Large-Scale Data Management for Earth Observation Data—Challenges and Opportunities

Marcus Paradies[1], Sirko Schindler[1], Stephan Kiemle[2], and Eberhard Mikusch[2]

[1] German Aerospace Center (DLR), Institute of Data Science, Jena
[2] German Aerospace Center (DLR), Earth Observation Center, Oberpfaffenhofen
{marcus.paradies,sirko.schindler,stephan.kiemle,eberhard.mikusch}@dlr.de

Abstract. Earth observation (EO) has witnessed a growing interest in research and industry, as it covers a wide range of different applications, ranging from land monitoring, climate change detection, and emergency management to atmosphere monitoring, among others. Due to the sheer size and heterogeneity of the data, EO poses tremendous challenges to the payload ground segment, to receive, store, process, and preserve the data for later investigation by end users.
In this paper we describe the challenges of large-scale data management based on observations from a real system employed for EO at the German Remote Sensing Data Center. We outline research opportunities, which can serve as starting points to spark new research efforts in the management of large volumes of scientific data.

1 Introduction

The data deluge induced by the pervasive availability and use of sensors gathering data about the planet's physical, chemical, and biological systems is one of the key innovation drivers of modern, large-scale data management. The sheer size and heterogeneity of the data poses tremendous challenges to all components of a big data platform, ranging from data storage organization & management to efficient yet intuitive user access to the data.

Earth Observation (EO) is a particularly challenging application domain, as it covers a wide range of different use cases, from environmental monitoring to disaster management [1–3]. This requires a large degree of flexibility to handle different EO missions within the same data management system. Besides the calibrated raw data, also value-added data products targeting specific use cases are generated on a regular basis, stored, and made available to end users.

The German Remote Sensing Data Center (DFD) is responsible for ground-based services in several national and international EO missions. Its operations are controlled by governmental contracts, that, e.g, put time constraints on the generation of certain data products and include rigid data preservation guidelines. Furthermore, in some missions there is a mandate to make the resulting data products available for end users. As a consequence, efficient data access on all levels as well as intuitive search capabilities are of utmost importance.

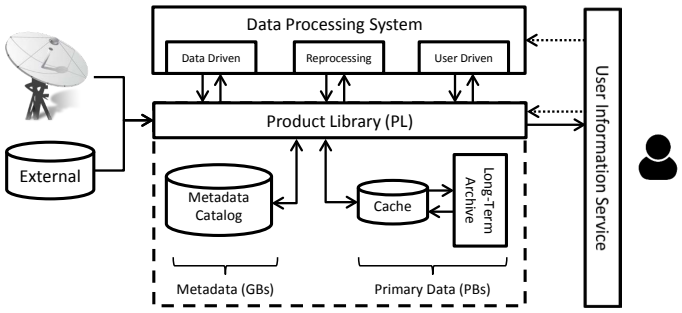

Fig. 1. EO data product generation and dissemination (simplified).

We give a (simplified) overview of the workflows of the DFD in Section 2. Based on our experience in operating and extending this platform we present challenges and research opportunities in Section 3, before concluding in Section 4.

2 System Overview

The *Data Information and Management System (DIMS)* [5] operated by the DFD is a multi-mission data management platform for EO data and accompanies all functions in the payload ground segment to produce, archive, order, and deliver EO products. Additionally, DIMS also provides near-realtime access to the latest EO products for time-critical applications, such as disaster management.

Data Reception & Ingestion: Ground stations receive data in batches of data frames. This data is calibrated, e.g., to account for atmospheric disturbances and augmented with metadata like sensing time and covered geodesic area.

Data Processing: Each data ingestion triggers a sequence of processing workflows to generate higher-level, value-added products, sometimes using additional auxiliary data. The final products are delivered to the respective customers and, in case of standard products, are also stored in the archive. Furthermore, so called reprocessing campaigns are conducted from time to time. Here, large units of raw data and their dependent data products are reprocessed using new and/or improved algorithms to further enhance the provided products. Finally, users can initiate the processing of specific products, which are only computed and delivered upon request.

Product Library (PL): The PL is the central management service and treats EO products as atomic units of metadata descriptions and primary data files. Besides extensive query capabilities, it also provides subscription services to both users and the processing system.

Long-Term Archive: The long-term archive is a robotic, active tape library with a current capacity of about 50 PByte [6]. To optimize access latency and throughput an additional online cache of about 170 TByte is used. Furthermore, all facilities are replicated on a different site to prevent catastrophic data loss in case of natural disasters.

Data Dissemination The EO products maintained within the archive are distributed on multiple channels: For example, a web gateway[3] provides

[3] `https://geoservice.dlr.de/egp/`

information access to both internal and external users. The actual data retrieval is managed via an order-based system, as some of the available EO products have limited access due to quota regulations and legal restrictions.

