Chapter 12: CYBELE: On the Convergence of HPC, Big Data Services and AI Technologies

Sophia Karagiorgou; https://orcid.org/0000-0002-1099-8463
Aikaterini Papapostolou; https://orcid.org/0000-0002-2950-1699
Zeginis Dimitris; https://orcid.org/0000-0002-5426-4031
Yannis Georgiou; https://orcid.org/0000-0003-1264-7234
Eugene Frimpong; https://orcid.org/0000-0002-4924-5258
Ioannis Tsapelas; https://orcid.org/0000-0002-0712-6226
Spiros Mouzakitis; https://orcid.org/0000-0001-9616-447X
Konstantinos Tarabanis; https://orcid.org/0000-0002-4663-2113

12.1 Introduction: Background and Driving Forces

With the earth's population approaching 8 billion, the United Nations (UN) estimate that global food production will need to increase by at least 60% to feed the world by 2050; and therefore, it is currently considered a daunting target [1]. This demand will also rise because of increase in people's wealth resulting in higher meat consumption plus the increasing use of cropland for biofuels [2]. Precision Agriculture (PA) and Precision Livestock Farming (PLF) promise both high quantity and quality of the products with minimum of resource usage, such as water, energy, fertilisers, and pesticides, promoting profitability, efficiency, and sustainability, while protecting the environment [3]. Despite the advancements in the field of applications, data and technologies, the adoption of novel products, services and tools by the farm operators has fallen short of expectations [4], while the agri- and aqua- food industry is turning to the ICT solution providers for the answer [5]. Towards this direction, big data and Artificial Intelligence (AI) have a pivotal role to play and, together with the High-Performance Computing (HPC) technology, they are already disrupting the agri- and aqua- food industry and pointing the way forward. In addition to this, as the data volumes and varieties increase with the expansion in sensor deployments, novel data engineering techniques are also instrumental in collecting, harmonising, enriching and processing distributed data from different sources [6]. These should

be conducted in a way that the latency, performance and precision requirements by the end users and applications are also satisfied.

The driving forces for empowering vegetable or livestock farmers, fish and seafood producers with digital agricultural and aquacultural innovation tools are leveraging the need for low-cost solutions using easy-to-deploy sensors, drones, computer vision and Machine Learning (ML) algorithms. Understanding how HPC, big data and AI technologies could improve farm or seafood productivity, it could significantly increase the world's food production by 2050 in the face of constrained arable land and with the water levels receding.

While much has been written about digital agriculture's potential, little is known about the economic costs and benefits of these emergent systems. In particular, the on-site decision-making processes, both in terms of adoption and optimal implementation, have not been adequately addressed. Besides, there are important questions to be answered related to technical viability, preparedness, and training of end users in such tools, economic feasibility, and data protection.

The biggest promise of digital agri- and aqua- culture technological advancements is the ability to evaluate the system on a holistic basis at multiple levels (individual, local, regional, and global) and generate tools that allow for improved decision making in every sub-process. Some of the applications of such tools target reduction of risks in agriculture and aquaculture production, such as predicting and detecting crop diseases early on in production [22]. For instance, the use of drones in arable frameworks or optimisations in fish feeding enables to create detailed maps for damage control, prevent waste of food and benefit the entire value chain [7]. Some other technologies target risks associated with extreme weather conditions and climate change, which impact the efficiency in yield production and society more broadly, including consumers and citizens [8].

In the CYBELE project (Fostering Precision Agriculture and Livestock Farming through Secure Access to Large-Scale HPC-Enabled Virtual Industrial Experimentation Environment Empowering Scalable Big Data Analytics)[1] [8], all these questions are addressed within a multi-modal and combinatorial approach. CYBELE demonstrates how the convergence of HPC, big data, AI, IoT and cloud computing could revolutionize farming, reduce scarcity, and increase food supply, bringing social, economic, and environmental benefits. The CYBELE framework coordinates unmediated access to huge amounts of datasets and their metadata from diverse data types including sensor data, textual data, spatiotemporal data, satellite and aerial image data, etc. from a multitude of distributed sources targeting the agriculture and aquaculture domain. The CYBELE approach is holistic in a sense that it guides end-to-end time-demanding and compute-hungry analytic pipelines by seamlessly converging HPC, big data and AI technologies under the umbrella of high performance e-infrastructures without requiring previous or extensive user experience and technical skills.

