D4.1: Evaluation of data cube software

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348.

Project no.	776348
Project acronym:	CoastObs
Project title:	Commercial service platform for user-relevant coastal water monitoring services based on Earth Observation
Instrument:	H2020-EO-2017
Start date of project:	01.11.2017
Duration:	36 months
Deliverable title:	D4.1: Evaluation of data cube software
Due date of deliverable:	Month 8

Organisation name of lead contractor for this deliverable: WI (Partner 1)

Author list:

Name	Organisation
Kathrin Poser	WI

Dissemination level		
PU	Public	x
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

History			
Version	**Date**	**Reason**	**Revised by**
01	13/11/2018	First version	Kathrin Poser (WI)
02	19/12/2018	Internal Review	Caitlin Riddick, Andrew Tyler, Evangelos Spyrakos (USTIR)
03	24/12/2018	Final version	Kathrin Poser (WI)

Please cite as:

Poser, K. (2018). Evaluation of data cube software. CoastObs Project Deliverable 4.1. Wageningen, NL.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

CoastObs Project

CoastObs is an EU H2020 funded project that aims at using satellite remote sensing to monitor coastal water environments and to develop a user-relevant platform that can offer validated products to users including monitoring of seagrass and macroalgae, phytoplankton size classes, primary production, and harmful algae as well as higher level products such as indicators and integration with predictive models.

phytoplankton

seagrass

harmful algal blooms

primary production

To fulfil this mission, we are in dialogue with users from various sectors including dredging companies, aquaculture businesses, national monitoring institutes, among others, in order to create tailored products at highly reduced costs per user that stick to their requirements.

With the synergistic use of Sentinel-3 and Sentinel-2, CoastObs aims at contributing to the sustainability of the Copernicus program and assisting in implementing and further fine-tuning of European Water Quality related directive.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Partnership

Water Insight BV. (WI)

The University of Stirling (USTIR)

Consiglio Nazionale Delle Ricerche (CNR)

Universite de Nantes (UN)

HZ University of Applied Sciences (HZ)

Universidad de Vigo (UVIGO)

Bio-Littoral (BL)

Geonardo Environmental Technologies Ltd. (GEO)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

TABLE OF CONTENTS

FIGURES

TABLES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

ABBREVIATIONS

List of abbreviations	
Abbreviation	**Explanation**
API	Application programming interface
AWS	Amazon Web Service
BLOB	Binary Large Object
DIAS	Data and information access service
EO	Earth Observation
GDAL	Geospatial Data Abstraction Library
GPL	GNU General Public License
LGPL	GNU Lesser General Public License
NetCDF	Network Common Data Format
OGC	Open Geospatial Consortium
S2	Sentinel-2
S3	Sentinel-3
SQL	Structured query language
WCS	Web Coverage Service
WMS	Web Map Service
WMTS	Web Map Tile Service

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

1 Summary

Objective:

The object of D4.1 is to evaluate software solutions available for raster data cube storage and manipulation, including Rasdaman, SciDB, Open Data Cube and Copernicus Data and Information Access Services (DIAS). The evaluation will be based on functional and non-functional criteria, including available features, interfaces, ease of integration with custom algorithms, efficiency and usability of query language, performance, ease of installation and configuration, documentation, community support and licensing.

Scope:

While multidimensional array data structures have long been used in the Earth observation (EO) domain and some software implementations have been available for many years, the topic has only gained momentum over the last few years. On the one hand, the need for efficient data structures has become more apparent with the recent sharp increase in free and open EO data (such as from the European Commission's Copernicus programme). On the other hand, the advent of cloud and big data technologies have enabled the development of new tools to help EO service providers manage, organise and share their data. Currently, many projects are researching, developing and implementing different data cubes based on a variety of software solutions and management approaches. Thus far, operational implementation of data cubes has mainly been restricted to large (national or international) organisations, however the CoastObs project aims to harness this technology also for small EO service providers. Thus, the first step is to review the most promising candidates of the currently available software solutions, which is documented in Chapter 2 of this deliverable. It should be noted that due to the very dynamic nature of the field, this document is only a snapshot as software and implementations are being actively developed. The aim of this document is therefore not to provide a comprehensive review of all technical and operational aspects of the different software solutions, but to provide a basis for an informed decision about the solution to be implemented in the CoastObs project.

