An Environmentally-sustainable Dimensioning Workbench towards Dynamic Resource Allocation in Cloud-computing Environments

Andreas Karabetian
Department of Digital Systems
University of Piraeus
Piraeus, Greece
adreaskar@unipi.gr

Athanasios Kiourtis
Department of Digital Systems
University of Piraeus
Piraeus, Greece
kiourtis@unipi.gr

Konstantinos Voulgaris
Department of Digital Systems
University of Piraeus
Piraeus, Greece
kvoulgaris@unipi.gr

Panagiotis Karamolegkos
Department of Digital Systems
University of Piraeus
Piraeus, Greece
pkaram@unipi.gr

Yannis Poulakis
Department of Digital Systems
University of Piraeus
Piraeus, Greece
gpoul@unipi.gr

Argyro Mavrogiorgou
Department of Digital Systems
University of Piraeus
Piraeus, Greece
margy@unipi.gr

Dimosthenis Kyriazis
Department of Digital Systems
University of Piraeus
Piraeus, Greece
dimos@unipi.gr

***Abstract*—With the exponential growth in data generated every year, Big Data has become one of the core research subjects in the overall computing domain. But when considering Big Data scenarios in a cloud-centric environment, the need for a resource management mechanism is of vital importance. Under those circumstances, intelligent allocation of resources can have a direct and noticeable impact on application performance. The aim of this paper is to present a solution on dynamic resource allocation for efficient cloud scalability. This is made possible by using machine learning algorithms as well as user feedback, in order to generate an adequate resource forecasting model. The efficiency of the tool is evaluated by repeatedly executing extensive analysis of various datasets provided by the end-users, exploiting the cloud computing paradigm for their analytic purposes. The given solution is able to learn and enhance its knowledge graph considering user feedback, as well as previously executed processes in our cloud environment. To this extent, the forecasting model will attempt to estimate optimal resource allocation for each user scenario.**

***Keywords*—cloud computing, horizontal scaling, Big Data, infrastructure as a service, knowledge-based systems, optimal allocation**

I. Introduction

Cloud computing is a way to access - on-demand - storing and computing resources in a managed infrastructure (i.e., Infrastructure as a Service, (IaaS)) IaaS [1]. Modern information systems in the field of Big Data are taking advantage of cloud computing so that they can exploit any resources required for analytic purposes. This access facilitates the execution of demanding analytic workflows without the need for an up-front financial investment. According to the study done by Stephan Höhl [2], worldwide storage sales will increase with an average growth rate per year of 5.4% from 2020 to 2026. The worldwide revenue is expected to increase from 41 US billion dollars in the year 2020, to 56 US billion dollars in the year 2026. It is undeniable and more than visible that extensive research needs to be performed in the field of cloud computing, to utilize its power on the accessibility of resources to multiple users at the same time.

In this paper, an architecture is being proposed, for an analytics service running on top of a cloud architecture, named *Dimensioning Workbench*. This Dimensioning Workbench includes two (2) components referred to as *Process Modelling* and *Dimensioning Tool*. Process Modelling consists of interfaces that aim to assist Data Scientists in making Big Data analytic workflows in a low code environment, thus enhancing the end user's experience. Meanwhile, the Dimensioning Tool is responsible for forecasting the resources of each analytic procedure before its execution. This forecast will be able to assist the orchestration services of the cloud architecture in being able to effectively scale horizontally in the computing cluster, in order to minimize the execution time of each analysis.

The rest of this paper is structured as follows. Section II provides the study of the related work, considering other research and innovation actions in the field of process modelling and dimensioning. Section III provides the initial architecture of the proposed Dimensioning Workbench, along with architectural and structural details of the Process Modelling and Dimensioning Tool components. Section IV includes an overall discussion of the aforementioned architectures, our concluding remarks, as well as our future goals, considering the overall implementation and exploitation of the Dimensioning Workbench.

II. Related Work

A. Process Modelling

In the field of Process Modelling, several research efforts have been provided. Naito et al. [3] present visual programming to support model testing. In more detail, this research focuses on the nuXmv model controller and suggests a visual programming language to describe an input model for the nuXmv. The visual programming language provides a node-graph interface, where the data flow of a target system is

represented as nodes and rows. It adopts the Node-RED optical action automation system to implement the visual programming language and to develop a programming environment that can be executed in a web browser.

Emerging workflow applications focus on data analysis and Machine Learning (ML). This has caused a change in the workflow management landscape, prompting the development of new data-driven Workflow Management Systems (WMSs) over the previous model of process-based WMSs. Mitchell et al. [4] studied such tools as Pegasus, Makeflow, Apache Airflow, and Pachyderm. Three (3) general cases of workflow usage are summarized and the unique requirements of each usage case are investigated in order to understand how the WMS meets the requirements of each case. For a more in-depth analysis of the four (4) WMSs examined in this study, three actual use cases are applied to highlight the specifications and features of each WMS. The evaluation for each WMS is presented after considering the following factors: usability, performance, ease of development, and relevance. The research's purpose is to provide information from the user's perspective on the challenges of WMS due to the evolving workflow landscape.

