Computer Networks 168 (2020) 107035

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

CYBELE – Fostering precision agriculture & livestock farming through secure access to large-scale HPC enabled virtual industrial experimentation environments fostering scalable big data analytics

Konstantinos Perakis[a,*], Fenareti Lampathaki[b], Konstantinos Nikas[c], Yiannis Georgiou[d], Oskar Marko[e], Jarissa Maselyne[f]

[a] UBITECH, Thessalias 8 & Etolias, Chalandri 15231, Greece
[b] SUITE5, Archiepiskopou Makariou III 95B, Limassol 3020, Cyprus
[c] Institute of Communication & Computer Systems, 42 Patission Str., Athens 10682, Greece
[d] Ryax Technologies, 7 rue robert et Reynier, Saint-Fons 69190, France
[e] BIOSENSE Institute, Dr Zorana Dindica str. 1, Novi Sad 21000, Serbia
[f] EV ILVO, Burgemeester van Gansberghelaan 96, Merelbeke 9820, Belgium

ARTICLE INFO

Article history:
Received 30 April 2019
Revised 27 September 2019
Accepted 26 November 2019
Available online 30 November 2019

Keywords:
Precision agriculture
Precision livestock farming
High performance computing
Big data analytics

ABSTRACT

According to McKinsey & Company, about a third of food produced is lost or wasted every year, amounting to a $940 billion economic hit. Inefficiencies in planting, harvesting, water use, reduced animal contributions, as well as uncertainty about weather, pests, consumer demand and other intangibles contribute to the loss. Precision Agriculture (PA) and Precision Livestock Farming (PLF) come to assist in optimizing agricultural and livestock production and minimizing the wastes and costs aforementioned. PA is a technology-enabled, data-driven approach to farming management that observes, measures, and analyzes the needs of individual fields and crops. PLF is also a technology-enabled, data-driven approach to livestock production management, which exploits technology to quantitatively measure the behavior, health and performance of animals. Big data delivered by a plethora of data sources related to these domains, has a multitude of payoffs including precision monitoring of fertilizer and fungicide levels to optimize crop yields, risk mitigation that results from monitoring when temperature and humidity levels reach dangerous levels for crops, increasing livestock production while minimizing the environmental footprint of livestock farming, ensuring high levels of welfare and health for animals, and more. By adding analytics to these sensor and image data, opportunities also exist to further optimize PA and PLF by having continuous data on how a field or the livestock is responding to a protocol. For these domains, two main challenges exist: 1) to exploit this multitude of data facilitating dedicated improvements in performance, and 2) to make available advanced infrastructure so as to harness the power of this information in order to benefit from the new insights, practices and products, efficiently time-wise, lowering responsiveness down to seconds so as to cater for time-critical decisions. The current paper aims to introduce CYBELE, a platform aspiring to safeguard that the stakeholders involved in the agri-food value chain (research community, SMEs, entrepreneurs, etc.) have integrated, unmediated access to a vast amount of very large scale datasets of diverse types and coming from a variety of sources, and that they are capable of actually generating value and extracting insights out of these data, by providing secure and unmediated access to large-scale High Performance Computing (HPC) infrastructures supporting advanced data discovery, processing, combination and visualization services, solving computationally-intensive challenges modelled as mathematical algorithms requiring very high computing power and capability.

1. Introduction

According to McKinsey & Company, about a third of food produced is lost or wasted every year, amounting to a $940 bil-

* Corresponding author.
E-mail addresses: kperakis@ubitech.eu (K. Perakis), fenareti@suite.eu (F. Lampathaki), knikas@cslab.ece.ntua.gr (K. Nikas), yiannis.georgiou@ryax.org (Y. Georgiou), jarissa.maselyne@ilvo.vlaanderen.be (J. Maselyne).

https://doi.org/10.1016/j.comnet.2019.107035

lion economic hit, at the same time when 50% more and better food will be needed over the next 20–30 years [1]. Inefficiencies in planting, harvesting, water use, reduced animal contributions, as well as uncertainty about weather, pests, consumer demand and other intangibles contribute to the loss. Precision Agriculture (PA) and Precision Livestock Farming (PLF) come to assist in optimizing agricultural and livestock production and minimizing the wastes and costs aforementioned. PA is a technology-enabled, data-driven approach to farming management that observes, measures, and analyzes the needs of individual fields and crops. Sensors on fields and crops are starting to provide literally granular data points on soil conditions, as well as detailed info on wind, fertilizer requirements, water availability and pest infestations, which in addition to aerial images captured by unmanned aerial vehicles, or drones, which can patrol fields, can alert farmers to crop ripeness or potential problems and provide early warnings of deviations from expected growth rates or quality. Satellites can be at the service of PA, too, facilitating detection of relevant changes in field by using satellite imagery, identification of crop threats as nutrients deficiency or insect damage etc. PLF is also a technology-enabled, data-driven approach to livestock production management, which exploits technology to quantitatively measure the behavior, health and performance of animals. PLF deals with the management of livestock by continuous, automated, real-time monitoring of (re)production, health and welfare of livestock and the corresponding environmental impact. The data sources utilized in it include amongst others on-line sound, video observations, feeding intake, drinking behavior data, data from sensors on the animals, and data from milking robots.

Big data delivered by all the aforementioned sources has a multitude of payoffs in the domains of PA and PLF, that include precision monitoring of fertilizer and fungicide levels to optimize crop yields as well as risk mitigation that results from monitoring when temperature and humidity levels reach dangerous levels for crops, increasing livestock production while minimizing the environmental footprint of livestock farming and ensuring high levels of welfare and health for animals, providing new efficient decision-making tools for helping agricultural development and livestock farming optimization [2]. By adding analytics (e.g. parallel matrix multiplication, deep convolutional neural networks) to these sensor and image data, opportunities also exist to further optimize these domains by having continuous data on how a field or the livestock is responding to a protocol. By allowing farmers to apply tailored care and manage resources more effectively, production is boosted, economic efficiency is improved, and waste and environmental impact is minimized.

In the domains of PA and PLF, two main challenges are identified: 1) to exploit this multitude of data in order to evaluate and benchmark the developed algorithms, thus facilitating dedicated improvements in performance, and thus advancements for new applications [3], and 2) to make available advanced infrastructure so as to harness the power of this information in order to benefit from the new insights, practices and products, and to not only facilitate the execution of these algorithms, but to do so efficiently time-wise.

Thus, PA and PLF are currently being shaped by two major technological trends: big-data and advanced-analytics capabilities on the one hand, and robotics—aerial imagery, feeding and milking robots, intelligent sensors, sophisticated local weather forecasts—on the other. Towards this end, PA and PLF are driven by the need of 1) fusing a vast plethora of data sources which through various technological breakthroughs are currently available, 2) algorithms optimized for parallel execution that can exploit this multitude of data and harness their power, and 3) advanced infrastructural capabilities to handle the execution of these big data enabled algorithms efficiently time-wise, lowering responsiveness down to seconds so as to cater for time-critical decisions.

The current paper aims to introduce CYBELE, a platform aspiring to safeguard that the stakeholders involved in the agri-food value chain (research community, SMEs, entrepreneurs, etc.) have integrated, unmediated access to a vast amount of very large scale datasets of diverse types and coming from a variety of sources, and that they are capable of actually generating value and extracting insights out of these data, by providing secure and unmediated access to large-scale High Performance Computing (HPC) infrastructures supporting advanced data discovery, processing, combination and visualization services, solving computationally-intensive challenges modelled as mathematical algorithms requiring very high computing power and capability.

