HPC-Enabled Geoprocessing Services
Cases: EUXDAT, EOPEN, and CYBELE European Frameworks

José Miguel Montañana
High Performance Computing Center Stuttgart (HLRS) University of Stuttgart,
Nobelstraße 19, 70569
Stuttgart, Germany
Email: montanana@hlrs.de

Antonio Hervás
Inst. Matemática Multidisciplinar (IMM)
Universitat Politècnica de València
Camino de Vera s/n, 46020
Valencia, Spain
Email: ahervas@mat.upv.es

Dennis Hoppe
High Performance Computing Center Stuttgart (HLRS) University of Stuttgart,
Nobelstraße 19, 70569
Stuttgart, Germany
Email: hoppe@hlrs.de

***Abstract*—There are big challenges with a great impact on the economy that can be addressed with geoprocessing such as the improvement in agricultural productivity, design of transport networks, prediction of natural disasters, or the study of climate change. This paper introduces recent developments in three European projects in High-Performance Computing (HPC)-Enabled geoprocessing Services applied to agricultural issues. The main goals of the European Union (EU) projects EUXDAT (*extreme data analytics in sustainable development*), CYBELE (*fostering precision agriculture and livestock farming through secure access to large-scale HPC-enabled virtual industrial experimentation environment empowering scalable big data analytics*), and EOPEN (*open interoperable platform for unified access and analysis of Earth observation data*) are, in general, to enable the use of large HPC systems, as well as big data management and user-friendly access and visualization of the results. In addition, these three projects focus on the development of software frameworks, develop Artificial Intelligence (AI) algorithms, and fuse Earth-Observation data, such as Copernicus data, and non-Earth-Observation data, such as weather, environmental and social media information. Finally, some initial results are shown.**

Keywords–High-Performance Computing; Cloud Computing; Big Data; Agriculture; Land Monitoring; Machine Learning.

I. INTRODUCTION

Geoprocessing is a tool that allows addressing important and complex challenges. It is understood as the mathematical processing done by geographic Information Systems (GIS). During the last decades, the results of geoprocessing have greatly improved thanks to the exponential technological progress in computational power as well as exponential decreases in costs. The GIS systems consist essentially of three parts, as shown in Figure 1. The first part is that of data storage, the second part of computational processing, and the third part of visualization or access to results.

However, challenges such improve the efficiency of agricultural productivity require increasing by several orders of magnitude both the amount of data to be stored, as well as computational load. Therefore, the improvement in each of the aspects of geoprocessing presents itself as a new challenge.

The rest of this paper is organized as follows: Section II provides a summary of the contributions of this paper. Section III describes the implementation and Section IV the pilots and use cases. Sections V and VI describe the testing environment and comments on the experiments executed, respectively. Finally, conclusions are provided in Section VII.

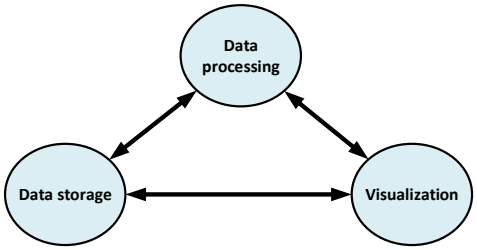

Figure 1. Fundamental components of geoprocessing systems.

II. CONTRIBUTIONS OF THIS PAPER

The main contribution of this paper is to present an innovative platform for solving multiple technological challenges. Such challenges are the integration of data from different origins in different formats, the definition of interfaces for geoprocessing applications, the capability to run such applications on computing resources like HPC and in the cloud. Additionally, the platform faces the challenges of huge data transfers as well as enforcing secure access and permission control for the data and the computation results.

These challenges are targeted by the EU projects EUXDAT [1], EOPEN [2] and CYBELE [3]. In particular, these projects focus on developing solutions for the collection of big-data from different sources, data transfer into large-scale High-Performance-Computing centers and Cloud Computing for processing, as well as visualization services and access to the results.