3 Challenges and Research Opportunities

The management of EO data faces unique challenges due to the data deluge produced by sensors, the need for efficient processing of complex, domain-specific, data-intensive workloads, and the ubiquitous necessity of enhancing the scientific data with descriptive metadata following a semantic metadata model. In the following, we describe four representative challenges from the EO domain and sketch research opportunities and directions for the database community.

Data Storage & Organization: EO data is often stored in robotic tape libraries with a large storage capacity and moderate operational costs. The unique properties of such cold data stores, such as the physical separation of tape drives and tape media and the repeated mounting/unmounting of tapes, demand a workload-aware data organization and placement strategy. There are potentially contradicting data organization strategies, such as maximizing the access locality for data products in the geospatial metadata domain. Besides reducing the data access latency, a secondary optimization goal is to reduce the overall wear of the tape media by avoiding unnecessary mount/unmount operations.

Another option to lower the data access latency is to use disk-based caches. However, the adequate configuration and provisioning of such a hierarchical storage management is still an open research question — especially, considering recent advances in storage technologies, such as decreasing prices for flash memory. We envision that the storage management configuration, i.e., storage sizing, number of storage layers, and cache eviction policies, of cold data stores will play a major role in big data management projects in the future.

Data Access: To identify long-term changes in the earth system, users often download and analyze EO data products covering large time intervals. Instead of downloading EO data, we envision an EO platform where users can upload their algorithm descriptions in their preferred programming language and the platform automatically generates efficient, executable code for it. This allows seamlessly scaling to new algorithms and other available compute resources, including specialized computing devices such as GPUs and FPGAs, and fosters *near-data processing* by identifying parts of the computation that can be pushed even closer to the data, potentially even into the storage layer.

Data Processing Pipelines: The three dominant data processing pipelines in EO are *data driven processing*, *reprocessing*, and *user driven processing*. Each of them exhibits different data access patterns, computational complexities, occurrence frequencies, and service-level agreements (SLAs). While some pipelines are predictable in their behavior and resource consumption, others, such as user-driven processing, are unpredictable with respect to required resources. Often, this also includes resource dependencies to external data resources (auxiliary data) provided by third-party data providers. To satisfy the mission-specific SLAs, the underlying hardware resources are over-provisioned to meet peak performance requirements. Providing efficient resource allocation and job scheduling in such a

cluster environment is a tedious task and often further complicated by black-box algorithm implementations.

We envision an elastic hybrid cloud infrastructure, which offers limited private resources to meet security requirements and allows extending the infrastructure elastically by allocating further resources from a public cloud infrastructure. This poses interesting research questions on an optimal data & operator placement, cost models for operational cost reduction, and data privacy & security concerns.

Semantic Enhancements: There is an increasing interest in EO data from domains outside of traditional remote sensing. However, these communities face a significant barrier of entry, as the terminology used to describe EO products deviates heavily from their own. We expect the use of semantic technologies to bridge this gap and allow a wider use of EO data across scientific domains.

Similarly, the vocabulary used within the metadata descriptions is in constant flux. This includes both the introduction of new terms as well as changes in their meaning. Besides catering to different audiences, a semantic metadata model has to be aware of these shifts and adapt to the current user's context.

The full potential of EO data is unlocked when integrating with other data sources like socio-economical statistics or volunteered geographic information (VGI) [4]. By establishing EO data as an integral part of the Linked Open Data cloud, we expect to further encourage use of EO data and open it to new audiences.

4 Conclusion and Outlook

EO data is crucial in answering important questions relevant to society like the impact of pollution on the Earth spheres. We have identified four different challenges in the areas of cloud computing, data storage & organization, efficient data access, and semantic enhancements of EO data that have to be addressed to unlock the potential for novel applications leveraging EO data. We believe that EO can serve as an initial application domain to develop novel methods for dealing with large volumes of scientific data from heterogeneous data sources with a strong demand for efficient post-processing capabilities. In the future, solutions can be transferred to other scientific domains, such as radio astronomy and high-energy physics.

References

1. BigGIS (2018), `https://www.fzi.de/forschung/projekt-details/biggis/`
2. EarthServer (2018), `http://www.earthserver.eu/`
3. GeoMultiSens (2018), `http://www.geomultisens.de/`
4. Goodchild, M.F.: Citizens as sensors: the world of volunteered geography. GeoJournal 69(4), 211–221 (Nov 2007)
5. Kiemle, S., Mikusch, E., Bilinski, C., Buckl, B., Dietrich, D., Kröger, S., Reck, C., Schroeder-Lanz, A.K., Wolfmüller, M.: Data Information and Management System for the DFD Multi-Mission Earth Observation Data. In: Digital Curation Center Conference Proceedings. The Royal Society Edinburgh (2005)
6. Kiemle, S., Molch, K., Schropp, S., Weiland, N., Mikusch, E.: Big Data Management in Earth Observation: the German Satellite Data Archive at DLR. In: Proceedings of the 2014 conference on Big Data from Space (BiDS'14). pp. 46–49 (2014)