CANDELA: A Cloud Platform for Copernicus Earth Observation Data Analytics

Jean-François Rolland, Fabien Castel, Anne Haugommard, Michelle Aubrun, Wei Yao, Corneliu Octavian Dumitru, Mihai Datcu, Michal Bylicki, Ba-Huy Tran, Nathalie Aussenac-Gilles, et al.

CANDELA: A CLOUD PLATFORM FOR COPERNICUS EARTH OBSERVATION DATA ANALYTICS

Jean-François Rolland, Fabien Castel, Anne Haugommard, Michelle Aubrun, Wei Yao, Corneliu Octavian Dumitru, Mihai Datcu, Michal Bylicki, Ba-Huy Tran, Nathalie Aussenac-Gillles, Catherine Comparot, Cassia Trojahn

AtoS, Thales Alenia Space, German Aerospace Center DLR, CloudFerro, IRIT, Université de Toulouse and CNRS

ABSTRACT

This article presents the achievements of the Candela project. This project aims to develop a platform and new algorithms for the handling, analysis and interpretation of earth observation data. The platform is hosted on the CREODIAS cloud ensuring the proximity of data and its processing. To ensure good performances the platform can scale up or down its computing resources. New algorithms based on machine learning methods for change detection and classification have been developed in the project. The results of these new algorithms are transformed into semantic data used to enrich earth observation products and provide new ways of exploitation. Finally, an end-to-end use of the platform is presented with a use case study of the impact of intense meteorological events on vineyards.

Index Terms— *Copernicus, cloud computing, big data, platform-as-a-Service, Earth observation, machine learning, semantic data, data analytics*

1. INTRODUCTION

All the collections of data provided by Copernicus satellites over the past recent years open the way for data scientists to develop hundreds of innovative use cases. This is true if they can manipulate this huge amount of complex data. The Copernicus Data and Information Access Systems (DIAS https://sentinel.esa.int/web/sentinel/sentinel-data-access), provided through an ESA program, are trying to solve the data storage and access issue, but data scientists need other tools to transform data into valuable information.

Candela, a research project funded by the European Commission through the H2020 program, intends to provide a generic platform to perform data analytics on Copernicus data with an online development environment providing built-in scalability management features, easy data access, edge-computing, and advanced geoprocessing tools.

Candela is implementing a hybrid Artificial Intelligence paradigm, integrating Machine learning, Deep Learning and semantic analysis in advanced data processing architecture for Big Data.

2. RELATED WORKS

The Candela platform technical foundations have been designed from a previous H2020 project, EO4wildlife [1], where the concept of data exploitation cloud platform was experimented in the domain of ecology and wildlife protection.

The data mining and data fusion tools presented in section 6.3 have been presented in more details in article [2] and [3].

Atos expertise on the development of cloud platforms for earth observation comes from the development of the Mundi DIAS (https://mundiwebservices.com/).

3. ONLINE DEVELOPMENT ENVIRONMENT

The development environment proposed to the user on the platform is based on Jupyter notebooks. This easy to use online development environment allows the user to access to the platform from anywhere.

The JupyterLab development environment for Candela is configured with a Python3 kernel and a set of standard libraries for data science and georeferenced data manipulation.

Dedicated libraries have been developed to ease the search and access to earth observation products hosted by CREODIAS and to facilitate the access to processing services. The Jupyter environment is the user entry point to the Candela platform: it is used to access earth observation products, to launch the processes developed in this project, to prototype new processing services.

When a Candela user launches the online development environment, a new instance of the Candela Jupyter environment is launched in a dedicated Kubernetes pod. The user benefits at the same time from an isolated instance and from the common environment configuration with geoscience libraries, easy access to EO data and processing services. The user has also access to a dedicated storage space, for upload or download.

4. SMART DATA ACCESS

Candela is hosted on CREODIAS, one of the DIAS platforms, and as such can access Copernicus data locally and perform data processing on the same infrastructure. Data is accessed through S3 protocol on the object storage buckets exposed by CREODIAS, ensuring a high level of performance.
In the context of exploiting earth observation data, users have to manipulate large set of data. Having the Candela platform hosted on CREODIAS allows an easy access to this data without the need to download or copy it.
Users have the possibility to run their own applications through a platform that uses the functionality of EO Finder – searching, ordering and processing through API or GUI (the user has access to this API from its notebook). Users can place an order by submitting a list of images that need processing alongside the required processing workflow.
Atos France has developed a Python library for searching, filtering and retrieving earth observation products. As the notebook is run on the CREODIAS platform the retrieving part is straightforward: the description of the product contains the path of the product in CREODIAS architecture. The product is directly accessible from the user's notebook.