Here, we provide a summary of the goals of these three projects:

A. EUXDAT

EUXDAT proposes an e-Infrastructure for enabling Large Data Analytics-as-a-Service, which addresses the problems related to the current and future huge amount of heterogeneous data to be managed and processed within the agricultural domain. EUXDAT builds on existing mature components by providing an advanced frontend, where users will develop applications on top of an infrastructure based on HPC and Cloud. The frontend provides monitoring information, visualization, different distributed data analytic tools, enhanced data and processes catalogs. EUXDAT includes a large set of

 ISBN: 978-1-61208-762-7

data connectors such as Unmanned Aerial Vehicles (UAVs), Copernicus, and field sensors for scalable analytics. Figure 2 shows the type of field sensors deployed for the EUXDAT project in farming areas. These weather stations [4] allow measuring a wide range of measurements on remote areas, like rain gauge, air temperature, air humidity, global radiation, wind speed, soil temperature, and leaf wetness.

Figure 2. Field sensors deployed in the farming areas.

As for the brokering infrastructure, EUXDAT aims at optimizing data and resource usage. In addition to a mechanism for supporting data management linked to data quality evaluation, EUXDAT proposes a way to orchestrate the execution of tasks, identifying whether the best target is HPC or Cloud. It uses monitoring and profiling information for making decisions based on trade-offs related to cost, data constraints, efficiency, and resource availability. During the project, EUXDAT is in contact with scientific communities, in order to identify new trends and datasets, for guiding the evolution of the e-Infrastructure. The result of the project will be an integrated e-Infrastructure, which encourages end-users to create new applications for sustainable development.

EUXDAT demonstrates real agriculture scenarios, land monitoring and energy efficiency for sustainable development, as a way to support planning policies.

B. CYBELE

CYBELE is a European research project combining Agriculture, HPC, and Big Data. It involves 31 research institutes and enterprises across EU countries. It stands for: Fostering Precision Agriculture and Livestock Farming through Secure Access to Large-Scale HPC-Enabled Virtual Industrial Experimentation Environment Empowering Scalable Big Data Analytics.

CYBELE generates innovation and creates value in the domain of agri-food, and its verticals in the sub-domains of Precision Agriculture (PA) and Precision Livestock Farming (PLF) specifically, as demonstrated by the real-life industrial cases to be supported, empowering capacity building within the industrial and research community. The project aspires at demonstrating how the convergence of HPC, Big Data, Cloud Computing, and the Internet of Things (IoT) can revolutionize farming, reduce scarcity and increase food supply, bringing social, economic, and environmental benefits. It develops large scale HPC-enabled testbeds and delivers a distributed big data management architecture and a data management strategy.

C. EOPEN

The objective of EOPEN is to fuse Earth Observation (EO) data with multiple, heterogeneous and big data sources, to improve the monitoring capabilities of the future EO downstream sector. The Earth Observation data consists of the Copernicus and Sentinel data, while the non-EO data is weather, environmental and social media information.

The fusion is done at the semantic level, to provide reasoning mechanisms and interoperable solutions, through the semantic linking of information. Processing of large streams of data is based on open-source and scalable algorithms in change detection, event detection, data clustering, which are built on High-Performance Computing infrastructures.

Alongside this enhanced data fusion, a new, innovative architecture, overarching Joint Decision & Information Governance, is combined with the technical solution to assist with decision making and visual analytics. EOPEN is demonstrated through real use case scenarios in flood risk monitoring, food security, and climate change monitoring.

III. IMPLEMENTATION

The main goal is to develop a sustainable approach that facilitates access to data and geoprocessing applications and, at the same time, using the state-of-the-art on big-data management, as well as computation resources from Cloud platforms to HPC centers.

The target of the implementation is to provide an open-source system that can be used by commercial products, as well as by other projects to run after being finalized. The reason for including support for accounting and billing is to facilitate the code actually being used in the future since it is necessary to consider the costs of using large computer systems, as well as the cost of data acquisition from proprietary sources.

Therefore, each of the components has a clearly defined User-Interface. Figure 3 shows the main components of the infrastructure platform. The first one is the User-Interface (UI) Application Programming Interface (API). This supports the development of applications such as mobile devices or web-interfaces, without knowing the complexity of the other components.

The portal provides users with a list of available applications and the data catalog available for them. The users do not need to consider the complexity or format of the data, neither the different data sources, because it is encapsulated by the platform and the applications internally.

