CO3D MISSION DIGITAL SURFACE MODEL PRODUCTION PIPELINE

Olivier Melet *, David Youssefi, Céline L'Helguen, Julien Michel, Emmanuelle Sarrazin, Florie Languille, Laurent Lebègue

Centre National d'Etudes Spatiales (CNES), 18 avenue E. Belin, Toulouse cedex 9, France

Commission WG II/2

KEY WORDS: DSM production pipeline, Big Data, Cloud, Optical Image processing, large scale production, processing architecture

ABSTRACT

Earth Observation (EO) remote sensing missions are producing an increasing volume of data due to higher spatial and spectral resolutions, and higher frequency of acquisitions. Thus, in order to prepare the future of image processing pipelines, CNES has carried out Research & Development studies related to the use of Big Data and Cloud technologies for image processing chains made. Since mid-2019, CNES in partnership with Airbus Defense & Space, has started a new High Resolution Optical EO mission dedicated to very high resolution 3D observation called CO3D (“Constellation Optique 3D”). To achieve those objectives, a new image processing pipeline prototype is being developed taking in consideration the lessons learned from the previous studies. The paper will introduce this new image processing pipeline, the processing paradigms used to take advantage of big data technologies and the results of production benchmarks at a large scale. The on-going works to optimize the processing pipeline and Cloud cluster will be also discussed.

1. INTRODUCTION

New Earth Observation (EO) remote sensing missions are dramatically increasing the volume of produced data due to higher resolutions and higher frequency of acquisitions, whereas architectures and technical solutions for ground processing designed in the 2000’s are now reaching their limits in terms of computing scalability, data reading/writing speeds, etc. Since 2010, new architecture paradigms have emerged and have been applied to "big data", focusing on horizontal scalability. One of the main challenges for image processing centers is the ability to manage larger volumes of data at lower costs. Thus, in order to prepare the future of image processing pipelines, CNES has carried out Research & Development studies related to the use of Big Data and Cloud technologies for image processing chains made. In this context, a Proof Of Concept (POC) was developed in order to explore the ability of massive image production regarding the flexibility, scalability and economic benefits in an environment similar to an operational image processing center (Melet et al. 2019).

Since mid-2019, CNES in partnership with Airbus Defense & Space, has started a new High Resolution Optical EO mission dedicated to very high resolution 3D observation called CO3D (“Constellation Optique 3D”). CO3D mission is based on a constellation of 4 small optical satellites (Lebègue et al. 2020). One of the goals of the mission is to generate worldwide, accurate, low cost Digital Surface Model (DSM) within a short time window (launch in mid-2023, disposal of the DSM in 2025). To achieve those objectives, a new image processing pipeline prototype is being developed taking in consideration the lessons learned from the POC.

The paper will introduce this new image processing pipeline, the processing paradigms used to take advantage of big data technologies and the results of production benchmarks at a large scale. The on-going works to optimize the processing pipeline and Cloud cluster will be also discussed.

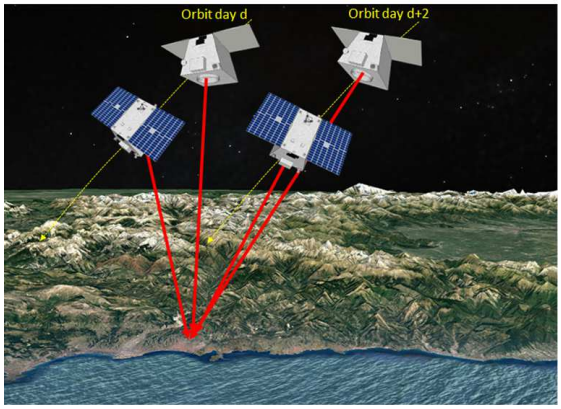

Figure 1. CO3D optical satellites constellation

2. DSM PRODUCTION PIPELINE PROTOTYPE USING SPARK AND DASK TECHNOLOGIES

* Corresponding author

2.1. Main challenges of the DSM production pipeline

The CO3D mission lifetime, expected to be at least 5 years, will be split in two main phases. The first one, called E2p, is the first year demonstration period dedicated to the assessment of the system performances over two areas: France and a 27 Mkm² covering the north of Africa and Middle-East countries.
During the E2p phase, 90 % of metropolitan France territory and 70 % of the world landmass has to be completed. Concerning the area covering north of Africa and Middle-East countries, 80 % has to be completed in less than 18 months.