12.2 Identified Gaps: Motivating the CYBELE Vision

PA and PLF come to assist in optimizing agricultural and livestock production and minimizing the wastes and costs. PA is a technology-enabled, data-driven approach to farming management that observes, measures, and analyses the needs of individual fields and crops. Typical factors affecting standard agriculture include amongst others soil, climate, seed, cultivation practices, irrigation facilities, fertilizers, pesticides, weeds, harvesting, post harvesting techniques, etc. These key enablers hold the potential of providing the concepts stemming from real-world cases, the information, the mathematical models, and the computational power required in order to make well-informed, optimal choices in various real-world driven PA and PLF verticals, and to ensure that the gaps currently encountered in these

[1] https://www.cybele-project.eu/

verticals are either: 1) due to the lack of information sources; or 2) due to the lack of proper mathematical models capable of generating value and extracting insights out of these data sources; or 3) due to the lack of the infrastructure capacity capable of handling the execution of the computationally demanding mathematical models harnessing the power of massive amounts of diverse types of big data. In addition to these, one of the key bottlenecks in leveraging the promise of digital agri-/aqua- culture is the lack of proven benefit from data shared by farmers or data owners in agri-/aqua- domains. There are also other factors that result in so-called identified gaps, namely: data ownership concerns, privacy issues, economics, financial incentives of data ownership and dissemination, anticipated and quantified return on investment, data aggregation and pipelining from the source(s) to the desired locations, to name a few. If even a subset of the gaps described above can be circumvented using a mix of technologies, policies, and awareness, the possible outcomes of digital agri-/aqua- culture can shine. Some of these include seed-variety mapping to performance characteristics resulting in better selection (e.g., soya yield and protein content prediction) [9], [10]; better understanding of regional and temporal conditions leading to sustainable and localized modelling of nutrition and supplementation needs (e.g., climate services for organic food production); and capability for statistical modelling of the on-site conditions (e.g., sustainable pig production, open sea fishing), leading to increased efficient hyperparameter tuning and optimal machine selection (e.g., pig weighing optimisation, aquaculture feeding optimisation).

Therefore, in CYBELE we are interested in both gaps and challenges on the one hand and some possible technological and policy solutions on the other to transform some of the promises of digital agri-/aqua- culture into a reality. This is the reason why we designed and implemented scalable Testbeds for efficient scheduling of HPC, big data and AI tasks with diverse capabilities including: 1) the flexible CYBELE data model; 2) data check-in, curation, alignment and integration with real agri-/aqua- culture multi datasets; 3) efficient information

retrieval from multiple sources with advanced data querying and exploration functionalities; 4) data science and business intelligent algorithms; 5) high performance computing; 6) high scalable storage, consistency, availability and partition tolerance; 7) security-by-design; 8) data governance and monitoring; 9) distributed and cloud deployments.

