Copernicus Access Platform Intermediate Layers Small Scale Demonstrator

D1.10 DataCube Integration Report v2

Document Identification			
Status	Final	**Due Date**	31/10/2019
Version	1.0	**Submission Date**	27/11/2019

Related WP	WP1	**Document Reference**	D1.10
Related Deliverable(s)	D1.9	**Dissemination Level (*)**	PU
Lead Participant	Atos FR	**Lead Author**	Jean-François ROLLAND (Atos FR)
Contributors	SGIS TAS FR DLR TerraNIS CloudFerro	**Reviewers**	Mihai Datcu (DLR) Romain Hugues (TAS FR)

Keywords:
Datacube, earth observations, analysis ready data, data processing

This document is issued within the frame and for the purpose of the CANDELA project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 776193. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission. The dissemination of this document reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

(*) Dissemination level: **PU**: Public, fully open, e.g. web; **CO:** Confidential, restricted under conditions set out in Model Grant Agreement; **CI:** Classified, **Int =** Internal Working Document, information as referred to in Commission Decision 2001/844/EC.

Document Information

List of Contributors	
Name	Partner
Jean-François ROLLAND Anne HAUGOMMARD François-Xavier STEMPFEL	ATOS FR
Anna PULAK	SGIS
Michelle AUBRUN Romain HUGUES	TAS FR
Michele IACOBELLI	TAS IT
Mihai DATCU	DLR
Mohanad ALBUGHDADI	TerraNIS

Document History			
Version	Date	Change editors	Changes
0.1	08/07/2019	Fabien CASTEL (ATOS FR)	Initial Table of Content
0.2	10/10/2019	All contributors	First version integrating contributions from partners
0.3	25/11/2019	Anne HAUGOMMARD (ATOS FR)	Version for QA
0.4	26/11/2019	Juan Alonso (ATOS ES)	Quality Assessment
1.0	26/11/2019	Jose Lorenzo (ATOS ES)	Final revision before submission

Quality Control		
Role	Who (Partner short name)	Approval Date
Deliverable leader	Jean-François Rolland (Atos FR)	25/11/2019
Quality manager	Juan Alonso (ATOS ES)	26/11/2019
Project Coordinator	Jose Lorenzo (ATOS ES)	27/11/2019

Table of Contents

List of Tables

List of Figures

List of Acronyms

Abbreviation / acronym	Description
API	Application Programming Interface
CRS	Coordinate Reference System
D1.10	Deliverable number 10 belonging to WP1
DIAS	Data and Information Access Services (ESA initiative to access Copernicus data)
EC	European Commission
GRD	Ground Range Detected
MEEO	MEteoreological Environmental Earth Observation
ODC	Open Data Cube
VHR	Very High Resolution
VM	Virtual Machine
WP	Work Package
S3	Simple Storage Service (developed by Amazon)
ETL	Extract, Transform and Load

Executive Summary

This document is the continuation of the deliverable D1.9 [1] which was a first version of Datacube integration report. In D1.10, we answer to the comments and questions raised by the previous version of this work and present the result of our research on data cube technologies.

This deliverable starts by presenting the concept of data cubes as one solution for accessing earth observation data.

In a second part, CANDELA partners from WP1 and WP2 express their needs and expectation in term of data access. We dissociate two types of needs in this chapter, first the WP2 members define what type of data their algorithms work with and define the constraints on the input for their tools. The second point of view is the one of the users of the platform represented by our partners from WP1. They define the geographical and temporal extent needed for their use cases.

In the next part we review and present different data cube solutions. These solutions can be separated in two major classes, the data cube software that could be deployed on a platform like CANDELA, and the emerging services providing access to the data through a data cube or a in a similar way.

Finally, a conclusion is presented on the use of the data cube solutions presented regarding both their status, functionalities and integration constraints, and CANDELA specific needs.

1 Introduction

1.1 Purpose of the document

The objective of this document is to provide a report on the investigation done on the topic of datacube systems and their integration to CANDELA platform. This document is the second version of such report, coming after D1.9 document delivered at M6. It shall serve as a basis for the decision on what data access system should be used and/or deployed on the CANDELA platform by providing all the technical and functional elements related to this topic.

1.2 Relation to other project work

This deliverable is part of the work performed on the Task “T1.3 Use case Dataset preparation" led by Atos FR with the collaboration of TAS FR, DLR, TerraNIS, SGIS and CloudFerro.

It is also related in some ways to the “WP3 Demonstrator Implementation" as all the information gathered on datacube systems will serve as an input to define what will be implemented into the CANDELA demonstrator.

1.3 Structure of the document

This document is structured into 4 major chapters

- **Chapter 2** presents the general concept of datacube
- **Chapter 3** presents the needs of CANDELA in terms of data access
- **Chapter 4** presents the solutions to fulfil these needs
- **Chapter 5** is the conclusion of the document

2 The Earth Observation Datacube concept

For an introduction on the Datacube concept, please refer to "D1.9 Datacube Integration Report v1" [1]. ETL (Extract, Transform and Load) functionalities for Earth Observation Data are crucial in order to ensure proper performance of Data analytics projects as shown by the examples of the two DIAS (Mundi & CreoDIAS). However, these functionalities can be covered by several types of technologies including Datacubes (which is just a part of the chain).

Considering the fast evolution of Storage and Data Access technologies such as object storage, Datacube might not be the only relevant technology.

