Earth Observation Data Center
for Water Resources Monitoring

Multi-cloud processing with Dask

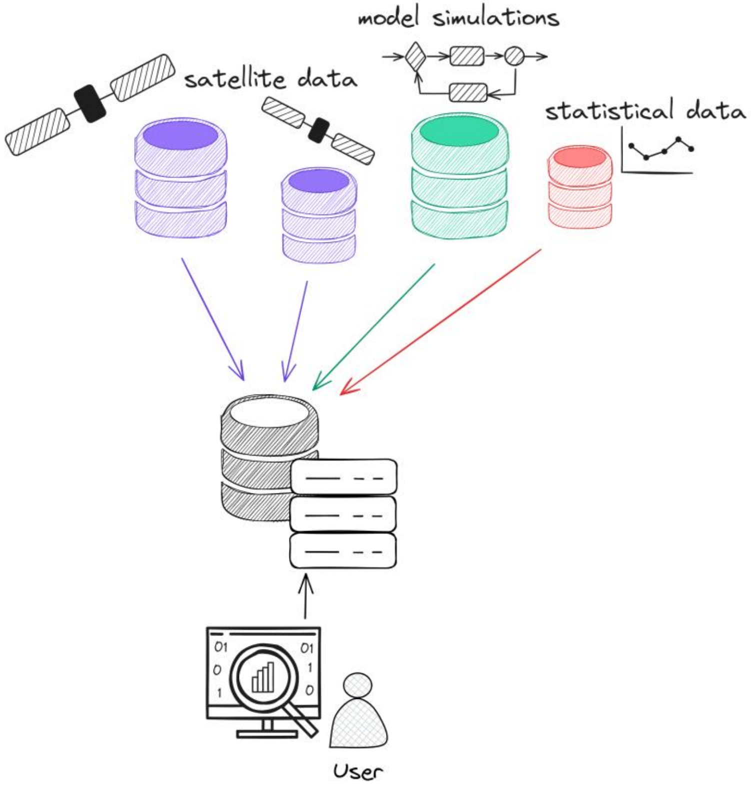
Demonstrating the capabilities of DestinE Data Lake (DEDL)

Christoph Reimer[1], Lukas Weidenholzer[1], Sean Hoyal[1], Bernhard Raml[2], Martin Schobben[2], Matthias Schramm[2], Wolfgang Wagner[1,2], Christian Briese[1]

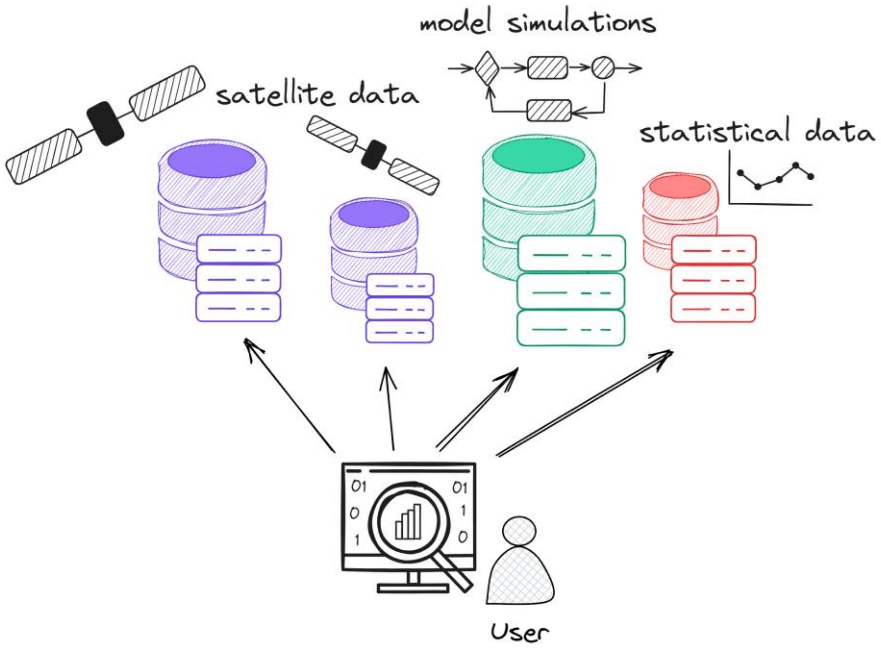
[1] EODC Earth Observation Data Centre for Water Resources Monitoring GmbH
[2] Technische Universität Wien, Department of Geodesy and Geoinformation