The data catalog collects data from different data sources. Some of these sources are free while others have an economic cost. Similarly, the catalog of applications can use free applications such as those developed in this project, or commercial applications. This is done to raise interest in using the platform by third parties that wish to commercialize applications or data.

Thus, once users select the task to perform, such as the prediction of temperature for a particular field on a particular date, they just wait for the result. Notice that the computation time is reduced by multiple orders of magnitude when using a large-scale HPC system.

The user request is submitted to the orchestrator, which is responsible for the transfer and execution of the applications on

 ISBN: 978-1-61208-762-7

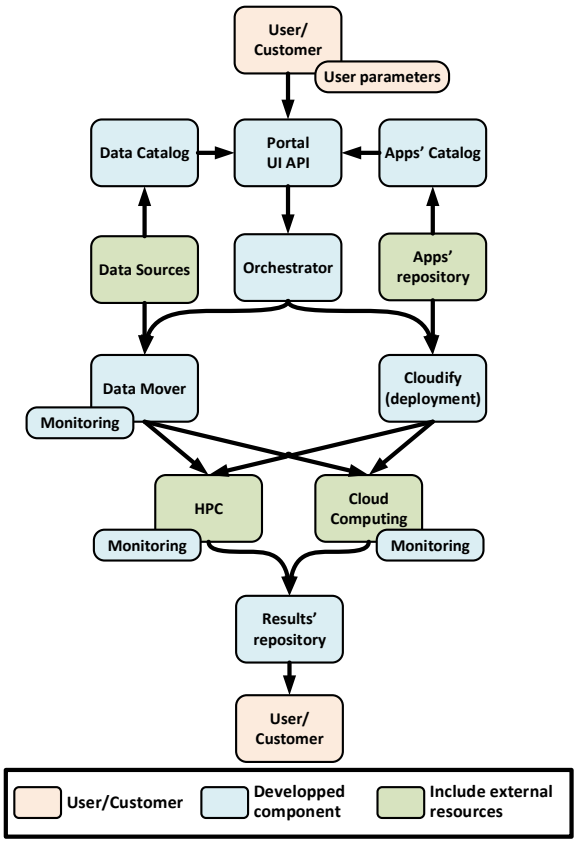

Figure 3. Infrastructure platform.

the computing resources. Basically, the orchestrator, based on the user request, selects the appropriate computation resource such as HPC or Cloud. Also, it queues a blueprint file [5] with the specifications of the user parameters, the input and output files, as well as the binary files to be transferred into the computation resources. A fragment of a blueprint is shown in Figure 4. Thus, the blueprint file allows the orchestrator which receives workload requests to delegate the required staging of input and output data to the Data-Mover component.

```
...
node_templates:
job:
        type: croupier.nodes.job
        properties:
        job_options:
                type: "SRUN"
                command: "olu coordinates.txt"
                nodes: 100
                max_time: "04:00:00"
        deployment:
                bootstrap: "bootstrap.sh"
                revert: "revert.sh"
                inputs:
                - "first_job"
                - {get_input: part_1}
...
```

Figure 4. Example of a fragment of a blueprint file.

The transfer of files is controlled by the Rucio server [6]. Rucio is the state of the art on large-scale data management. It is open-source and developed by ATLAS [7] for managing big-data at the European Organization for Nuclear Research (CERN); it is currently used to move more than 1 petabyte per day, and more than one million files per day [8][9].

Rucio allows defining three levels of architecture of file access in the DataMover. The lowest level is the storage of data in physical storage systems, These storage systems are referred to as Rucio-Storage-Elements (RSEs). The intermediate level corresponds to the logical access to the files. The physical location of the file is obtained from a database based on the logical identifier of the file, which consists of a text label. Thus, it is not necessary to provide the physical location of the requested files. At the highest level, datasets or sets of files are defined. Note that physical files can be included logically in different datasets without the need to be physically replicated. This makes it possible to avoid transmitting the same file multiple times to the same destination, for example, if an input file was copied to a certain computer system, it would not need to be transmitted again for any other application that needs it. In particular, we set up an RSE on the computation side. Rucio allows uploading data there and, at the same time enforces secured access and permission control for those files.

