BIG EARTH SCIENCES & THE NEW 'PLATFORM ECONOMY'

Hervé Caumont[1], *Fabrice Brito*[1] *and Enguerran Boissier*[2]

[1] Terradue Srl, Roma, Italy
[2] Terradue UK Ltd, Oxford Harwell, UK

ABSTRACT

We present here the Cloud Platform Operations conducted by Terradue for serving Earth Sciences practitioners in deploying scalable data processing applications, and for helping them make extensive use of Earth observations (EO) data. The associated business model is based on several enablers that pertain to the new born 'platform economy', and are applied to the earth sciences domain.
Performances analysis embedded in the daily operations, combined to a cost effective software development environment on the Cloud, are enablers for creating trustworthy partnerships on the Platform, with two-sided networks involving Open Science practices and Business-to-Business agreements to deliver innovative EO services.

***Index Terms*—** cloud platform, earth sciences, DevOps, big data analytics, computational science, hosted processing, operations support, partners ecosystem, open science.

1. INTRODUCTION

Earth observation sensors are continuously feeding the "data deluge", a phenomenon that is experienced by practitioners in numerous domains of business, government and science, and that is emphasizing cycles of data management where gigabytes of new information flow daily through the operations networks. For example, the ESA Sentinels missions, developed by the European Space Agency for the space component of the Copernicus programme, are starting to deliver unmatched amounts of Earth Observation data, and will drive the community to transition towards innovative solutions, that aim at better prepared people for the new digital era of computational science. This transition is impacting user services, but most of all, the way people and organizations work and collaborate. Considering more automated usages of Earth observation data, stakeholders in key applications domains are closely looking at means to keep up with this flood and at how the current platforms and technologies will evolve to serve Earth science research under such new constraints. Cloud operations offering partner agreement flexibility and deployment cost efficiency are part of the new 'platform economy' that is now reaching the Earth observations community.

2. THE GROWTH OF GLOBAL DATA ACCESS

In the last decade, we have helped to build successful EO e-Infrastructures and user services such as G-POD [1] that was precursor in bringing users and processing to the massive ESA data archives, and GENESI-DEC [2], extending data exploitation to e.g. seafloor, sea surface or global atmospheric observations. But in the coming years, continuously growing data archives will provide tens of terabytes of earth observation data, through tens of thousands registry entries representing tens of millions of data records, sitting on the Web and Cloud Data Centers. Partners from research and businesses now expect to find viable ways to make sense of it. This will require new business & science services to allow a range of matchmaking operations, with key aspects such as data savvy support teams, solid partner programs for data staging on the cloud and compute resources provisioning, easy to understand service level agreements for cloud-based operations, and technical teams able to grab innovations of the open web to enable fast data sharing and cross-community fertilization scenarios at the global scale.

We have evolved our processes to address many of these challenges, as the EO community needs more than ever to reach out with the way data is handled by the rest of the world. Evolutions in distributed Cloud storage and computing, lean software engineering, social web platforms, and web APIs are the ingredients to deliver to everyone on-demand data processing services, with reliable access to data sources from distributed, multi-tenant data archives, globally. Spatial and time dimensions of Earth observations datasets are fundamental, like addressed with standards such as the OGC® OpenSearch Geo & Time extensions, elaborated and implemented by Terradue through ESA funding [3] that can be largely adopted within multiple software development communities. But there is also a critical need for easy to understand semantics of all the sensor-generated data, for dimensions like the observed electromagnetic spectrum, the measured physical properties of the earth surface, or the sounding of the earth atmosphere dynamics. Standards elaboration and implementation are the way to go for such data access improvements. Keeping pace with standards maturation and adoption requires corporate processes that bring value to the customers: proven interoperability, effective lower costs, easy to deploy solutions. We had to improve the engineering processes and the design of our Cloud Platform to be natively ready for standards-driven data processing workflows that bring such value to users. It implied a new organizational process of continuous integration, ranging from the software development team (including platform partners) to the IT operations team in charge of the daily delivery of improved user experience. The result is a Cloud Platform model able to leverage a tremendous pool of distributed data repositories, built over the last decade through Spatial Data Infrastructure efforts [4], system of systems initiatives and EO ground segments harmonization [5], Grid Computing learning's [6], research e-Infrastructures developments [7], and more recently through the common use of Cloud storage technology and distributed services for EO data management and exploitation [8].