Traditional big data processing

- Access to a diverse set of data required to address todays challenges
- Aggregation of data into a single data repository (data fortress)
- Costs of data duplication
  - Storage
  - Ingress/Egress
- Data management burden
  - healthy and up-to-date repository needed



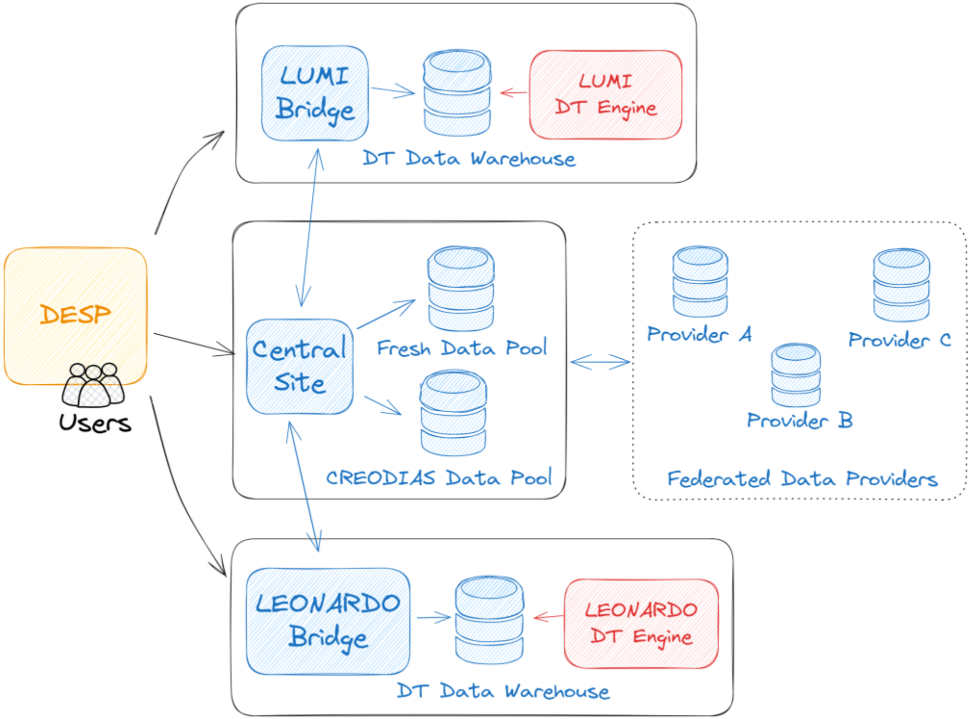
Data proximate computation



- High quality data comes with an increased data volume
- Moving computation to the data
- Transfer data only when needed
- Lower costs for storage and network
- Access always the latest available data