Another recent development that is aimed at facilitating access to and use of Copernicus data is the deployment of five cloud-based platforms (Data and Information Access Services – DIAS), providing centralised access to Copernicus data and information, as well as processing tools. While the DIAS are not directly related to the concept of data cubes, they are another important building block for setting up an efficient infrastructure to acquire, process and distribute EO data for service providers. Therefore, as with data cubes, this deliverable is not intended as a comprehensive review of DIAS platforms, but rather a comparison of the available

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

DIAS offers for use with EO data that would be most suitable for implementation in the CoastObs project.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

2 Data cube software

Until recently, Earth observation (EO) datasets have been typically handled as collection of scenes, with individual scenes representing a temporal snapshot and a particular region on the Earth's surface. Using these data in complex spatio-temporal analyses becomes difficult when they include many scenes that may spatially overlap, and data volumes often exceed capacity. In particular, with higher resolution EO data increasingly available (in particular Sentinel-2), datasets can easily exceed memory and storage capacities of single computers (Lewis et al., 2017). Also, the organisation of EO datasets as a collection of scenes can be inefficient. For example, accessing and analysing a time series of high resolution imagery may yield nonoptimal data access patterns, because values of a single time-series come from many files. As a result, EO data users spend an increasing amount of time on data management instead of developing and validating algorithms and analysing the results (Appel et al., 2018).

To facilitate the use and exploitation of large EO datasets, an alternative approach is to represent EO datasets as multidimensional arrays ('data cubes') and to use array databases or other array data structures for storage and analysis. There is no agreed-upon definition of a data cube yet, but a working definition has been proposed by Baumann (2017) as

> *A data cube is a massive multi-dimensional array, also called "raster data" or "gridded data"; "massive" entails that we talk about sizes significantly beyond the main memory resources of the server hardware. Data values, all of the same data type, sit at grid points as defined by the d axes of the d-dimensional datacube. Coordinates along these axes allow addressing data values unambiguously.*

Baumann (2017) also defines six principles of data cube services, which are further elaborated by Strobl et al. (2017). These principles include aspects of data representation, organisation and access.

An important element for the CoastObs project is that data cubes allow efficient trimming and slicing along any number of axes in a single request, thus enabling spatial, temporal and spatio-temporal statistics to be performed easily. Another important feature of data cubes for CoastObs is adaptive partitioning that is invisible to the user when performing access and analysis. Finally, data cubes should feature a query language allowing clients to submit simple as well as composite extraction, processing, filtering, and fusion tasks in an ad-hoc fashion. For CoastObs, such a query language will facilitate the implementation of higher-level products and integration of EO data with predictive models.

So far, data cubes have mainly been implemented by larger organisations serving a variety of users (such as the Australian geoscience data cube (Lewis et al., 2017) or the Swiss Data Cube (Giuliani et al., 2017)). However, the CoastObs project consortium believes that also smaller

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

projects and organisations can benefit from using data cube technology. It is an explicit aim of the project to implement a data cube, integrate it into the operational processing infrastructure for the project case studies and evaluate the drawbacks and benefits for smaller-scale operations.

In the following sections (2.1 to 2.5), four different data cube software packages are introduced and compared. These include Rasdaman (Baumann et al., 1998) and SciDB (Stonebraker et al., 2013), which are popular array databases within the EO community (e.g. Wagemann et al., 2017, Planthaber et al., 2012). The Open Data Cube, an initiative founded under the auspices of the Committee on Earth Observation Satellites (CEOS), is the infrastructure mainly used by larger national data cube projects (e.g. Lewis et al., 2017, Guiliani et al., 2017). EODataBee is a data cube solution in development by the Horizon 2020 DataCube Service for Copernicus (DCS4COP) project led by Brockmann Consult. A scan of the market did not yield any further software solutions that seemed useful and feasible for CoastObs.

2.1 Rasdaman

Rasdaman (http://www.rasdaman.org/) is one of the pioneers in array database systems. It is a domain-independent database management system (DBMS) which supports multidimensional arrays of any size and dimension. On storage, arrays get partitioned ("tiled") into sub-arrays, which are the basic unit for data storage and access. Rasdaman can store array data in different ways:

- Arrays in a file system directory, array metadata in SQLite; this is default.
- Everything in PostgreSQL: arrays in binary large objects (BLOBs), array metadata in tables.
- Access to pre-existing archives of any structure (enterprise edition only)

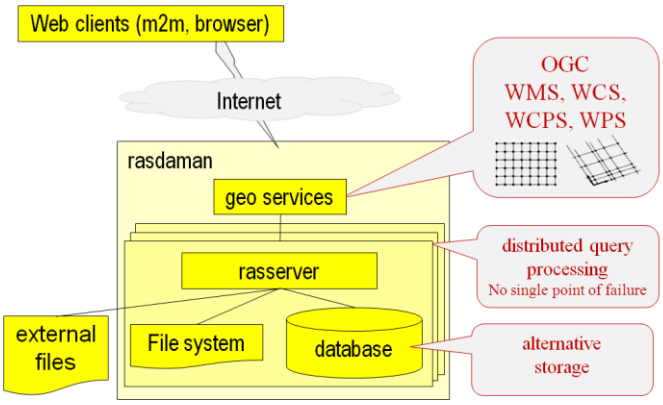

Figure 1: Rasdaman overall architecture (source: rasdaman.com)

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 776348

It follows the classical client/server architecture with queries processed on the server side. The Rasdaman server, as a middleware, maps the array semantics into the relational table semantics.