To achieve those objectives, the DSM production pipeline aims at producing DSM from multi-stereo images and will have to process at less 40 000 images a day which is a massive production challenge.

The pipeline must be deployed on cloud technologies in order to benefit from the flexibility of use of its IT resources. Moreover, the architecture has to be agnostic from cloud providers and private IT infrastructures in order to be able to optimize operation costs.

To prepare the development of this pipeline and validate its performance goals, the DSM production pipeline prototype is being tested at large scale with numerous stereo and tri-stereo acquisitions from Pléiades satellites covering a large area in southeastern France simulating future CO3D acquisition strategy.

2.2. DSM algorithms pipeline principle

The production strategy of the future DSM production line is to generate independently DSM tiles that are geometrically consistent with each other especially at the tiles border.
Moreover, it aims also at providing an accurate absolute location and altimetry precision of the 3D data at a large scale.
This strategy takes into account, for each tile, the processing of massive bundle-adjustment centered on the tile, and covering margins on neighboring tiles.

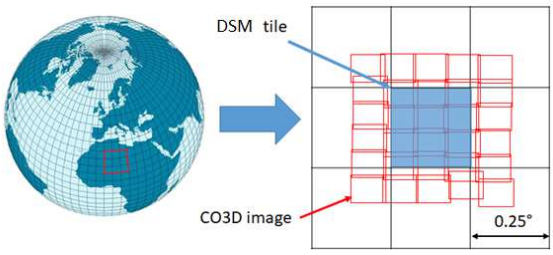

Figure 2. GLOBAL-DEM tiles production strategy (example)

So each instance of the DSM production pipeline has to process numerous stereo acquisitions to generate a DSM tile (all images that intersect the tile and margins).
The DSM production pipeline is designed as a sequence of different main algorithmic steps according to parallelization that can be applied:

1- Context generation step: computation of all the stereo pairs that have to be processed (one instance)
2- Absolute refining step: refining each product towards on absolute reference (one instance by acquisition)
3- Stereo refining step: refining geometry between the images of the same stereo pair (one instance by stereo pair)
4- Global refining step: refining geometry between all the images (one instance)
5- Perfect sensor step: processing each acquisition to generate single perfect sensor geometry product (one instance by acquisition)
6- Generate DSM: processing each stereo pair to generate a DSM (one instance by stereo pair)
7- Merging step: merging the DSM generated by stereo pair into a single stereo tile (one instance)

The DSM processing pipeline prototype does not implement all the steps yet but is already enough representative for the goals of this prototype. Indeed, the Context generation step, Generate DSM step and Merging step are implemented, so this steps are the one that need the more IT resources and synchronization too which are the main issues that have be addressed to implement such image processing pipeline with Big Data paradigms.

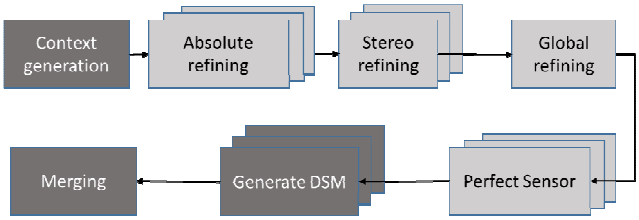

Figure 3. DSM production line (steps in dark grey are implemented in the processing pipeline prototype)

The DSM processing pipeline prototype is based on new and existing algorithm software steps combined together and adapted for Big Data technologies.

In order to develop the image processing pipeline, the Apache Spark Big Data Framework identified in our previous study has been selected as it combines maturity, robustness and ease of use. The software design for the 3D production pipeline was adapted to take full advantage of Spark technology and Cloud resources. Also, the optimization of image algorithm software in order to maximize the exchanges in memory is also a point which was deeply studied.

