EUXDAT

www.EUXDAT.eu

European e-Infrastructure
for Extreme Data Analytics
in Sustainable Development

HPC-Enabled Geoprocessing Services

Cases: EUXDAT, EOPEN, and CYBELE European Frameworks

José Miguel Montañana Aliaga, HLRS, Germany
jmmontanana@gmail.com
Antonio Hervás, UPV, Spain
Dennis Hoppe, HLRS, Germany

GEOProcessing 2020
The Twelfth International Conference on Advanced
Geographic Information Systems, Applications, and Services

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777549

Dr. José Miguel Montañana Aliaga

Education:

PhD. in Computer Science (2008), Univèrsitat Politècnica de Valencia, Spain.

Thesis: Efficient Mechanisms to Provide Fault Tolerance in Interconnection Networks.

B.Sc. Electronic Engineering (2003), University of Valencia, Spain.

Final project: Study of Recursive Algorithms for Training Neural Networks.

B.Sc. Physics (1997), University of Valencia, Spain

Work Experience

2017-2020 Senior researcher at HLRS, Germany

2015-2016 Researcher at University of York, UK

2011-2015 Visiting Lecturer at UCM, Madrid, Spain

2009-2010 Researcher at ITACA, UPV, Spain

2008-2009 Researcher at National Institute of Informatics (NII), Japan

Research interest and experience:

IT Professional with over 10 years of experience. I can implement a wide range of effective solutions. Experienced with computation and big data management solutions in large supercomputers and computation in the Cloud, as well as in small embedded systems. I used the SoC Samsung ARM S3CEV40 board in my courses. I developed a tool for monitoring devices and applications, tested with the Raspberry-Pi 3 and Beagle boards.

5 years teaching CPU architecture and programming in assembler in the UCM University, provides me valuable knowledge for writing eficient and optimized code.

Research experience in networks inside and outside the chip, in Mathematical Optimization. And, experience in cryptography and blockchain valuable for Cybersecurity.

Programming: Multi-thread applications, C, C++, VHDL, X86 and ARM assembler, Python, Java, R, SQL and relational ElasticSearch databases, Docker, MPI, Rucio, OpenMP, restful APIs with NodeJS, Initiated on Machine Learning with CUDA.

Project management: Plannication and scheduling of development and integration tasks.

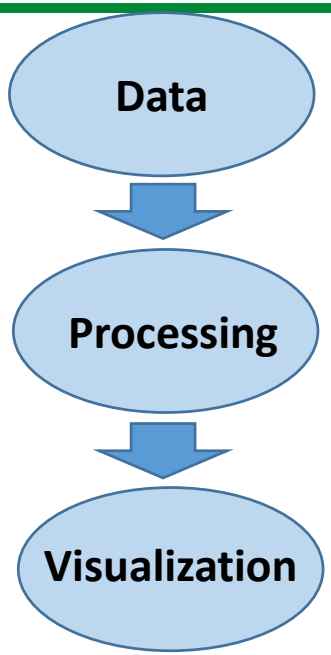
CloudWave

Outline

1. Introduction
2. Objectives
3. Implementation
4. Use cases
5. Experiments
6. Analysis of the experience
7. Conclusions

1. Introduction

- Geoprocessing is a set of tools for Geographic Information System (GIS). These tools consist essentially of three parts: data storage, **computational processing**, and visualization.



- Improving the efficiency of agricultural productivity requires reduce the **computational time** by several orders of magnitude.

- Reduction of execution time through the use of **parallelization of code** libraries. The main tools used for parallelization are **MPI and OpenMP**.

2. Objectives

The European research projects EUXDAT, EOPEN and CYBELE target combining Agriculture, HPC, and Big Data

An innovative platform for solving multiple technological **challenges**:

- **Integration of data** from different sources, in different formats.
- **Definition of interfaces** for geoprocessing applications.
- Capability to run such applications on computing resources like **HPC and Cloud.**
- **Big data transfer** into large-scale High-Performance-Computing centers and Cloud Computing for processing.
- **Enforcing secure access** and permission control for the data and the computation results.

