GERMAN COPERNICUS DATA ACCESS AND EXPLOITATION INFRASTRUCTURE

Christoph Reck, Gina Campuzano, Klaus Dengler, Torsten Heinen, Mario Winkler

German Aerospace Center (DLR), Earth Observation Center (EOC), Oberpfaffenhofen, Germany

ABSTRACT

We present the architecture for the planned German Copernicus collaborative data access and exploitation infrastructure. It shall enable the use of operational earth observation data from the Copernicus ground segment for applications and services developed by commercial, public and scientific institutions in Germany.

Copernicus – the European Earth Observation programme Global Monitoring for Environment and Security – is an ambitious undertaking of the European Commission, the European Space Agency, EUMETSAT and all its member states, to support European citizens, decision makers, scientists and industry with a constant and reliable stream of up-to-date information measured by Earth orbiting satellites.

Due to the high data volume of Sentinel and other future missions, access to Sentinel data in form of download and data distribution service is not enough. The infrastructure also has to provide hosted processing capacity following the Big-Data paradigm of "transferring the software to the data" that can be cost-effectively used by many projects and services in Germany and beyond.

Index Terms – Earth Observation, Online Data Access, Copernicus, Sentinel mission datasets, Big-Data, CODE-DE

1. INTRODUCTION

In 2013, the Space Administration of the German Aerospace Center (DLR) asked the DLR Earth Observation Center (EOC) to conceive an architectural concept for a national platform to mirror incoming Sentinel data for use and offering computing services to its users.

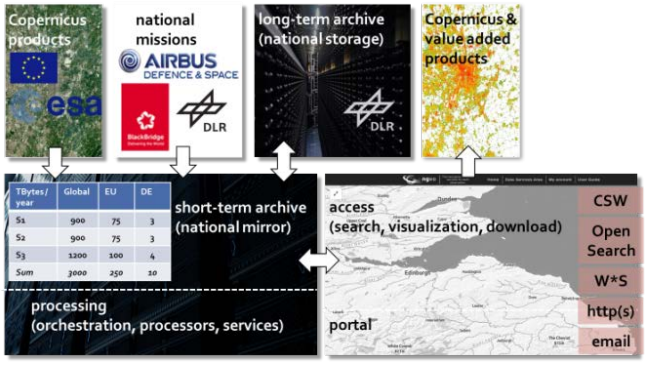

Figure 1: Context of the German national Copernicus data access and exploitation collaborative infrastructure

The primary goal of the Copernicus collaborative data access and exploitation infrastructure is to provide a platform to exploit the possibilities of the continuous data stream of "free, full and open" Copernicus Sentinel data with the following elements:

- Ingestion and Archive
- Search and Access
- Portal and User Management
- Processing and Value Added Products
- Monitoring and Reporting

2. CHALLENGE

Data rates of all incoming Sentinel-1, Sentinel-2, Sentinel-3 and Sentinel-5p earth observation user products itself pose the biggest challenge:

	2014	2015	2016	2017	2018	2019	2020
Yearly volume [TB]	180	966	4 490	6 591	7 250	7 469	8 127
Average Data Rate [Mbit/s]	194	257	1 194	1 753	1 928	1 987	2 162

Figure 2: Estimated data volume and rates for ingesting all Sentinel user products

However, the outbound rate is far larger when this data is systematically processed with different algorithms and accessed by several end users. This requires extremely performant infrastructure and data access methods.

3. ARCHITECTURE

The proposed architecture is highly modular being composed of individual services, which form a general-purpose system, largely based on existing software components, put together in an innovative manner to serve as a high-performant and scalable platform. Figure 3 below depicts the complete architectural concept that is briefly described in the next paragraphs.

