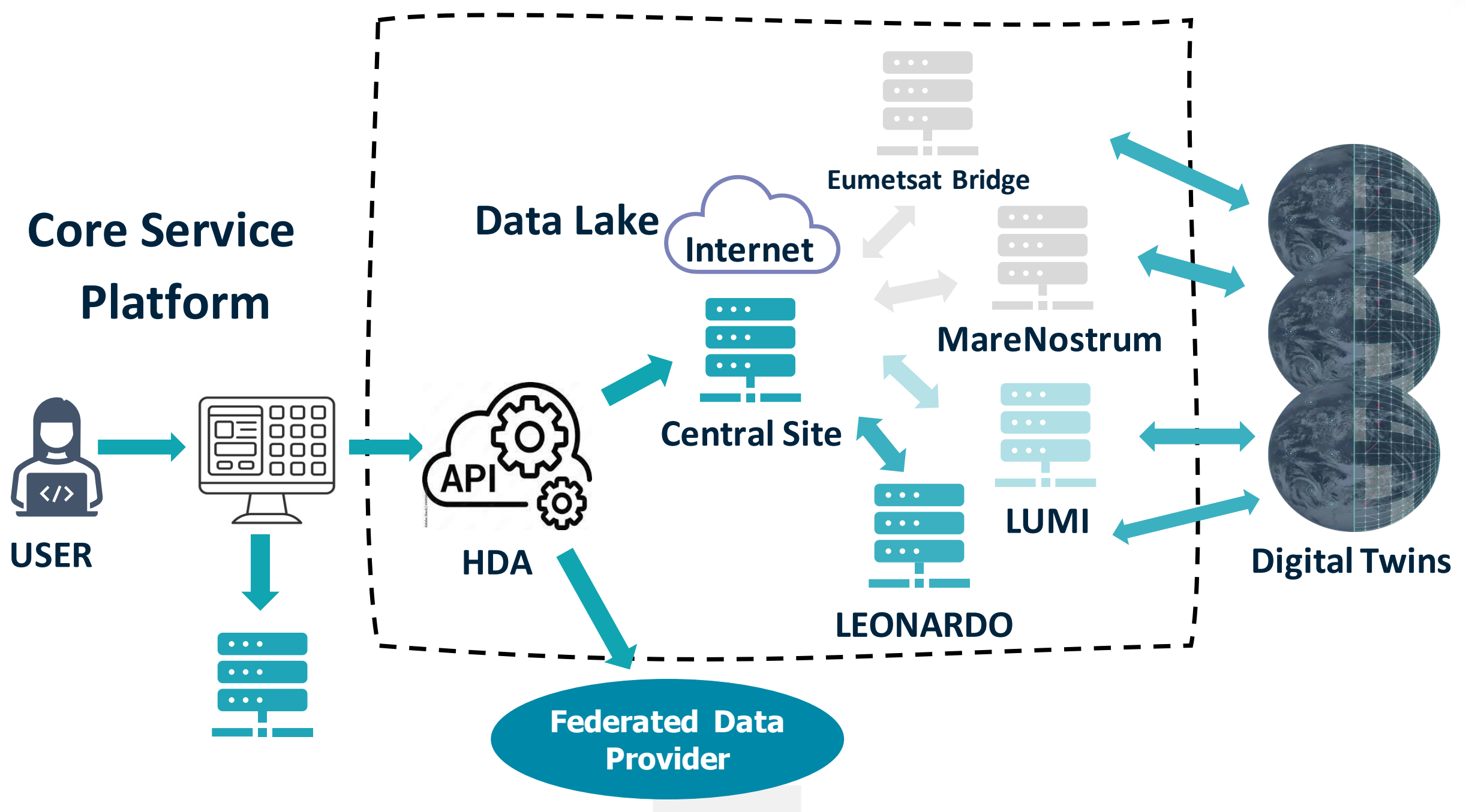
- **The Geohazard / Extreme Weather Digital Twin** that will allow anticipating the occurrence and impacts of extreme natural events (e.g. flooding, droughts, forest fires).
- **The Climate Change Adaptation Digital Twin** that will be used to predict the impact of climate change with unprecedented reliability at regional and national scales.