The structure of the paper is as follows: Section 1 introduced the scope of the paper and highlighted the rationale behind the proposed research. Section 2 analyzes the methodology followed, graphically illustrating and explaining the proposed architectural approach for detaching the design, development and execution of HPC empowered big data analysis processes. Section 3 outlines the anticipated results, to be achieved, evaluated and technically validated through a series of 9 demonstrators, 5 from the domain of PA, and 4 from the domain of PLF, which are also presented in the same section. Section 4 concludes the current paper.

2. Materials and methods

Few approaches have started being developed during the last few years mainly in the field of PA, but also targeting the domain of PLF. With regards to PLF, scientific approaches since 2010 focused mainly on the theoretical aspects of the challenge, explaining mainly how to implement PLF into practice [4], how IoT or even how audio-visual monitoring and analysis could also be exploited towards recognizing group behavioral patterns, identifying individual animals, detecting the occurrence of fertility, disease and discomfort, as well as to measure changes between individuals and groups of animals over time, including for example the approaches presented by Romeo S. [5], Terrasson G., Llaria A. et al. [6], Terrasson G., Villeneuve E. et al. [7], Andonovic I. et al. [8], while the first platforms claiming to support PLF have already been developed, including for example the Precision Livestock Farming system by Bosch [9]. PA on the other hand is a more mature domain, and various commercial solutions currently exist, offering advanced services including real time insights, yield monitoring, built-in accounting, field management and more, including amongst others Trimble [10], AgDNA [11], Sentera [12], AgroSense [13], Fasal [14], Agricolus [15], OneSoil [16], ProAgrica [17] and more. These have been developed over the years building upon the technological advancements mainly in the technological domains of IoT, Unmanned Aerial Vehicles and Satellite and Image Processing and analysis, and upon the scientific literature made available during the last decade, including (yet not limited) for example the approaches by Lin J-S. and Liu C-Z [18], van Henten E. J. et al. [19], Ge Y. et al. [20], Primicerio J. et al. [21], Zhang C. and Kovacs J. M. [22], Ye J. et al. [23], Srbinovska M. et al. [24], Ferrandez-Pastor F. J. et al. [25], Popovic T. et al. [26], and Sawant S.et. al. [27].

Nevertheless, CYBELE aspires to reach far beyond the currently offered services, and offer a holistic platform targeting both domains of PA and PLF, and offering services not only to farmers, but also to a plethora of stakeholders involved in the agri-food value chain (research community, SMEs, entrepreneurs, etc.). CYBELE also innovates in providing to these stakeholders integrated access to a vast amount of very large scale open and proprietary datasets of diverse types and coming from a variety of sources (including sensor data, environmental and climate historical data, satellite and aerial images). Last but not least CYBELE offers the technolog-

ical tools to the aforementioned stakeholders for experimentation with these data assets, and for the composition of novel services through the corresponding modelling environments, while abstracting the infrastructure required to support the development and the delivery of these resource demanding services.

CYBELE will capitalize on: 1) The computing capacity and efficiency potential delivered by HPC e-infrastructure and HPC-empowered services, enabling the processing of large amounts of heterogeneous data, and boosting modern scientific discovery, solving computationally-intensive challenges modelled as mathematical algorithms requiring real HPC architectures to achieve the required efficiency. 2) Technological advancements in big data technologies and related services, a) facilitating the aggregation of very large scale datasets (and/or their metadata) of diverse types (sensor data, satellite and aerial image data, etc.) from a multitude of distributed data sources; b) allowing the (semantic) alignment of the aggregated data and metadata to a common schema and data model; c) empowering the execution of advanced data analytics extracting hidden knowledge and insights and d) empowering the delivery of intuitive and adaptive data visualization services, providing to the CYBELE stakeholders with a more understandable and more easily evaluated interface to the results of the complex simulations. 3) The potential of cloud services for delivering simple and secure service provisioning, thus providing a bouquet of domain specific and generic services on top of the CYBELE virtual, industrial experimentation environment to both research organizations but mainly to industrial communities with focus on SMEs lacking access to HPC infrastructures and the competences necessary to fully exploit them, facilitating the elicitation of knowledge from big agri-food related data, addressing the issue of increasing responsiveness and empowering semi-automated / automation-assisted decision making when the prediction window is narrow.

The proposed architectural approach for detaching the design, development and execution of HPC empowered big data analysis processes is depicted in Fig. 1. This layered approach aims at ensuring interoperability among all involved components, putting emphasis on the way that pipelining of information (from data query, to simulation formulation, to analysis and to visualization) is supported, safeguarding smooth interoperation of the aspired services. In order to achieve this, CYBELE consortium aims at designing and delivering 1) on the one hand standardized interfaces putting emphasis on exposing reusable functional primitives for the HPC and Big Data frameworks integrated within CYBELE, so that pipeline reusability is supported, minimizing the need of reprograming core integration engines, and 2) on the other hand normative schemes and common data models featuring common semantics, annotating the information to be exchanged between components, thus ensuring information harmonization and enabling seamless communication amongst the various heterogeneous components.

Big, heterogeneous data (to be also made available to the CYBELE industrial test beds) are made available through HPC powered repositories. Prior to the check-in and storage, as demonstrated on the middle left part of the architecture, the data pass through a data quality check pipeline in order to address the data veracity and timeliness challenges that are typically associated with big data. From the moment that data are collected in CYBELE, quality checks are performed to discover inconsistencies and other anomalies in the data and eventually ensure their integrity and completeness, to be followed by a number of steps associated with data cleansing procedure (ranging from data filtering and cleaning to normalization). Finally, the CYBELE Data Provenance Service puts in place the necessary mechanisms to record all relevant information that influence the "incoming" data of interest. In CYBELE, the Data Provenance Service is intrinsically linked to the Data Policy and Assets Brokerage Engine that facilitates the data sharing and trading features that will be offered by the platform to link data providers and data consumers. Checked-In data will also be semantically annotated and harmonized so as to promote data interoperability and reuse. Since the data will come from a multitude of physically distributed data sources, a common semantic data model will be created which will be used to semantically describe and annotate the data. The model will be used as a common language to annotate data and the messages exchanged between the components, so as to facilitate the pipelining and enable the seamless communication of the various heterogeneous components. The clean and semantically uplifted (open and proprietary data) will be then made available for querying, analysis and visualization.

In order to enable simulation execution, a dedicated Experiment Composition Environment will be designed and delivered, as demonstrated on the upper right part of the architecture. The Experiment Composition Environment aims to facilitate the detaching of the design, development and execution of the big data analysis processes, supporting embedded scientific computing and reproducible research. The analysis process will be based on the selection of an analysis template, where each analysis template will represent a specific algorithm with the associated software and execution endpoint, and will provide to the user the flexibility to adjust the relevant configuration parameters, including input parameters for the algorithm, execution parameters, parameters associated with networking and computing resources constraints, as well as output parameters. The Experiment Composition Environment will support the design and implementation of data analysis workflows, consisted of a series of data analysis processes, interconnected among each other in terms of input/output data streams/objects. Upon the execution of an analysis template, the outcome could constitute the input for another analysis template. The output of the analysis template execution will be a session object that contains on memory all output values.