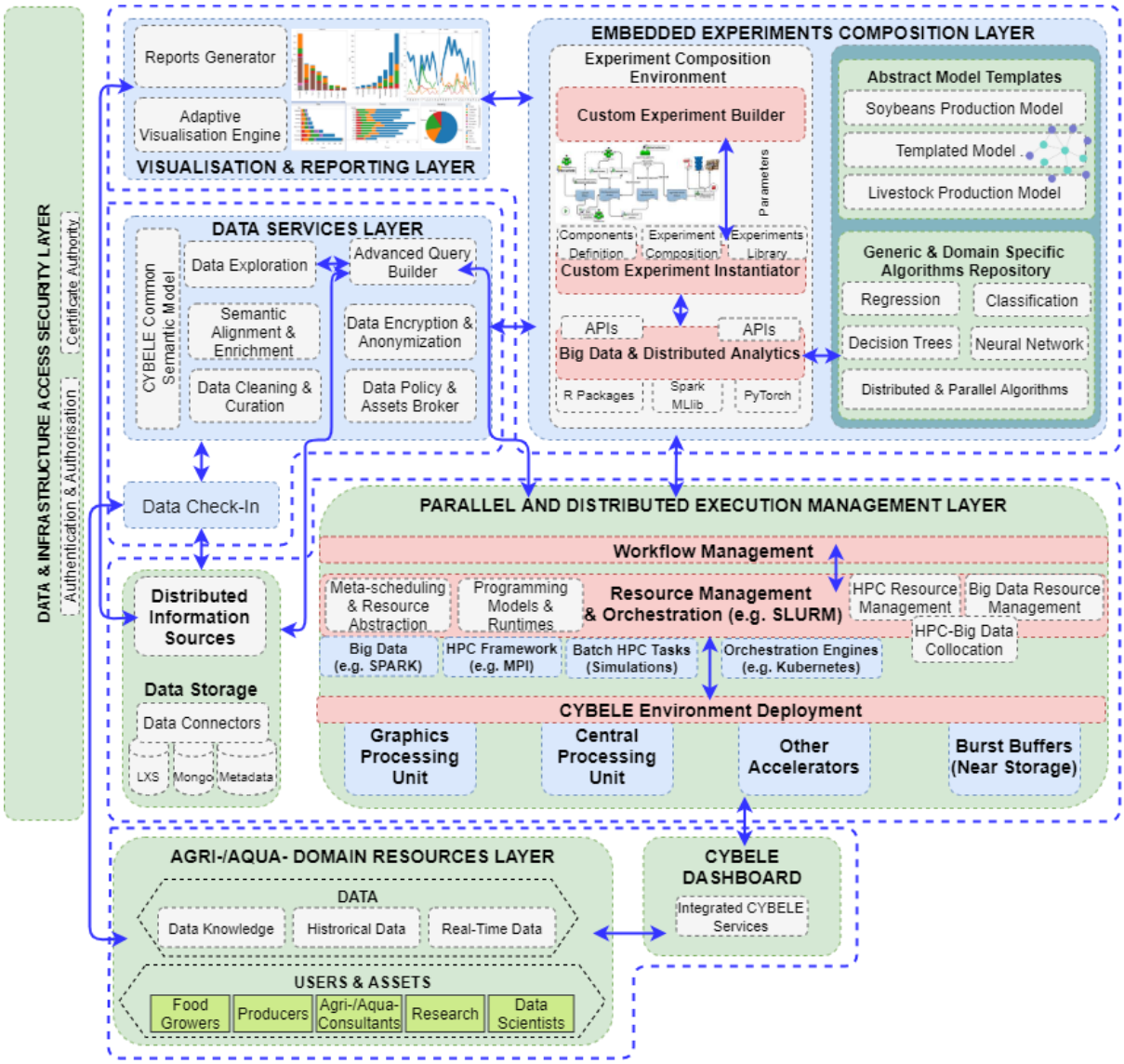

Figure 1. The CYBELE Solution towards the Convergence of HPC, Big Data and AI technologies

12.3 Materialising the Solution: Convergence of HPC, Big Data and AI

On the convergence of HPC, big data and AI technologies and in order to materialise the solution, the CYBELE framework follows a layered approach which aims at ensuring interoperability among all involved components, putting emphasis on the way that pipelining

of information (from data queries to simulations formulation, to data analytics and to visualisations) is supported, safeguarding smooth interoperation among the different services. Figure 1 presents the high-level architecture of the CYBELE framework.

More analytical information per component of the framework, as well as the perceived information flows, the technical interfaces (APIs) and the interaction amongst them are presented in the following figure (Figure 1).