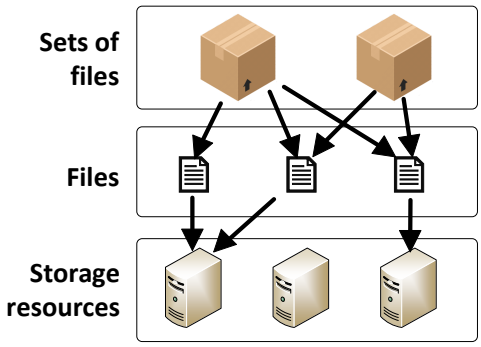

Figure 5. Three layers for data access.

In order to improve future application executions, the utilization metrics of the different resources are registered into the monitoring Prometheus server [10]. This will help with the decision on where to allocate the next requests depending on the user constraints, such as reducing computation time or reducing computation cost. Once the computation is completed, the results are moved into an accessible repository by the end-user, and the user is notified.

IV. USE CASES

The three projects presented above are focused on the development and test solutions for the agriculture field. Agriculture is a key aspect of economic and political stability. Because of its importance, governments are funding the development of solutions for those challenges based on data access systems, geoprocessing, and support for decision making.

The different uses cases will demonstrate the capacity of the HPC solutions proposed in the projects. They will be eventually open for end-users communities in the last phase of the projects, but currently, only consortium partners have access to the pilots' implementation.

The use cases cover a wide range of scenarios from detecting weather conditions, humidity or crop diseases up

 ISBN: 978-1-61208-762-7

to Precision Agriculture, Livestock Farming, and exploration. Here, we provide a brief description of some of them.

Pilot for Open-Land Monitoring and Sustainable Management Implementation: It targets on developing a deep learning algorithm which correlates input spectral data with ground truth, to be used for prediction of soil and crop status. To achieve it, multi-rotor UAV systems with a hyperspectral camera combined with earth-observation and meteorological data will be used for classification of crop status.

Pilot for Energy Efficiency Implementation: It focuses on developing analytics algorithms in order to obtain models of processes cost and profits to support energy-efficiency in agriculture.

Pilot for 3D Farming Implementation: It focuses on analytics models, mainly, on spatial analysis, for locating the highest productivity zones. It will provide 3D visualization for the obtained results, which especially help to understand the conditions of water, soil particles, and nutrients.

Organic Soya yield and protein-content prediction: There is an interest in the prediction of the soybean cultivation, due to the EU is strongly dependent on other continents for plant-based proteins. For that reason, this use case develops methods for predicting yield and protein-content maps based on crowdsourced data, satellite imagery and additional information, when available, such as electromagnetic soil scans, and other sensory data.

Climate-Smart Predictive Models for Viticulture: It targets the development of complex, highly-nonlinear models for vine and grape growth, which rely on a large number of variables that have been shown to affect the quality and quantity of the produced yields. The range of data includes earth observations, soil/elevation maps, genomics data, chemical analysis data, environmental and climatic data.

Climate services for organic fruit production: The goal is to help with the prevention of damage effects due to frost and hail. The solution under development focuses on providing risk probability mapping calculated based on models obtained by machine learning techniques. To do that, a wide range of data sources is used including but not limited to climate instability indices, digital terrain models, in-situ environmental and climatic data, and satellite images.

Optimizing computations for crop yield forecasting: Crop yield monitoring can be used as a tool for agricultural monitoring (e.g. early warning & anomaly detection), index-based insurance (index estimates) and farmer advisory services. Its goal is to compute a productivity estimation based on cropping systems model and a combination of different datasets, such as ingest crop, soil, historic weather data, weather forecasts data. However, it becomes a challenge to do that computation as the amount of available data keeps increasing as well as it is resolution.

V. EXPERIMENTS

In order to test the platform, we have been testing the deployment of the EUXDAT software platform in Hazelhen supercomputer at HLRS. Table I shows the details of the current supercomputer [11], and the new one to be installed on Q1 2020 [12].