DestinE Data Lake Architecture

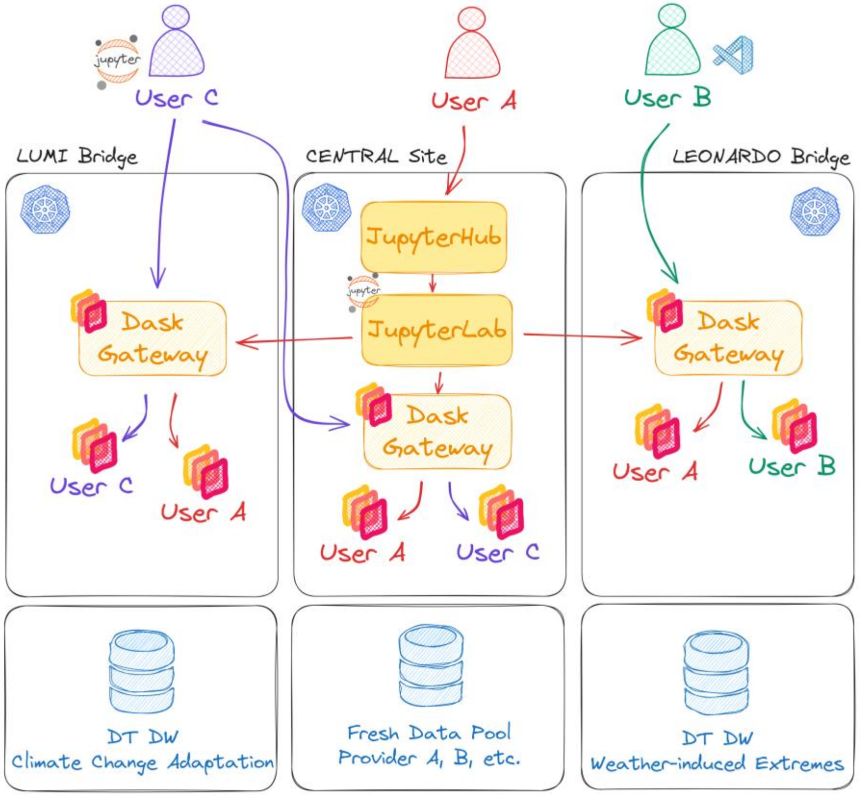
- Geographically distributed cloud infrastructures
- Direct access to DTE outputs
- Data federation via HDA
  - EO data and statistical data



DEDL Stack Service

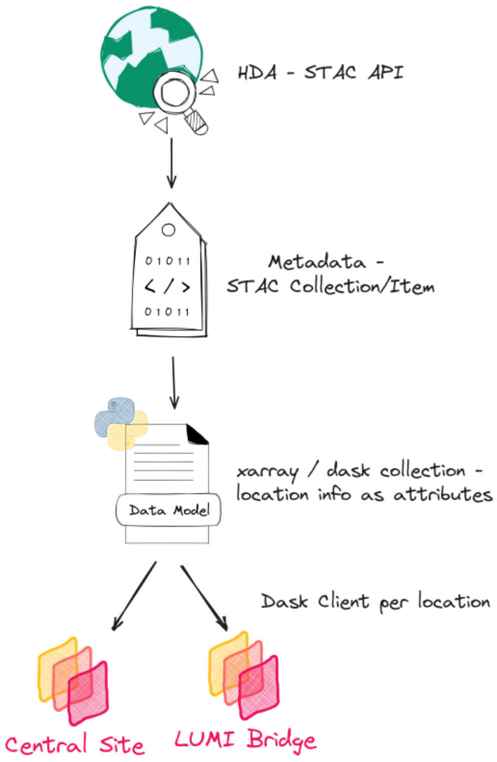
- Big data processing services (Islet, Stack, Hook)
- Stack Service (Application Service) targeting a specific user community such as:
  - Scientific programmers, data analysts / scientist, etc.
- Python as programming language of choice
- Managed service build on top of
  - Jupyter Ecosystem
    [JupyterHub, JupyterLab, Jupyter Enterprise Gateway]
  - Dask Ecosystem [Dask Gateway, dask-kubernetes]

DEDL Dask Usage Scenarios



[2] T. Augspurger et al.: Multi-Cloud workflows with Pangeo and Dask Gateway. February 05, 2022, https://doi.org/10.1002/essoar.10510416.1