Note that the DSM production pipeline used in a previous POC was based on the open-source tool S2P (De Franchis et al. 2014). However, the software architecture of S2P was not well adapted for integration in Big Data framework as it was mainly developed to be used on High Performance Computing architecture or on single computer. CNES therefore decided to develop a new tool from scratch in order to implement the DSM production pipeline. Composed of well-known state of the art algorithms, this new tool called CARS (“Chaine Automatique de Restitution par Stéréoscopie”) is a stereovision pipeline for satellite images tailored for scalability and robustness. CARS is used to handle the Generate DSM step of the production line.

2.3. Generate DSM step

CARS (Youssefi et al., 2020) is a new stereo pipeline developed by CNES. It is designed for failsafe. It is composed of two parts in

order to improve its chances of success and the performances: a preparation part, the quicker one, that allows to get all the necessary information for the compute digital surface model part which is the most time consuming one.

The preparation part classically uses a sparse matching for the pre-estimation of the disparity range and to correct epipolar error. This sparse matching is performed in epipolar geometry using a low resolution DTM to improve computation speed.

The compute DSM part starts by resampling the stereo images using epipolar grid. Then, CARS uses PANDORA (Cournet et al. 2020), developed by CNES, to produce disparity map from these resample images, using a cost volume minimization with Semi-Global Matching (Hirschmuller H., 2008). The disparity map is converted to a list of homologous points to draw lines of sight with sensor model. Finally, the intersection of these lines gives a 3D points cloud that is projected in a regular terrain grid to produce the Digital Surface Model.

The previous subsection briefly sets out the steps of CARS: note that the algorithms are fully described in (Michel et al, 2020).

2.4. Parallelization paradigms

Different languages were considered to develop the processing chain: Java, Scala and Python (pySpark). The Scala language was a good candidate as the native language of Apache Spark, however PySpark was chosen because the CARS orchestration module is written in Python, so it is very easy to adapt the CARS orchestrator in consistence with the official branch of this tool.

Previous study (Melet et al. 2018) identified that the principle of shuffle in map/reduce paradigm is problematic for image processing which potentially requires contextual information and thus access to spatially related data. For this new pipeline, the choice has been done to avoid the partitioning of input data and process each instance of the processing pipeline steps by each Spark process.

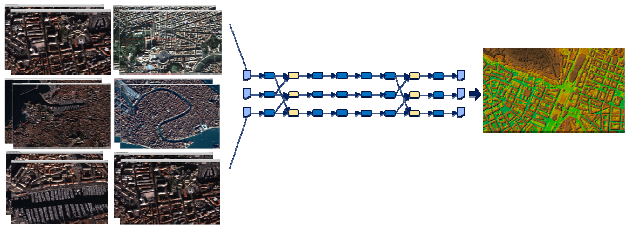

Figure 4: Paradigm for Spark parallelization.

However, each Spark job uses the native multi-threading parallelization capabilities of algorithmic software that compose the pipeline (i.e. stereo image processed by CARS in a Spark job is cut into sub-tiles which are themselves processed by CARS threads using Dask local distribution). In this way, each Spark job uses the entire CPU resources of a calculation node. This paradigm has the advantage to greatly simplify data management and drastically reduce exchanges between computing nodes.

3. BENCHMARKS

Benchmarking activity is underway to validate the massive processing capability of the CO3D pipeline, its operability in an environment representative of an operational center and the economic benefit of using an external Cloud.

The criteria for validating the mass production capacity cover performance aspects (performance behavior in production), costs (overall cost and production cost of a scene) and operability.
The benchmarks are performed on several configurations depending on the most critical parameters (volume of images, number of computation nodes, number of parallel productions, etc.).
The performance measures cover, for each configuration, unit performances (for the production of a product) and load performance (mass production capacity). Production costs are systematically evaluated during benchmarks.