3. DEVELOPER CLOUD SANDBOXES

The cornerstone of Terradue's cloud platform is the Developer Cloud Sandboxes service, a virtual laboratory component where scalable processing chains are prepared and validated by data scientists. For each user or team of users, the service delivers custom Linux Virtual Machines hosted on Terradue's private Cloud. It consists in a Platform as a Service (PaaS) environment for the development of processing chains with the flexibility and scalability of the Hadoop framework. As a baseline approach, executable programs can be embedded as-is to become the processing units of a target application, but the PaaS environment also supports developers to build fully scalable applications in R or Python. The laboratory of our Cloud Platform complements this mechanism with data casting and data staging facilities natively designed for Cloud operations.

Numerous processor or model integration and exploitation activities have been performed in the last two years. They cover many Earth Sciences domains such as Geohazards (e.g. SBAS and ROI_PAC processors), environmental studies (e.g. MOHID, MyROMS models), or Climate Change (specific developments for Ice Sheets – Ice Velocity ECV productions, or for spatial and temporal variability in the onset of the growing season in Arctic regions).

The encapsulation of a processing chain within Hadoop follows the Hadoop Streaming programming framework, and benefits natively of a "Cluster Simulation" test mode (Hadoop pseudo-cluster). The processing chains can therefore be tested and validated with a cost effective model on Terradue's Cloud Platform. Once validated for data access and distributed processing, the Hadoop-enabled chain is natively ready for "scaling out" on nearly any commercial public Cloud provider, or through research agreements, on academic resources like the ones provided by EGI.eu.

The cost effectiveness of the Sandbox approach opens the door for more and new types of processor integration frameworks. Recent examples are the Argans' Toucan framework, a Open Source Python development for the ingestion of both satellite and in-situ measurements and their exploitation for inter-sensor calibration and analysis, and the Silk MapReduce framework, a Link Discovery Framework for the Web of Data, used to generate RDF links between datasets using a cluster of multiple machines to process Linked Data at the scale of the Web.
The cost effectiveness of the Cloud Platform also applies to the phase of scale-out operations, for which thousands of processing hours running on a large commercial Cloud cluster means a lead time, from order to delivery, in the range of a few days or hours, for a cost most often below the thousand euros threshold.

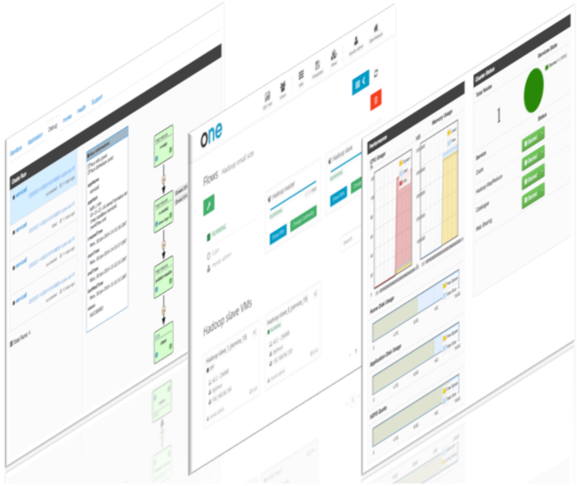

Figure 1: Terradue's Developer Cloud Sandboxes service

Overall, shortened time and costs from development to operations is a key aspect for enabling Platform operations and creating the synergies were partners join a Cloud Platform to innovate and bring new solutions to their customers or research partners. Practices for continuous integration, testing and monitoring are part of that equation.

4. PERFORMANCES ASSESSMENTS

Earth Sciences applications most often require large storage and compute-intensive processing. Performances assessment for distributed data processing jobs, in particular with respect to "big data issues", is key for scaling-out architectures. Terradue's Cloud Platform is continuously monitored and optimized for data staging and compute load operations. Amongst the typical 'Developments to Operations' (DevOps) activities that we perform are load testing of our Hadoop Clusters, and applications build with automated tests using Jenkins continuous integration server. Powerful deployment and monitoring tools like Puppet and Ganglia are part of our DevOps culture. We also regularly publish performance assessment results (both in terms of simplified tasks for the user and of computing resources exploited) through our participation in R&D projects and testbed initiatives, in which Terradue's Cloud Platform has been supporting user workflow improvements, system load testing and parallel computing stress tests.