In order for the Experiment Composition Environment to run, it requires input datasets (training and/or evaluation datasets). For this reason, an Advanced Query Builder will be designed and developed which will provide users an intuitive environment to define and execute queries on data available in the CYBELE platform.

After input datasets have been selected and the workflows have been designed, the advanced analytics on top of big data need to be executed. In the case of CYBELE, advanced analytics algorithms will be provided to the stakeholders with the ability to visually explore the different kinds of data, while discovering and addressing new patterns. Machine learning and predictive modelling techniques will be updated in order to be able to manage the predictive life cycle of data preparation, exploration and analysis, for achieving better deployment and monitoring. However, the execution of advanced analytics on top of big, diverse data, raises the need for strong computational power and increased computing memory so as to be able to not only extract insights, but to do this within a reasonable time frame. The execution of the demonstration cases selected raises the need for several HPC attributes including Storage Intensity, Computing Intensity, Memory Intensity, Throughput Intensity and Short Turnaround Time.

Towards this end, analytics workflows designed, will be sent for execution on well-known HPC and Big Data frameworks, which will run on HPC resources abstracted to the user, as demonstrated on the lower part of the architecture. CYBELE relies heavily on HPC infrastructure, to provide the compute power required to advance models and methods and will take a two-fold approach; a) focus on tuning the HPC software stack to allow for efficient execution of Big Data processing frameworks on top of HPC resources, b) target the HPC resource management layer and its interface with Big Data processing frameworks and orchestration engines, thus bridging the gap between the HPC and Big Data worlds. In addition to that, the efficient execution of analytics workflows and the man-

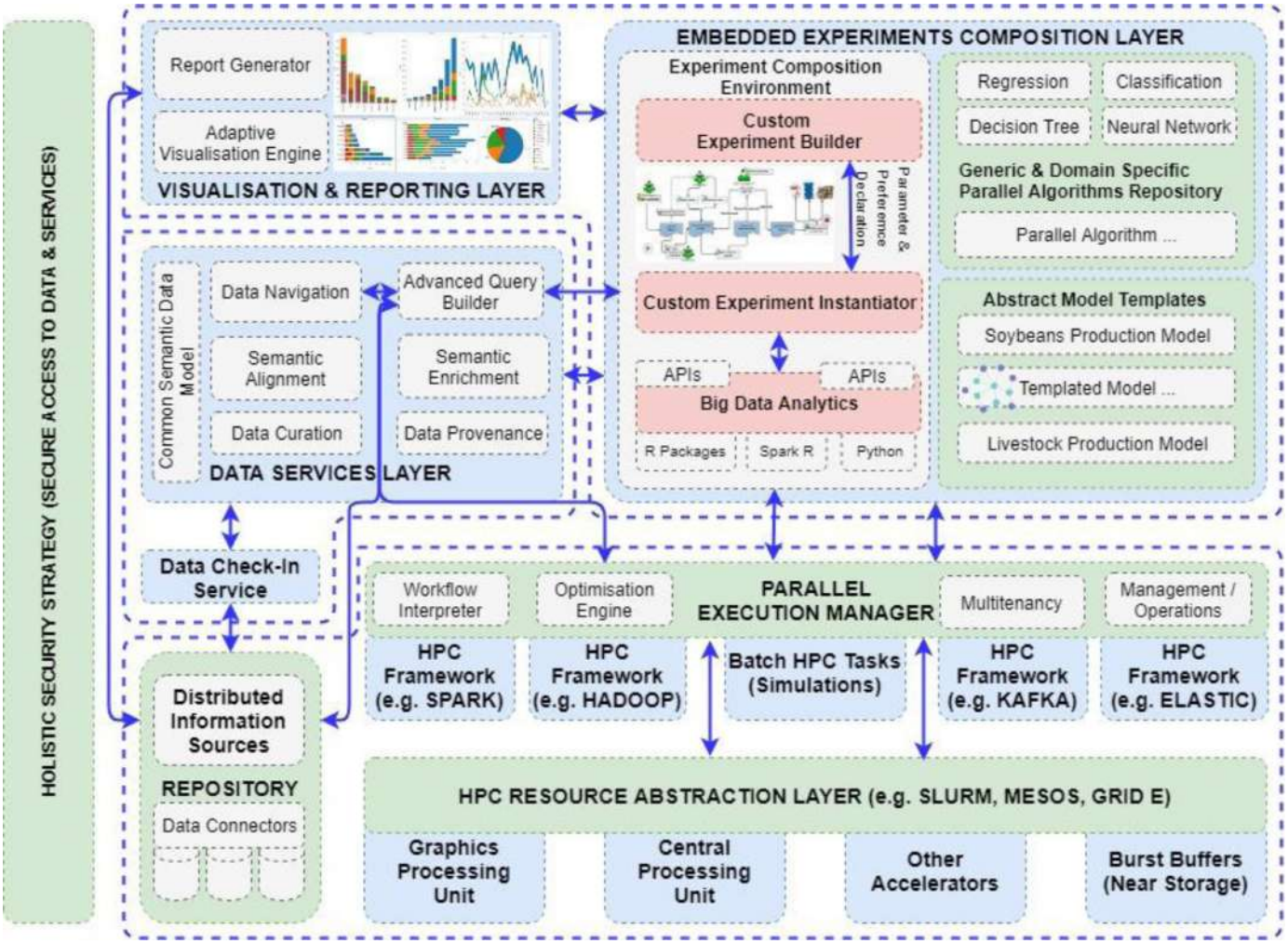

Fig. 1. CYBELE conceptual architecture.

agement of resources comes with multiple challenges. In the context of CYBELE, optimization of workflows on different frameworks will be examined for performance, along with tasks scheduling on resources to meet resource constraints/performance constraints/time constraints. Depending on the workflow and the frameworks in use, CYBELE consortium will explore the abilities of the framework's scheduler to optimize the workflow and guarantee performance, as well as the deployment through an orchestrator such as Kubernetes, which can optimize the performance of a workflow over multiple frameworks with its highly modular architecture and ability to run multiple and customized schedulers.

The analysis results will then need to be visualized and presented to the end users in a precise, coherent, and user-friendly way. Towards this end, as demonstrated on the upper left part of the architecture, adaptive visualization / user interfaces will be delivered, improving the way in which information is presented, for example by converting the raw data into interactive visualizations and dosing the information available at any given time. Synchronized views can be included within a customizable dashboard that allows adaption of the visualization and the analysis of dynamic data to an appropriate level according to the profile and knowledge of the end user.

The following sub-sections go into more depth for some of the aforementioned core components of the CYBELE conceptual architecture, to better familiarize the user with the technical approach to be followed in order to yield the aspired results.

2.1. (HPC & Big Data storage & computing) resource abstraction

CYBELE relies on HPC infrastructure, to provide the compute power required to advance models and methods for PA. In this direction, CYBELE will take a two-fold approach; first, we will focus on tuning the HPC software stack to allow for efficient execution of Big Data processing frameworks on top of HPC resources. Second, we will target the HPC resource management layer and its interface with Big Data processing frameworks and orchestration engines. This approach is necessary to bridge the gap between the worlds of HPC and Big Data, which reflects to the system architecture, the software stacks and the resource abstraction, and stems from having their roots in different classes of problems. We should note that HPC infrastructure used in CYBELE includes both typical HPC resources, as well as partitions of resources built for data analytics and other data-intensive applications. However, within CYBELE, we envision to explore the full scaling and computing capabilities of all available HPC resources.