2.1 Comparison criteria for Earth Observation ETL solutions

The following list defines the ETL characteristics specific to Earth Observation Data needs sorted by decreasing importance:

- **Scalability:** Capacity to operate datasets with large size. Copernicus data scale is around 50 PB at the moment.
- **Evolutivity:** Apart from legacy datasets such as Meris or Old Landsat, EO datasets are constantly growing. 10PB / year for Copernicus.
- **Throughput:** Data Volume which can be downloaded/uploaded per unit of time
- **Searchability:** Given metadata such as product type, location and date at minima (possibly more), the system should be able to retrieve the corresponding data. No less and No More.
- **Parallel accessibility:** Capacity for several users to access the same piece of data and for one user to access several pieces of data without loss of performance
- **Data footprint:** storage overhead due to redundancy and duplications caused by the exploitation of data.
- **Latency:** Time between data request and data reception. Some solution provides it as a best effort, and some have a predicable latency.

And of course, **Cost**.

2.2 The components of data access

Data access solutions combine different components: hardware, storage, search and pre-processing, described in the following sections.

2.2.1 Hardware

Different types of hardware come into play for accessing the data on a large scale:

- The type of memory (Hard Disk Drive, Flash, Bands)
- The type of server hosting the drives
- The network connecting the servers.
- Etc.…

For CreoDIAS Local storage is based on a Ceph solution with around 10,5 PByte capacity. Resources are connected with 100 GBit/s SDN network switching.

2.2.2 Storage System

Different types of storage system are available the most commonly used are file, block and object storage. They have different costs and characteristics.

File Storage

File storage is the old way to store data but can still be relevant today. The data is accessed by file name and is organised in folder. Files have attributes like type, size, creation date, permissions. This storage system can have advantages when it comes to share data or to store a mix of structured and unstructured data. This kind of storage shows its limit when it comes to rescale and their inability to spread the workload across multiple server.

Block storage

In block storage the data is organized through blocks, each block has a unique block Id. Each file is splitted into a set of blocks of equal size. There is no metadata associated with blocks and a server-based operating system manages these volumes. This solution can scale up and often can guaranty the replication of the data. One of its main advantages is the low latency to access to blocks.

Object storage

Object storage is based on key-value addressing. The user usually accesses to the data through an application over HTTP. Each file also has a set of associated metadata.

Object storage is now the baseline system for cloud due to its advantages in terms of scalability, parallel accessibility, throughput although a sometimes.

Object Storage allow some in-place processing yielding very reduced data footprint and latencies. However, it can't be applied for any applications at the time.

2.2.3 Storage Format

Earth Observation data including Copernicus is mostly N-dimensional arrays with at least 2 (sometimes 3) spatial dimensions, and a spectral dimension. Additional dimensions are time, product level, sensor type etc…

- The way data is stored will induct the access performances.
- Some solutions like Datacube chose to add the temporal dimension to spatial/spectral dimensions with significant advantages in terms of latency for time series request since it is natively included in the storage. On the other hand, it requests (sometimes heavy) product pre-processing
- Some other solutions work on the spatial dimension optimization. It is known that Google Earth Engine stores Copernicus Sentinel 1 & 2 in a tiled and multi-zoom level format allowing low-latency on-the-fly images processing and visualization.
- Other solutions will keep data product under the form which was provided by the ground and rely on a combination of Hardware / Storage system + Search tool for accessibility and reduced latency.
- Some storage formats have affinities with Storage systems: Zarr, Cloud Optimized Geotiff

2.2.4 Search tool

For requesting data with metadata field corresponding to dimensions which are not natively included in the storage format or other, a search tool is necessary.

Different solutions can be used, in an object storage, the search tool can be based on the key value associated to the object or on its metadata.

Another solution is to use a relational database to store the list and properties of the data.

2.2.5 Pre-processing

Last step of data access, pre-processing is probably the first step of application.

- Pre-processing can be done once at ingestion phase of the data in the storage format or on the fly during data loading which can increase latency and reduce scalability.
- Decompression: some data are compressed either with generic compression like zip or Image specific like JPEG 2000. Data decompression always takes a significant amount of time and highly increases latency.
- Geometric alignment: certain applications can't
- Radiometric correction: correct the radiometric variations between two images
- Additional formatting: tiling, zooming

2.3 CANDELA ETL solution

2.3.1 CANDELA solution

The CANDELA platform runs on the CreoDIAS architecture and can use the different types of access that CreoDIAS offers.

The simplest way to access to a product on CreoDIAS is to perform an OGC OpenSearch research on its Rest API. The response of a search is a set of products corresponding to the search criteria.

For each product in this result we can have two way to access to the data:

- download
- and direct access.

A product description contains a download link to retrieve a zip files containing all the data and metadata of the product. This method is quite slow and requires a lot of interaction. It is never used on CANDELA.

The second solution consist to use the product identifier. In the CreoDIAS system the product identifier corresponds to the path of the product on the S3 object storage system containing all the EO data proposed. As this storage system is directly accessible from CANDELA, any process on the platform can access to a product directly once it knows its identifier. It's very efficient and fast. It allows to access to the raw data and all its metadata. The main drawback of this solution is that the user can not have a granularity smaller than one product.

The third solution is to use the OGC standard WMS/WCS services provided by Sentinel-Hub. This solution allows the user to select precisely an area of interest. But this access time is slower than the direct access when the objective is to access the full content of the product. This solution is well suited for usage of OGC standards with reduced area of interest, or for the display of a product in a web application.

2.3.2 Datacube solution

A datacube is a multidimensional array of values. When applied to geoscience this term describes both a data structure and a set of tools used to create, manipulate, extract data from the datacube. The data stored in the datacube is pre-processed and calibrated in order to align different sources of data. The same pixel in different layers of the datacube represents the same location. This unified representation allows to reduce the processing time of analysis of the data. But on the other hand, it needs more time for the ingestion of the data. One of the main advantages of a datacube is the simplicity to generate time series of observation on an area.

Another drawback of this solution is that the data is duplicated from the data sources (DIAS products for example), to the datacube.