The Rasdaman server configuration consists of one dispatcher process per computer, called rasmgr, and server processes, called rasserver, of which at a given time none, one, or several processes can be running. All server processes are under control of the manager. The server manager (rasmgr) and Rasdaman server(s) (rasserver) all run on the same physical hardware, the Rasdaman host. The manager accepts client requests and assigns server instances to them, taking them from the pool of server processes it maintains. The servers resolve requests, thereby generating calls to the relational database system which in turn accesses its database files. Rasdaman servers running on different computers can be linked to form one single server network. In distributed installations, a rasmgr process keeps contact to the managers on other machines to further dispatch client requests across all the rasdaman servers available.

With regard to client interfaces, Rasdaman offers a SQL-like language named Rasdaman Query Language (RasQL) to manipulate raster data. RasQL queries are parsed and executed by Rasdaman servers, which retrieve data from the base RDBSM. Client development is supported by a C++ API (raslib) and a Java API (rasj). In addition, an R package (RRasdaman) and a python library (RasdaPy) are available. RasdaPy is, according to its own documentation, not as mature as the C++ and Java implementations.

In addition, Rasdaman provides a component called petascope for geo data management, which implements a number of Open Geospatial Consortium (OGC) standards for web service interfaces including Web Coverage Service (WCS) and Web Coverage Processing Service (WCPS).

- Web Coverage Service (WCS 2.0.1) and Web Coverage Service Transactional (WCS-T 2.0)
- Web Coverage Processing Service (WCPS 1.0)
- Web Map Service (WMS 1.3.0)

With this application, both geospatial raster data and geoprocessing functions can be shared on the Internet.

Rasdaman is continuously tested on several Linux platforms: Debian, Ubuntu and CentOS. It is recommended to have at least 8 GB main memory. Disk space depends on the size of the databases, as well as the requirements of the base DBMS of rasdaman chosen. The footprint of the rasdaman installation itself is around 400 MB.

The Rasdaman technology is available in two variants: as free Rasdaman community software and commercially supported Rasdaman enterprise. The Rasdaman community license releases

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

the server as General Public License (GPL) and all client parts as Lesser General Public License (LGPL), thereby allowing to use the system in any kind of license environment. The enterprise version is developed, distributed and supported by the company Rasdaman GmbH. The community open-source project is governed by its Project Steering Committee (PSC), which is in principle open to anyone, but is in practice also controlled by Rasdaman GmbH. Rasdaman has been developed since 1996 and has reached a high level of maturity.

2.2 SciDB

SciDB is another open source, dual-license data management system that has been developed primarily for scientific domains producing very large array data (https://www.paradigm4.com/technology/) (Stonebraker et al., 2013). It uses a nested-array data model, mapping array data into multi-dimensional arrays with strongly typed attributes within each cell. Internally, attributes of the cell are stored separately such that each cell only contains one value. The arrays are segmented into fixed chunks as specified in the schema definition (Cudre-Mauroux et al., 2009). SciDB features science-specific operations, uncertainty, lineage, and named versions (Paradigm4, 2017).

SciDB focusses on efficient processing of massive data by using a distributed parallel architecture. SciDB is deployed on a cluster of servers, each with processing, memory, and local storage, interconnected over a network. A PostgreSQL database is used to store the SciDB catalogue of array schema and the distribution of data in the cluster. While all instances in the SciDB cluster participate in query execution and data storage, one server is the coordinator and orchestrates query execution and result fetching. It is the responsibility of the coordinator instance to mediate all communication between the SciDB external client and the entire SciDB database. The rest of the system instances are referred to as worker instances and work on behalf of the coordinator for query processing (Paradigm4, 2017).