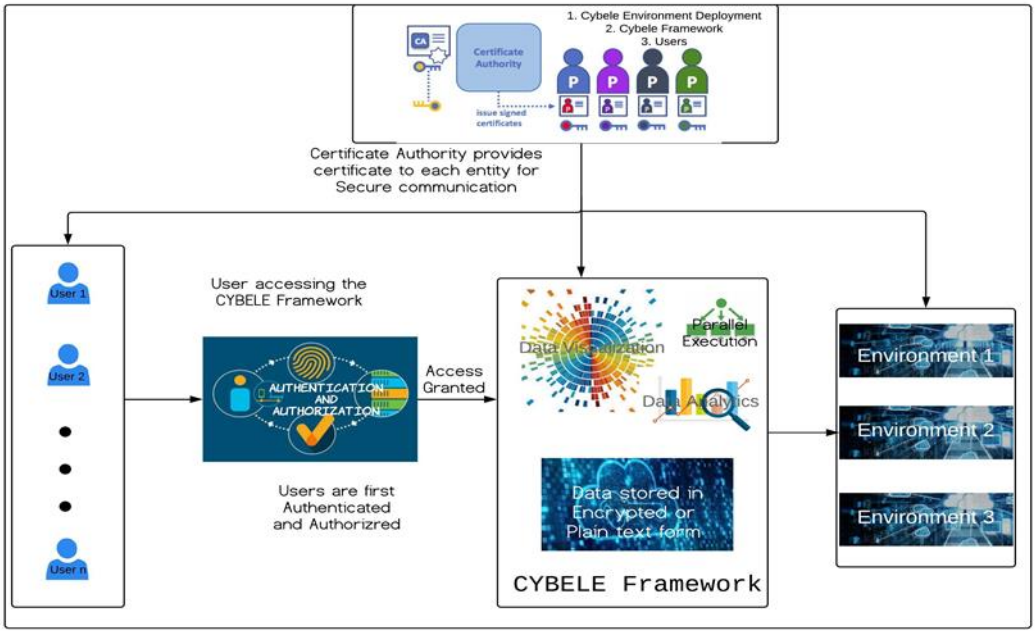

Figure 2. User Interaction with the Data & Infrastructure Access Security Layer

12.3.1 Data & Infrastructure Access Security Layer

CYBELE implements an integrated *data & infrastructure access security layer* which is spread along the whole Framework and e-infrastructure. Following this approach, for all the designed components and the described workflows, security mechanisms and protocols are used towards a Framework of enhanced security capabilities. Taking this approach into consideration, every component has been implemented with the appropriate security functionalities in mind. These include security of data-at-rest, data-in-motion and data-in-use. The security layer consists of four (4) security modules; i) Certificate Authority (CA), ii) User Authentication and Authorization (UAA) server, iii) Vulnerability Assessment (VA) toolkit, and iv) Anomaly

Detection service. Figure 2 presents the user interactions with the data & infrastructure access security layer.

12.3.2 Embedded Experiments Composition Layer

The *embedded experiments composition layer* comprises of two components: a) The *Experiment Composition Environment (ECE)* which automates the design, development and execution of the big data analysis and simulation processes; and b) The *generic & domain specific analytic algorithms* which supports the methods for the descriptive, predictive, and prescriptive analytics in the frame of the CYBELE industrial applications. An analysis process or simulation through the Experiment Composition Environment includes retrieving as input the datasets from the *advanced query builder* and defining a new or selecting an existing analysis template from the *abstract model templates*. Then, the end user is able to i) set end-to-end objectives for the experiment to be conducted (i.e. time performance, algorithm's accuracy, time constraints, etc.), ii) select a specific algorithm with the associated software and execution endpoint, iii) 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 scheduled or automated way, as well as the periodicity factor for the latter case, parameters associated with networking, storage or computing resources, and iv) adjust the output parameters along with their type (text, image, data, html). The implementation of the experiment composition environment is based on open source including custom UIs in React [11] and data pipelines engines including the Spring Cloud Data Flow [12] to facilitate workflows authoring as directed acyclic graphs (DAGs) of analytic processes. In this case, each process within a workflow is represented by a node which is fully parameterizable through the above-mentioned attributes (ii), (iii) and (iv) and the intercommunication of the processes is represented by an edge which is also parameterizable through the above-mentioned attribute (i). The user is able to save her end-to-

end analyses by appending all the specifications of an experiment or simulation for future use and reuse in a YAML file [13]. The file can be stored to the *data storage* in a structure namely the *experiments' library* for future reuse. Figure 3 presents the user interaction with the embedded experiments composition layer.