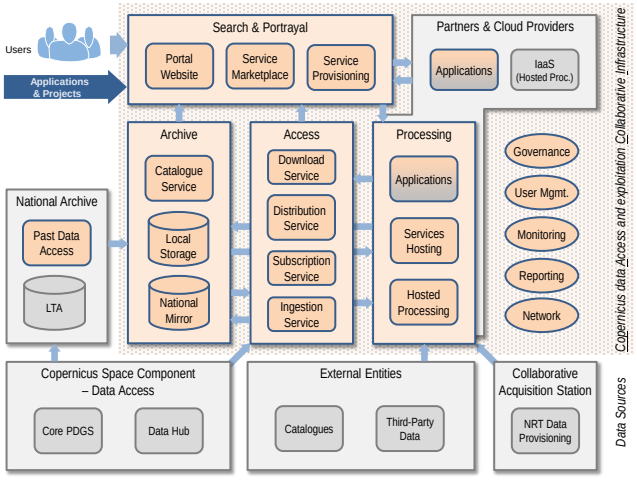

Figure 3: Architectural Concept

Search and Portrayal

Portal Website provides the access point for the collaborative infrastructure elements, portrayed from the registered services, such as the catalogue of the archived products, application provisioning monitoring, etc.

A *Service Marketplace* allows the collaborative infrastructure, external users and projects to register and advertise their accessible services (e.g. a third-party catalogue or an algorithm).

The *Service Provisioning* enables the users to deploy and launch their applications.

Archive

National Mirror provides fast access to data products for global, regional or time-series processing and re-processing. The products provided over the national mirror are the same as disseminated by ESA to the member states.

Local Storage archives the value-added and convenience products of the collaborative infrastructure.

Catalogue Service for the inventory and metadata of the Copernicus Core Services and convenience products available in the Archive. It shall permit dataset discovery visualization and selection, with pointers to the download location.

Access

Ingestion Service is in charge of populating the National Mirror from the Data Hub.

Subscription Service manages registered filters on incoming datasets to provide triggering capabilities for data-driven applications. A trigger may fire an email, dissemination to the users, or drive a data driven workflow.

Distribution Service is in charge of dissemination with equal priority the Sentinel data to the application modules (push service). This contrasts to the capability for direct access to the National Mirror or Local Storage which is subject to polling and workflow delays (pull service).

Download Service provides the means for the users to access products stored in the online archive. The download will provide direct HTTP access to the data products [OGC 13-043] or tailoring and subsetting capabilities via the OGC WCS standard.

Processing

Applications are deployed and run within the collaborative infrastructure. Among them are services and processors for the systematic and on-demand generation of convenience products from Sentinel data.

Services Hosting provides the collaborative infrastructure or deploying and running system applications (e.g. all of the above listed services) and hosted applications for the users. In principle this is a cloud environment with virtual machines and suitable software tools and web interface.

Hosted Processing provides a platform for massive-parallel earth observation data processing, for deploying processors in a cost- and resource-effective way on a cluster close to the data, and job handling with the data selection and processor execution. Also monitoring and accounting functionality is attributed to this element.

Supporting Services

Other cross-cutting services like *Governance*, *User Management*, *Monitoring* and *Reporting*, *Backup* and *Network* infrastructure complement the architecture to its completeness.

4. SOFTWARE SYSTEMS

DLR EOC designed a system utilizing known best breed software and hardware systems assembled to a streamlined simple, scalable and performant architecture covering all interfaces from **Discovery** over **Visualization** to **Download** for users with novel clients:

- Fast catalogue with HMA CSW and OGC OpenSearch interfaces
- Flexible dataset browsing with OGC Web Map Service (WMS)
- High performance data access using HTTP protocol
- Advanced data access using OGC Web Coverage Service (WCS)
- Parallel file system on an online storage attached network (SAN)
- Redundant hardware

The system shall additionally enable retrieval of historical data from the archive.

Discovery

The *EOWEB®* catalogue is based on a database, a metadata-model, ingestion and operation interfaces and user service interfaces compiled for performance [1]. It is configured to hold OGC EOP metadata [4] and provides an HMA CSW standard compliant interface [5] that allows novel clients and user interfaces to comfortably search for data (*EOWEB GeoPortal*, *FEDEO*, *mapshup*, *EOxClient*).