2. Objectives

EUXDAT's objective is an integrated Infrastructure for:

- **Data management linked to data quality evaluation**
  - Aims at optimizing data and resource usage.
  - Includes a large set of data connectors such as Unmanned Aerial Vehicles (UAVs), Copernicus, and field sensors for scalable analytics.
  - Enabling Large Data Analytics-as-a-Service, solving how huge amount of heterogeneous data to be managed and processed within the agricultural domain.
- **An orchestrator for execution of tasks**
  - Identifies whether best target is HPC or Cloud.
  - Monitories information for making decisions based on trade-offs related to cost, data constraints, and resource availability.
- **An advanced frontend**
  - Where users will develop applications on top of an infrastructure based on HPC and Cloud.
  - It provides monitoring information, visualization, distributed data analytic tools, enhanced data and processes catalogs.

Field sensors deployed in farming areas.

2. Contributions of this paper

CYBELE involves 31 research institutes and enterprises across EU countries.

It targets:

- Demonstrate 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.
- Fostering **Precision Agriculture (PA)** and **Livestock Farming (PLF)** through Secure Access to Large-Scale HPC-Enabled Industrial Environment Empowering Scalable Big Data Analytics.

EOPEN targets:

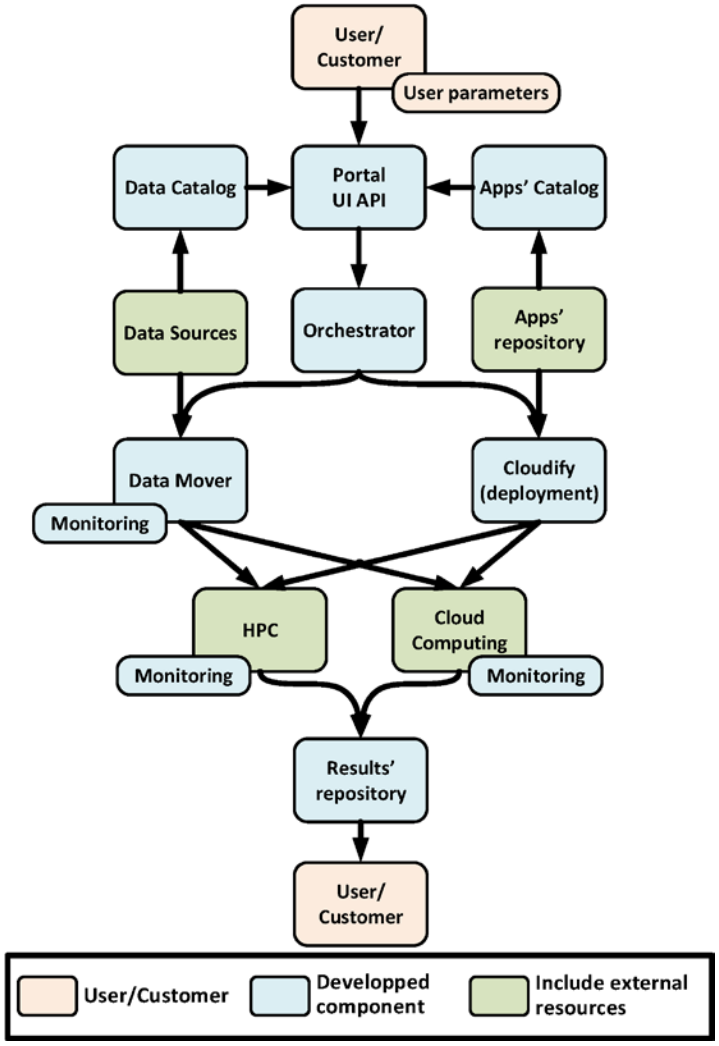
- **to fuse Earth Observation (EO) data with multiple, heterogeneous and big data sources,** to improve the monitoring capabilities.