B. Dimensioning Tool

Regarding the vision of the Dimensioning Tool, several research works have been introduced. In this context, a relevant and recent software developed for resource forecasting, such as the Dimensioning Tool, is Hydra [5]. Hydra is a multi-agent system that makes resource management predictions based on some restrictions given as input. The proposed architecture uses artificial intelligence algorithms, to predict its ability to improve the use of electronic resources in state-of-the-art vehicles. Also, another relevant software deals with the results of the research done by Gurleen, Anju and Inderveer [6]. The authors in this work have designed a system that absorbs workload and then predicts the amount of CPU usage in the environment. ML algorithms are used, which are already pre-trained and then the above-mentioned workloads are executed on them. Furthermore, Elhoseny et al. [7], proposed a model for cloud computing to manage Big Data in health services applications. This model aims to optimize the production of virtual machines within the built-in cloud, in order to better process health data from Internet of Things (IoT) devices. This effort exploits three optimization algorithms (GA [8], PSO [9] and Parallel PSO) and manages to surpass existing methodologies by 50% in terms of performing the experiments set in the mentioned model. However, the system that has been designed assumes an infinite number of processing units and is therefore difficult to implement in real environments.

C. Advancements beyond the Related Work

It is clear that multiple research efforts have been already provided in the context of the Dimensioning Workbench, considering the Process Modelling and Dimensioning Tool components. These efforts however lack genericity, since they are tailored to be sector-specific and use case-specific, without considering the requirements of different correlating domains (e.g., correlations between the IoT and the environmental domain to consider the carbon footprint of the performance optimization algorithms). Moreover, another drawback of the aforementioned research studies, is the amount of knowledge that the end-user must have in order to use this service. Basic analytics algorithms should suffice for the Data Scientist to create a workflow and eventually get the expected results. Also, the mentioned systems require input by the user in order to proceed to resource prediction. Such a process could be avoided by designing with a different approach. To this end, the proposed Dimensioning Workbench goes beyond the current related work, since it has been designed to provide a pleasant and user-friendly way for the stakeholders to design their analytics workflow. No programming expertise is required for a task as simple as connecting the available nodes to form an analysis workflow. Additionally, the users are not asked to provide the required resources, only to submit constructive feedback at the end of the procedure. The platform can evolve at resource prediction by using two different kinds of evaluation, as well as making a classification between resource-demanding users and not. To this degree it is able to facilitate the end-user with their analytics requirements, without expecting crucial - in terms of technicalities - input on their behalf.

III. DIMENSIONING WORKBENCH

The Dimensioning Workbench is a service responsible for designing the analysis data flow and forecasting the required resources of the process that will follow, for Data Analysts and Data Scientists. The proposed Workbench will make the cloud on which it is deployed, able to respond with its computing cluster, and scale according to forecasted resources. A Data Analytics software can benefit from a service like this, by giving to each user only the required number of resources to execute the preferred analytics. In this way, the system will be able to have resources available for other possible users who want to access its computing power at the same time. Furthermore, thanks to the carefully designed user interfaces, the user can maximize his productivity in this Big Data ecosystem. The Dimensioning Workbench, gives the users faster results, effectively reducing standby time. Fig. 1 shows the Architecture of the Dimensioning Workbench as a whole.

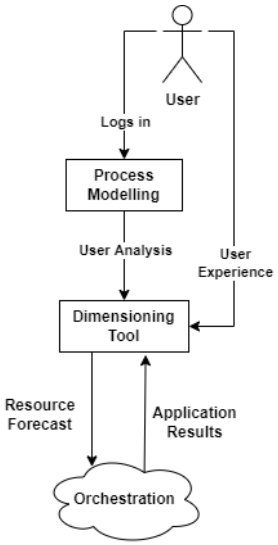

Fig. 1. Dimensioning Workbench Architecture

The users' interaction with the with the Dimensioning Workbench occurs when they want to perform an analysis on a dataset that have access to. At this point, they meet the Process Modelling component of the architecture. Here, they have access to all the necessarytools to make any complex analytic workflow that their data require. This component aids them without the need for any coding expertise to make the desired analysis, by creating an analysis graph. The graph contains information based on the datasets used, the execution