Three Components of DestinE

1. **Core Service Platform (DESP)** - entry point for user. It will provide evidence-based decision-making tools, applications and services, based on an open, flexible, and secure cloud-based computing infrastructure managed by ESA.
2. **Data Lake (DEDL)** - provides discovery, data access as well as near data processing to explore and process simulations from DTs and observations managed by EUMETSAT
3. **Digital Twins (DTs)** - highly complex models of Earth's environment, managed by ECMWF, to simulate and predict potential impact of human activity on many thematic domains.







Unified Data Access and Processing Capabilities of DEDL

The Data Lake (DEDL) brings together data from ESA, EUMETSAT, ECMWF, Copernicus, DTs, and many other federated repositories. All the data sources are available via the **Harmonized Data Access (HDA)** interface that allows users to discover, access and process data using a single endpoint without knowing where the data originated from. **Apart from the HDA**, DEDL will offer near data processing using **cloud computing via the ISLET** service (IaaS**), STACK application development environment** such as JupyterHub/DASK based on Python, and **Hook service for workflows** to run algorithms and/ or routine tasks (e.g. generate temporal composites).

User Access and Computation Management in the DestinE Enviroment

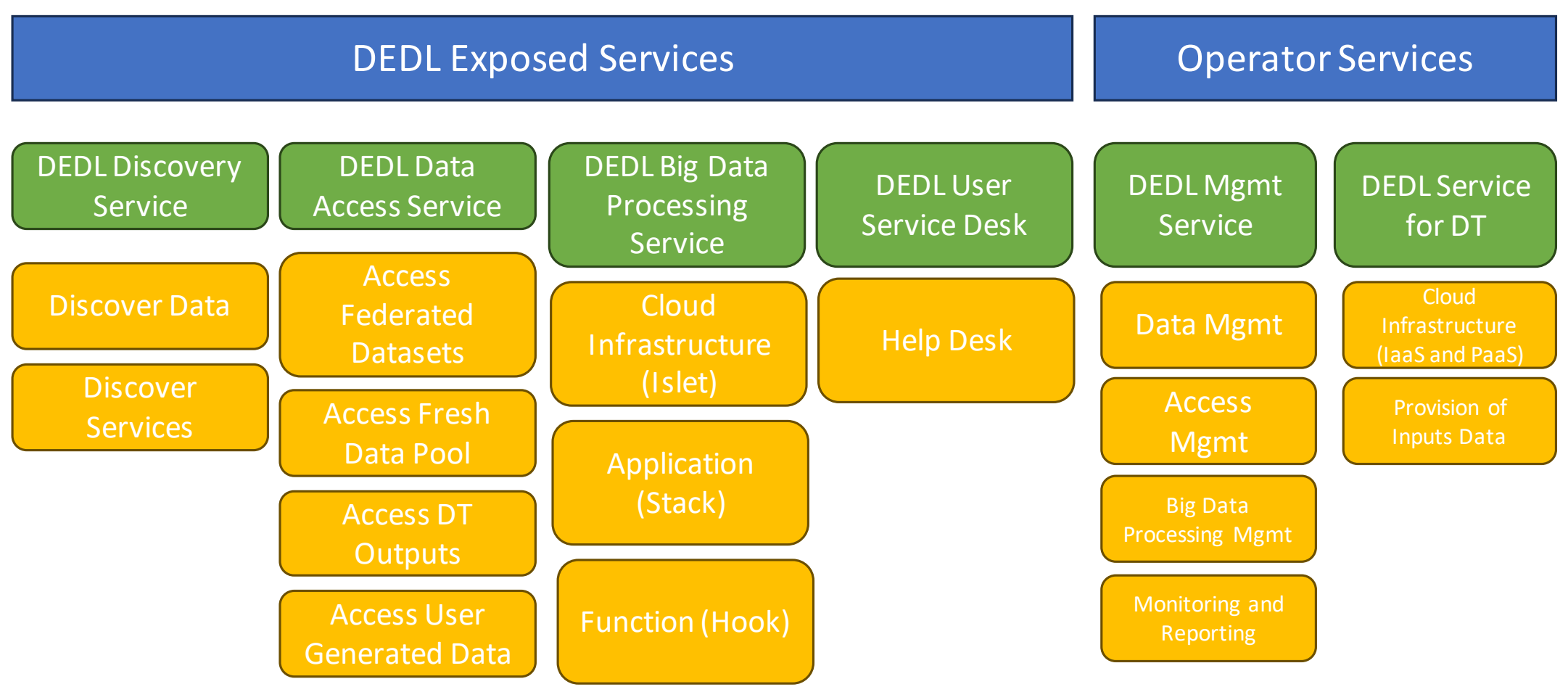
Within the DestinE environment, user will have access to services and data via Destination Earth Core Service Platform (DESP) based on DestinE Access Governance. Users can request access to near data processing services and once granted they can compute near the data such as DT outputs and fuse this with federated datasets.
The data placement can be performed:
1) Central Site with users' local storage and "Fresh Data Pool" acting as intermediate storage
2) Data Bridges (Lumi, Leonardo, EUMETSAT & Mare Nostrum), where large volume of data is available
3) Federated Data Providers
The data bridges have also processing capabilities including AI/ ML capabilities.