The EO 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 solutions through the semantic linking of information.

All the EU projects demonstrates their achievements through real agriculture scenarios, land monitoring and energy efficiency for sustainable development.

3. Implementation



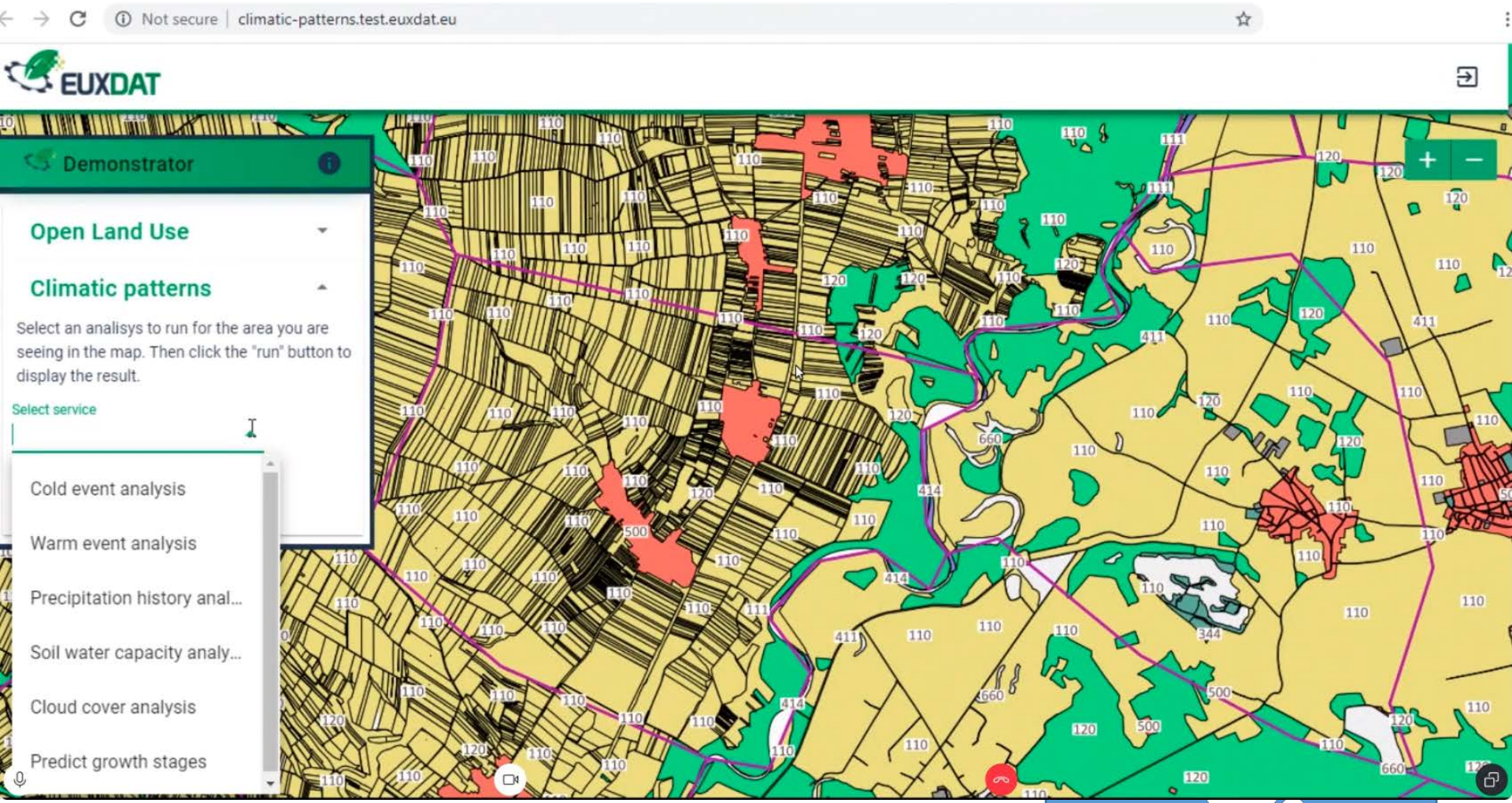
Infrastructure platform.