A core issue to be addressed within CYBELE is efficient I/O and data movement, to handle the volume and velocity of data. On HPC infrastructure, storage resources and compute resources are decoupled, and compute resources use a parallel file system, such as LUSTRE [28] or General Parallel File System (GPFS) [29], and Portable Operating System Interface (POSIX) [30] semantics to perform I/O operations, or a network file system (NFS) and name-space sharing. As a result, storage is optimized for infre-

quent I/O of large batches of data, typically to instantiate simulations or store results or checkpoints. On the other hand, traditional Big Data infrastructure relies on low-latency and reliable local storage, with a distributed file system, such as the Hadoop Distributed File System (HDFS) [31], handling the batches of data and guaranteeing fault-tolerance. The mismatch between the two architectures can cause severe performance degradation when executing big data frameworks over HPC resources, as big data frameworks are designed with the principle of data being close to the processing unit. We will work to alleviate this mismatch using local file systems mapped by the shared file system and will explore software layers that eliminate overheads related to POSIX-like semantics [32]. We will additionally enable efficient mechanisms for Big Data frameworks to use near storage, such as Solid State Drives (SSDs), to enable scaling up to multiple nodes avoiding the latencies involved from frequently accessing the shared file system [33]. Besides techniques for efficient "vertical" data movement, within CYBELE, we will procure Remote Direct Memory Access [34] (RDMA)-enabled data analytics frameworks [35], to enable efficient "horizontal" data movement, i.e. efficient communication between nodes, since modern HPC infrastructure that allows for such deployments. Container technologies (e.g. Docker) will allow for seamless adaptation of the software stack to the needs of the application, without affecting the HPC applications executing on the HPC systems.

CYBELE will also work on efficiently coupling the resource management for Big Data frameworks over HPC resources. HPC resource managers (SLURM [36], Torque [37], etc.) interact with the user, who submits a "job", only once, to grant the requested resource allocation that remains unmodified until completion of the application. In contrast, Big Data resources managers (MESOS [38], YARN [39], Kubernetes [40], etc.) embrace the concept of elastic resource allocation: they either allocate resources to meet application demands, or negotiate with applications for resources. Also, Big Data frameworks implement their own, standalone schedulers to manage their tasks on a set of resources. To enable the execution of Big Data frameworks over HPC resources, within CYBELE, we will enable the execution of Big Data resource managers over resource allocations provided by the traditional resource managers of HPC systems. To achieve this, we will deploy advanced orchestration engines, like Kubernetes, alongside traditional HPC resource managers. We will extend the HPC resource managers with the ability to interface with orchestration engines, to allow for the creation and deployment of YARN, Hadoop, Spark, and other clusters on resource allocations granted by HPC resource managers. This approach has manifold benefits: i) the operation of HPC systems is not disrupted, ii) any Big Data framework can run over HPC resources, iii) any Big Data scheduler can be deployed, iv) the HPC resource manager dedicated a set of isolated resources (CPUs, memory, GPUs, FPGAs, etc.) to the executing frameworks, v) the HPC resource manager can implement topology-aware and I/O aware policies, vi) the Big Data resource manager/orchestrator (as a second level of scheduling) can implement resource elasticity within the given resources and locality-aware policies.

2.2. Parallel execution management

Within CYBELE, the data preparation and modeling tasks, which fall under the category of Big Data processing and analytics, need to run efficiently in its ecosystem, for a variety of data and model sources. The CYBELE architecture will support a variety of frameworks, for batch processing, e.g. Hadoop, for streaming processing, e.g. Apache Storm and Apache Kafka, for iterative processing, e.g. Apache Mahout and Apache Spark. Several of these frameworks come with components that will serve the needs of CYBELE demonstrators for machine learning (e.g. MLBase), deep learning (e.g. DeepLearning4j), predictive analytics (e.g. Oxdata H2O), graph analytics (e.g. GraphX). In addition, many frameworks offer support for transparent execution of workloads on accelerators, namely GPUs. The efficient execution of analytics workflows and the management of resources comes with multiple challenges. In the context of CYBELE, we will work on the following aspects: i) how to optimize workflows on different frameworks for performance, ii) how to schedule tasks on resources to meet resource constraints/performance constraints/ time constraints.

For workflows with time constraints, e.g. stream processing, CYBELE will rely on dedicated data analytics platforms. Depending on the workflow and the frameworks in use, we will explore the abilities of the framework's scheduler to optimize the workflow and guarantee performance, as well as deployment through an orchestrator such as Kubernetes, which can optimize the performance of a workflow over multiple frameworks with its highly modular architecture and ability to run multiple and customized schedulers. The challenges are different for compute-intensive workflows, as are deep learning tasks on high-dimensional data: such workloads require high computational power and large scale. To this direction, CYBELE will work with best practices and optimized software, coming from the world of HPC, to enhance performance through scaling and efficient utilization of all available compute resources (CPUs, GPUs, FPGAs, etc.). For non-critical data processing tasks with low demands on resources, we will explore the potential of using YARN or we will directly use the built-in schedulers of frameworks.

An important issue CYBELE will need to address is workflow scheduling across different system partitions and/or resource allocations, as provided by the HPC resource manager. For example, a workflow may include some data streaming tasks and some data analytics tasks, which, for reasons of performance, should run on different partitions of a system and be composed of different frameworks. Within CYBELE, we will develop those necessary components, interfaces and tools for HPC resource managers and the orchestrator, that will allow communication of events and data between frameworks through the resource manager, for the execution of a workload.

Efficient workflow execution heavily relies on selecting the right framework and the best set of resources for the various tasks. A compute or data operator within a workflow can be expressed on top of various engines, and some may be more efficient than others. Also, for the efficient execution of the workflow, each task needs to be assigned the right amount of resources. To optimize workflow execution on HPC systems, we will explore the usage of IReS, an open-source, multi-criteria meta-scheduler of tasks onto analytics engines and resources, developed for the EU-funded project ASAP. In CYBELE, we will focus on integrating IReS with the orchestrators and resource management tools used within the project.

2.3. Workflow & experiment composition & instantiation

The Experiment Composition Environment aims to facilitate the detaching of the design, development and execution of the big data analysis processes (including descriptive, predictive, classification, clustering, and prescriptive analytics) aspired and expected by the demonstrators for the execution of their scenarios, supporting embedded scientific computing and reproducible research.

The basic requirement driving the implementation of the Experiment Composition Environment regards the need to provide simple and homogeneous access to a variety of algorithm packages, without the necessity of having deep knowledge of the execution requirements of each algorithm. This requirement will be supported through the provision of access to set of registered algorithms along with the provision of user-friendly interfaces for the

specification of the main execution parameters. In order to support the flexibility on realizing part or an overall analysis workflow, the design and implementation of the Experiment Composition Environment will be based on a microservices-based architecture enabling in a modular way both registering new algorithms at the analysis engine, as well as invoking the execution of the analytic processes.

The Experiment Composition Environment will expose a set of a set of open APIs facilitating the access to the analytics mechanisms and the design, development and execution of analytics, empowering software developers to develop analysis scripts without restrictions in the programming language (e.g. R, Python, Java) and data scientists to design analytic workflows, consisting of sets of processes and related input and output parameters in a user friendly and intuitive way. The environment will also support the application of different execution modes, e.g. sequential or parallel execution of R algorithms for smaller data streams or in clustering mode by the HPC and big data frameworks (e.g. Spark, Hadoop, Kafka etc.) provided through the parallel execution management layer over the managed HPC clusters abstracted through the Resource Abstraction Layer.