2.4 Conclusion

Setting up and maintaining a Datacube instance on the CANDELA platform seems not efficient in terms of resource usage and invested time. Deploying a datacube is a complex task, the ingestion part of different data sources requires a lot of preparation to align the different types of data. It is also a process that may take a lot of time and resource storage.

Relying on efficient data access services (such a WMS/WCS services and POSIX type file access provided by CreoDIAS or Mundi for instance) is more efficient for a data analytics platform such as CANDELA.

A third solution would be to use an external Datacube service such as the Datacube Facility Service that will be offered soon in the frame of an ESA ITT.

3 CANDELA specific needs and expectations

To apply the CANDELA data analytics algorithms on satellite imagery data there is a clear need for multi-sensor data pre-processing and multi-temporal data stacks preparation. The following use cases are considered:

Retrieval of time series for user-defined specific spatial areas

Access to multi-source data in an efficient manner (reprojection, co-registration...)

Preparation of the data as input to the processing libraries

The data should be made available for search, processing and download via queries to a Restful interface to ensure full interoperability between the data provider service and the algorithms.

The following chapters detail the specific needs related to each use case and class of algorithm currently planned to be implemented on the platform.

3.1 Change Detection

The optical change detection tool is designed to work only on Seninel-2 data products available for general users that are generated by the ground segment or by the SENTINEL-2 Toolbox (see Figure 3). Indeed, the module of pre-processing of the change detection tool is based on folder architecture and file names of these products. For instance, the algorithm searches a folder name “IMG_DATA” inside each input product to extract the bands of interest.

Name	High-Level Description	Production & Distribution	Data Volume
Level-1C	Top-Of-Atmosphere reflectances in cartographic geometry	Systematic generation and online distribution	~600 MB (each 100km x 100km^2)
Level-2A	Bottom-Of-Atmosphere reflectances in cartographic geometry	Systematic and on-User side (using Sentinel-2 Toolbox)	~800 MB (each 100km x 100km^2)

Figure 1: SENTINEL-2 product types ***(source: https://sentinel.esa.int/web/sentinel/missions/sentinel-2/data-products)***

Moreover, these Sentinel-2 data products have some useful features that are exploited by the change detection tool, like:

- Fixed tiles of 100x100 km^2 in UTM WGS84 projection. Thus, the tool looks only for the coordinates of the top left pixel in order to know if two images are aligned and represent exactly the same area.
- A single orbit for each tile. Thus, the tool registers ids and acquisition dates of both tiles used inside the metadata of the change detection map
- All the Sentinel-2 bands provided are ortho-images. Thus, the tool does not take care of problems related to viewing angle.

As Sentinel-2 satellites provide an important quantity of data with their wide swath width (290 km) and high revisit time (5 days), the optical change detection tool has been developed in order to be

compatible with big data applications, that is means to process 1 year of data upon 1Mkm2, what represents a few thousand data.

The SAR change detection tool is intended to work on standard Sentinel-1 Ground Range Detected (GRD) Level-1 data products available for general users and generated by the ground segment or the Sentinel-1 toolbox. The pre-processing module searches for all Sentinel-1 products in the specified input folder in the SAFE format (either zipped or extracted), as downloaded from the database and converts them in the format needed by the Change Detection module.

Table 1: Sentinel-1 product types

Name	High-Level Description	Production & Distribution	Data Volume
Level-1 IW GRDH	Focused, multi-looked and projected to ground range backscattering amplitude data acquired in Interferometric Wide-swath mode	Systematic generation and online distribution	~1.65 GB (each 2580 x 16690 km^2)
Level-1 SM GRDH	Focused, multi-looked and projected to ground range backscattering amplitude data acquired in Strip Map mode	Systematic generation and online distribution	~240 MB (80 x 165 km^2)

These products are characterized by the following features exploited by the change detection tool:

- Squared 10 m pixels in UTM WGS84 projection for geocoding and extraction of the intersected area between different acquisitions.
- Standard naming convention allowing chronological sorting of images by simply comparing the name of the SAFE or zip archive.
- Each image is the result of a single acquisition (i.e. orbit), allowing registration of *ids* and sensing date in the metadata of the change detection map.
- Dual polarization (*VV* and *VH*). Thus, each product is made of two images: both are processed by the tool if no preference is specified by the user.

With a 250 km swath in the main IW mode and a 6 days revisit time at the equator, the Sentinel-1 mission provides up to 2 TB of data per day (potentially 730 TB per year), available to general users within 24 hours of observation and it is thus suitable for big data applications.

3.2 Macro-economics and agriculture

Two sub-use cases are proposed by TerraNIS in the framework of macro-economics and agriculture. The first sub-use case is related to urban expansion and agriculture and targets metropolitan areas (Bordeaux, Milan and Sicoval). The objective is to conduct a temporal analysis over these areas using 2 or more years. The second sub-use case is dedicated to change detection on vineyards, where a pair of images are compared (before and after a natural hazard). For these sub-use cases Sentinel-1 Ground Range Detected – Interferometric Width (GRD-IW) and Sentinel-2 Level 2 bottom of the atmosphere (L2A) products are used. Additionally, very high resolution (VHR) datasets coming from Pleadies, SPOT6-7 and TerraSAR-x will be used for validation purposes.

3.3 Forest health monitoring

This use case targets specific forest areas and issues related to them. First sub-use-case is focused on abrupt natural disasters for instance windthrows which cause a lot of damage to the European forests. The test area is a forest of Regional Directorates of the State Forests in Toruń where in 2017 windfall cover almost 1,000 square kilometres of trees. The second sub-use case is related with forest heath monitoring to detect long-term changes in the forest cover on the example of the Białowieża primeval forest. For both use case, Sentinel 1 and 2 products are used. In the first use case pair of images before and after windthrown is needed. For second scenario data should be compered in an annual basis. For validation phase VHR datasets (Pleadies, SPOT6-7, TerraSAR-) will be used.