5. FROM INFRASTRUCTURE TO PROCESSING-AS-A-SERVICE

Candela provides a geoprocessing environment for data scientists, on top of a cloud infrastructure with built-in scalability. Instead of an Infrastructure as a service, in which application developers must have an expertise on deployment configuration, Candela provides Processing as a Service, in which the application developer can let the platform orchestrator deploy the service in a transparent way. Candela components are shown in Figure 1, and described in the following sections.

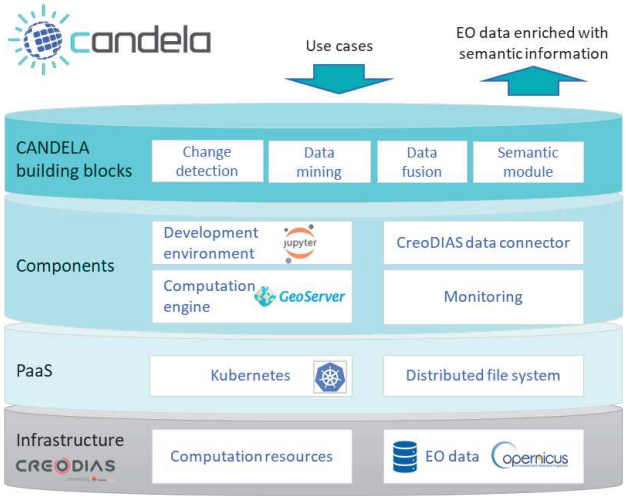

Figure 1 : Candela platform component overview

5.1. Infrastructure

The CREODIAS infrastructure is build based on opensource software. The main component is OpenStack (computing resources) and CEPH (highly scalable storage solution) with access to 14.5 PB of EO data. Ceph is a distributed object - based storage system capable of delivering storage services in the form of object repositories, block devices and storage volumes. Ceph is fully integrated with Openstack and can provide persistent block storage for VM-s, storage volume services, object buckets and filesystem services.
On top of these resource, CloudFerro provides functionalities to meet the specific requirements of EO Search Engine. Overall infrastructure consists of more than 14 000 virtual Cores (virtual CPUs), GPUs available on demand and 15 TB of RAM.
Candela platform is deployed on four virtual machines totalizing 20 virtual CPUS and 72GB of memory. A fifth machine is dedicated to resource consuming algorithms, it has 40 virtual CPUs, 128GB of RAM and is equipped with GPU. The GPU is used for machine learning algorithms based on the Tensorflow library.

5.2. Platform as a service

On top of these nodes, Docker and Kubernetes technologies are used respectively as containerization and orchestration solutions. The Docker technical foundation allows to standardize the deployment of heterogenous components in terms of technologies used, language or origin (off-the-shelf components, developed in research institute or in an industrial context). The architecture solution based on containers also allows to enrich the platform building blocks easily with a simplified and standardized integration process. The Kubernetes container orchestrator is the chosen solution to make the most benefit of the processing power provided by the cloud. It allows building a cluster of workers and easily pilot the resources to run dynamic processing tasks on the best available nodes, with built-in scalability.

5.3. Computation engine

The computation engine deployed on Candela platform is Geoserver. All processing services are packaged into Docker containers and made available as standard OGC WPS services. From Candela development environment, the user can access to the catalogue of available services and get the description and parameters of each service. The user can then request the execution of a service with required input parameters. When the Geoserver receives the WPS request, the Docker container corresponding to the processing service is launched on a dedicated Kubernetes pod, and the computation benefits from both the containerization offered by Docker, the smart EO data access offered by CREODIAS, the built-in scalability offered by the Platform-as-a-service and the user private storage area offered to Candela user for providing private input data or obtaining private computation

results. If the user works on a large area or a time series, a set of computations can be run in parallel.
Geoserver also allows to offer standard OGC access points to the products of the processing services (WMS, WFS, WCS).