Together with the scientific partner CNR-IREA, we investigated through the OGC® OWS-10 testbed [9] how large deployment and cluster processing applications operate in a multi-tenant Cloud environment. The testbed used Cloud resources provisioned on Amazon Web Services (AWS) for compute, Interoute for Cloud-based data access service, and Terradue's Cloud Platform for Applications bursting. CNR-IREA is an early adopter of the Developer Cloud Sandbox service, exploited to research performances improvements for its Small Baseline Subset (SBAS) processing technique. Hours of processing time were gained compared to previous local runs, and further parallelization improvements were identified in analyzing the testbed logs.

Another example is the successful data processing campaign for the ESA Climate Change Initiative (CCI), Ice Velocity Essential Climate Variable productions, where the partners DTU and S&T Corporation worked collaboratively on our Cloud Platform to develop, test intensively and significantly improve their workflow processing times and costs, to finally confidently use the Platform's API to burst their production chain on Amazon Web Services clusters, and monitor their performances according to the estimates.

5. EMPOWER USERS FOR CLOUD OPERATIONS

Once a processing chain is validated for distributed processing, it is also available and reusable, including shareable by a group of partners within our Community Hub service. This Hub is operated according to either an "Open Store" concept of operation (Open Science purposes) or according to a "Marketplace" concept of operation (Business to Business transactions), and is powered by a set of partner programs. For these improvements, we defined user workflows that integrate the APIs of powerful Cloud Services as various as e.g. OpenNebula, Github or Zenodo. Cloud operations at the Producer level are also leveraging the processing power of Cloud Computing providers, leading to unmatched costs in the realization of data processing or re-processing campaigns.

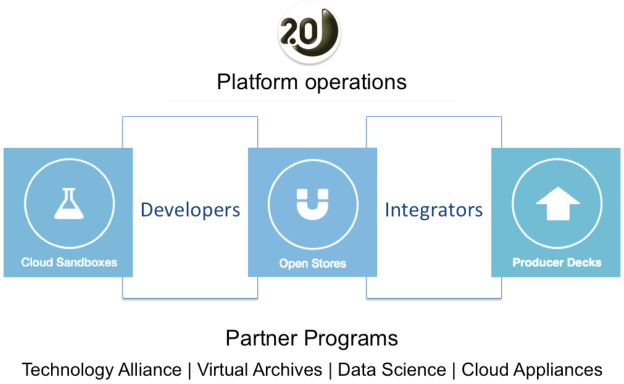

Figure 2: the value chain involves open source tools, open science practices, and community management in order to leverage synergies, from developers' environments to marketplaces for integrators and producers.

Started in spring 2014 for the GEO communities dealing with Geohazards [10], we are applying this approach for new Cloud provisioning capabilities of the ESA Geohazards thematic exploitation platform [11], in order to support users in deploying on-demand hosted processing. We are developing the tools that will accompany users in the selection and dimensioning of a production environment on a Cloud provider. The provisioning of a production cluster can be done either by users themselves from a community portal, or can be delegated to Terradue, when special cases need to be handled, via our capacity to operate for our partners as a Cloud broker.

6. NEW MODELS OF COLLABORATION

Our Cloud Platform partners are forming an ecosystem of producers and consumers sharing a common goal. As we need to build trust in our platform to bootstrap new value chains making use of Earth observations, our technology selection process was focused on supporting the collaborations that enact and sustain such ecosystems. An example is the "e-Collaboration for Earth observations" service deployed for ESA [12], a collaborative environment supporting sponsors to express a data challenge to solve, invite data scientists to participate, and provide them with the framework services to implement their solutions. Here the operations consist in running focused contests around such dynamically formed communities. The model is fully supported by our Developer Cloud Sandboxes service, and is an example of the many innovations that can be built from there. Having our Cloud Platform allowing operations based on a two-sided network model provides our partners with opportunities to gain from network effects on the Platform.

7. ENTER THE PLATFORM ECONOMY

We see a whole set of benefits for space agencies and industry in their current transition to cloud computing for the exploitation of Earth data. This is a move to the platform economy. Decades of scientific data gathered from Earth observation satellites proved useful to society. Upcoming are citizen services built by an ecosystem of value-adding providers. New jobs are needed for creating these innovative apps and we already engaged with several of them. The current 'digital era' market conditions require companies to simplify their operations and embrace continuous innovation. This allows us to get ourselves in position to provide low cost, scalable and reliable services.