Multi-cloud processing with Dask

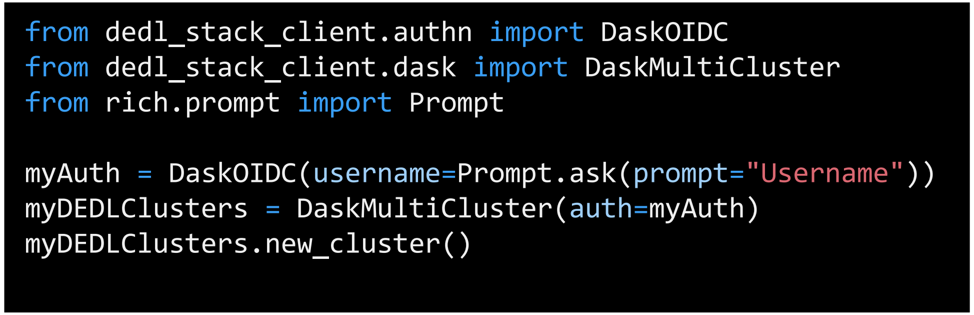


- Make use of data discovery services to get metadata
- Metadata (STAC collection) holds information about the data host
- Propagate the data host / location information further to Python data model
- Route computations via multiple Dask clients to the right cluster

https://github.com/eodcgmbh/dedl-stack-client

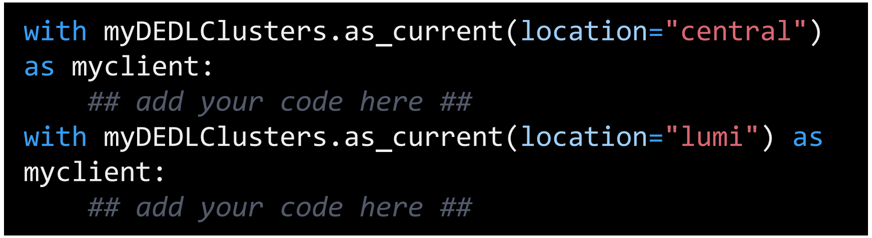
DEDL Stack client

- Implementation to support multi-cloud processing with Dask
- Manage multiple Dask clusters, as if one would manage a single cluster
  - Automatically creates Dask clusters based on a cluster registry object
- Tailored to the needs of DEDL by providing an OIDC authentication layer implementation
- Context manager to interact with different Dask clusters



```python
from dedl_stack_client.authn import DaskOIDC
from dedl_stack_client.dask import DaskMultiCluster
from rich.prompt import Prompt

myAuth = DaskOIDC(username=Prompt.ask(prompt="Username"))
myDEDLClusters = DaskMultiCluster(auth=myAuth)
myDEDLClusters.new_cluster()
```



```python
with myDEDLClusters.as_current(location="central")
as myclient:
    ## add your code here ##
with myDEDLClusters.as_current(location="lumi") as
myclient:
    ## add your code here ##
```

Use Cases

- Disaster events in 2022 with high socio-economic impact
  - Pakistan Flood
  - Drought in the Po Valley in Italy
- What “IF” scenario
  1. Policy maker needs to take decisions about risks
  2. Analysis needed → Experts are tasks to generate information for decision making
  3. What is the situation today? (forecast and observation)
  4. Which areas are under risk? (forecast)
  5. Does the alert system work correctly? (forecast vs. observation)