5.4. Scalability management

Candela platform needs to adapt to an increasing number of parallel users, or to increasing resource consumption.
The real-time monitoring of available resources is used to automatically scale up or down the platform. When a threshold is reached, because lot of users are using the platform at the same time for example, a corrective action is triggered: a new node needs to be added to the Kubernetes cluster. The OpenStack API is used to spawn a new virtual machine. When the machine is ready, it is connected to the Kubernetes cluster and new Docker containers can be deployed on this node.
In order to monitor the state of the platform in real-time, a set of components based on influxdata have been deployed on the platform. These components retrieve and store metrics for each computing nodes, Kubernetes pod or service. The availability status of critical services is logged.
For each node, basic metrics including CPU and RAM usage, disk space available and uptime are monitored. For pods, CPU and RAM usage are stored as well as execution duration. This monitoring data is also used to inform Candela users of their resource consumption.

6. GEOPROCESSING TOOLS

Several geoprocessing building blocks have been integrated into Candela platform, including standard libraries like GDAL, geopandas and dedicated building blocks for change detection, semantic search, data mining and data fusion developed in the frame of the project.
Specific libraries have also been integrated to ease EO data access: CreoDIAS data access library developed for the project, but also sentinelhub, eolearn.

6.1. Change Detection

Thales Alenia Space France has implemented a change detection building block [4], which is based on an unsupervised approach to detect generic changes and which is designed to run on Sentinel-2 data time series. Using the fine temporal (5 days) and spatial (10 meters per pixel) resolution of these Earth Observation optical satellites, a lot of use cases can be carried out in many diverse sectors.
First, the analytic tool defines the different time series thanks to the image metadata (acquisition date and geographic scene coordinates). Then, it transforms each Sentinel-2 image of a time series into a more representative feature space thanks to an encoder model based on a neural network and calculates the distance metric between the Sentinel-2 images in this space to generate the change detection maps. Smart feature spaces allow to be more robust to the lighting and atmospheric condition differences during the data acquisitions that do not represent changes of interest.
This analytic tool has been developed in Python using TensorFlow/Keras and GDAL libraries and encapsulated in Docker containers in order to be properly integrated on Candela platform (section 5.2). One of the objectives of this analytic tool is to run on a big amount of data. Thus, the change detection building block requires fast data access, good storage capacity and intensive processing resources.

6.2. Semantic module

The Candela platform proposes a semantic search module on the top of a semantic knowledge base. This knowledge base has been constructed from an integration process that relies on a modular ontology for integrating EO and contextual data. The semantic search module is a means to promote the use of EO image data together with their contextual data (weather information, land cover, etc.). There are three categories of datasets of interest:

- Sentinel image metadata: these metadata are available together with the Sentinel images.
- Sentinel image related data: the datasets are provided by partners of the project and are extracted from Sentinel images through image processing tools.
- Contextual datasets: the datasets come from open data sources, such as: administrative units, weather measures or land cover.

The task requires first to perform semantic data integration, based on the spatial and temporal dimensions of these datasets. It involves the design of ontologies needed for integrating the datasets and then integrate the sources through ETL process based on the developed ontologies (data extraction, processing and transformation). We have then developed an interface that helps to query the knowledge base and to display the results in a user-friendly way. Each subtask is encapsulated by a Docker as shown in Figure 2.

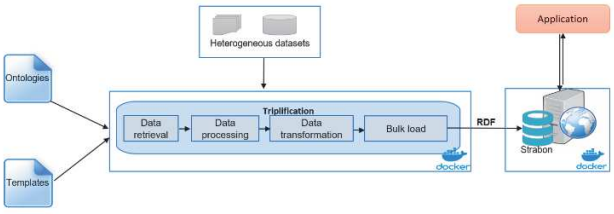

Figure 2 : System architecture of the semantic search tool.