Visualization

Geospatial Data Access Service (GDAS) provides the tools and components to register, describe, access, search and retrieve geospatial data. It is manly composed of:

- *GeoServer* (http://geoserver.org) is an open source server for sharing, processing and editing geospatial data. It offers all major OGC standards services needed for this project (WMS, WFS, WCS, CSW), and also can be setup for INSPIRE conformance.
- *GeoWebCache* (http://geowebcache.org/) is used to cache geospatial data (e.g. for WMS) and therefore speed up the access of this data for the clients.
- *PostgreSQL* (http://www.postgresql.org) is an object-relational database with the PostGIS extension it handles spatial-referenced data.
- *Nginx* (http://nginx.org/) is a web server with a strong focus on high concurrency, performance and low memory usage. It is also used for the proxy and download server.

Catalog Client

The envisaged EOX catalog client provides a as “zoom-in on the data” solution to discover, view, and download available EO data. A timeline and advanced search capabilities allow additional filtering, e.g. by sun angle or cloud coverage.

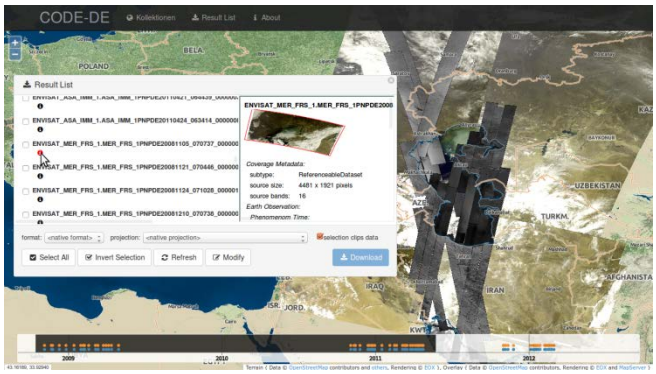

Figure 4: Catalog Client with result details and download links

Download

The large storage allows on-line access to the data products. We have benchmarked the Nginx server and the underlying filesystem to prove the capability of serving files at a total rate of more than 2 GBytes/sec using parallel transfers. *Nginx* was chosen for its performance as well as its simple configuration model that provides existing extension modules as load-balancing, access control and on-the-fly unzipping and content retrieval.

Historical data from the long term archive will be made equally accessible such that these appear to be nearly on-line.

Distribution

To minimize the load accessing recently ingested data, we are designing an efficient data distribution service based on *UFTP* (http://uftp-multicast.sourceforge.net/) for an multicast file transfer mechanism. Each distributed file is preceded by a metadata record that allows the client side filter element to decide whether a data product is needed. At the end of the reception, the client library launches a command that has been configured for the filter condition.

Processing

Bringing the processing to the data is enabled by directly interfacing a processing cluster to the storage area. Processing control and the user interface is allocated in the virtualized or cloud environments. Projects and users can process on the platform interacting on different levels ranging from

- Use of data processors and generic processing workflows and toolbox operators available via the Web GUI, producing products from Copernicus data.
- Simple data access from their VMs, interactive work with tools available (e.g. Sentinel Toolbox) or provided by the project.
- Submission of processing jobs to the cluster, control of processing workflows (orchestration) with platform tools (scripting) or project-specific workflow engines; use of infrastructure services (e.g. subscription) for systematic data-driven processing of new input data; small-scale processing on the VM or (bulk) processing on the cluster.
- Deployment of own processors in Docker containers, or as a service in a VM.
- Provision of data or processing service to other users with publishing in the Service Marketplace.

The processing environment is based on Apache Hadoop using the *Calvalus* environment designed by Brockman Consult GmbH, enhanced with *Docker* (deployment), *Apache Mesos* (resource manager), *Marathon* (service control and scheduling) and *Chronos* (batch processing).

User Management

A central use database and authentication system avoids redundancy and provides single sign on (SSO) capability to the users for accessing infrastructure services after a single login. The user management service consists of the following subsystems:

- LDAP Administration Service, which allows an operator to manage user accounts.
- The User Registration Service for new users.
- The User Self-Management Service allowing a user to manage his own data.
- The CAS Server providing interfaces for authentication of users and services, as well as issuing and validating tickets for the SSO process.