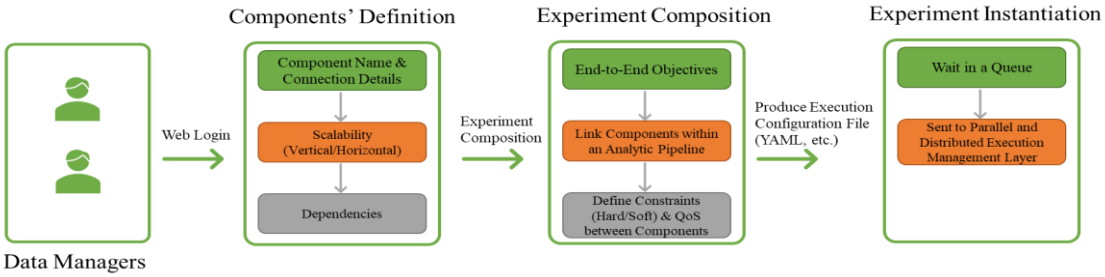

Figure 3. User Interaction with the Embedded Experiments Composition Layer

Within CYBELE, the applications of ML for PA and crop management include yield prediction, disease or insect damage detection, weed detection crop quality, crop yield monitoring while the respective ones for PLF include animal welfare, and livestock production, appetite detection, feed optimisation and biomass estimation. With regards to the aforementioned applications, the part of *domain specific analytic algorithms* coupled with ECE are presented in Figure 4, where based on user needs an analytic pipeline, also called *custom workflow*, is created consisting by both generic and custom algorithms. Most of the *domain specific analytic algorithms* are built over *distributed processing and machine learning environments* (e.g., Apache Spark MLlib [14], Distributed TensorFlow [15], etc.) exploiting their capabilities for scalable cluster computing on executing advanced analytics.

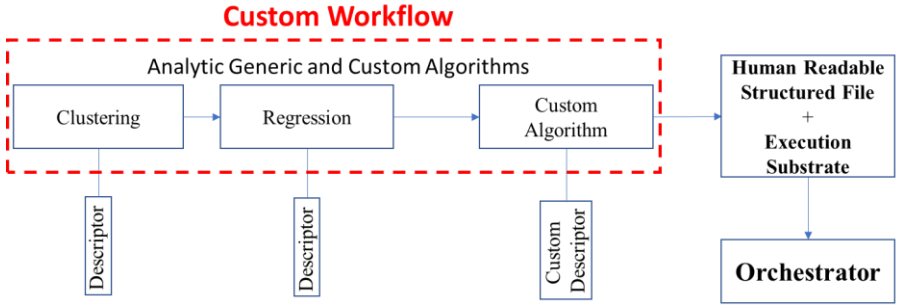

Figure 4. Algorithm Implementation Pipeline Workflow Diagram

12.3.3 Parallel and Distributed Execution Management Layer

The data analysis workflows are deployed for execution to the *parallel and distributed execution management layer*. These workflows are instantiated on both HPC and Big Data resources abstracted to the end user, as depicted in the lower part of the CYBELE Framework. The technological approach relies heavily on HPC e-infrastructure, to provide the compute power required to advance models and methods. The *workflow management* component is responsible for interpreting workflows designed with the experiment composition environment and forwarding them to the component responsible for orchestration, which performs the execution of the task collections upon the computational resources. A common pattern in scientific and cloud computing involves the execution of many computational and data manipulation tasks which are usually coupled, i.e., the output of one task is used as an input for another task. Hence non-trivial coordination is required to satisfy data dependencies. The workload of task execution needs to be directed to the available distributed computational resources [12].

Therefore, a tight integration among ECE workflows and the orchestration component is required to guarantee execution and it is performed through the *workflow management* component. Individual tasks supported by CYBELE workflows are HPC simulations, big data analytics jobs or even simple data transfer or data transformation tasks. Once a workflow is designed, the *workflow management* component interprets the workflow in the language of the orchestrator and through the orchestration component, it then proceeds with its deployment upon the adapted computational e-infrastructure consisting of HPC and big data partitions.

The *resource management & orchestration* component holds a very important place in the software stack of distributed systems since it is responsible for providing the necessary compute power to the executed tasks based on their requirements and the availability of resources. The component consists of five different modules. It combines resource

management features, such as providing fine control of hardware resources, mapping tasks upon resources, and enabling isolation of tasks upon allocated resources, along with orchestration features, such as environment provisioning and applications' life-cycle management. Traditional state-of-the-art HPC resource managers, such as Slurm [16] and Torque [17], which have been designed with performance in mind, provide optimisations for resource management and job scheduling; however, they do not provide any additional orchestration features. On the other hand, new-generation resource managers developed originally for Cloud and Big Data, such as Mesos [18], and Kubernetes [19], have been designed with elasticity in mind; hence, they give more importance to orchestration and less to performance. As the CYBELE workflows typically consist of both HPC and big data analytic tasks, as presented in Figure 5, the CYBELE Framework includes programming models and runtimes from both fields. Specifically, the *Programming Models & Runtimes* component, consists of three modules:

- HPC Programming Models & Runtimes: These include the programming models & runtimes typically employed by scientific tasks executed on HPC resources, such as MPI [20].
- Big Data & AI Programming Models & Runtimes: These include the programming models & runtimes typically employed by data analytic tasks executed on cloud & big data e-infrastructures, such as Apache Spark [14], TensorFlow [15], etc.
- HPC-enabled Big Data & AI Programming Models & Runtimes: To support multiple Big Data and AI runtimes to be deployed on HPC clusters, we deliver specific modules featuring optimized versions of the runtimes, referred as *HPC-big data collocation*, tightly integrated with the HPC computational resources [20], [21].