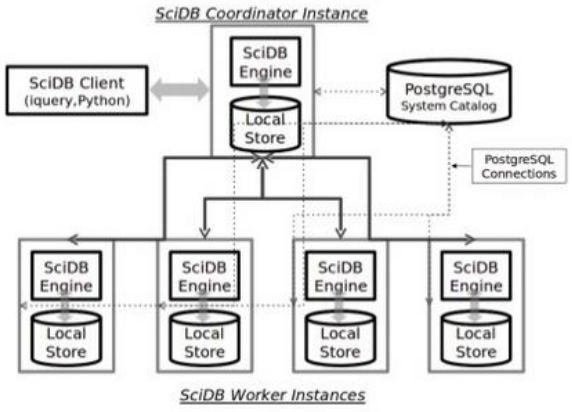

Figure 2: SciDB architecture (copyright: Paradigm4)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Regarding query languages, SciDB currently provides two distinct query languages: the Array Query Language (AQL) and the Array Functional Language (AFL). AQL is very similar to SQL, whereas AFL is a functional description of array operations, using a function syntax which allows for nesting of operations. In addition, Python, R and C++ libraries are available for interfacing with SciDB.

SciDB is developed and maintained by paradigm4. It is licensed under a dual license: the Community Edition, which comes under the GNU Affero license, allows loading unlimited data, using all the core functionality, and running on a machine or cluster. The Enterprise Edition (Paradigm4 license) includes additional software for commercial deployments such as high availability, cluster management or data access control. Although relatively new, SciDB has already been used in different scientific domains such as astronomy and genetics, including some EO applications (e.g. Planthaber et al., 2012, Tan et al., 2017). Tan et al. (2017) propose an extension to address the specific needs of handling EO data.

2.3 Open Data Cube

The Open Data Cube (ODC) (https://www.opendatacube.org/), created and facilitated by the Committee on Earth Observation Satellites (CEOS), is a global initiative to increase the value and use of satellite data by providing users with access to free and open data management technologies and analysis platforms. It is based on an initial development by Geoscience Australia, the Australian Geoscience Data Cube (AGDC). It is mainly aimed at national and international organisations (currently there are 3 national data cubes implemented in Australia, Switzerland and Columbia, with more planned). The ODC is a common analytical framework that includes API development, cloud integration, a web-based user interface, and data analytics to facilitate the organization and analysis of large, gridded data collections. Data is usually file based, either in local directories of GeoTIFFs or NetCDF files, but data can be anything that GDAL can read, including Cloud Optimised GeoTIFFs stored on Amazon Web Services' (AWS) S3. For the Index, the ODC uses PostgreSQL as a database to store a list of Products and Datasets. The index enables a user to ask for data at a time and location, without needing to know specifically where the required files are stored and how to access them. The Software at the core of the ODC is a Python library that enables a user to index data (add records to the Index), ingest data (optimise indexed data for performance), query data (returning data in a standard data format) and carry out a wide range of other functions related to managing data (Leith, 2018).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

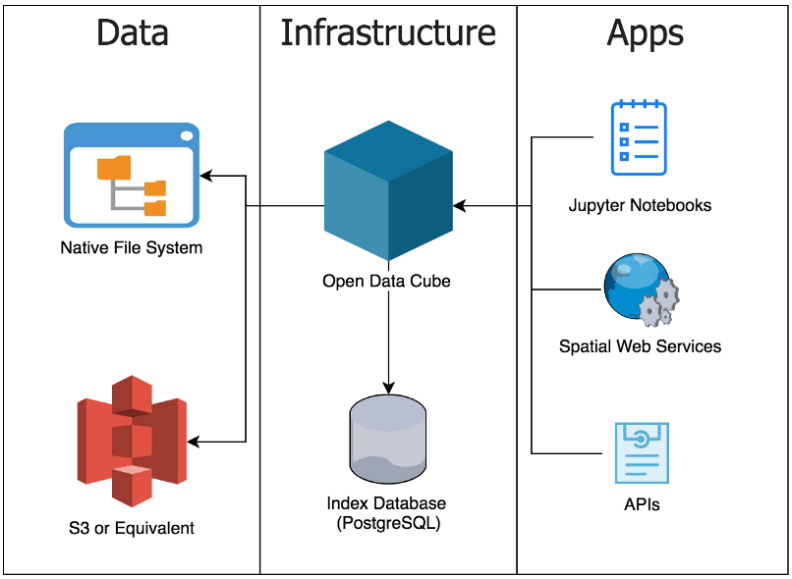

Figure 3: The architecture of the ODC (from Leith, 2018)

The interface language is Python, and a module for an OGC WMS is available. In principle, ODC can be set up on many platforms, however the recommended installation method is to use a container and package manager such as Miniconda. Specific documentation exists for Windows, Ubuntu and MacOS. In terms of system requirements, the recommendation is at least one core per desired concurrent user and at least one gigabyte of memory per core as well as shared storage with capacity for all ingested data, original data and analysis results.

ODC is available under Apache License Version 2, and the project is shepherded by CEOS.