3.4 Foreseen use cases

The foreseen use case that best illustrated the ETL functionalities is the Natura 2000 use case that is presented in the deliverable D1.5[20]. This use case regroups more than twenty thousand sites spread in all European Union and that represent an area of little more than 1Mkm^2. The idea of this use case is to run both optical and SAR change detection pipelines on the Natura 2000 areas and to store the results in a semantic way thanks to the semantic indexation tool.

3.5 Conclusion

The persistent scenario seems well suited for the "Macro-economics and Agriculture" and "Forest Health Monitoring" uses cases, as a specific set of Sentinel 1 and 2 data targeting the needed areas could be pre-indexed and pre-ingested to enable faster analysis.

The "Change Detection" algorithm should be used in the frame of the 2 presented use cases so a persistent scenario could be sufficient. However, setting up an on-demand or hybrid scenario could be used for the testing and benchmarking of the algorithm on wider areas.

The need for VHR data for these 2 use cases has to be kept in mind as the ability by a data cube solution to ingest high resolution images is not ensured. This criterion has to be taken into account for the choice of the solution.

4 Datacube Solutions

Two different paradigms are proposed to users willing to integrate a datacube solution to their earth observation processing chain.

The first one is to setup and operate a datacube instance in its own infrastructure, managing the data indexing and ingestion, as well as the storage of the metadata and reprocessed data. The requirements in term of processing capabilities (for the ingestion part mainly) and storage is not negligible but can be afforded if a trade-off is done between the available resources and the area and time of interest the installed datacube provides.

The second one is to rely on an external service providing datacube features, i.e. an API that can serves earth observation data on user defined area and time of interest, from the pixel level to wider area and time period. This kind of solution offers a lot of benefits for the users, who do not need any more to deal with low level issues and infrastructure resources. The availability and reliability of this kind of service is the problematic point though. Initiatives are starting to emerge from European institutions (ESA mainly) but no operational service is available now. The cost for this service will also be an important matter, as it is not sure at all that a free service will be available any time soon.

4.1 Datacube Solutions with Full Installation

4.1.1 Rasdaman

Please refer to “D1.9 Datacube Integration Report v1” [1] for a presentation of Rasdaman.

4.1.2 Open Data Cube

Open Data Cube (ODC) [4] is an open source project that was born from an Australian spatial agency initiative, the "Australian Geoscience Data Cube". The core code base moved to the Open Data Cube open source project in 2017. The willingness expressed by the project is to always remain 100% open source, free for all to use and release under an Apache 2.0 license.

Many national initiatives using this code base exist over the world. The core code is totally open and can be installed on personal computers or cloud facility.

An ecosystem of tools for data access, processing and visualization is under development but not yet guaranteed to be stable. These various projects gravitating around ODC Core can be found in the Open Data Cube GitHub repository [5].

4.1.2.1 API

The core ODC system provides a Python API. Various Jupyter notebooks are provided using this API to perform many use cases. No web service API is provided at all. A project to provide an OGC standards interface exists but might not be stable yet.

4.1.2.2 Data Ingestion

Open Data Cube is a physical datacube system. It indexes earth observation product in a relational postgres database. Then an ingestion process retiles the earth products into ODC specific NetCDF files that are used to serve subset data.

Support for specific earth product exists. The document [4] the developing team of the Swiss Datacube presents how to setup a datacube and ingest products from landsat and sentinel 2 missions. In [5],the authors present different workflow to prepare and ingest sentinel 1 products into an open datacube.

4.1.2.3 Testing

The Open Data Cube team provides two ways of testing their product:

- A sandbox packaged with a set of algorithms to demonstrate the power of ODC [6]
- A docker image containing a ready-to-use ODC instance, called “Cube in a Box” [7]

4.1.3 Pangeo

Pangeo is more a community for geoscience than a real datacube solution. The community develops maintain and document a set of tools dedicated to scientific research. There is no single package called “Pangeo”, it is more a software ecosystem involving open source tools. The goals of the Pangeo project are:

- Foster collaboration around the open source scientific python ecosystem for ocean / atmosphere / land / climate science.
- Support the development with domain-specific geoscience packages.
- Improve scalability of these tools to handle petabyte-scale datasets on HPC and cloud platforms.

4.1.3.1 Technical architecture

Due to the nature of the project there is no fixed architecture but an ecosystem of tools that can be used together.

The key concepts and tools we envisioned in the Pangeo ecosystem were:

- Ability to use high-level data models
- Ability to leverage distributed parallel computing on HPC systems or on cloud computing systems
- Ability to work interactively or using batch processing

The core packages of Pangeo reflects these key concepts:

Xarray

Xarray is an open source project and Python package that provides a toolkit for working with labeled multi-dimensional arrays of data. Xarray dataset is an in-memory representation of a netCDF file. Xarray provides the basic data structures used by many other Pangeo packages, as well as powerful tools for computation and visualization.

IRIS

Iris seeks to provide a powerful, easy to use, and community-driven Python library for analysing and visualising meteorological and oceanographic data sets.

With Iris you can:

- Use a single API to work on your data, irrespective of its original format.
- Read and write (CF-)netCDF, GRIB, and PP files.
- Easily produce graphs and maps via integration with matplotlib and cartopy.

Iris is an alternative to Xarray.

DASK

Dask is a flexible parallel computing library for analytics. Dask is the key to the scalability of the Pangeo platform; its data structures can represent extremely large datasets without actually loading them in memory, and its distributed schedulers permit supercomputers and cloud computing clusters to efficiently parallelize computations across many nodes.

JUPYTER

Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages. Jupyter provides the interactive layer to the Pangeo platform, allowing scientists to interact with remote systems where data and computing resources live.