The analysis process will be based on the selection of an analysis template after the input datasets (training and/or evaluation datasets) have been queried and provided to the Experiment Composition Environment by the Advanced Query Builder. Each analysis template will represent a specific algorithm with the associated software and execution endpoint, and will provide to the user the flexibility to adjust the relevant configuration parameters, including input parameters for the algorithm along with their description and their default value, execution parameters that denote whether an analysis should be realized in a manual or automated way, as well as the periodicity factor for the latter case, parameters associated with networking and computing resources constraints, as well as output parameters along with their type (text, image, data, html).

The proposed approach also supports the design and implementation of data analysis workflows, consisted of a series of data analysis processes, interconnected among each other in terms of input/output data streams/objects. Upon the execution of an analysis template, the outcome could constitute the input for another analysis template. In this way, complex analysis processes can be broken down in smaller processes interlinked in the form of a workflow.

The output of the analysis template execution will be a session object that contains on memory all output values (e.g. set of URLs providing access to the set of results). Session object values returned by a template call can feed as arguments a subsequent template call, without ever retrieving the object. Thus, analysis templates chaining will become even more powerful and pave the way for greater experimentation flexibility, considering a code snippet as an input value parameter, enabling injecting raw code into the function call. The implementation of the Experiment Composition Environment will be based on open source tools such as Apache Airflow enabling authoring workflows as Directed Acyclic Graphs of tasks.

2.4. Advanced analytics

In order to perform advanced analytics to the provided data through either its autonomous or semi-autonomous examination, several sophisticated techniques or tools will be used in order to discover deeper insights, make predictions, or generate recommendations. The main object is to focus on forecasting future events and behaviors, enabling businesses to conduct what-if analyses to predict the effects of potential changes in business strategies. For that reason, classical statistical methods, as well as newer, more machine-driven techniques will be used, such as deep learning, for identifying patterns, correlations and groupings in data sets.

In more details, in the case of CYBELE, advanced analytics algorithms will be provided to the stakeholders with the ability to visually explore the different kinds of data, while discovering and addressing new patterns. What is more, machine learning and predictive modelling techniques will be updated in order to be able to manage the predictive life cycle of data preparation, exploration and analysis, for achieving better deployment and monitoring. Through CYBELE, the provided advanced analytics algorithms will semi-automate, or even automate the already developed processes that exist to improve data performance, thus gaining quicker and more efficient results. Moreover, the data mining process will be streamlined, creating both high-performance and accurate models, while predictive models will be possible to be built using traditional statistical, data mining or text mining algorithms. As a result, the latter will lead to cost and latency reduction, since CYBELE will automate in-database scoring to improve model performance and get faster results. To this end, another innovation of CYBELE will be to design machine learning and prototyping algorithms for production that will be able to automatically update themselves by constantly retaining data sets, performing cross-validation, refining and discovering new rules. The work that will be done will be to first build incremental algorithms that will be then parallelized. In this case, there will be a strong need for building accurate local knowledge in order to optimize the work of aggregating nodes. Our goal will be to consider the most suitable existing algorithm in the incremental case, and make it distributed. For instance, the FP-Growth principle may be used, in order to evaluate the performance when the computing nodes divide their input streams into batches and send regular updates to the intermediate nodes. A major issue will be to deal with the trade-off between computation time and results accuracy, thus approximate algorithms will be considered since they are able to perform in real time.

3. Results

The CYBELE project officially started on January 2019, thus no scientific results have yet been made available. Nevertheless, the CYBELE concept, approach and technical solution will be evaluated and technically validated through a series of 9 demonstrators, 5 from the domain of PA, and 4 from the domain of PLF, which are briefly described in the forthcoming sections.

3.1. Organic Soya yield and protein-content prediction

The EU is strongly dependent on other continents for plant-based proteins. Plant proteins are mainly used for animal husbandry and by far the majority of the need is covered by soybean. Currently, the EU imports the equivalent of about 32 M tons of soya, mainly as processed soya-meal, from Brazil, Argentina and the US. Unlike the US and the rest of the world, GMO products are banned in the EU, meaning that we must rely on our own production and increase its efficiency. The constrains are only partially due to market trends while there is a large room for technical improvement in cultivation and processing phases. Those improvements should be aiming at increasing the efficiency of production and, at the same time, decreasing the environmental impact. An innovative concept that well summarises the need for improvement is to optimise the protein production instead of optimising the production of soybean in general, meaning that the inputs brought into the system should aim at producing as much protein as possible. Using the dataset acquired through crowdsourcing, we will derive methods for predicting yield and protein-content maps based on satellite imagery and additional information (if available) concerning electromagnetic soil scans, drone images and sensory data.

Time-series of satellite images are very indicative of the relative yield and protein content on the field, i.e. they can pinpoint the areas in which soybean grows better and in which it grows worse. By knowing the absolute value of the yield and protein content on the whole farm, we can "reverse-engineer" these traits in these specific areas of the field and derive the corresponding maps. We will train state-of-the-art satellite image processing techniques for delineation of different zones inside the fields and advanced machine learning algorithms for prediction of yield and protein content based on the crowdsourced data.

3.2. Climate smart predictive models for viticulture

Climate change has a profound impact on the growth rate and growth patterns of plants and crops. More specifically, different crop growth and development processes are affected by climatic variability via linear or nonlinear relationships resulting in complex and unexpected responses. It has been argued that such responses can best be captured by process-based crop simulation models that quantitatively represent the interaction and feedback responses of crops to their environments. The purpose of this Use Case is to demonstrate the capacity of the HPC solutions proposed in the project for supporting complex highly-nonlinear models for vine and grape growth with respect to the extreme number of variables (data types) that have been shown to affect the quality and quantity of the produced yields. Such crop models could estimate vine and grape growth and crop yield at larger scales, with spatial sources of information on soils, water, land use, and other factors. This way, much larger predictions of yield could be achieved across regional scales. This could also allow researchers to look at different scenarios of land use change, water, and climate change. In the context of the proposed Use Case, we aim to examine the efficiency of different optimization techniques on both directions. Indicative examples for direction (a) is the usage of quasi-Newton optimization techniques like BFGS and its memory optimized L-BGFS variation, where an approximation of the Hessian matrix of the cost function is used and, in the latter case, is reduced to vector-matrix multiplications. Similarly, for direction (b) we will examine the applicability of different parallel matrix multiplication algorithms like the Coppersmith–Winograd algorithm which has been shown to achieve the best complexity thus far ($O(n^{2.376})$) or the memory-optimal Cannon's algorithm.

