CloudFerro

Lessons learned from operating DIAS platform

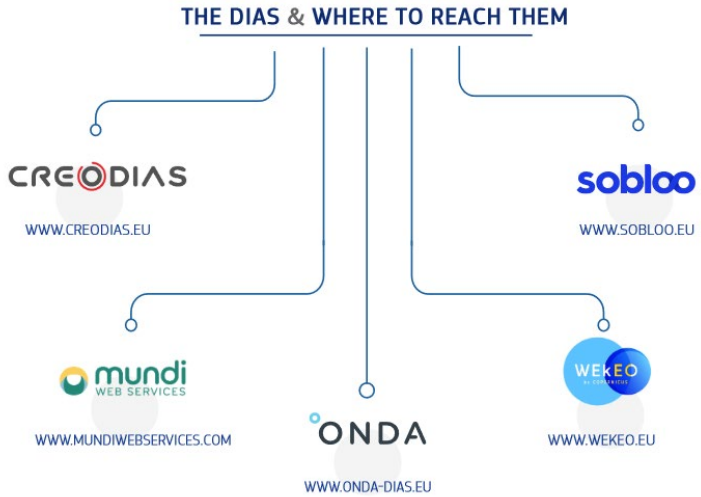
| 25th May 2022

Maciej Krzyżanowski
mkrzyzanowski@cloudferro.com
CloudFerro

Copernicus data access: Open Hub and Copernicus DIASes

Copernicus Open Hub – data download

Copernicus DIASes – cloud processing and download



Copernicus Data Access System (DAS)
download, cloud, streamlined access, Copericus Data Space

What else? If any...

CloudFerro

CloudFerro – who we are

- **European, private, technological company** founded in 2015, leveraging over 20 years of experience of its team members
- Provider of **cloud computing services** - open source based, tailor made
- Delivers and operates cloud platforms for **demanding markets** - European space sector, climate research, science…
- Specialized at storing and processing **big data sets**, like multipetabyte repositories of Earth Observation satellite data
- Achieved **technological autonomy** and guarantees reliability and independence of delivered services thanks to a full control of technology stack

EO cloud platforms built and operated by CloudFerro

CREODIAS commissioned by ESA and European Commission – a public cloud computing platform enabling immediate and free access to >28PB online EO satellite data, together with user tools and resources for its processing; built and operated by CloudFerro under the Copernicus program as one of five DIAS platforms; ESA's independent benchmarking: „**Excellent overall performances coupled with a very large online Copernicus data collection and complete service offering**....” – we believe – **best of ESA DIASes**.

WEkEO contracted by **EUMETSAT** – second DIAS built by CloudFerro, where we deliver cloud computing and storage services; platform provides meteorological satellite data, provided by the key EU entities: EUMETSAT, ECMWF, Mercator Ocean