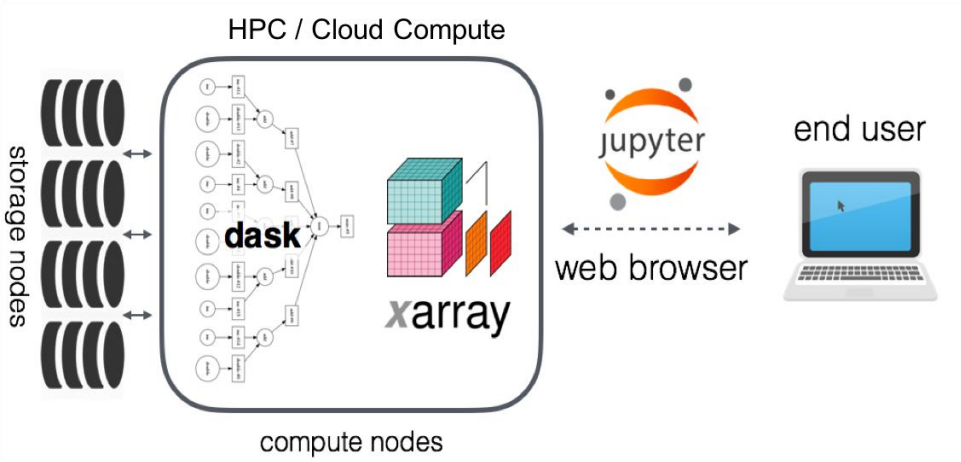

Figure 2:Technical architecture of Pangeo – (source: http://pangeo.io/)

4.1.3.2 Data storage

The main storage formats used in Pangeo are The Hierarchical Data Format (HDF) and the Network Common Data Format (NetCDF), two of the most common on-disk storage layers across the geosciences

The preference for storing multidimensional array data in the cloud is the Zarr format. Zarr is a new storage format which, thanks to its simple yet well-designed specification, makes large datasets easily accessible to distributed computing. In Zarr datasets, the arrays are divided into chunks and compressed. These individual chunks can be stored as files on a filesystem or as objects in a cloud storage bucket. The metadata are stored in lightweight .json files. Zarr works well on both local filesystems and cloud-based object stores. Existing datasets can easily be converted to Zarr via xarray's zarr functions. An ingestion phase is needed to parse the data in its original format and to convert it to Zarr.

At this time Pangeo is focused on HDF and NetCDF file format. Xarray as functions to read geotiff images but these functions are deemed experimental.

4.1.3.3 Use cases

The main use cases presented by Pangeo are focused on climate modeling, oceanography and meteorology:

- Sea Surface Altimetry Data Analysis
- CESM Large Ensemble Tracer Budget
- US Precipitation and Temperature Analysis

4.1.3.4 Deployment

By its nature, the deployment of a DataCube based on Pangeo requires some work to:

- choose tools from those which are suggested,
- design the architecture of the components,
- implement the solution.

4.1.4 MEEO

In the last 10 years MEEO (MEteoreological Environmental Earth Observation) [2] has developed a platform named Eodataservice [18] that implements the Digital Earth concept: eodataservice is a generic platform that allows accessing to large set of multi-year global geospatial collections allowing data discovery, visualization, combination, processing and download, providing OGC OpenSearch and WCS(Web coverage service) endpoints .

The MEEO datacube tool alone is a full commercial and licensed virtual datacube solution and provides limited documentation. Its integration into CANDELA platform would be highly complex but this datacube solution is the one chosen by Mundi webservices to provide a “Datacube as a service” (see section 4.2.2).

4.1.5 Xcube

Xcube [17] is an open source project proposing a data cube toolkit based on Xarray, dask, and zarr. The project is based on the same technologies as Pangeo but with the goal to provide Earth observation data in an analysis-ready form to users.

Xcube achieves this by converting EO data sources into self-contained data cubes that can be published in the cloud.

4.1.5.1 Data Model

All Xcube datasets are structured in the same way following a common data model. They are also self-describing by providing metadata for the cube and all cube’s variables following the CF conventions.

An Xcube dataset is self-describing, it contains metadata that define the type of data and its organisation in the cube. When an Xcube is opened, only the structure and this metadata are loaded in memory. The actual data is loaded on demand when a process needs it. A process can load only a part of the data for its needs: a subset of the bands available, a certain area…

If new data is generated by a process it can be added to the DataCube.

Xcube uses the Zarr format as physical representation, the zarr format support parallel processing and is optimized to be used on cloud storage such as amazon s3.

The main drawback of using the zarr format is that all the EO must be imported and duplicated in this format.

4.1.5.2 Workflows

The first part of the workflow is to generate an xcube and deploy it:

1. generate an xcube dataset from some EO data sources. Each data source needs a specific input processor.
2. Next the datacube can be optimised for better performance and for reducing its size
3. Before being exploited by the user a datacube usually needs to be deployed on a shared filesystem.

Then users can:

4. access, analyse, modify, transform, visualise the data using the Python API and xarray API through Python programs or JupyterLab, or
5. extract data points by coordinates from a cube using the xcube extract tool, or
6. resample the cube in time to generate temporal aggregations using the xcube resample tool.

4.1.5.3 Toolkit

On top of its base packages (xarray, dask, zarr), xcube includes several high-level tools to exploit the data:

- CLI - access, generate, modify, and analyse xcube datasets using the xcube tool;
- Python API - access, generate, modify, and analyse xcube datasets via Python programs and notebooks;
- Web API and Server - access, analyse, visualize xcube datasets via an xcube server;
- Viewer App – publish and visualise xcube datasets using maps and time-series charts.