2.4 EODataBee

A novel data cube service called EODataBee is currently being developed within the H2020 DCS4COP (DataCube Service for Copernicus) project (https://dcs4cop.eu/) as a commercial service with a portfolio comprising a data cube providing environmental information layers, easy-to-integrate user interfaces and applications, tailored trainings and consultancy services. The offered service types are Processing as a Service (PaaS) and Software as a Service (SaaS). It is targeting the value-adding Earth Observation industry and public organisations. The service is based on the Python xarray technology and provides a Python interface as well as supports a tile map service, with some OGC Web Map Tile Service (WMTS) 1.0 compatibility. It is intended to be deployable in a range of environments (local server, Cloud, DIAS, cluster).

EODataBee is being developed by Brockmann Consult and will be rolled out as a commercial service offering.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

2.5 Comparison

Table 1 compares some of the features of the different data cube solutions. EODataBee is not included as the service is not operational yet. Service maturity is judged on length of time the service has been available and number of successful use cases/examples; Documentation/community is judged on the amount of documentation and user guides available; and the availability and activity of community support, e.g. via forums and mailing lists.

Table 1: Comparison of data cube solutions

Comparison of data cube solutions			
	Rasdaman	SciDB	Open Data Cube
License	GPL 3.0 Dual License	AFFERO GNU GPL v3 Dual License	Apache License Version 2
Supported platforms	Debian, Ubuntu, CentOS	Ubuntu, RHEL, CentOS	Windows, Ubuntu, MacOS
Query language / APIs	RasQL / Java, C++, R, Python	AQL, AFL / C++, R, Python	Python
OGC interfaces	WCS 2.0.1 WCS-T 2.0 WCPS 1.0 WMS 1.3.0	N/A	WMS
Maturity	+ (since 1996)	+- (since 2008)	- (since 2016)
Documentation / Community	+ (Extensive documentation, active mailing list)	+- (Comprehensive documentation, active mailing list)	+- (Comprehensive documentation, no mailng list or forum)

A number of comparisons have been performed between SciDB and Rasdaman (sometimes also including other solutions) (e.g. Kovanen et al., 2018, Tan et al, 2017, Tan and Yue, 2016). Overall, they conclude that both have achieved a good level of functionality and performance. As a shortcoming Kovanen et al. (2018) point out documentation that is out of date and inconsistent, regarding both the languages and capabilities of their model. While speed is often an important factor in these comparisons, for CoastObs we do not consider the speed of data ingestion or query performance by the data cube software as a limiting factor as we do not expect end users to directly interact with the data cube. Tan and Yue (2016) point out that Rasdaman is more tailored to geospatial raster data implementing OGC standards, while SciDB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

is more popular in the life sciences and financial markets, and provides a set of solutions in these areas. The Open Data Cube has not been compared with other solutions.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

3 Copernicus DIAS

The Copernicus programme of the European Union makes satellite data from the Sentinel constellation available on a free, full and open basis. When all Sentinel satellites are operational, they will deliver more than 10 petabytes of data each year. On top of that, thematic information is made available from the six Copernicus core services, adding to the total amount of geospatial data generated or made available by the Copernicus programme. This makes Copernicus the third largest data provider in the world, creating great opportunities, but also presenting great challenges.

Sentinel satellite data are distributed by the European Space Agency (ESA) and the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT). There are several access mechanisms to the Sentinel data, tailored to the purpose they will be used for. Both ESA and EUMETSAT operate data access 'hubs' for on-demand, open access to Sentinel data, and additionally some other data access mechanisms such as EUMETCAST. Over the last years it has become obvious that the central infrastructure for the access hubs is currently not able to efficiently handle the amount of requests that are made, resulting in very long download times and limited access per user.

To facilitate access to Copernicus data, a Collaborative Ground Segment has also been set up, which consists of national access mechanisms for Copernicus data in some European Countries. These national access points focus on data and information that is considered as particularly useful for national users, often also providing a user interface in the national language.

No cloud processing service is currently offered by Copernicus to its users. However, some commercial initiatives have emerged and some national access points plan to offer processing services or infrastructure for Sentinel satellite data in the cloud.