The simultaneous use of large systems by a large number of users requires that each user execution request has to specify the number of computing nodes and software to be used. The requested executions will keep waiting in a queue until there will be free resources to fit the particular requirements of each one. Notice that the waiting time can be from a few minutes to a few days depending on the load on the supercomputer. Obviously, the cost to bill the user will be based only on the effective computation time. The cost never includes the queue waiting time. Therefore, the required computation time is an important aspect of using geoprocessing applications in real cases.

TABLE I. CHARACTERISTICS OF THE HLRS SUPERCOMPUTERS.

Name	Cray XC40(HazelHen)	HPE Apollo 9000(Hawk)
Number of cores	185,088	720,896
Storage capabilities	10 PB*	25 PB*
Interconnection network	Ariel	InfiniBand HDR (200Gbit/s)
Power consumption	3200 KW	Initially 3200 KW, but planned to be increased

*: 1PB = 1024 TB = 1,048,576 GB = 1,073,741,824 MB

The platform proposed in this paper uploads the required data in advance and not during the computation time. Because uploading data in advance can save significant computation time for geoprocessing applications. And therefore, it saves a significant cost. Notice that each of the use cases is composed of a series of geoprocessing applications. Thus, the computation time and data storage requirements of each use case will be the accumulation of the requirements of those applications. We can not provide the final requirements of the geoprocessing applications in our projects, because their implementation and parallelization are still not finalized. For that reason, we provide here only the preliminary requirements of the two applications which are commonly shared in almost all of the listed use cases. Table II shows the current data size to transfer and the required computation load of the applications for calculation of the land morphometry characteristics and weather predictions.

TABLE II. REQUIREMENTS OF TWO DIFFERENT APPLICATIONS.

Applications	Agroclimatic zones Frost date calculation	Morphometry characteristics calculation
Storage requirements	316 MB (ERA5-land Czech Rep)	25 GB (Austria Area) 1 TB (Full Europe)
Computation time in core-hours	70 (Czech Republic)	3000 (Full Europe)

The computation time is also an important aspect to take into account when there is a need to have the result at a certain time. For instance, a farmer needs to know if the next morning's temperature is below 27 degrees, because "at full bloom, the blossoms are usually killed by temperatures around 27 degrees" [13]. We consider that in these cases is preferred to use a computer center. Because in our experience, computing on the Cloud takes an immense amount of time when compared with the computation at the HLRS supercomputer.

The benefit of time-efficient computing on a supercomputer requires that the application be prepared to run in parallel. However, there was not needed a big effort to prepare the first parallelization of geoprocessing applications like the morphometry characteristics calculation. Because in this particular case, the load was easily distributed among computing nodes

 ISBN: 978-1-61208-762-7

just by dividing the computation load by geographical areas to be processed. Currently, considering that the applications are implemented with python, the developers' team is using Message Passing Interface (MPI) for Python [14], since it seems the most efficient way to obtain the best performance on these systems.

VI. Analysis of the experience

The proposed platform currently satisfies all use case requirements, and there were not deficiencies detected. The proposal simplifies the deployment and execution of geoprocessing tasks. It helps to do a more efficient deployment of data and computation, both in terms of time. The experience shows that the proposal seems to be the best cost-effectiveness for geoprocessing, especially for big projects, in particular for governmental large scale studies. In addition, it seems that the proposal is a cost-effective solution for companies interested in selling results of geoprocessing to small customers that do not have access to the data or the software to do the computation by themselves.

VII. Conclusions

In this paper, the infrastructure for HPC and Cloud computing of geoprocessing services has been described. The infrastructure is running and the use cases are under the last development stages in the last year of the projects EUXDAT and EOPEN, while CYBELE will keep running until the end of 2021.

The solutions being developed will greatly support improving farming performance and competitiveness, not only providing access to the tools, but also because the tools will run on most time-efficient computation resources. They will simplify the access for non-technical users, such as farmers who may access the services through their mobile phones. The developed platforms are expected to keep running after the end of the project. The partners in the projects are interested to use them for selling their products, such as datasets and weather forecasting services. For that reason, the EUXDAT consortium is looking for attracting service providers that sell the final products and services directly to the farmers. The consortium can potentially take the roles of software and cloud platform provider in order to support the ASPs.