Benchmarks are based on batches of 25 stereo and tri-stereo images from the High Resolution Pléiades satellites carried out on a public Cloud. Two kinds of benchmarks are carried out:

- Unit tests that are performed on isolated n-uplets (stereo or tri-stereo). This kind of benchmarks is used for functional validation and tuning parallelization processing according to the technical features of the computer cluster (number of nodes, node features, memory of each node, …),
- Mass production tests are carried out on the set of 25 Pléiades n-uplets (converted in more than 500 CO3D image products) covering a large area of south-eastern France (PACA set). The PACA set cover about 10 000 km². The covered area has the advantage of containing a wide variety of reliefs: coastline edges, plain, urban areas (even very urban with Monaco and Nice) and mountains (Alps). This diversity is very interesting to test the performances of the chain on different situations. Mass production tests are carried out to validate the ability of the prototype to process n-uplets at a large scale.

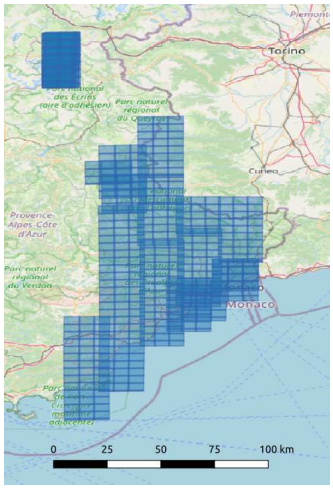

Figure 5: Footprints of CO3D image products covering PACA

Benchmarks are carried out on the public infrastructure of OVH Cloud provider (French Cloud Provider). The processing pipeline

is also tested on the CNES computer cluster in its "Big Data / Cloud" version in order to validate its functioning and performance on the private IT infrastructure of CNES.

3.1. Unit tests results

Unit tests allowed the first DSM to be achieved with the new processing pipeline. The results obtained are as good as expected from an image quality point of view.

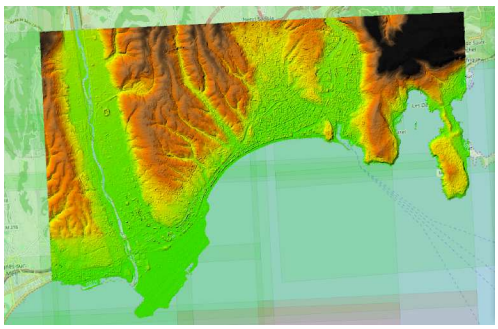

Figure 6: DSM of Nice generated in unit tests

The results in terms of performance in the Cloud environment are also consistent with the estimations made from the measurements made on the HPC (High Performance Computing) version of CARS (carried out on the CNES cluster). The CPU load required to produce the DSM in a cloud infrastructure is equivalent to that required to produce a DSM on the CNES HPC cluster with DASK cluster distribution. This validates that the use of Apache Spark technology on a Cloud does not bring overhead in terms of CPU load. In addition, these tests allowed us to identify optimization options.

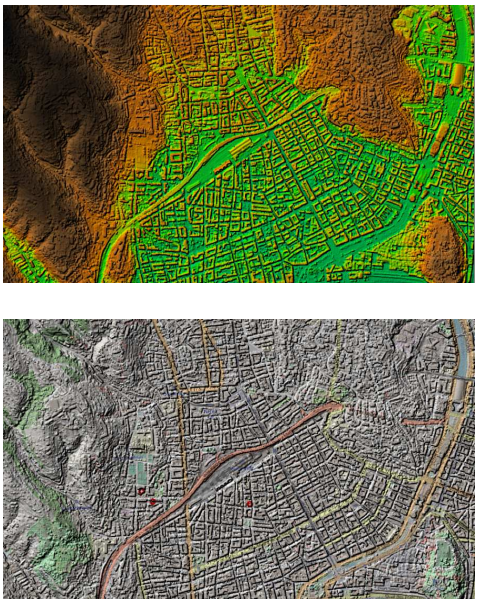

Figure 7: DSM overview of Nice (raw and mapped on Open Streep Map) generated in unit tests

Different unit benchmarks have also shown that the most efficient and cost-effective cloud cluster configurations do not consist of the most powerful compute nodes (ie, 16-cores compute nodes perform better than 32-cores compute nodes and the 4-cores compute nodes are more economical than the more powerful compute nodes even for a large computing volume).