Accessing External and Internal DEDL Services via DESP Interface

DEDL services will be available to users via the Core Service Platform:

- **HDA API** for data discovery and access
- **Storage services** including Fresh Data Pool for intermediate data, DT outputs and user's private data
- **Near Data Processing services**: cloud computing, JupyterHub/DASK and workflows.
- **Service Desk –** via Core Platform Help Desk

Internal DEDL services include maintenance of the entire service (hardware/ software infrastructure) under responsibility of CloudFerro.



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 Overview

- Data Lake (DEDL) service is operated by **CloudFerro** consortia under responsibility by **EUMETSAT** and is one of the three components of **DestinE.**
- Data Lake Services will be available from **DESP**, which is under responsibility by ESA
- **DEDL** will **host data** from the Digital Twins (ECMWF), **EUMETSAT, ESA, ECMWF, Copernicus** and other federated repositories
- The DEDL infrastructure is distributed across Europe with the Central Site in Poland and data bridges in Germany, Finland, Italy and Spain

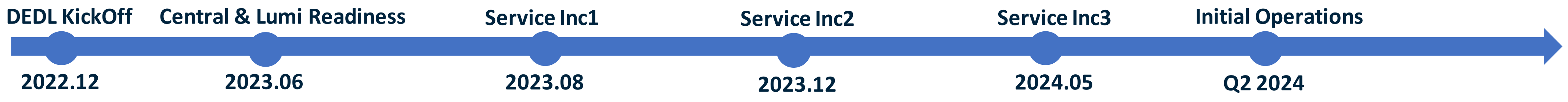
The DEDL offers not only storage but near data processing capabilities through:

1. **Islet Service** – grants access to Virtual Environment based on the OpenStack
2. **Stack Service** – grants access to JupyterHub with Python and DASK
3. **Hook Service** – grants access to processors to execute workflows

Although the **DEDL** is still under development, many of its functionalities have already been **implemented**.

DEDL Big Data Processing Services

- The Islet service is deployed on OpenStack with the Horizon interface, which allows users to manage virtual machines and local users' storage using a GUI.
- The Stack service based on the JupyterHub/ DASK with Python allow users to develop their own applications and services based on DestinE datasets.
- The hook service will allow users to execute workflow to for example derive information (e.g. temporal composites, Sentinel-1 coherence and backscatter) using DestinE data.



References

[1] CloudFerro S.A. Nowogrodzka 31, Warsaw Poland
[2] CS Group, avenue Galilée, LE Plessis Robinson, France
[3] EODC, Franz-Grill-Straße 9, Vienna, Austria
[4] EUMETSAT, Eumetsat-Allee 1, Darmstadt, Germany

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