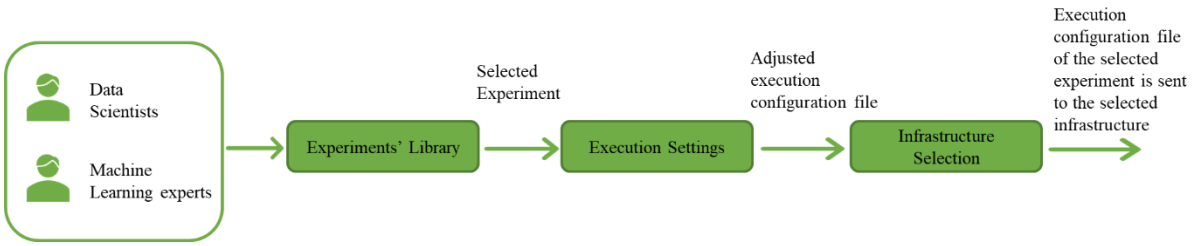

Figure 5. User Interaction with the Parallel and Distributed Execution Management Layer

12.3.4 Data Services Layer

The *data services layer* is composed of a collection of services which facilitate data check-in, cleaning, enrichment & alignment, storage, querying and controlled proprietary data sharing. The data are ingested through the *data check-in* service and are stored in the distributed *data storage*. The data check-in is an umbrella of services that ensure the veracity, timeliness, transparency and legacy characteristics coupled with the big data. The *data cleaning & curation* service performs a set of quality checks to discover inconsistencies, missing values and other anomalies in the data and eventually ensure their integrity and completeness by following several data cleaning procedures. The *data policy and assets brokerage* service facilitates data sharing and offers IPR features to link data managers (i.e., agri-/aqua- tech providers, data providers and data consumers). The *data encryption & anonymization* service ensures the preservation of the private information coming with data having intellectual property rights and is integrated in the *data check-in*.

Checked-in data are semantically annotated and harmonized through the *semantic alignment & enrichment* service to promote data interoperability and reuse. Data-oriented enrichment helps to develop robust and flexible annotations and provide a valuable source for common representation of similar concepts for disambiguation purposes. Since the data are coming from a multitude of physically distributed data sources, the *common semantic model* serves as a reference model to semantically align, describe, annotate and share these diverse data collections. Thus, the model enables the on-demand data discovery, exploration and querying. The clean and semantically enriched data are stored in the *data storage* while the data

annotations are stored at the *CYBELE metadata repository*. Both the data and annotations are made available to the *advanced query builder* for further exploration, analysis and visualisation. The *advanced query builder* provides to end users an intuitive environment to select the preferable datasets, combine them, define and execute queries on the available/combined data in the distributed *data storage*. The user interactions with the different data services are depicted in Figure 6.