4. HARDWARE INFRASTRUCTURE

The infrastructure hardware configuration consists of a virtualized environment, a high performance computing cluster, and the storage server:

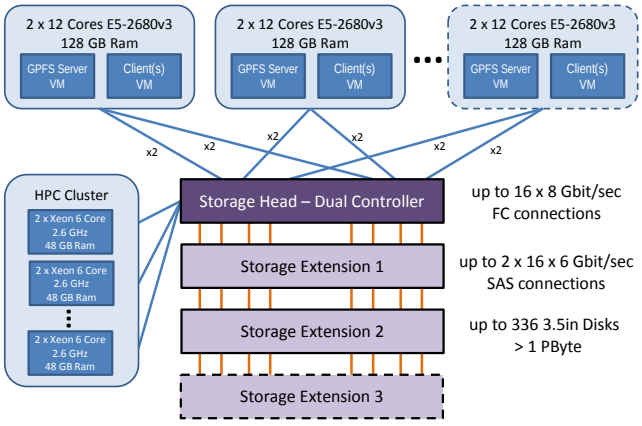

Figure 5: High-Performance server environment

- 3 servers with 2 Intel Xeon E5-2680v3 12 Core 2.5GHz, 128 GB RAM, 2 * 300 GB 10k SAS Disks, 2 FC HBA 8 Gbit/s, Quad Port 10 GE NIC, Quad Port 1 GE NIC, 2 hot swappable 750 watt power supplies.
- 7 HPC servers with 2 Intel Xeon 6 Core 2.6 GHz, 48 GB RAM, 500 GB SATA Disk, Dual Port 10 GE NIC, Dual Port 1 GE NIC, 2 hot swappable power supplies.
- Storage using a GPFS file system with 835 TB in RAID 6 over 282 4 TB NL-SAS disks, dual controllers in one head. Expandable to 12 Petabyte with 4 heads.
- Public Cloud within the DSI vCloud environment.

Network and interfacing components are not depicted. The hosts are connected via 10 Gbit/s network segments. The external connection starts with a dedicated 3.5 Gbit/s connection interfaced to the Internet and GEANT networks by the DFN X-Win provider.

5. CONCLUSION

The German Copernicus Data Access and Exploitation Infrastructure is based on the architecture, software elements and hardware infrastructure outlined in this paper. Most of the system components are re-using existing open-source software, components developed for the DLR multi-mission German Remote Sensing Data Center and the evolution complemented by the OPUS-GMES project that prototyped the:

- utilization of near real time and offline (catalogued) earth observation data
- development and implementation of end to end value adding services
- establishment of state of the art product and service-access
- demonstration of representative production ready service chains

The hardware infrastructure has been setup and tested within the ESA Data Hub Relay project to ensure its capacity and suitability for the high throughput of the envisaged system. Benchmarks showed the capability of serving data to more than 50 clients with a total data rate beyond 2 GBytes/second in the initial minimal setup.

Therefore the presented infrastructure elements have been tested and the risk has been minimized in preparation for the upcoming official German CODE-DE (Copernicus Data and Exploitation Platform – Deutschland) project. Remaining challenges in setting up the operational infrastructure are harmonizing the means for configuration handling of the individual elements, interfacing with the SSO, and the additional effort on the extensions for quota management, prioritization, and ensuring that the security requirements are fulfilled.

SELECTED REFERENCES

[1] P. Harms, S. Kiemle, D. Dietrich, *Extensible Earth Observation Data Catalogues with multiple Interfaces*, PV2007

[2] K. Dengler, T. Heinen, A. Huber, K. Molch, E. Mikusch, *The EOC Geoservice: Standardized Access to Earth Observation Data Sets and Value Added Products* , PV2013

[3] C. Reck, K. Molch, E. Mikusch, J. Hoffmann, *Copernicus Data Access and Exploitation Collaborative Infrastructure – COPACI System Concept*, DLR 2014

[4] *OGC Earth Observation Metadata profile of Observations & Measurements (EOP O&M)*, OGC 10-157r3, Version 1.0.0, 2012-06-12

[5] European Space Agency (ESA), (2012) *Heterogeneous Missions Accessibility (HMA)*, ISBN 978-92-9221-883-6, http://esamultimedia.esa.int/multimedia/publications/TM-21/TM-21.pdf