A sustainable **open-source infrastructure** platform have been developed.
It facilitates access to data and geoprocessing applications and using the state-of-the-art on big-data management, as well as computation resources from Cloud to HPC.

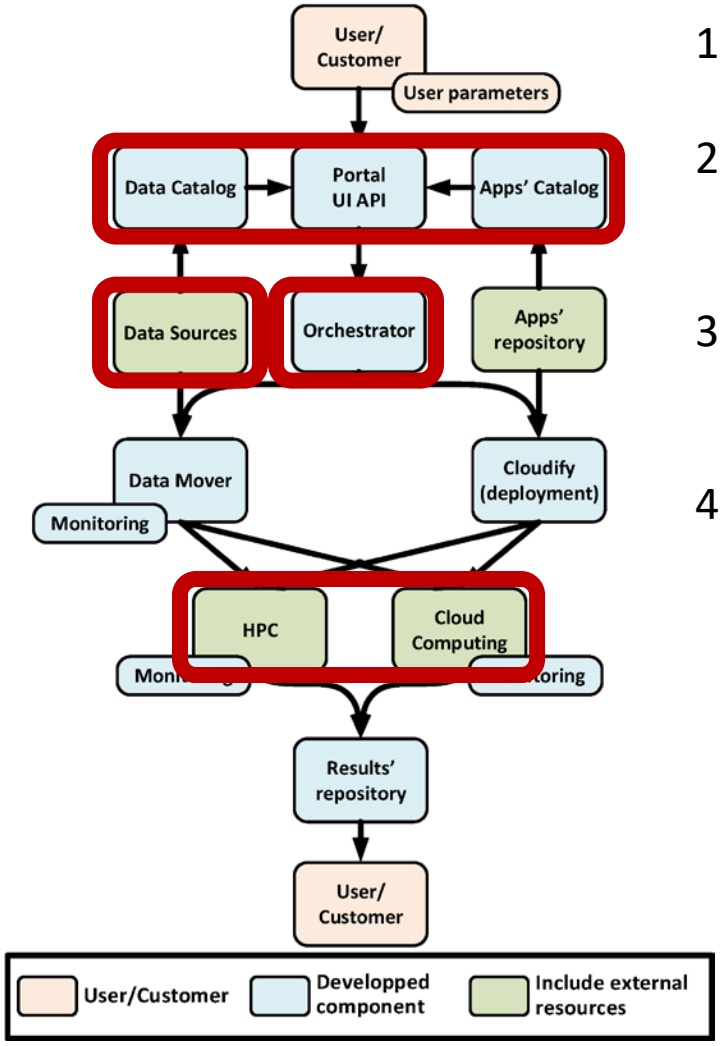
Each of the components has a clearly defined **User-Interface (UI) Application Programming Interface (API).**
This supports the development of mobile and web-interfaces, without knowing the complexity of the other components.

The platform includes support for **accounting and billing** to support commercial products.

3. Implementation



3. Implementation



Infrastructure platform.

1. The **portal** provides a list of available applications and the data catalog available for them.
2. The **data catalog** collects data from different data sources. The users do not need to consider the complexity, source, or format of the data.
3. The **orchestrator** is responsible for the transfer and execution of the applications on the computing resources.
4. The **computation time** is reduced by multiple orders of magnitude when using a HPC system.

It supports free and commercial data and applications, raising interest by third parties that wish to commercialize applications or data.

3. Implementation

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

Example of blueprint in yaml format

Each **blueprint** file contains the user parameters, the path of binary, input, and output files to be transferred into the computation resources.