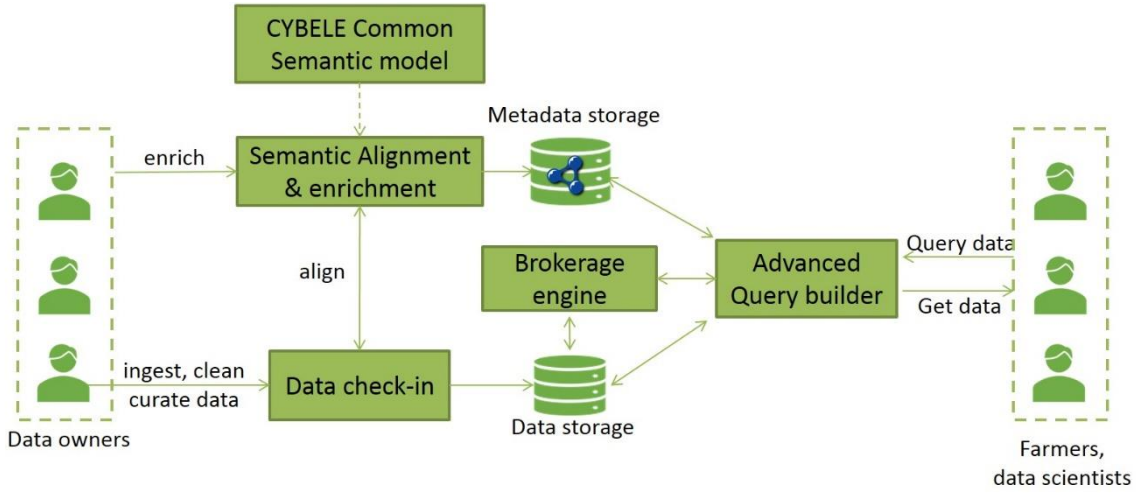

Figure 6. User Interaction with the Data Services Layer

12.3.5 Visualisation and Reporting Layer

The *visualisation and reporting layer* is responsible for the visual representation and reporting of the results produced from the other functional components of the CYBELE framework. This layer consists of an adaptive visualisation tool which follows a user centred design approach. It facilitates end users to generate or use beautiful and appealing, as well as scientifically correct and relevant visualisations. Users are able to explore data, dynamics (i.e., evolving weather conditions, prices prediction, etc.), draw conclusions, and create reports. The *visualisation and reporting layer* apart from applying the user interfaces, it is able to exploit large datasets resulting from computer simulations that use HPC resources provided that the data follow a machine-readable format. It is also capable of extracting insights from large and complex data coming from combined structured (e.g., simulations, sensors) or unstructured (free text,

images) sources and presenting them in the most useful manner interacting with the *data storage*, *ECE* and *advanced query builder*. The user interaction with the *visualisation and reporting layer* is presented in Figure 7.

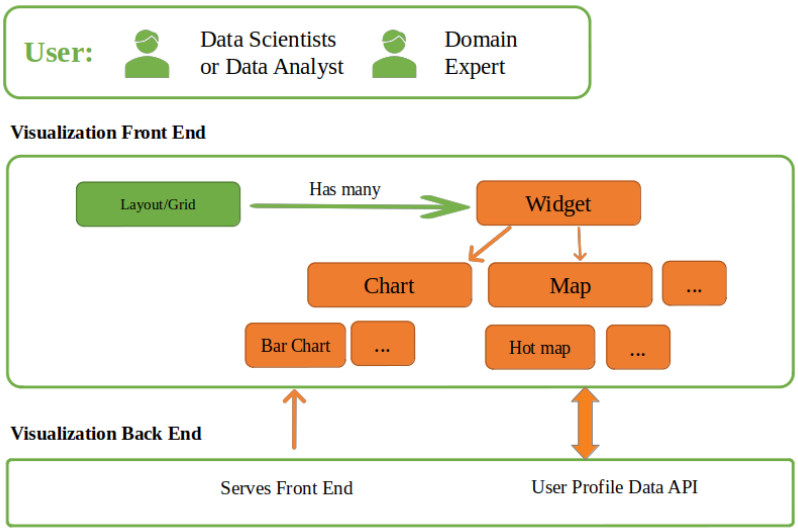

Figure 7. User Interaction with the Visualisation and Reporting Layer

12.4 Key Takeaways and Conclusions

The CYBELE Framework facilitates the execution of different scenarios coming from the agri-/aqua- culture domains by enabling the execution of batch, micro-batch and streaming processes. Following a layered approach, each layer serves to abstract to the end user the technical details and eases the design, configuration and enactment of complex big data, HPC and AI applications. The *experiments composition environment* facilitates the detaching of the design, development and execution of the big data, AI and HPC tasks, supporting embedded scientific computing and reproducible frameworks. A set of *generic & domain specific Analytic algorithms* have been developed, stored and fetched in the definition of data analysis workflows, consisted of a series of data analysis processes, interconnected among each other in terms of input/output data streams/objects. The *parallel and distributed execution management layer* focuses on tuning the HPC software stack to allow for efficient distributed

execution of big data processing frameworks and AI algorithms on top of parallel HPC resources and enriches with programmable mechanisms the *resource management & orchestration* and its interface with big data processing frameworks and orchestration engines, thus bridging the gap between the *HPC and big data worlds*. The *data services layer* takes care of the entire data lifecycle from ingestion and integration to semantic alignment and querying. The results of queries and analytics are exposed to the *visualisation and reporting layer* with the ability to visually explore the different kinds of data, while discovering and addressing new patterns and insights. The analysis results use adaptive visualisations and user-friendly interfaces, improving the way in which information is presented.

