FACILITATING COOPERATIVE RESEARCH, DEVELOPMENT AND OPERATIONS IN EARTH OBSERVATION

Christian Briese[1], Werner Mücke[1] & Michael Aspetsberger[2], Wolfgang Wagner[3,1]
[1]EODC Earth Observation Data Centre for Water Resources Monitoring GmbH, Vienna, Austria
[2]Catalysts GmbH, Linz, Austria
[3] Department of Geodesy and Geoinformation, Vienna University of Technology, Vienna, Austria

ABSTRACT

The Sentinels generate huge amounts of data with a high spatio-temporal resolution. This leads to new challenges in research and in the processing and archiving of big EO data. Based on the EODC concept this contribution provides an insight in the practical realisation of the ideas within the EODC framework. Next to the aspect of collaboration it focuses on the information exploitation within EODC and the distribution of results to the users. Finally, in order to fully exploit the richness of EO data, the necessary federation of data centres is discussed.

Index Terms— Big EO data, processing, collaboration, virtual research environment, federation of data centres

1. CHANGING THE PARADIGM

With the advent of the Copernicus space programme and its novel Sentinel satellite missions, both the temporal and spatial resolutions of the produced Earth Observation (EO) data are dramatically increasing, which consequently leads to Big Data volumes and unprecedented challenges in their exploitation. In the medium-term, a paradigm change in the EO landscape is to be expected: there is not going to be a single data centre for processing, storing and distributing of EO data, instead we are moving towards a federation of multiple interconnected data centres spreading and handling the workload among each other. Additionally, due to the inherently increasing complexity of observed phenomena and retrieval algorithms, the association of interdisciplinary working groups will be advantageous. Furthermore, contrary to current practice, the methods, algorithms and software will have to be transferred to where the data are located, in order to avoid bottlenecks due to inadequate data transfer capabilities. This contribution introduces the EODC concept and focuses on its realisation. EODC aims at addressing the above mentioned challenges, thereby facilitating cooperation between science, industry and public sectors, giving the users the means to access Sentinel data and work with them in an efficient and economic way.

2. RESEARCH, DEVELOPMENT AND OPERATIONS ON THE EODC PLATFORM

2.1. Introducing the basic concept

EODC builds its IT capacities on three basic pillars (see Figure 1): (1) NORA, which stands for “Near real-time operations and rolling archive”, (2) SIDP, the “Science Integration and Development Platform”, and (3) GTR, the “Global testing and reprocessing facility”. NORA offers four services: data input from external satellite data archives, output of these data to EODC storage, status monitoring of system performance and data transfer, as well as near real-time processing of selected products. SIDP is EODC’s fully equipped and flexible cloud infrastructure, where tailor-made pre-configured virtual machines and other services (e.g. continuous integration and deployment) are being hosted, supporting the remote development and testing of methods and source code. GTR is a top500-ranking supercomputer, namely the Vienna Scientific Cluster 3 (VSC-3), intended for large scale processing. These three components are complemented by a Petabyte-scale data archive (i.e. the EODC data pool), which is physically co-located with SIDP and GTR and connected via InfiniBand network technology to minimise transfer times and increase I/O.

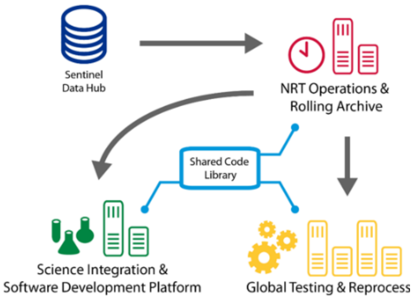
Sentinel Data Hub
NRT Operations & Rolling Archive
Shared Code Library
Science Integration & Software Development Platform
Global Testing & Reprocess

Figure 1: The three pillars of the EODC IT concept. [1]

2.2 Connecting the community

Within the EODC, individuals join together into communities sharing similar foci or research and development goals. The idea is for the single researcher or organisation to take advantage of the experience and skills of all involved partners and to participate in collaborative development processes, while at the same time gaining increased external visibility via a larger group. In order to support community building, EODC organises regular community events for information exchange. In an effort to pro-actively stimulate cooperation, several tools are

doi: 10.2788/854791

implemented on the EODC IT platform, such as a shared code library (based on GitLab) or a knowledge base (similar to wiki).