3. Implementation

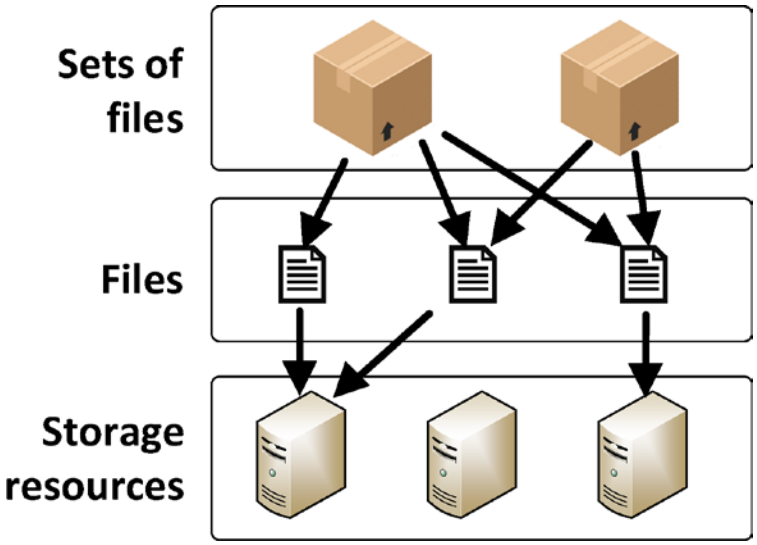
The **transfer of files** is controlled by Rucio.
It is open-source and developed by ATLAS for managing big-data at the European Organization for Nuclear Research (CERN); it is currently used to move more than 1 petabyte per day, and more than one million files per day.

The levels of file access:

- **Physical storage systems**, are referred to as Rucio-Storage-Elements (RSEs).
- **Logical access** to the files. It is not necessary to provide the physical location of the files.
- **Datasets or sets of files**. Files can be included logically in different datasets without the need to be physically replicated.

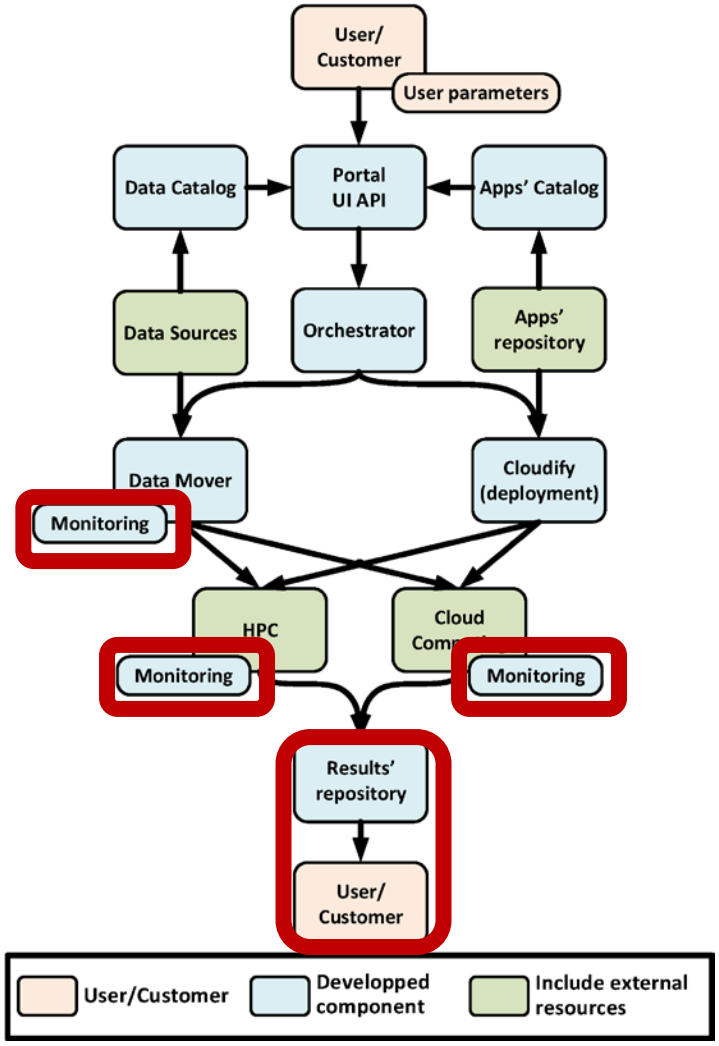
This allows avoiding transmitting the same file multiple times to the same destination. If a file was copied any other application that needs can use it.