References

[1] Growing at a slower pace, world population is expected to reach 9.7 billion in 2050 and could peak at nearly 11 billion around 2100. Available Online: https://www.un.org/development/desa/en/news/population/world-population-prospects-2019.html

[2] Valin, Hugo, Daan Peters, Maarten Van den Berg, Stefan Frank, Petr Havlik, Nicklas Forsell, Carlo Hamelinck et al. "The land use change impact of biofuels consumed in the EU: Quantification of area and greenhouse gas impacts." (2015).

[3] Schrijver, R., et al.: Precision agriculture and the future of farming in Europe. Report of STOA, Science Foresight Unit, European Union (2016).

[4] Schimmelpfennig, David. Farm profits and adoption of precision agriculture. No. 1477-2016-121190. 2016.

[5] Evans, Dean. Precision Farming with Big Data Analytics: The future of farming will rely on big data to give farmers new insights into how they can grow crops more efficiently and sustainably. Available Online: https://www.intel.co.uk/content/www/uk/en/it-management/cloud-analytic-hub/big-data-helps-farmers.html

[6] Eurobarometer team: Europeans, agriculture and the common agricultural policy. Special Eurobarometer 440, The European Commission (2016).

[7] Mouzakitis S., Tsapelas G., Pelekis S., Ntanopoulos S., Askounis D., Osinga S., Athanasiadis I.N. (2020). Investigation of common big data analytics and decision-making requirements across diverse precision agriculture and livestock farming use cases. Proceedings of 13th International Symposium on Environmental Software Systems, February 5-7, 2020, Wageningen, The Netherlands. https://link.springer.com/book/10.1007/978-3-030-39815-6

[8] Perakis K., Lampathaki F., Nikas K., Georgiou Y., Marko O., Maselyne J. (2020). CYBELE – Fostering precision agriculture & livestock farming through secure access to large-scale HPC enabled virtual industrial experimentation environments fostering scalable big data analytics. Computer Networks, Vol. 168, 107035. https://doi.org/10.1016/j.comnet.2019.107035

[9] Fabiyi S. D., Vu H., Tachtatzis C., Murray P., Harle D., Dao T. K., Andonovic I., Ren J., Marshall S. (2020). Varietal Classification of Rice Seeds Using RGB and Hyperspectral Images. IEEE Access, Vol. 8, pp. 22493-22505. https://doi.org/10.1109/ACCESS.2020.2969847.

[10] Paudel D., Boogaard H., De Wit A., Janssen S., Osinga S., Pylianidis C., Athanasiadis I. N. (2020). Machine learning for large-scale crop yield forecasting, Agricultural Systems, 103016, ISSN 0308-521X, https://doi.org/10.1016/j.agsy.2020.103016.

[11] React. Available Online: https://reactjs.org/

[12] Spring Cloud Data Flow. Available Online: https://spring.io/projects/spring-cloud-dataflow

[13] YAML file. Available Online: https://en.wikipedia.org/wiki/YAML

[14] MLlib. Available Online: https://spark.apache.org/mllib/

[15] Distributed TensorFlow. Available Online: https://www.tensorflow.org/guide/distribute_strategy

[16] Slurm Workload Manager. Available Online: https://slurm.schedmd.com/documentation.html

[17] Torque Resource Manager. Available Online: https://www.adaptivecomputing.com/products/torque/

[18] Apache MESOS. Available Online: http://mesos.apache.org/

[19] Kubernetes (K8s). Available Online: https://kubernetes.io/

[20] Open MPI. Available Online: https://www.open-mpi.org/

[21] Horovod. Available Online: https://github.com/horovod/horovod

[22] Georgiou, Y., Zhou, N., Zhong, L., Hoppe, D., Pospieszny, M., Papadopoulou, N., Nikas, K., Nikolos, O.L., Kranas, P., Karagiorgou, S. and Pascolo, E., 2020, June. Converging HPC, Big Data and Cloud Technologies for Precision Agriculture Data Analytics on Supercomputers. In *International Conference on High Performance Computing* (pp. 368-379). Springer, Cham.