4.1.6 Conclusion on datacube solutions with full installation

Deploying a datacube instance on CANDELA does not seems to be a relevant solution regarding the work needed to set it up and subsequent costs of infrastructure. Indeed, all above solutions require to deploy and maintain an infrastructure that must be hosted somewhere, implying storage costs. Data to be ingested also needs to be preprocessed and a workflow must be designed to this end, implying processing costs too. The needed amount of data put forward by some use cases, a time range of about 2 years and 1 million km², largely exceeds CANDELA's scope in terms of both storage and processing resources.

More promising solutions lie in Datacubes as a Service.

4.2 Datacube as a Service

4.2.1 CreoDIAS platform

The CANDELA platform is deployed into the CreoDIAS infrastructure. CreoDIAS is hold by CloudFerro and is one instance of DIAS (Data and Information Access Service), an ESA program willing to provide an easy access to Copernicus data to the European earth observation ecosystem of applications. CreoDIAS hosts the CANDELA platform, i.e. VMs and storage are provided by the CloudFerro infrastructure, but it also provides access to Sentinel data.

CreoDIAS proposes an APIs to search for Sentinel products based on metadata (through CSW and OpenSearch standards) and access to the actual data on their storage facility. This way of accessing data forces user to deal with Sentinel tiles as a whole.

In parallel, CreoDIAS also proposes APIs supporting OGC standards like WMS and WCS [8] [9]. Through these standards, users are able to retrieve only the data corresponding to their area and time of interest. This kind of API, while missing the capability to build time series on a single request, could be a workaround acting as a "poor" datacube but efficient enough for some use cases. Those API are implemented by Sinergise.

CreoDIAS data sources are available through an OGC endpoint with commercial access provided by Sinergise through the Sentinel-hub [17], including the following datasets, for global area:

- Sentinel-1 GRD, since 2014-10
- Sentinel-2 L1C, since 2015-11
- Sentinel-3 OLCI L1, since 2016-05
- Sentinel-5P L2, since 2018-04.

The Sentinel hub capacities are discussed in chapter 4.2.3.

4.2.2 Mundi DIAS platform

Mundi offers an access to earth observation products through several services:

- A download service that allow the user to retrieve entire products
- A web service based on OGC standards
- A datacube implementation based on MEEO tools.

The access to Mundi-webservices through Sentinel Hub is currently under implementation and will be available with commercial access in a few months.

The following services will be provided:

- OGC WMS (13 unprocessed raw bands for Sentinel-2: B01-B12)
- OGC WMTS (Web Map Tile Service)
- OGC WCS
- OGC WFS
- Statistical information (FIS service)

The datacube deployed on mundi is based on Meeo and is still in development at the time this document is written. At this point Sentinel 1 and Sentinel 2 observations since the start of 2018. This datacube is not open for public use currently.

It would be interesting to have a deeper look at the Datacube as a service functionality provided by Mundi platform when integration is complete on their side.

4.2.3 BigDataCube

The BigDataCube [19] project is developing flexible and scalable services for Earth Observation (EO) data, offered as datacubes under the lead of Jacobs University.

The project is based on Rasdaman datacube solution (see section 4.1.1) and is deployed in two infrastructures:

- The commercial hosted processing environment of cloudeo
- The public service of CODE-DE, the German Copernicus hub

Both version of the service should be federated to allow users to combine datacubes from both services without the need for downloading them first.

All the Sentinel 2 (L1C and L2A) products and Sentinel 1 GRD products over Germany are available. Sentinel 1 GRD product are also available over the North Sea.

It seems that this solution is limited in terms of geographical area, and not extendable in terms of ingested data. Consequently, this solution does not fulfill CANDELA needs.

4.2.4 Sentinel Hub

4.2.4.1 OGC Standards Functionalities

Sentinel Hub is a multi-spectral and multi-temporal big data satellite imagery service, which provides

- Access to a large set of data
- A RESTful API (SentinelHub API) to access raw satellite data, rendered images and statistical analysis [12]
- OGC API to access EO products through WMS/WFS endpoints for a standard integration into external applications [13]
- A Graphical user interface to browse the EO products available, to configure custom layers and to preview online the WMS visualization.
- Capability to define custom layer processing through Javascript code

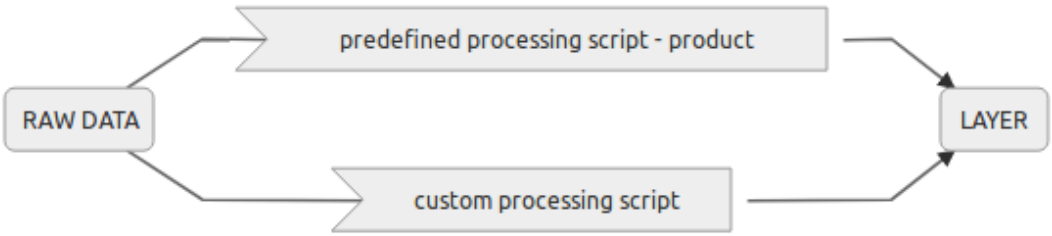

Figure 3: Sentinel Hub evalscript functionality (custom layer processing) – Source: sentinel-hub.com

A set of predefined custom scripts are available on an open-source repository [14].

The Sentinel Hub allows the support any S3 endpoint, including AmazonWebService and CreoDIAS, already available but also Mundi Web services or any custom S3 endpoint [15][16].

4.2.4.2 Data cube facility service

In parallel to the initiatives related to DIAS, ESA has launched in September 2018 an invitation to tender for a "Data Cube Facility Service". The project was kicked-off on April 23 for a duration of 3 years. The idea behind is to promote the setup of an online service providing datacube access to a wide range of data, from free Copernicus data (Sentinel, Landsat, core services…) to commercial data (VHR data, private company data…), through a marketplace system where any kind of user can access, buy and sell data. This service is planned to be built on top of a DIAS, i.e. it will not duplicate and store on its own all the data but rather relies on a DIAS to access the raw format data. The following figure illustrates this vision with DIAS living in the bottom resources tier, the datacube facility on the centre platform service tier and applications like CANDELA on the top exploitation tier.