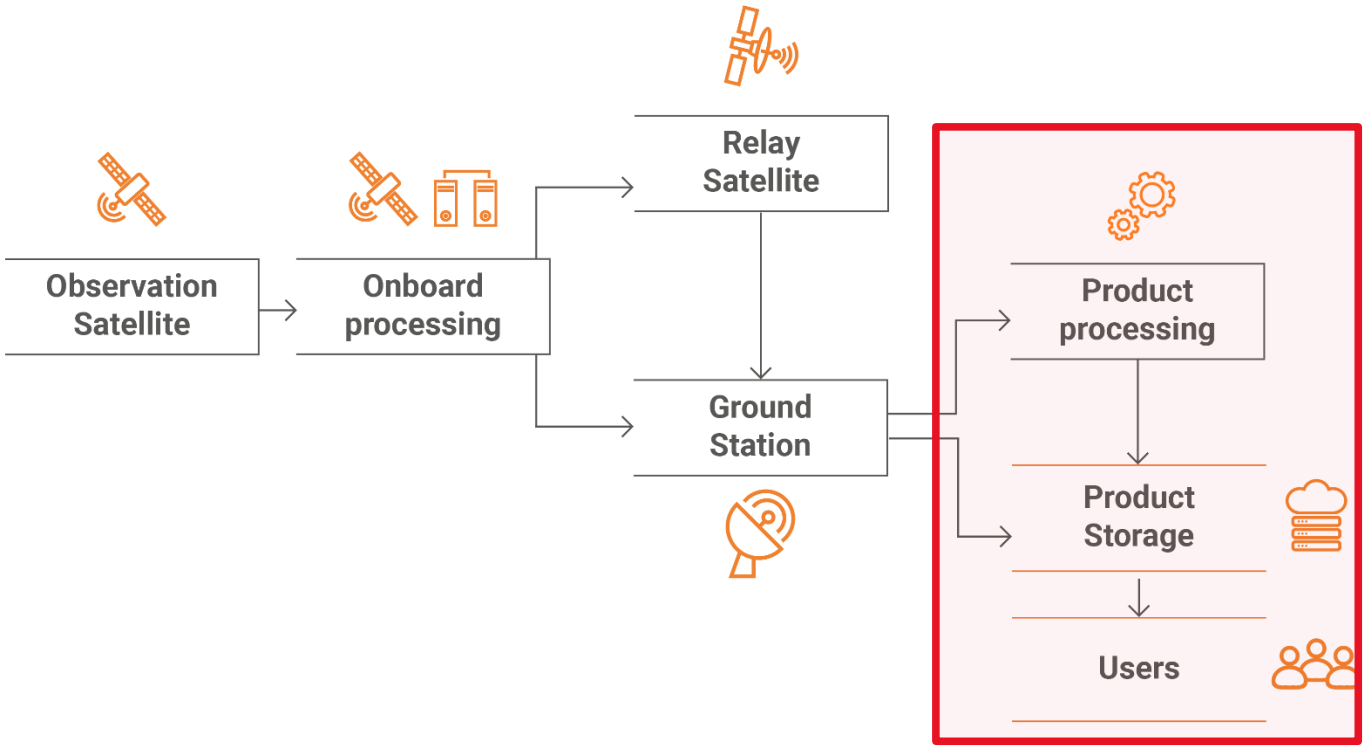
EO Cloud for **ESA** – Earth Observation Innovation Platform Testbed

CDS ordered by **ECMWF** – hybrid cloud computing providing access to climate data, analyzes and forecasts in time and space scales, built and operated by CloudFerro

CODE-DE commissioned by **DLR** – a national EO platform providing easy and free access to EO data for Germany and efficient processing environment; users can benefit from a synergy with the CREODIAS platform that ensures efficient use of resources as well as an autonomy for key national processes

CloudFerro

The Earth observation data path



CloudFerro

Challenges in acquiring, storing and disseminating EO data

As a cloud platform operator, CloudFerro addresses the needs of data providers, users and policy makers.

Policy Makers/Society:

- Strategical autonomy
- Sustainability
- Standardization
- Market competitivity
- Usage promotion

Users:

- Data availability
- Several platforms, clouds and data sources
- Data consistency (temporal and spatial)
- Diverse access methods
- Download times/reliability
- Limited processing capacity
- Tools availability
- Vendor lock-in
- Complex data licensing

Data providers:

- Growing data sizes
- Growing number of datasets
- Downlink bandwidth
- Reach customers
- Keep costs reasonable

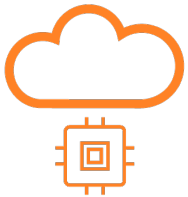
CloudFerro

EO Data consumption models

Browse & View

Download (or push) & process locally

Process in cloud using IaaS infrastructure



Process in cloud using predefined processors

Present and sell results

Best practices/Lessons learned

Store online, immediately available (almost) everything you have observed or generated

Distribute data in standard, easily accessible formats

Provide lots of scalable bandwidth and processing power

Federate users and data access

Use and contribute to open-source

Save Energy and resources

CloudFerro

Store (almost) everything you have observed or generated

- Storage and processing are relatively cheap compared to acquisition and downlink
- Storage is cheap compared to processing
  The cost of generating a product may be similar to its storage cost over 10 years
- Product availability boosts usage and enables new use-cases.
  For a product used 10 times in average, one access 'costs' 9,1 EUR in terms of mission cost share.
  Preproducing and storing a valuable higher-level product at the cost of 0,90+0,98=1,88EUR, could significantly boost its usage.
  A 30% boost would reduce mission cost share to 7 EUR for that product.