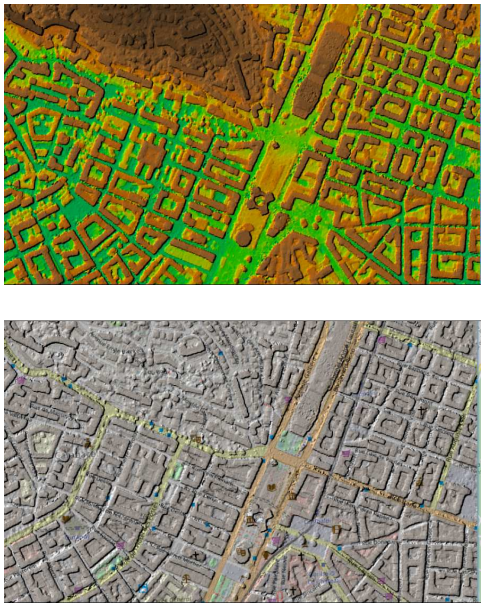

Figure 8: DSM of Nice-Acropolis Congress
(raw and mapped on Open Streep Map) generated in unit tests

3.2. Mass production results

Mass production tests have been started. Better performances are observed than with the previous POC in term of processing performances as CARS has a stable processing time regarding the differences in landscapes in the n-tuples (coastline, city, mountain). This topic was a performance issue in the previous POC.

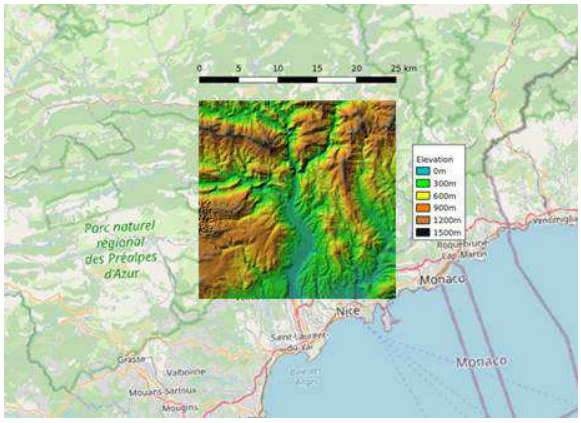

Figure 9: DSM tile generated in mass production tests

The Cloud cluster used in mass production tests is composed of 26 computer nodes with 16vCPU and 60GB each. Each Spark job downloads needed data for computation from a cloud object storage and uploads the result in the same object storage. Same computing resources are used for Spark job on the CNES HPC cluster to compare the performances with Cloud tests (excepted for storage which is a centralized GPFS storage).

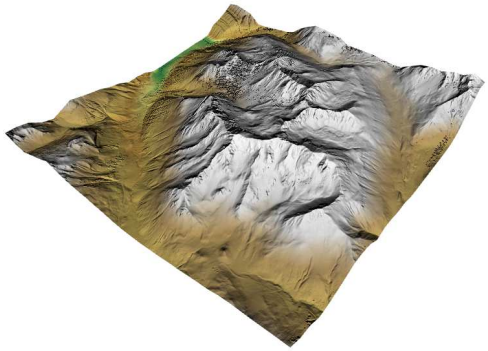

Figure 10: 3D View of the Combeynot Peaks

The first results confirm a significant improvement of the processing time performance. The total computation time to generate the DSM tiles of the entire PACA set is about 14 hours on the Cloud cluster.

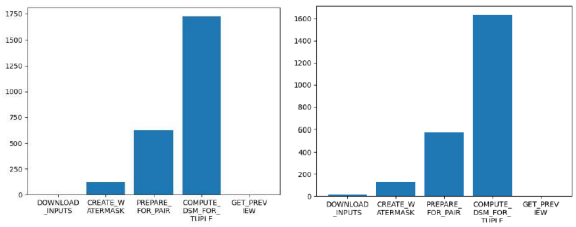

Figure 11: Average time by step in seconds:
CNES HPC (left) and cloud (right)

Figure 11 shows average time by step: downloading, watermask creation, prepare step (sparse matching), compute DSM step (dense matching) and preview creation. There is almost no difference in processing time between CNES HPC and Cloud. The only difference is the download step: the data is not downloaded on CNES HPC, furthermore, it can be observed that this duration is insignificant considering processing steps.