Another aspect is that the developed platform for agriculture geoprocessing is also suitable for other purposes than agriculture, such as providing optimum paths through transportation networks, or predicting disasters like wildfire, flooding, or effects of a storm. Potential users can also include local authorities interested in Urban and Regional Planning and water management, or insurance companies interested in risk prevention or disaster resilience.

Acknowledgment

This work has been done within the projects *European e-infrastructure for extreme data analytics in sustainable development* (EUXDAT), *fostering precision agriculture and livestock farming through secure access to large-scale HPC-enabled virtual industrial experimentation environment empowering scalable big data analytic* (CYBELE), and *Open interoperable platform for unified access and analysis of Earth observation data* (EOPEN). See the projects' web pages [1][2][3] for further information. The research leading to these results has received funding from the European Unions Horizon 2020 Research and Innovation Programme, grant agreements n. [777549, 825355, 776019], respectively.

We wish to especially thank the partners who have collaborated with this paper providing their applications, as well as the estimation of their requirements. Those are Dimitrij Kozuch (P4ALL), Pavel Hájek (WRLS), and Dr. Karl G. Gutbrod (CEO at Meteoblue AG). We would also like to thank the work and collaboration of the rest of more than 31 research institutes and enterprises across EU countries partners in these projects. The list is too long to mention everyone.

References

[1] F. J. Nieto et al. (EUXDAT consortium), "EUXDAT *European e-Infrastructure for Extreme Data Analytics in Sustainable Development*," [online]: https://www.euxdat.eu/ [retrieved: Mar-2020].

[2] G. Vingione et al. (EOPEN consortium), "EOPEN *Open interoperable platform for unified access and analysis of earth observation data*," [online]: https://eopen-project.eu/ [retrieved: Mar-2020].

[3] S. Davy et al. (CYBELE consortium), "CYBELE *Fostering Precision Agriculture And Livestock Farming Through Secure Access To Large-Scale Hpc-Enabled Virtual Industrial Experimentation Environment Empowering Scalable Big Data Analytic*," [online]: https://www.cybele-project.eu/ [retrieved: Mar-2020].

[4] Pessl Instruments GmbH, "Stations and dataloggers," 2020, [online]: http://metos.at/micrometos-clima/ [retrieved: Mar-2020].

[5] J. Carnero, "Example of blue-print, available on the github project," 2019, [online]: https://github.com/hlrs-121991-germany/croupier/blob/master/croupier_plugin/tests/blueprints/blueprint_four.yaml [retrieved: Mar-2020].

[6] "Rucio: scientific data management," 2020, [online]: https://rucio.cern.ch/ [retrieved: Mar-2020].

[7] "Atlas experiment," [online]: https://atlas.cern/discover [retrieved: Mar-2020].

[8] C. Serfon et al., "Rucio, the next-generation Data Management system in ATLAS," Nuclear and Particle Physics Proceedings, vol. 273–275, 2019, p. 969–975, [retrieved: Mar-2020].

[9] R. Gardner, B. Riedel, and M. Lassnig, "Rucio concepts and principles," Dec. 2017, Presentation available at URL: https://indico.fnal.gov/event/15861/session/0/contribution/2/material/slides/0.pdf [retrieved: Mar-2020].

[10] "Prometheus sorftware: free software application for event monitoring and real-time alerting," 2020, [online]: https://prometheus.io/ [retrieved: Mar-2020].

[11] University of Stuttgart, HLRS, "Technical Description of the Hazelhen Supercomputer," 2020, [online]: https://www.hlrs.de/systems/cray-xc40-hazel-hen/ [retrieved: Mar-2020].

[12] ——, "Technical Description of the Hawk Supercomputer," 2020, [online]: https://www.hlrs.de/systems/hpe-apollo-9000-hawk/ [retrieved: Mar-2020].

[13] J. Muhollem, "Warm winter has put state's apple crop at risk, expert warns," 2017, Pennsylvania State University, [online]: https://phys.org/news/2017-03-winter-state-apple-crop-expert.html [retrieved: Mar-2020].

[14] L. Dalcin, "MPI for Python," 2019, [online]: https://mpi4py.readthedocs.io/en/stable/intro.html [retrieved: Mar-2020].

 ISBN: 978-1-61208-762-7