To facilitate and standardise access to Copernicus data, the European Commission is funding the deployment of five cloud-based platforms (Data and Information Access Services – DIAS) providing centralised access to Copernicus data and information, as well as to processing tools (https://www.copernicus.eu/en/upcoming-copernicus-data-and-information-access-services-dias). The offer of the five DIAS is considered complementary to the existing data access portals that will continue to operate.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

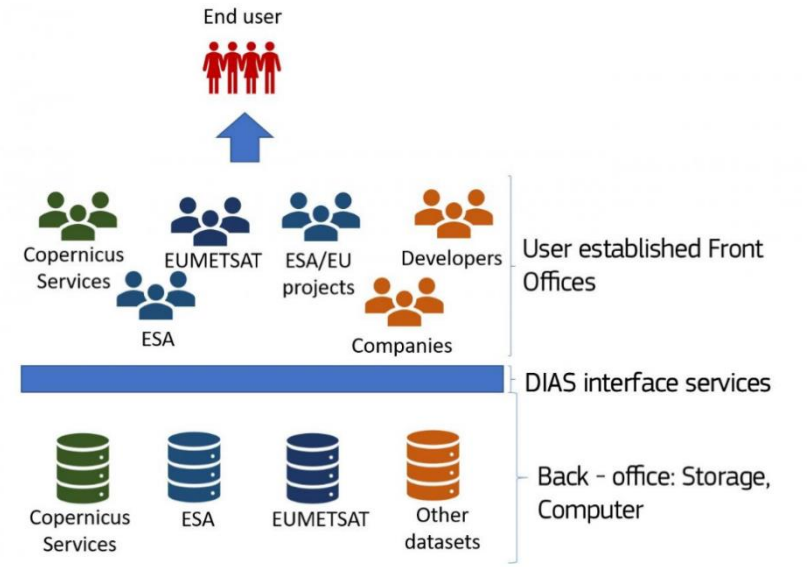

Figure 4: The role of the DIAS in the Copernicus infrastructure (from Copernicus: https://www.copernicus.eu/en/upcoming-copernicus-data-and-information-access-services-dias)

Based on requirements laid down by the EC (European Commission, 2016) and approved by Participating Countries, ESA launched a call for services to establish the DIAS with the aim to deploy operational access platforms in early 2018. Four commercial consortia were selected, and in parallel, EUMETSAT are building up a DIAS in a stepwise approach. The official launch of the DIAS was in June 2018.

The DIAS have to fulfil a number of different requirements with respect to data, service, performance and legal issues. The full list of criteria can be found in European Commission (2016), however a selection is provided below:

- Access to Copernicus data and information virtually collocated with computing resources
- Operational in terms of reliability, robustness, performance
- All data accessible (but not all locally available), including Sentinels + Services information
- Full, free and open access to the Copernicus data and information
- Non-discriminatory access and use
- Include processing services, viewing, discovery & download
- Interoperability features (Data/Services)
- Offers storage and processing under commercial conditions

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

- Protection of data and IPR, different licences (management, free, paid, etc.), and IPR belongs to the creators
- Provide development environment solutions for applications building, including chaining of services/software/API/etc
- Third parties can offer their front office services on top of DIAS back office.

Through the DIAS, users will have full and free access to Copernicus data and services and will, at commercial conditions determined by the DIAS providers, be able to process the data and information to create services for their end users. It is expected that this will stimulate the downstream industry as service providers will be able to more easily develop their own applications based on the free Copernicus data and any other data made available through the system, and to build flexible value chains based on the DIAS computing and storage resources at a competitive cost.

The five DIAS are:

- CREODIAS (https://creodias.eu/), led by Creotech Instruments, consortium includes cloud provider Cloud Ferro, Sinergise Ltd, Geomatis SAS, Outsourcing Partner Sp. z o.o., Wroclaw Institute of Spatial Information and Artificial Intelligence Sp. z o.o.
- Mundi (https://mundiwebservices.com/), led by ATOS Integration, consortium includes cloud provider T-SYSTEM International, DLR, eGEOS, EOX, GAF, Sinergise Ltd, Spacemetric, and Thales Alenia Space.
- ONDA (https://www.onda-dias.eu/cms/), led by Serco Europe, consortium includes cloud provider OVH, Gael Systems and Sinergise Ltd.
- sobloo (https://sobloo.eu/), led by Airbus Defence and Space, consortium includes cloud provider Orange SA, Airbus Defence and Space, Geo SA, Capgemini Technology Services SAS, CLS and VITO.
- WEkEO (https://www.wekeo.eu/), led by EUMETSAT, consortium includes ECMWF and MERCATOR OCEAN.

Wekeo is different from the other DIAS providers in that it is not an industry-led consortium, but coordinated by EUMETSAT. The concept is based on an initial phase in which the existing private clouds of EUMETSAT, ECMWF and Mercator Ocean are being exploited, while setting up a public cloud to that is open to third parties and can gradually be scaled up. Currently, access is granted to third party users free of charge after evaluation of the request, while for the next phase, provision of cloud services is being tendered out to market parties.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

3.1 Comparison

In the following, the DIAS systems are compared with respect to their data content and pricing model storage and processing services (volume storage and virtual machine) based on the information made available on their web sites. For the data content, we investigated the data sets that are relevant for CoastObs (based on sensor, processing level and area coverage), and the results are listed in Table 2. The general cloud service providers Google and Amazon were included as a benchmark.