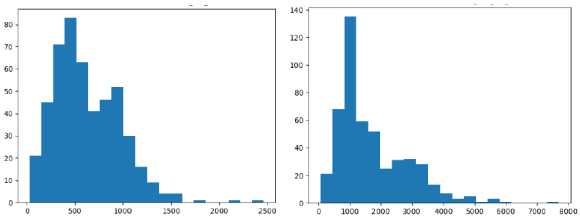

Figure 12: Histograms of the two most time consuming steps (in seconds): prepare (left), compute_dsm (right)

We also present here duration time of the two most time consuming steps as histogram (see Figure 12). There are two blobs in each histogram. This can be explained by the fact that stereo and tri-stereo are processed: the prepare step for a pair takes on average 500s on average (1000s for a triplet), and compute DSM step takes on average 1500s (3000s for a triplet).
Moreover, there is no calculation on sea area in the images according to water masks that explains why the computation time for several tuples is very short.
Finally, the time spread is due to the disparity range: knowing that a low resolution model is used for resample in epipolar geometry, it corresponds to the difference between this initial DTM and the final DSM.

4. WORK IN PROGRESS

Unit benchmarks and mass production have highlighted a number of possible optimizations. The optimization of the processing pipeline is in progress especially on two axes: optimization of the strategy of parallelization of the entire processing pipeline (example: reduce the number of synchronization point between the different steps) and the optimization of the parallelization of CARS (linked to the way CARS is implemented in the pipeline). New performance tests will be conducted at the end of these optimization works.

The choice of the Big Data framework was focused on Apache Spark for its maturity, it is now planned to test the processing chain with micro services technologies to fit with future processing center architecture.

The next main step is the development of the operational DSM production pipeline based on this processing pipeline prototype that will start at end 2020 according to the CO3D mission planning.

5. CONCLUSION

The CO3D image processing pipeline prototype is fully operational for the processing steps implemented. First benchmarks showed that the DSM generated are as good as expected from an image quality point of view. The parallelization paradigm implemented using Big Data technologies makes the processing pipeline enough scalable for a mass production. Moreover, the processing time are consistent with the specification of the CO3D mission.
This processing pipeline prototype will be the first release of the DSM production chain of the CO3D program.

REFERENCES

Cournet, M., Sarrazin, E., Dumas, L., Michel, J., Guinet, J., Youssefi, D., Defonte, V., Fardet, Q., 2020. Ground-truth generation and disparity estimation for optical satellite imagery. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

De Franchis, C., Meinhardt-Llopis, E., Michel, J., Morel, J-M., Facciolo, G., 2014: On Stereo-Rectification of Pushbroom Images, ICIP 2014

Hirschmuller, H., 2008, Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2), 328-341.

Lebègue, L., Cazala-Hourcade, E., Languille, F., Artigues, S., Melet, O., 2020: CO3D a worldwide one-meter accuracy DEM for 2025, ISPRS 2020, Nice

Melet, O., Youssefi, D., Michel, J., Cournet, M., Languille, F., Lebegue, L., Bouisson, C., Paccini, A., 2019: Large Scale Digital Surface Model production on Cloud using Big Data technologies for future EO mission, IGARSS 2019, Yokohama

Melet, O., Masse, A., Ott, Y., Lassalle, P., 2018: A new architecture paradigm for image processing pipeline applied to massive remote sensing data production, SPIE 2018, Berlin

Michel, J., Sarrazin, E., Youssefi, D., Cournet, M., Buffe, F., Delvit, J., Emilien, A., Bosman, J., Melet, O., L'Helguen, C., 2020. A new satellite imagery stereo pipeline designed for scalability, robustness and performance. ISPRS – International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

Youssefi, D., Michel, J., 2020: CARS: A photogrammetry pipeline using Dask graphs to construct a global 3D model, IGARSS 2020, Honolulu