3.3. Climate services for organic fruit production

The increased occurrence of extreme weather events due to climate change has heightened the need to develop support decision systems that can help farmers to mitigate losses in agriculture. Environmental hazards, such as frost and hail, have a relevant economy impact on crops since they may cause several damages and injuries in sensitive crops and, therefore, production losses. Horticultural crops, such as apple trees, are sensitive to frost and hail events. and protecting them from the effects of low temperature and hail damage is crucial. Frost is a serious problem for horticultural / fruit-trees production both early and late in the season since water within the plants may freeze during a frost event, while the damage caused by hail depends on the hailstone size, number per unit area and kinetic energy. In both cases, climate conditions influence the occurrence probability of these events, together with other issues such as vegetation present, topography and soil type with relevance at local scale. Passive and active protection methods for frost and hail exist in the market with their different characteristics, effects and costs. So, early warning systems at local scale with a suitable spatial resolution on frost and hail occurrence and their associated risks are relevant for agriculture. Frost and hail forecasts may help farmers to reduce any possible injuries to their crops since protection methods can be used. Integration and comparison of estimated stage of fruit bud development models with temperature and air humidity forecasts and other ancillary data can be used for risk probability mapping in order to establish an early warning system that can help farms to prevent damage effects through the use of protection methods for frost and hail. This pilot demonstrator will be focused on climate predictors that are either correlated with frost or hail occurrence and then can be used for planning risk prevention operations. satellite-derived earth observation data together with climate forecasts approaches will be joined for development and validation of the climate services. Moreover, earth observation time series based on validated data-sets from internationally leading organizations can be used for validation procedures. The aim is to explore the potential added value of novel earth observation satellites for climate services as a tool for horticultural crop / fruit-trees management. Climate projections, crop growth models, soil parameters and satellite based time series of observations will be integrated to produce added value indicators for organic fruit producers.

3.4. Autonomous robotic systems within arable frameworks

Dictated by the weather, farming tasks have often to be carried out within a short time window. Consequently, equipment has increased in size to complete the work rapidly. One alternative solution is for farmers to manage fleets of smaller, autonomous vehicles and carry out the tasks as required. The range of operations to be delivered include soil chemical analysis, hyperspectral imaging (HSI) of soil/crop condition, real time object level (plant/weed) identification, individual plant harvest readiness assessment (particularly for soft fruits) and plant level automated harvesting, currently not possible because it would be massively labor intensive. The ultimate goal is for minimally sized equipment e.g. small tractor or scouting vehicle to carry the sensory devices e.g. spectral analysis equipment, imaging (visible and HSI). Such sensor 'transporters' can be combined with a network of 'actuator' devices such as plant level harvesters, precision soil enrichment vehicles or cultivation/planting equipment. It is envisaged that a pair (at least) of systems can operate in tandem on a given task with the sensory elements passing over the crop relaying measurement data to a central location. The data can then be processed to identify plant, weed, readiness for harvest etc., generating the inputs for the actuator to harvest the appropriate plant.

3.5. 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 for various stakeholders. These crop yield monitoring solutions ingest crop, soil & historic weather data, while also 10–25 weather data forecasts are used to provide a distribution of potential forecasts, using a cropping systems model to provide the productivity estimate. In this use case the parcel specific data associated with advanced weather forecasts and computations (weather data interpolation, crop growth model) will be prepared for computations on an HPC, while also considering the addition of a third data source, data processed for Sentinel Satellite Imagery for validation of the parcel specific estimates. First step is to reproduce the currently used application for Europe to function as a baseline, second step is to use the same system for parcel specific estimates, at least for the Netherlands and potentially another Member State. There are three scientific challenges in this use case: 1) Data preparation for efficient computation on the HPC. Currently an individual computation for one grid cell/polygon is optimized for efficient computation. 2) The added value of producing crop productivity estimates

on a farmer's parcel needs to be explored. 3) With faster and more computations, the possibilities for easily ingesting additional data (e.g. NDVI/WDVI derived from Satellite Imagery) increase dramatically. This requires scientific advances to ingest this data in a smart and automated way.

3.6. Pig weighing optimization

An accurate estimate of the live weight of slaughter pigs is useful to the farmer in several different ways. First of all, knowing the weight of the pigs in a pen allows the farmer to know the optimal time to send his pigs to the slaughter house. Second, an accurate estimate of the weights of the pigs can be used for more accurate dosing of medicine, which can potentially lead to a lower use of e.g. antibiotics, which is important for combating the spread of multi-resistant bacterial strains in farm animals as well as humans. Knowing the optimal slaughter time alone is of such great value that some big pig producers have staff employed for the sole task of performing manual weightings. This practice is very laborious and time consuming, making it unfeasible for most producers. On these more common herds, being able to infer the live weight of the pigs indirectly via e.g. video images would be optimal. A number of studies have been published which attempt to achieve this using traditional image processing. The demonstrator has three main goals: (1) To estimate the mean and standard deviation of the live weight of grower/finisher pigs in a pen based on video images; (2) To track the weight of individual pigs in a pen based on video images; (3) To incorporate the growth curve estimated by the CNNs in previously developed models for early warning of diarrhea.

3.7. Sustainable pig production

Sustainable pig production and global food challenges require producing with optimal productivity, health and welfare of the pigs. The pig farmer is becoming a manager of growing farms with several thousands of fattening pigs. There are large efforts being done to improve genetics, improve feeding, etc. to maximize the productivity of the pigs. However, the usage and fusion of all data generated throughout the lifetime and after slaughter is the future way to be able to really fully exploit the potential of each fattening pig and remains a relatively uncultivated field of innovation. This demonstrator wants to improve the health and welfare of the pigs, and work on fulfilling the potential of each pig through its life and increase the quality of the end-product for the market and the consumers. This will be done through data fusion of various data sources coming from multiple on-farm sensors and software systems, image analysis, management data and slaughterhouse records. The impact of the data fusion and analytics will be demonstrated for the purpose of health and welfare warnings, boar taint and meat quality assessment. The demonstrator has two main goals: (1) improve the detection of health, welfare and performance problems at fattening pig farms through better use of available sensor and farm data; (2) reduce boar taint and improve carcass and meat quality by linking on-farm related factors and slaughterhouse data at a large scale. In general, the demonstrator aims to bring data and techniques together to enlarge the impact.

3.8. Open sea fishing

During the last decade, fisheries management in the EU increasingly succeeded in rebuilding overfished stocks and preventing overfishing. These successes stem mainly from the increased availability of data and better analysis methods that enabled to assess, and thus provide more precise management for an increasing number of commercially exploited fish stocks. Despite this positive trend, the state of the largest part of the marine ecosystem, including most fish stocks, remains largely unknown causing that little ecosystem-based management has been put in place. An important reason for this is that most marine data is collected by means of scientific surveys on research vessels. Such surveys are expensive, and consequently, it is practically impossible to provide a full spatiotemporal data-coverage of the seas. Due to the lack of sufficient processing capacity and adequate database systems, nor fishers e.g. to optimise their operational decisions, nor fisheries managers make optimal use of these data. Solutions to be explored within the context of CYBELE include: 1) Hidden Markov Modelling combined with nonparametric methods (e.g. interpolation of vessel tracks) will be used to analyse vessel trajectories, whereupon the various states will be coupled to landing data. This will provide information about the occurrence of hotspots and provide better insights into the targeting behaviour of the fleet, as well as the spatial distribution of fish. 2) The data of all the sensors on board of a fishing vessels will be merged whereupon multivariate analysis will be used to increase the value of the information and provide advice to skippers. This demonstrator's objective is to reduce the fuel usage per kg of landed fish, and to reduce the amount of bycatches. 3) Images of the hyperspectral and RGB cameras will be analysed using Deep Convolutional Neural Networks. Adapted segmentation and classification networks such as YOLO, U-Net, Mask RCNN and others will be used to obtain accurate fish species masks during the haul. As deep learning architectures are data-hungry, autoencoder networks will be used to apply data augmentation in a more efficient way.