It also enforces **secured access and permission control**.



Layers for data access

3. Implementation



Infrastructure platform.

In order to improve future application executions, the metrics of the different resources are registered into the monitoring Prometheus server.

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.

4. Use cases

The uses cases will demonstrate the capacity of the proposal solving the challenges based on data access systems, geoprocessing.
They will be eventually open for end-users communities in the last phase of the projects.
The use cases cover a wide range of scenarios from detecting weather conditions, humidity or crop diseases up to Precision Agriculture, Livestock Farming, and exploration.

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 affect the quality and quantity of the produced yields. The range of data includes earth observations, soil/elevation maps, genomics data, chemical analysis, environmental and climatic data.

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 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.

4. Use cases

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.

Organic Soya yield and protein-content prediction:
The EU is interested in the prediction of soybean cultivation, due to its strongly dependent on other continents for plant-based proteins. The methods for predicting yield and protein-content maps based on crowdsourced data, satellite imagery, electromagnetic soil scans, and other sensory 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 provides risk probability mapping calculated based on machine learning techniques, and 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 for agricultural monitoring (e.g. early warning & anomaly detection), index-based insurance, and farmer advisory services. Its goal is to estimate the productivity based on cropping systems model and datasets such ingest crop, soil, historic weather, and weather forecasts.

5. Experiments

	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

CHARACTERISTICS OF THE HLRS SUPERCOMPUTERS.

The conducted tests and development have been done with the **supercomputer** at HLRS.

The simultaneous use of large systems by a large number of users requires that the **requested executions will keep waiting in a queue** until there will be free resources to fit the particular requirements of each one.

The required computation time is an important aspect of using geoprocessing applications in real cases, because **the bill will be based on the computation time.**

5. Experiments

The platform uploads the required data in advance and not during the computation time. In order to save a significant cost.

The computation time is an important aspect to take into account when there is a need to have the result at a certain time.

To achieve time-efficient computing is required that the application be prepared to run in parallel. It does not needed a big effort to have a first parallelization of geoprocessing applications like the morphometry characteristics calculation. Because the load was easily distributed among computing nodes just by dividing the computation load by geographical areas to be processed.

In our experience, there is a trade-off on computing on the Cloud takes larger amount of time for a lower price when compared with the computation on HPC.

	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)

REQUIREMENTS OF TWO DIFFERENT APPLICATIONS.

6. 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.

7. Conclusions

Current status of the proposed platform:

- **It simplifies the access for non-technical users**, such as farmers who may access the services through their mobile phones.
- The solutions **support improving farming performance and competitiveness**, providing access to the tools, and using the most time-efficient computation resources.
- The developed platform for agriculture geoprocessing is also **suitable for other purposes**, 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 and disaster prevention.

Next steps:

The project partners will use the developed platforms **for selling their products,** such as datasets and weather forecasting services.

The consortium take the roles of software and cloud platform providers.

The consortium is looking for **service providers** that sell the final products and services directly to the farmers.

EUXDAT

www.EUXDAT.eu

European e-Infrastructure
for Extreme Data Analytics
in Sustainable Development

Thank you for your attention

Jose Miguel Montañana

jmmontanana@gmail.com

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777549