Table 2: Overview of data availability for Europe at the different DIAS and two general cloud service providers according to their respective websites

Data availability overview							
Data set	CREODIAS	MUNDI	ONDA	SOBLOO	WEKEO	Google	Amazon
SENTINEL-3 L1 OLCI EFR	05/2016 - present	-	- (L2 only)	04/2016 - present	12/2016 - present	-	-
SENTINEL-2 L1C	07/2015 - present	06/2017 – present	07/2015 - present	04/2016 - present	06/2015 – present (?) status n/a	2015 - present	2015 - present
SENTINEL-2 L2	07/2015 - present	06/2017 - present	07/2015 - present	-	-	-	2015 - present
LANDSAT-8 L1	03/2013 - present	07/2018 - present	Last 12 months	-	-	2013 - present	2013 - present
ENVISAT MERIS L1B FRS	05/2002 – 04/2012	-	- (announced for 06/2019)	-	-	-	-
MODIS AQUA L1	-	-	-	-	-	-	-

For the storage and processing offer, we took two examples (volume storage and a virtual server with the same or similar specifications), with the results shown in Table 3. Wekeo is not included in this comparison because of their specific set up. They offer their services free of charge (after approval of the request by a board) until April 2019. For the period after April 2019, they will tender out these services to market parties, which means that at this point no

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

information is available on the pricing. We included Google Cloud services in the comparison as a benchmark for general-purpose cloud services.

Table 3: Comparison of example cloud services

Comparison of two example cloud service offers from the DIAS systems and Google					
Offering	CREODIAS	MUNDI	ONDA	SOBLOO	Google
Virtual server 2 core, 8GB RAM / month	67.326 €	72.45 €	from 22.00 € (7GB RAM)	53.00 €	$48.55 (approx. 42.69 €) (7.5GB RAM)
Volume backup storage per GB / month	0.046 €	0.01 €	0.04 €	0.05 €	$0.02 - $0.035 (approx. 0.018-0.031 €)

It should be noted that the offers of the DIAS systems are still to some extent in development. For example, not all data that are required by the commission are available from all providers yet, and additional data sources are also being added. In this respect, ONDA is the most transparent as they have a roadmap on their website informing the users of the expected release of all upcoming datasets.

Due to the recent launch of the DIAS systems, little practical experience is available yet on their usability and whether all functionality is available as expected and works out-of-the-box or whether a lot of extra effort is required to set up a custom processing infrastructure within the systems. As it is not feasible to thoroughly test all five systems, a first decision on which system to use at this point is made based on the published specifications (data offering, storage and processing offering).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

4 Discussion

The field of data cubes and big data analytics for EO data is currently very dynamic and it is clear that they will play an important role in the future of Earth observation. The advantages are obvious, but due to the dynamic nature of the field and the overall not so high level of maturity, the transition to using these technologies operationally is not easily made. So far, it has been mainly large national or international organisations that have invested in setting up operational data cubes. CoastObs provides a good opportunity to evaluate and test the current solutions for use in a small project or company with limited resources. Ideally, the system set up in the project can be used in operational service provision afterwards. Should it turn out that the solution implemented in the project does not fulfil the needs of the service provider, at least the first steps have been made and the next iteration with a different solution can be pursued on a more experienced and better informed basis.

There is a certain risk involved in setting up a processing infrastructure aimed to be operational and sustainable in one of the DIAS, as there is no guarantee that all of them will exist beyond the period of funding through the EC, which ends in about three years. Even during this period, funding by the EC is tied to fulfilling their requirements, so a DIAS could be terminated at any time. Using WEkEO might pose somewhat less of a risk as it is coordinated by EUMETSAT, which means that it is not aimed at making a profit, and so is expected to continue as long as it is business neutral. On the other hand, all of the DIAS offer similar interfaces and processing infrastructure, so in theory, it should be relatively easy to switch from one to another should the need arise. Despite this risk, the use of the DIAS infrastructure is also seen as a great opportunity by the CoastObs project, as it is expected to allow for more efficient data acquisition and processing, as well as easy upscaling when required to enable seamless graduation of applications from research into operational use.

Thus, the CoastObs project will aim to use rasdaman and Creodias in the first instance. The decision for Rasdaman is mainly based on the higher level of maturity and documentation as compared to other solutions. The Creodias is chosen for use in CoastObs mainly due to the most extensive data offering. Finally, we note that this will be the initial approach tested for CoastObs and that should the need arise to make amendments to the infrastructure this will be considered as part of an adaptive process.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

5 References

Appel, M., Lahn, F., Buytaert, W. and Pebesma, E. (2018): Open and scalable analytics of large Earth observation datasets: From scenes to multidimensional arrays using SciDB and GDAL. ISPRS J. Photogramm. Remote Sens. 138: 47–56.