3.9. Aquaculture monitoring and feeding optimization

Aquaculture is probably the fastest growing food-producing sector and now accounts for more than 50 percent of the world's fish that is used for food. With the world population expected to reach nine billion by 2050, the aquaculture sector will play a key role in ensuring food and nutrition security. However, this growth is not without challenges; in order to satisfy the demand and minimize the impact on the environment, the sector has to use new technologies to intensify, diversify and produce in a more efficient, sustainable and environmental friendly way. One of the main issues in commercial aquaculture is the lost food when the fish are fed. This not only increases the cost of the produced fish (feed cost is a major cost component that accounts for approximately 70% of the OPEX of the farm) but furthermore, this wasted food is deposited in the seabed and generates an environmental impact on the surrounding area. Another challenge is maintaining the farm in a good condition. If the cages are not in the correct positions, have deformations, anti-bird nets not placed correctly, etc. this usually leads to damages, financial losses and uncontrolled escapes to the environment. The project will make use of drones, image processing and data mining to optimize feeding, evaluate impact on the environment and evaluate the status of the infrastructure in open sea aquaculture. Within the context of CYBELE, we will use methods like segmentation and region proposal and object tracking, in order to analyze water movements from color, problems in nets and cages, fish positions, etc., up to video analysis and machine learning, in order to investigate fish behavior in a deeper level. This information will be combined with other data such as weather information and sensor measurements (mainly related to Oxygen and current speed) in order to develop an efficient feed management system that can help companies to make optimum use of the feed, reduce costs and also reduce the impact on the environment.

Table 1
Necessity of integrating and making available HPC and Big Data infrastructures.

HPC Attributes	Derm. #1	Derm. #2	Derm. #3	Derm. #4	Derm. #5	Derm. #6	Derm. #7	Derm. #8	Derm. #9
Storage intensity	√	√	√		√		√	√	√
Computing intensity	√	√	√	√	√	√	√	√	√
Memory intensity		√			√	√		√	√
Throughput intensity	√	√	√	√	√			√	√
Short turnaround time				√			√	√	√

3.10. Necessity of HPC and Big Data infrastructures

The following table summarises how aligned is the set of demonstrators organized by CYBELE consortium partners, with the necessity of integrating and making available HPC and Big Data infrastructures as opposed to trying to achieve the demonstrator goals through conventional infrastructures, (Table 1).

4. Conclusions

The scope of the current paper is to introduce CYBELE, a platform aspiring to safeguard that the stakeholders involved in the agri-food value chain (research community, SMEs, entrepreneurs, etc.) have integrated, unmediated access to a vast amount of very large scale datasets of diverse types and coming from a variety of sources, and that they are capable of actually generating value and extracting insights out of these data, by providing secure and unmediated access to large-scale HPC infrastructures supporting advanced data discovery, processing, combination and visualization services, solving computationally-intensive challenges modelled as mathematical algorithms requiring very high computing power and capability. The CYBELE project officially started on January 2019, thus no scientific results have yet been made available. Nevertheless, the CYBELE concept, approach and technical solution will be evaluated and technically validated through a series of 9 demonstrators, 5 from the domain of PA, and 4 from the domain of PLF, as briefly presented.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

CYBELE project is being funded by the European Commission under the Horizon 2020 Programme (Grant Agreement No. 825355).

References

[1] Available at: https://www.forbes.com/sites/timsparapani/2017/03/23/how-big-data-and-tech-will-improve-agriculture-from-farm-to-table/#70fc3ef75989, Accessed on 30/04/2019.
[2] Available at: http://aims.fao.org/activity/blog/big-data-unlocking-future-agriculture, Accessed on 30/04/2019.
[3] Available at: https://www.big-data-europe.eu/six-challenges-for-agriculture/, Accessed on 30/04/2019.
[4] Peyraud J.-L., How to implement precision livestock farming into practice, EU-PLF, 2016, Available at: http://www.eu-plf.eu/wp-content/uploads/11_PLF_-_ATF_vision-Eu-PLF1-Sept2016_JL_Peyraud.pdf
[5] S. Romeo, Enabling smart farming through the internet of thingscurrent status and trends, Sensing Technologies for Effective Land Management Workshop, June 2016.
[6] G. Terrasson, A. Llaria, et al., Accelerometer based solution for precision livestock farming: geolocation enhancement and animal activity identification, Mater. Sci. Eng. 138 (1) (2016).
[7] G. Terrasson, E. Villeneuve, et al., Precision Livestock Farming: A Multidisciplinary Paradigm, SMART 2017: The Sixth International Conference on Smart Cities, Systems, Devices and Technologies, 2017.
[8] I. Andonovic, C. Michie, et al., Precision Livestock Farming Technologies, 2018 Global Internet of Things Summit (GIoTS), 2018 Available at https://doi.org/10.1109/GIOTS.2018.8534572 .
[9] A closer look at precision livestock farming, Bosch ConnectedWorld Blog, Available at: https://blog.bosch-si.com/agriculture/a-closer-look-at-precision-livestock-farming/.
[10] https://agriculture.trimble.com/solutions/data-management/
[11] https://agdna.com/
[12] https://sentera.com/
[13] https://agrosense.eu/
[14] https://fasal.co/
[15] https://www.agricolus.com/en/
[16] https://onesoil.ai/en/
[17] https://proagrica.com/
[18] J.-S. Lin, C.-Z. Liu, A monitoring system based on wireless sensor network and an soc platform in precision agriculture, 11th IEEE International Conference on Communication Technology, 2008.
[19] van Henten E.J., Goense D. and Lokhorst C., Precision agriculture '09, https://doi.org/10.3920/978-90-8686-664-9
[20] Y. Ge, A. Thomasson, R. Sui, Remote sensing of soil properties in precision agriculture: a review, Front. Earth Sci. 5 (3) (2011) 229–238 Springer LinkSepember.
[21] J. Primicerio, S.F. di Gennaro, et al., A flexible unmanned aerial vehicle for precision agriculture, Precis. Agric. 13 (4) (2012) August.
[22] C. Zhang, J.M. Kovacs, The application of small unmanned aerial systems for precision agriculture: a review, Precis. Agric. 13 (6) (2012) December.
[23] J. Ye, B. Chen, Q. Liu, Y. Fang, A precision agriculture management system based on internet of things and webgis, IEEE 21st International Conference on Geoinformatics, 2013.
[24] M. Srbinovska, C. Gavrovski, et al., Environmental parameters monitoring in precision agriculture using wireless sensor networks, J. Clean. Prod. 88 (February) (2015) 297–307.
[25] F.J. Ferrandez-Pastor, J.M. García-Chamizo, Developing ubiquitous sensor network platform using internet of things: application in precision agriculture, Sensors 16 (7) (2016) 1141 https://doi.org/10.3390/s16071141.
[26] T. Popovic, N. Latinovic, Architecting an iot-enabled platform for precision agriculture and ecological monitoring: a case study, Comput. Electron. Agric. 140 (August) (2017) 255–265.
[27] S. Sawant, S.S. Durbha, J. Adinarayana, Interoperable agro-meteorological observation and analysis platform for precision agriculture: a case study in citrus crop water requirement estimation, Comput. Electron. Agric. 138 (June) (2017) 175–187.
[28] Available at: http://lustre.org/, Accessed on 25/09/2019.
[29] Available at: https://www.ibm.com/support/knowledgecenter/en/SSPT3X_3.0.0/com.ibm.swg.im. infosphere.biginsights.product.doc/doc/bi_gpfs_overview.html, Accessed on 25/09/2019.
[30] Available at: https://linuxhint.com/posix-standard/, Accessed on 25/09/2019.
[31] Available at: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, Accessed on 25/09/2019.
[32] N. Chaimov, A. Malony, S. Canon, C. Iancu, K.Z. Ibrahim, J. Srinivasan, Scaling spark on hpc systems, in: Proceedings of the 25th ACM International Symposium on High-Performance Parallel and Distributed Computing, ACM, 2016, pp. 97–110.
[33] Y. Wang, R. Goldstone, W. Yu, T. Wang, Characterization and optimization of memory-resident mapreduce on HPC systems, in: Parallel and Distributed Processing Symposium, 2014 IEEE 28th International, IEEE, 2014, pp. 799–808.
[34] Available at: https://en.wikipedia.org/wiki/Remote_direct_memory_access, Accessed on 25/09/2019.
[35] Available at: http://hibd.cse.ohio-state.edu, Accessed on 30/04/2019.
[36] Available at: https://slurm.schedmd.com/documentation.html, Accessed on 25/09/2019.
[37] Available at: https://en.wikipedia.org/wiki/TORQUE, Accessed on 25/09/2019.
[38] Available at: http://mesos.apache.org/, Accessed on 25/09/2019.
[39] Available at: https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html, Accessed on 25/09/2019.
[40] Available at: https://kubernetes.io/, Accessed on 25/09/2019.