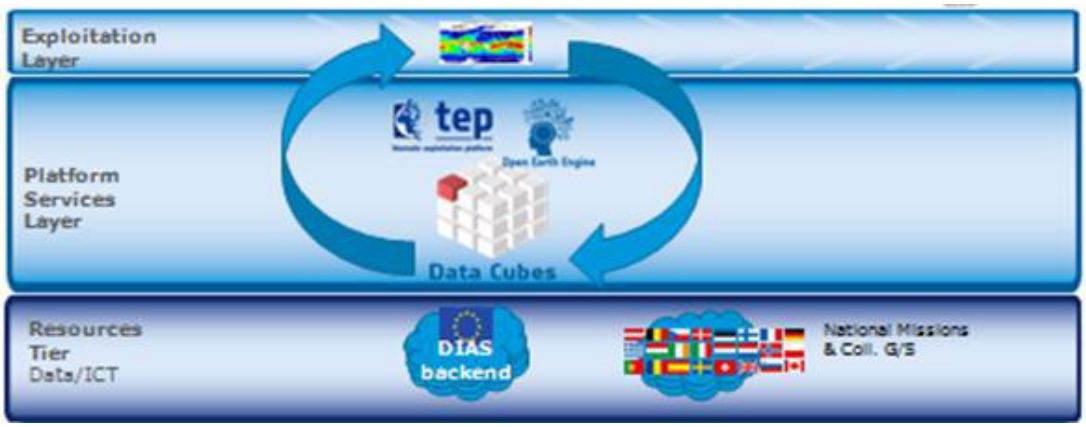

Figure 4: ESA vision on the Data Cube Facility Service

An initial and partial version of the service should be available 6 months after the start of the project, at the end of October and a complete offer after 12 months.

The European Data Cube Facility service provides different types of entry points:

- Client Data Processing Engine (CDPE) (planned for late 2019)
- On the fly data cube access service based on Sentinel Hub (online in November 2019)
- Mass processing Sentinel Hub service with asynchronous response (planned for late 2019)
- Pre-generated data cubes based on Xcube (no date communicated)

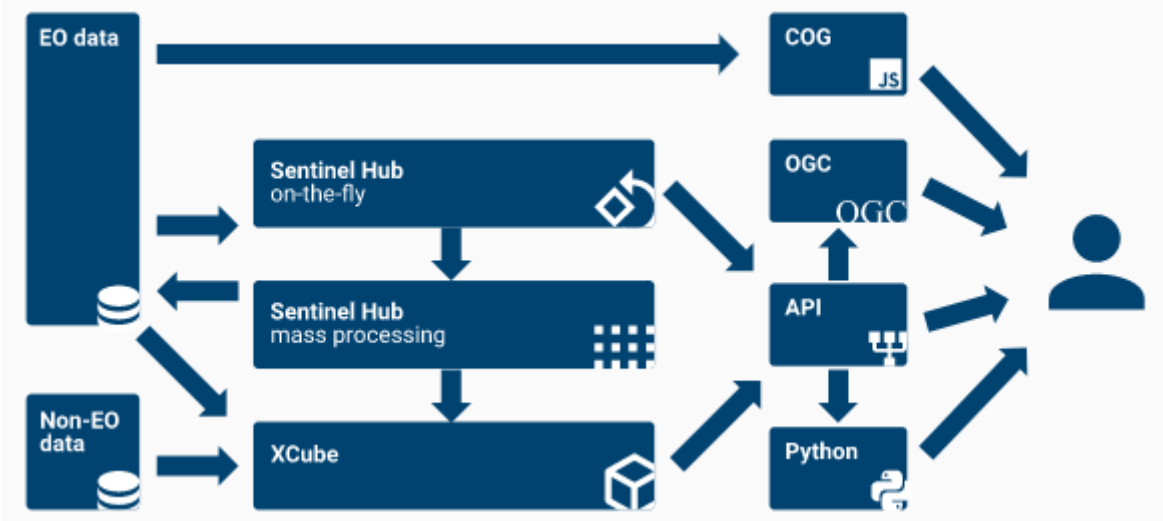

Figure 5: The European Data Cube Facility Service high level architecture. Source : https://eurodatacube.com/

The CDPE is due for late 2019. It is a web client for processing user-provided data. It is meant to leverage the COG format, therefore the client only downloads portions of data needed for the processing which is done in the user's web browser. When the data is downloaded the processing can be done offline. Various features are available: band arithmetic, filtering and subsetting on pixel values, colour correction like gamma correction and sigmoid contrast,

On-the-fly data cube access is implemented as a RESTful API that provides ARD data. The user can bring their own dataset as well as have access to archives of Sentinel, Landsat and Modis data. They can develop and run their own algorithms by packaging them with requests. The processing is then executed on an optimized system reducing processing time and costs, since there is no need for super

computer or huge storage on the user's side. Data from different sources are available and can be accessed the same way.

The mass processing service can be used to request data at large scale (for instance a whole continent). The user can define a job and its entries (a set of bands on an area over a period) and run it asynchronously.

The last service allows the user to configure and store an Xcube datacube on the cloud. They still describe their area of interest, a time span and the set of bands they want. The system instantiates an Xcube and the user can access it through a CLI, a Python API and a web API. These data cubes can be shared with others if they are published in object storage. Xcube viewer can be used to visualize data.