	Sentinel 1 (A and B)
Mission Cost	420 000 000
Lifetime (years)	10
Products produced per year (L0 NTC)	460 000
Average size of product in GB (L0 NTC)	1,30
Cost of generating 1 product	91
Estimated downlink cost (AWS)	3,34
Cost of 1 hour processing (16vCPU, 128GB RAM)	0,90
Cost of 1 year of storage (HDD, LTA-type, 10 EUR/TB*month)	0,16
Cost of 10 years of storage (HDD, LTA-type, degressive 13% per year)	0,98

CloudFerro

Distribute data in standard, easily accessible formats

- Different users need different access methods:
  - » HTTP/ object (S3) for remote,
  - » filesystem for local access (NFS, CIFS)
  - » OGC WMS/WMTS for tiled access
  - » zarr backed by object storage for multidimensional datacubes
- Store data in uncompressed formats,
  use cloud-optimized geotiffs for fast sub-granule access
- Avoid unnecessary steps – download, copy, decompress
- Make use of special libraries optional
- Provide fast, homogenous catalogue tool with API and GUI;
  preferrably with multiple sources

CloudFerro

Provide scalable bandwidth and processing power

This is big data!

- Carrier-grade, scalable, redundant Internet access is mandatory for both acquisition and dissemination
- Redundant, scalable storage at 10-s of PB scale
- Provide scalable processing power to enable easy, repeatable, fast, large scale on-demand product generation
- Pre-generate useful datasets up-front to boost their usage
- Avoid multi-step pipelines and bottlenecks in the systems architecture

Big Data - Copernicus / Creodias example

	Copernicus technical budget estimate	Current Creodias capabilities
Storage	90 PB	50 PB
Dissemination	400 TB/day	70 TB / day
Tested throughput		2000 TB/ day

Federate users and data access

- Users often need to **combine datasets from different sources**
- Considering data sizes, it is **ineffective to keep more copies** of big data than necessary for redundancy
- **Federate** to provide users with transparent access to a large number of datasets
- Provide **homogenous interfaces**, **high bandwidth and low latency** (requires inter-operator cooperation)
- Ideally, keep **a common catalogue with references** to multiple data sources

Use and contribute to open-source projects

- **By using open source**:
  - » Leverage common knowledge, standards and investments
  - » Flexibly extend and adapt the software to your needs
  - » Access expert communities
  - » Avoid user lock-in
- **By contributing to open source**:
  - » Work on a standard version
  - » Gain community recognition and expert status
  - » Create and keep local competencies and work places
  - » Ensure development and long term viability of your projects

Save energy and resources

Good news – Energy efficiency ≈ Cost Efficiency

- Use effective storage – large HDD-s, erasure-coding, compression, cold storage
- Energy efficient computing – low datacenter PUE, effective cooling, energy-efficient processors
- If you can – optimize over the whole technology stack
- Optimize and keep selected hardware for a long time
- Adapt processing timing to make use of renewables

What’s next?

- Greater federation of data, resources and users
- Renewable Energy for processing
- Expansion from EO data to all spatial information
- Semantic layers supplementing data for AI
- Most data immediately available online
- Optimization of storage, computing and network consumption in federated environment

CloudFerro

Thank you for your attention

Maciej Krzyżanowski
mkrzyzanowski@cloudferro.com

For more information, please visit: www.cloudferro.com or link up with us on:

facebook.com/cloudferro

linkedin.com/company/clfr/

twitter.com/CloudFerro