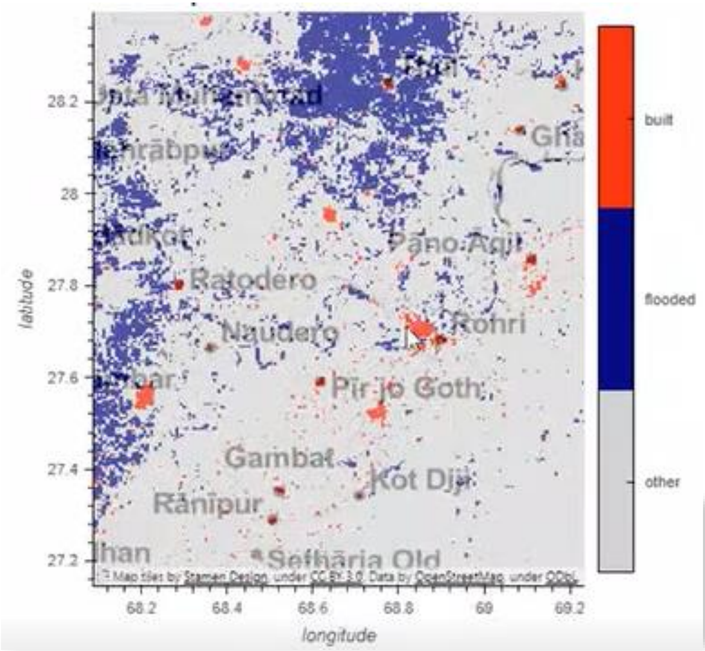
©Andrea Carrubba | Anadolu Agency | Getty Images

Use Case Details

- Combining EO data with DT outputs
  - ERA-5 to simulate Digital Twin outputs
    - rainfall and soil moisture
  - Flood UC
    - Global Flood Monitoring (GFM) data
    - Global Human Settlement (GHS) layer
  - Drought UC
    - Copernicus Land Monitoring data
      - Sentinel-1 SSM
      - Land Cover
- Use case workflow
  - Data discovery and access
  - Pre-processing and information retrieval close to the data
  - Interactive visualisation of remotely computed data

STAC

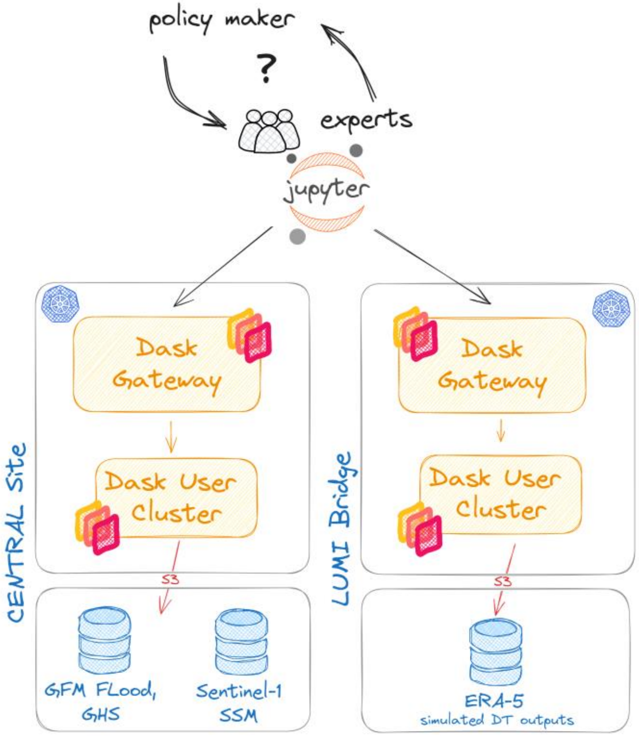
xarray



HoloViews

Panel

User Journey



- Expert creates Jupyter notebook to provide required information
- Connecting to DEDL Dask service via client library
- Run computation on Dask cluster next to the data
  - Extraction, resampling, aggregation, etc.
- Fetch only data needed for visualisation, interpretation supporting the decision making process

https://github.com/eodcgmbh/DEDL-Demonstrator

Conclusion

Simple but powerful concept for processing in multi-cloud environments

✓ Use case demonstrated the successful integration of the concept in DEDL

Thank you

✉ christoph.reimer@eodc.eu

https://github.com/christophreimer

EUMETSAT

CloudFerro

CS a Sopra Steria company

eodc