2.3. Accessing the data pool

EODC has been acquiring Sentinel-1 (S-1) from day one and is providing them through its long-term data archive. On average, the data are available in the archive approximately 2.5 hours after their processing time and 6.25 hours after acquisition time. Currently, a total of 66556 S-1 acquisitions are hosted in the data pool (status October 2015). Once a user is logged on to the EODC network, he may access the data either through SIDP or GTR. Search and discovery are provided via a meta database that supports spatio-temporal search functions and scripting. Conventional access is established by means of file system access and http-export. Evaluations are ongoing to provide OGC WM(T)S and WCS interfaces for selected products.

2.4. Exploiting the information

The workflow from raw or pre-processed to refined and value-added data is as follows: Starting on SIDP, users or communities develop their methods, while testing it on small data sets (typically hundreds of Mega- and up to few Gigabytes). As soon as their code base has reached a certain state of maturity, they transfer to GTR, where the same codebase is available for testing on larger areas, usually regional or even global scale. In order to exploit the computational capacities of the super computer to its full potential, programmers are advised to either provide parallelized code or achieve parallelization through the respective application, e.g. parallel processing of many data files at once. If a near real-time product is targeted, the developers move to NORA for the establishment of an operational NRT service. While working with SIDP and NORA is basically similar to working on personal computers, as both offer the possibility to host user-definable virtual machines, accessing the supercomputer GTR is a completely different environment. It involves a minimum level of understanding for high performance computing, parallelisation and certain programming experience. To create an environment that supports a variety of users with differing skill levels, EODC offers a simplified processing submission and task scheduling interface (see Figure 2). A browser-based graphical user interface allows the users to select pre-defined processing chains from the EODC code library, use their own configuration files to define settings for included algorithms, and select the number of computing nodes they would like to employ. After the processing is triggered by the user, everything else is done automatically: the respective codebase is pulled from the library and together with the configuration files processing containers are built, representing sand-boxed environments including all software packages necessary for successful computation. During data processing on GTR, which is handled by the open source workload manager SLURM, the browser-based job monitor gives access to principle information of the tasks, such as estimated run time. On job completion, the users are notified and may inspect their results using the available data viewer.

2.5. Distributing the results

Value-added data produced on the EODC platform may be stored in the archive for later usage, and additionally published or redistributed. The EODC interactive delivery platform provides multi-channel online access and basic analysis features for various kinds of EO data and is intended to offer users tailor-made solutions to deliver their information products to a wider audience.

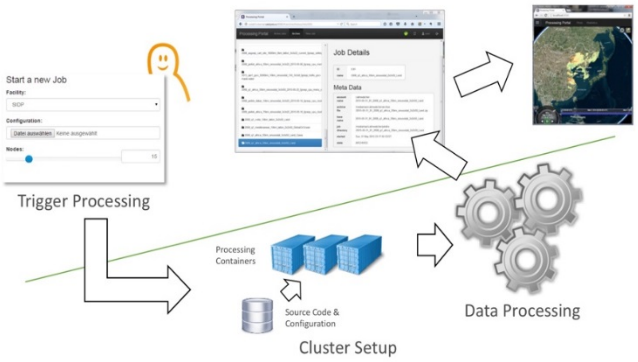

Figure 2: Illustration of simplified scheduling of processing jobs on the VSC-3 supercomputer.

3. OUTLOOK

The ever-increasing amounts of EO data to be expected in the upcoming years demand for a sophisticated expansion strategy for computational power, as well as for storage space. EODC is planning to receive and provide Sentinel 1, 2 and 3 data and therefore aiming to significantly extend its storage. In parallel, the virtual research environment SIDP is being equipped to host 100+ users in the same time frame. Given the diversity of EO products in Europe, and the data centres producing them, EODC is strongly working towards a federation of data centres or other similar initiatives across Europe, in order to be able to exploit the information richness of up-to-date EO data to its full potential.

4. REFERENCES

[1] Wagner, W. et al., 2014: Addressing grand challenges in earth observation science: The Earth Observation Data Centre for Water Resources Monitoring. In: ISPRS Technical Commission VII Mid-term Symposium 2014, Volume II-7, Istanbul, Turkey, pp. 81-88.

doi: 10.2788/854791