Dr Konstantinos Perakis (male) was born in Athens, Greece in 1979. He received his diploma in Electrical & Computer Engineering from the National Technical University of Athens in October 2003. He received his M.Sc. in Techno-Economical Systems in 2005 and his Ph.D. degree in Medical Informatics in 2009. Currently, Dr Perakis is the Head of the Data Science Research Group at UBITECH. Since 2004, he has been active in a number of European and National R&D programs, through which he has gained considerable experience in the field of e-Health and m-Health (indicative projects include Linked2Safety ICT and MARK1 respectively), Cloud Computing & Cloud interoperability (indicative projects include PaaSword H2020-ICT), Big Data Analytics (indicative projects include OPTIMUM, AEGIS, BigDataOcean and ICARUS H2020), Data Analytics and Mining (indicative projects include LinDA ICT) and Security (indicative projects include Linked2Safety ICT, SHIELD FI-STAR and PaaSword H2020-ICT). Dr Perakis has excellent communication and mediation skills, and the ability to deal well with people in many different contexts, which he has gained through his participation in multinational consortia within the context of European projects. He has published more than 20 scientific papers in international journals and conferences as well as two book chapters, all in the field of Information and Telecommunications Technologies, and has served as reviewer and chairman in international journals and conferences. Dr Perakis has also served as a reviewer for R&D projects for the European Commission, and is a Member of the Institute of Electrical and Electronics Engineers (IEEE) - EMBS Society and Computer Society, and a member of the Technical Chamber of Greece and the Greek Society of Biomedical Technology.

Dr. Fenareti Lampathaki (female) holds a Ph.D. Degree in Information Systems' Semantic Interoperability (2012) and a Diploma - M.Eng. Degree in Electrical and Computer Engineering (2005), as well as an MBA Degree in Techno-Economics (2009). Prior to co-founding Suite5, she worked as a R&D Project Manager at the School of Electrical and Computer Engineering in the National Technical University of Athens (NTUA) and acted as an adjunct lecturer in the post-graduate programmes of NTUA and the University of Aegean. During the last 12 years, she has successfully led the team's research and management activities in a series of EU-funded R&D projects in multiple domains (e.g. Big Data, Factories of the Future, Cloud Computing, eGovernance) related to data interoperability, modelling and analytics (e.g. ICARUS H2020 Technical Coordinator, EOSChub H2020, UTILITEE H2020, UPTIME H2020, AEGIS H2020, UNICORN H2020, PSYMBIOSYS H2020, FITMAN FI-PPP, OPENi FP7), managed 3 Coordination and Support Actions (as the overall project manager for FutureEnterprise FP7, ENSEMBLE FP7 and CROSSROAD FP7) and was involved in the research activities of numerous initiatives (e.g. CloudTeams H2020, LinDA FP7, SONNETS H2020, PADGETS FP7, COCKPIT FP7, webinos FP7, LEXIS FP6, GENESIS FP6, and the Greek Interoperability Centre, G.I.C.). Fenareti has also acquired significant experience in a large number of domestic R&D, and commercial projects related to semantic interoperability and data analytics. Her research results have appeared in over 75 publications in international journals, edited books and conference proceedings while she has co-edited 1 book (on interoperability). Finally, she has been serving as a reviewer for R&D projects and evaluator for the European Commission since 2012, as well as a peer reviewer in academic journals and conferences.

Dr. Konstantinos Nikas received his Diploma in Electrical & Computer Engineering from the National Technical University of Athens, Greece and his Ph.D. in Computer Science from the University of Manchester, UK. He is a senior researcher at the Computing Systems Laboratory (CSLab) in the School of Electrical and Computer Engineering of NTUA. His research interests include high-performance computing, parallel and high performance computer architectures, parallel programming models for shared memory and distributed platforms, multithreaded and multicore processors, memory hierarchies and resource-aware scheduling. He has participated in several research projects funded by the Greek Government as well as the EC (ACTiCLOUD, EuroEXA, Bonseyes, Grid4ALL, EGEE III, PRACE-1IP, PRACE-2IP, PRACE-3IP, HP-SEE, EGI Inspire, CELAR). He is a member of IEEE, of the Technical Chamber of Greece and HiPEAC.

Dr. Yiannis Georgiou (male) is working as the CTO at Ryax Technologies. He holds a PhD and a MsC from University Grenoble-Alpes, France upon resource and job management for High Performance Computing systems. He holds a bachelor engineering degree from the Technical University of Crete. He has various publications in scheduling and resource management for HPC and Big Data systems in international conferences. He has worked as technical leader and software architect in the R&D group of Bull Atos Technologies company where he was responsible for the research and developments of the supercomputers' resource management systems. During that time he was developer of Slurm HPC scheduler and maintainer of resource management and energy efficiency group of LinuxFoundation OpenHPC project. He has participated as collaborator or principal investigator for Bull in different funded projects during his career.

Oskar Marko (male) is a researcher at BioSense Institute highly interested in data analytics and image processing in agriculture. He is currently finishing PhD studies of electrical engineering at University of Novi Sad, Serbia, where he received his bachelor's and master's degrees in the same area. He spent the 3rd year of his undergraduate studies at City University London, where he did his final BEng project in signal processing. He was the leader of BioSense's team, which proposed a novel technical solution for yield prediction and seed variety selection that got them the 1st prize at Syngenta Crop Challenge 2017.

Jarissa Maselyne, PhD MSc (female), is an electromechanical engineer with a PhD on 'Automated monitoring of feeding and drinking patterns in growing-finishing pigs: towards a warning system for performance, health and welfare in individual pigs'. She is working as a researcher in the group of Agricultural Engineering, with the main focus on Precision Pig Farming and IoT activities. She is also vice-president of the Precision Livestock Farming committee at EAAP.