Baumann, P. (2017): The Datacube Manifesto. http://www.earthserver.eu/tech/datacube-manifesto. Retrieved 2018-04-21.

Baumann, P., Dehmel, A., Furtado, P., Ritsch, R. and Widmann, N. (1998): The multidimensional database system RasDaMan. SIGMOD Rec. 27, 2 (June 1998), 575-577. DOI: https://doi.org/10.1145/276305.276386

Cudre-Mauroux, P., Kimura, H., Lim, K.-T., Rogers, J., Simakov, R., Soroush, E., Velikhov, P., Wang, D.L., Balazinska, M., Becla, J. , DeWitt, D., Heath, B., Maier, D., Madden, S., Patel, J., Stonebraker, M. and Zdonik, S. (2009): A demonstration of scidb: A science-oriented DBMS," Proc. VLDB Endow., vol. 2, no. 2, pp. 1534–1537.

European Commission (2016): Functional Requirements for the Copernicus Distribution Services and the Data and Information Access Services (DIAS). Ref. Ares(2016)6887639 - 09/12/2016 [Online] http://copernicus.eu/sites/default/files/documents/News/Data_Access_Functional_Requirements_Dec2016.pdf Accessed 2/11/2018

Giuliani, G., Chatenoux, B., De Bono, A., Rodila, D., Richard, J.-P., Allenbach, K., Dao, H. and Peduzzi, P. (2017): Building an Earth Observations Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD), Big Earth Data, 1:1-2, 100-117, DOI: 10.1080/20964471.2017.1398903

Kovanen, J., Mäkinen, V. and Sarjakoski, T. (2018): An Approach for Assessing Array DBMSs for Geospatial Raster Data. Proceedings of GEOProcessing 2018: The Tenth International Conference on Advanced Geographic Information Systems, Applications, and Services, 71-76.

Leith, A. (2018): What is the Open Data Cube? Medium. [Online] https://medium.com/opendatacube/what-is-open-data-cube-805af60820d7 Accessed 1/11/2018.

Lewis, A., Oliver, S., Lymburner, L., Evans, B., Wyborn, L., Mueller, N., Raevski, G., Hooke, J., Woodcock, R., Sixsmith, J., Wu, W., Tan, P., Li, F., Killough, B., Minchin, S., Roberts, D., Ayers, D., Bala, B., Dwyer, J., Dekker, A., Dhu, T., Hicks, A., Ip, A., Purss, M., Richards, C., Sagar, S., Trenham, C., Wang, P. and Wang, L.-W. (2017): The Australian geoscience data cube - Foundations and lessons learned Remote Sens. Environ., 202: 276-292.

ODC contributors (2018): Open Data Cube Manual [Online] https://datacube-core.readthedocs.io/en/latest/index.html Accessed 1/11/2018

Paradigm4 (2017): SciDB Documentation 18.1 [Online] https://paradigm4.atlassian.net/wiki/spaces/scidb/overview. Accessed 1/11/2018.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348

Planthaber, G., Stonebraker, M. and Frew, J. (2012): EarthDB: Scalable analysis of MODIS data using SciDB. In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, Redondo Beach, CA, USA, 6 November 2012.

Rasdaman team (2018): rasdaman v9.7 documentation [Online] http://doc.rasdaman.org/index.html. Accessed 1/11/2018.

Stonebraker, M., Brown, P., Zhang, D. and Becla, J. (2013): SciDB: A database management system for applications with complex analytics. Comput. Sci. Eng., 15: 54–62.

Tan, Z. and Yue, P. (2016): A comparative analysis to the array database technology and its use in flexible VCI derivation. In Proceedings of the 2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Tianjin, China, 18–20 July 2016; pp. 1–5.

Tan, Z., Yue, P. and Gong, J. (2017): An Array Database Approach for Earth Observation Data Management and Processing. ISPRS Int. J. Geo-Inf. 6(7: , 220; https://doi.org/10.3390/ijgi6070220

Wagemann, J., Clements, O., Figuera, R.M., Rossi, A.P. and Mantovani, S. (2017): Geospatial web services pave new ways for server-based on-demand access and processing of Big Earth Data, International Journal of Digital Earth, 11(1): 7-25, DOI: 10.1080/17538947.2017.1351583

Wagner, W., 2015. Big Data Infrastructures for Processing Sentinel Data. Photogrammetric Week 2015, pp. 93–104.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776348