All sentinel data can be accessed with this system. The data itself is stored either on amazon, CreoDIAS or mundi. The user can choose one of these backends as the source for its datacube. If the user requests an Xcube object and wants to run their processes either on CreoDIAS or mundi it can be interesting to choose these DIAS as the source of data.

The price model for these services are based on the usage of the user. The price is independent of the type of usage (commercial/non-commercial) and will be the same for any user. Sinergise claims that their price cost less than the cost of building and using a private datacube (cost for development and resources).

At the time this document is written, the service could not be tested yet. A first version should be deployed by mid-2020.

4.2.4.3 Other SentinelHub services

Sinergise offers a data science Python library: eo-learn. Coming with machine learning features, it is also useful to request time series of Earth observation products. To that regard and although it obviously comes as a Python API and not a REST API, it fulfills one of the features demanded by some use cases, which require time series retrieval.

The module uses SentinelHub's OGC APIs under the hood. It creates OGC requests based on a set of parameters that include bounding box, time interval, cloud coverage and others. Eo-learn model is based on tasks (EOTasks) that can be chained to build workflows and achieve more complex tasks than a simple data request. By combining tasks, resulting data which contain too many nodata pixels (or which match any other criteria) can be filtered out.

The result of a WCS-based task is an EOPatch instance, EOPatch being an eo-learn data object containing multi-temporal data of a single area of interest. Data is stored as NumPy arrays inside EOPatches, as well as metadata, masks, etc.

A Jupyter notebook with a code snippet showing how to get time series with this module is available on CANDELA platform, at *public/SHARED/SentinelHub/SentinelHub.ipynb*.

5 Conclusions

At this point of the project (November 2019), the integration of a datacube component into CANDELA is not a priority. However, there is a need for some extended functionalities in terms of data access, that could be covered by a datacube solution or another kind of tool.

As described in section 3, CANDELA needs in terms of datacube concern time-series retrieval, efficient access to multi-source data and pre-processing of the data. According to section 2 conclusions, the solutions to take into account could be datacubes as a service or access to similar functionalities through the usage of standard OGC endpoints and APIs. Different DataCube solutions have been presented in section 4. Datacube solutions with full installation are not a realistic target for CANDELA platform in terms of resources, storage and integration costs.

Providing access to data as datacubes can be an additional access point to the data, which means that existing processing services do not necessarily need to be redesigned if they do not need datacube functionalities.

The datacube solution proposed by Sentinel-Hub as “Eurodatacube service” might be a good solution for future use cases in CANDELA. This service could be used by a user to generate a datacube specific to their use case. The user could then make the most of the datacube from CANDELA platform. This solution needs to be evaluated again in mid-2020.

The datacube as a service solution provided by MundiWebservices, and based on MEEO datacube, currently under integration, also needs to be evaluated once integrated in 2020.

The conclusion of this analysis is that we should wait for these two deeper evaluations before integrating a DataCube solution into CANDELA platform.

An alternative solution already available now as an additional access point offering a subset of the needed functionalities (time-series retrieval and pre-processing of data) can be the usage of SentinelHub on top of CreoDIAS datasets, associated to EoLearn library, for which a usage example notebook has been provided.

References

[1] F. Castel; 2018, “D1.9 DataCube Integration Report v1”; Deliverable of the CANDELA project.

[2] MEEO official website, http://www.meeo.it/wp/, retrieved 2018-10-08

[3] CEOS Open Data Cube national initiatives, https://www.opendatacube.org/ceos, retrieved 2018-10-08

[4] Open Data Cube official website, https://www.opendatacube.org/, retrieved 2018-10-08

[5] Open Data Cube GitHub, https://github.com/opendatacube/, 2019-10-24

[6] Open Data Cube Sandbox, https://www.opendatacube.org/sandbox, 2019-10-24

[7] Open Data Cube In a Box, https://www.opendatacube.org/ciab, 2019-10-24

[8] CreoDIAS documentation, data access interfaces, https://creodias.eu/data-access-interfaces, retrieved 2018-10-08

[9] CreoDIAS documentation, Data Related Services, https://creodias.eu/data-related-services, retrieved 2018-10-08

[10] Chatenoux, B & Richard, J.-P & Poussin, C & Guigoz, Yaniss & Giuliani, Gregory. (2019). Bringing Open Data Cube into Practice - Workshop Material. 10.13140/RG.2.2.17703.91044.

[11] Truckenbrodt, John & Freemantle, Terri & Williams, Chris & Jones, Tom & Small, David & Dubois, Clémence & Thiel, Christian & Rossi, Cristian & Syriou, Asimina & Giuliani, Gregory. (2019). Towards Sentinel-1 SAR Analysis-Ready Data: A Best Practices Assessment on Preparing Backscatter Data for the Cube. 4. 93. 10.3390/data4030093

[12] SentinelHub API documentation, https://docs.sentinel-hub.com/api/latest/#/API/, retrieved 2019-10-01

[13] SentinelHub OGC API documentation, https://sentinel-hub.com/develop/documentation/api/ogc retrieved 2019-10-01

[14] Sentinel Hub custom scripts repository, https://github.com/sentinel-hub/custom-scripts, retrieved 2019-10-01

[15] Sentinel Hub Bring your own data functionality, https://www.sentinel-hub.com/bring-your-own-data , retrieved 2019-10-01

[16] CreoDIAS SentinelHub deployment status, https://docs.sentinel-hub.com/api/latest/#/data/?id=creodias-sentinel-hub-deployment, 2019-10-01

[17] XCube documentation, https://xcube.readthedocs.io/, 2019-10-01

[18] EODataservice, http://www.eodataservice.org, 2019-11-14

[19] Big data cube website, http://www.bigdatacube.org/, 2019-11-25

[20] M. Albughdadi, 2019, D1.5 Use Case #1 Requirements v2 v1.0; Deliverable of the CANDELA project