High-performance Online Data Access for Collaborative Infrastructures

Christoph Reck, Gina Campuzano, Klaus Dengler, Torsten Heinen

German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Oberpfaffenhofen, D-82234 Weßling, Germany
Email: Christoph.Reck@dlr.de, Gina.Campuzano@dlr.de, Klaus.Dengler@dlr.de, Torsten.Heinen@dlr.de

Earth Observation Center

Background and Motivation

The value of Earth Observation data is increased by accessibility. While catalogues improve finding datasets, retrieval of the physical data products is still cumbersome and often slow. DLR's German Remote Sensing Data Center (DFD) archives petabytes of data, and it interfaces internal processing chains to the EO Product Library, enabling semi-parallel and performant data driven processing or re-processing to generate value added products. With the new huge datasets, as those accumulated over years or from the new era of Sentinel satellites, new data access and exploitation mechanisms need to be devised.

Data rates of all incoming Sentinel-1, Sentinel-2, Sentinel-3 and Sentinel-5p EO products are estimated to be:

ESA Data Hub Sentinels user products	2014	2015	2016	2017	2018	2019	2020
Yearly volume [TB]	180	966	4490	6591	7250	7469	8127
Average Data Rate [Mbit/s]	194	257	1194	1753	1928	1987	2162

Figure 1. Estimated data volume and rates for ingesting all Sentinel user products

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.

DLR DFD 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.

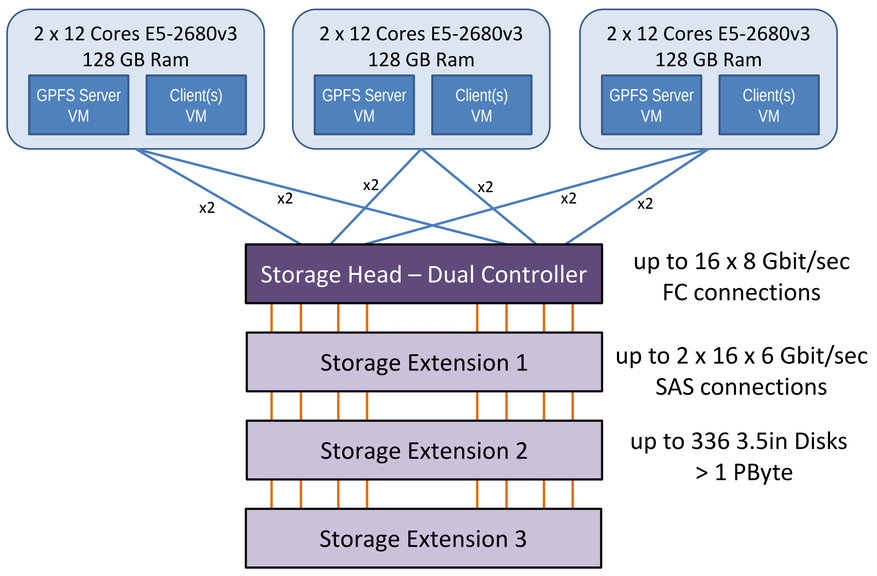

Figure 2. Redundant hardware configuration with virtualized storage.

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 [2] and provides an HMA CSW standard compliant interface [3] that allows novel clients and user interfaces to comfortably search for data (EGP, FEDEO, mapshup, ngEO).

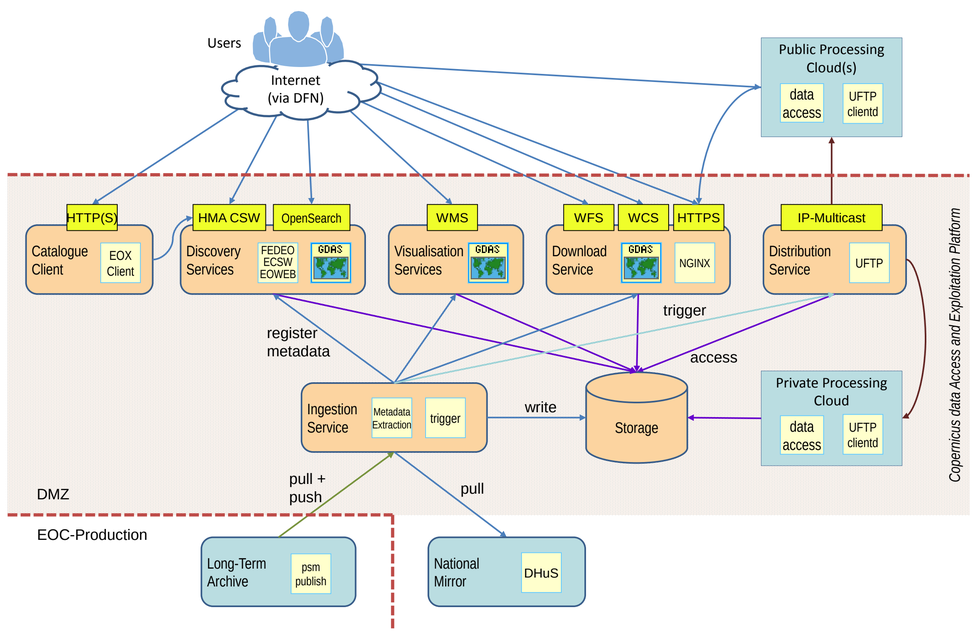

Figure 3. Components of the online data access system

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 implements all major OGC Standards needed for this project (WMS, WFS, WCS, CSW), as well as providing INSPIRE conform metadata.
- 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.

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 when using 3 or more 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.

Selected References

[1] P. Harms, S. Kiemle, D. Dietrich, (2007) Extensible Earth Observation Data Catalogues with multiple Interfaces, PV2007
[2] OGC Earth Observation Metadata profile of Observations & Measurements (EOP O&M), OGC 10-157r3, Version 1.0.0, 2012-06-12
[3] 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
[4] Dengler K., Heinen T., Huber A., Molch K., Mikusch E., (2013) The EOC Geoservice: Standardized Access to Earth Observation Data Sets and Value Added Products. Ensuring Long-term Preservation and Adding Value to Scientific and Technica