6.3. Data mining and data fusion

The Data Mining module implements interactive and iterative functions for discovering useful information in collections of the Sentinel-1/Sentinel-2 products and generate semantic annotations. The Earth observation (EO) product is processed and the extracted image features and metadata are structured

in a Data Base (DB). The Data Mining is operated in 2 modes. The first is *EO Image Mining*; the users via an interactive GUI operate a machine learning (ML) module which is using the actionable information. The functions are search, browse, query for image patches of interest for the user and their semantic labelling. This is the actual EO image semantics adapted to the user conjecture and application. It is an Active Learning process achieving: 1) learn the targeted image category as accurately and as exhaustively as possible and 2) minimize the number of iterations in the relevance feedback loop. This is particularly important for EO since labeled data are very expensive and little available and is also increasing the trust in the learning process. The second function is *EO Data Mining*; this is performed via SQL search, queries, browse extracting the data analytical information from the DB. It uses image features, image semantics, and selected EO product metadata.

The Data Fusion module provides functions for land cover classification using jointly the complementary information from the Sentinel-1 and Sentinel-2. It enhances the classification accuracy, extends the use of the Sentinel-2 data in case of atmospheric effects or is enlarging the variety of land cover classes which can be recognized.

7. USE CASE

Once the change detection building block has been integrated on Candela, it was applied on a region of interest near Bordeaux in France to detect the crops affected by the frost that happened on 27th April 2017. After such meteorological events, winemaking farmers need to evaluate the level of damage that occurred in their vineyards in order to receive subsidies from the state insurances. Additionally, insurance companies must also estimate the damage levels to check the information provided. Currently, field visits are used to do the estimations, which is non-trivial and requires a huge budget and workforce.

Using a notebook from the JupyterLab user interface (section 3), it was easy to execute the following steps:

- The first one consists in accessing the Sentinel-2 data of interest via the CREODIAS data access functions (section 4). Thus, two optical Sentinel-2 Level-2 (atmospherically corrected) data products of the T30TYQ tile with low cloud cover on the region of interest have been uploaded on the user workspace. One acquired during 19th April 2017 (before the frost) and the other during 29th April 2017 (after the frost).
- The second step consist in preprocessing the Sentinel-2 data. In our case, all the bands at 10 and 20 meters of resolution have been extracted, resampled according to the blue band and concatenated into a single image for each data product. To do this, Thales Alenia Space France has also developed and encapsulated a preprocessing tool in a Docker container.
- The last step consists in running the change detection building block to generate the change detection map, whose pixel values represent the probability that a change has occurred between the acquisition dates. Several neural network models are available on Candela platform.

The entire pipeline has taken less than one hour to run with a consumption of 100% of one virtual CPU and 20% in GPU and has provided results very close to the ones of the operator.

8. CONCLUSION AND PERSPECTIVES

In this article we have presented the Candela platform which provides an easy-to-use development environment for geoscience. This environment is characterized by a smart data access to earth observation products, a scalable computing platform for performance processing and the publication of results as semantic information. The data access allows to search for earth observation products easily and to retrieve them efficiently. The scalable platform ensures high performances when dealing with processing of large amount of data. The creation of semantic data as a result of the processing algorithms enables the dissemination of earth observation data enriched with semantic information.

An illustration of the use of the platform has been presented in section 7. This use case shows how the platform can be used for detection of the effect of high intensity meteorological events.

The next steps around Candela project are:

- to industrialize this prototype on one or more DIAS platforms and propose it as a value-added service;
- to develop furthermore the integration of machine learning and semantic on earth observation data.

9. ACKNOWLEDGEMENT

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 776193.

10. REFERENCES

[1] Fabien Castel, Gianluca Correndo, Alan F. Rees, *"EO4Wildlife: A cloud platform to exploit satellite data for animal protection",* Earth Obs Phi-week, Frascatti, 2018.

[2] M. Datcu, C.O. Dumitru, G. Schwarz, F. Castel, and J. Lorenzo, *"Data Science Workflows for the CANDELA Project*", Big Data from Space, Munich, Germany, 19-21 February 2019, online.

[3] C.O. Dumitru, G. Schwarz, F. Castel, J. Lorenzo, M. Datcu, "*Artificial Intelligence Data Science Methodology for Earth Observation*," in Advanced Analytics and Artificial Intelligence Applications, InTech Publishing, 2019

[4] Aubrun, M., Troya-Galvis, A., Albughdadi, M., Hugues, R., Spigai, M.: Unsupervised Learning of Robust Representations for Change Detection on Sentinel-2 Earth Observation Images. ISESS 2020, IFIP AICT 554, pp. 1-6, 2020