As we receive a growing number of signs from research organizations in search of improved work processes, our answer is to embrace the new platform economy. Transitioning Terradue's technology stack into services for the Platform operations was a long endeavor. It first required us to deeply consider other corporations through their primary needs for value-added. Our business foundations are now made of core services that sustain a full stack of critical EO data management functions. This model is fueled by our corporate processes for strengthening the core functions having validated value for partners, and then finding the proper channels to develop network effects via user engagement on the platform. It builds on decentralized value-creation, matching well the Big Data exploitation challenges that are ahead for the earth sciences sector.

8. CONCLUSIONS

Terradue's Cloud Platform for big earth sciences is supporting community-driven research, collaboration and innovation. This approach is delivered at different scales: as an Open Source backbone supporting the ESA Geohazards Thematic Exploitation Platform (TEP) Pilot; as a turn-key Data Challenges Platform for the ESA E-Collaboration for Earth Observation (E-CEO) project; and as a commercial Cloud Platform through which Terradue operates several business partnerships in support of Earth Observation data exploitation goals.

The next logical step is to implement platform-based analytics to better understand the community trends, improve the user experience and maintain a focused set of user-approved services: what are the data needed for an area of interest at a given time span ? What are the data archive gaps compared to the platform usages ? Analytics can answer such questions, from assets such as volume of information generated from the platform user logs, platform searches and tickets issuing, website browsing patterns, open data available on the Web and within OpenSearch-enabled catalogs, scientific blogs, twitter feeds… There is a whole new landscape of innovations to explore.

9. REFERENCES

[1] G-POD, Grid-Processing on Demand for the ESA user community, http://bit.ly/1jLyL48

[2] The GENESI-DEC communities, http://www.genesi-dec.eu/images/img2.png

[3] Pedro Gonçalves (Ed.), "OGC OpenSearch Geo and Time extensions", *Open Geospatial Consortium*, OGC® Implementation Standard ref. OGC 10-032r8 version 1.0, 14th April 2014, http://www.opengeospatial.org/standards/opensearchgeo

[4] Francisco J. Lopez-Pellicer, Rubén Béjar, Aneta J. Florczyk, Pedro R. Muro-Medrano, F. Javier Zarazaga-Soria, A review of the implementation of OGC Web Services across Europe, Intl. Journal of Spatial Data Infrastructures Research, 2011, Vol.6, 168-186

[5] Gian Maria Pinna, Eberhard Mikusch, Manfred Bollner, Bernard Pruin, Earth Observation Payload Data Long Term Archiving The ESA's Multi-Mission Facility Infrastructure

[6] Niranjan Suri, Jeffrey M. Bradshaw, Marco M. Carvalho, Thomas B. Cowin, Maggie R. Breedy, Paul T. Groth, and Raul Saavedra, Agile Computing: Bridging the Gap between Grid Computing and Ad-hoc Peer-to-Peer Resource Sharing, Proceedings of the 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRIDí03)

[7] Jean-Luc Dorel, e-Infrastructures in Horizon 2020, Vision, approach, drivers, policy background, challenges, WP structure INFODAY, European Commission DG CNECT, 25th march 2014

[8] James Williams, NASA Nebula in Action: Cloud Computing Case Examples,
http://www.techrepublic.com/resource-library/whitepapers/nasa-nebula-in-action-cloud-computing-case-examples/

[9] Edric Keighan, Benjamin Pross, Hervé Caumont, OGC® Web Services Testbed 10 - Performance of OGC Services in the Cloud: the WMS, WMTS and WPS cases, 14-028r1, 26th June 2014,
http://www.opengeospatial.org/projects/initiatives/ows-10

[10] Philippe Bally (Ed.), Scientific and Technical Memorandum of The International Forum on Satellite EO and Geohazards, 21-23 May 2012, Santorini Greece,
http://esamultimedia.esa.int/docs/EarthObservation/Geohazards/esa-geo-hzrd-2012.pdf

[11] The Geohazards Thematic Exploitation Platform, https://geohazards-tep.eo.esa.int/

[12] E-collaboration for Earth observations – Data Challenges Platform, http://challenges.esa.int