order of each step of the workflow, and every piece of information needed to execute each analytics service. When its composition has been completed, it can proceed to deployment. Before the orchestration service of the cloud executes the analysis described in the graph it is received by the Dimensioning Tool to be further analyzed. This analysis should be automated to occur every time the user wants to make a new workflow. The Dimensioning Tool will save and try to find patterns in the graph according to already executed analytics from the past, thus gathering knowledge and using it on every analysis execution. The described tool will try to make a forecast of the required resources that each analysis needs in order to be executed from the cloud's orchestration system. The major variable that needs to be taken into consideration is the required time to boot new Virtual Machines so that they can be allocated to the computing cluster of the cloud, making them available for the user as resources. Besides that, the resources that are already in use by other users must also be taken into account, as well as the past analytics done by previous users. After the analysis execution conclusion, the user encounters the Dimensioning Workbench for one last time. This interaction is the evaluation of the Dimensioning Tool given by the user, by enriching the component's knowledge base with useful insights about the cloud's performance. Hence, the cloud gathers metadata about resource demanding users. By repeating this process every time a new analysis is completed, it gets easier to separate the two user classes and allocate the required resources. Now the user experience is personalized, and the application can adapt to how it distributes cloud resources to the end-users.

A. Process Modelling

Process Modelling provides a clear overview of the services that need to be performed in an analysis procedure, all done in a visual programming environment. The Process Modelling component aims at modelling data analysis processes in a user-friendly and low-code way. More specifically, it allows users to easily create a graph describing the services, so-called steps, that they want to perform for a given set of data. The analysis process is divided into several sub-steps. Each sub-step is a part of the user's analysis. For example, the designed workflow could begin with a data ingestion service, then followed by a data cleaning job, and concluding with a classification algorithm, to run on the cleaned dataset. After entering the application, the user meets the Process Modelling interface which includes a custom graph creation tool. This tool is responsible for creating the graph that represents the data flow. The graph creation tool provides a plethora of tools available to the user, to compose the analysis that needs to be executed on their dataset. These tools contain analytic jobs, in the form of drag and drop components, with the ability to connect using a line, in order to declare a complete data flow. The added benefit is the recommendations that the platform provides, to complete the user's graph with the best-suited services that match her data. During the graph creation, it is simultaneously translated into JSON format under the hood but also validated by custom scripts, providing user-friendly messages and recommendations on how to make the current workflow more efficient. When the analysis is ready, the now called analysis graph, contains all the information encoded so that it can be later used by the rest of the platform. The automatically generated JSON file will be delivered before the beginning of the analysis to the Dimensioning Tool, to be analyzed and get a forecast on the resources needed for the graph execution. Fig. 2 shows the component's architecture.

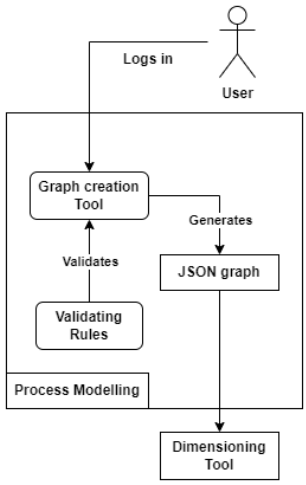

Fig. 2. Process Modelling Architecture

B. Dimensioning Tool

In a cloud-centric system [10], even if there is a large number of resources in the infrastructure, the resources need time to be delivered to the end-user. This is because in cloud computing environments, according to how the system is managed by their manufacturers, for optimal use of resources it is necessary to perform actions of building virtual machines and execution environments. However, these actions require time, so a dilemma arises as to how many resources the user wants to allocate as a total, compared to the amount of resources they want to start performing their analytics directly. This dilemma exists because of the need to minimize the waiting time for the construction of resources, but at the same time, maximize their amount as much as possible, to speed up the analytic processes. The Dimensioning Tool is a component with the main purpose of answering the above dilemma before every analytic workflow. The Dimensioning Tool consists of three basic functionalities: (i) Prediction of resources for the analysis of an end-user, (ii) Self-evaluation through the speed of the analysis performed based on the prediction forecasted, and (iii) Hetero-evaluation that gets received by the end-user.

In terms of resource forecasting, the tool receives the analysis graph that the end-user wants to perform, as well as the created metadata for the analytics process. It then compares the analysis obtained with the history of the predictions it has made using ML algorithms [11]. After the comparison is executed, it makes a forecast for the user's resources and passes the analysis graph to the orchestration service of the said cloud. The work of the tool, however, is not completed in the forecast. Operations are performed to continuously improve the accuracy of its predictions. The Dimensioning Tool uses a combination of Supervised and Unsupervised Learning [12]. Initially, the tool awaits the results of the analysis which is carried out through its forecasting and evaluates it, with the aim that in the future if it encounters a similar scenario, it will predict it accordingly. In addition, the tool can receive ratings (i.e., feedback) from end-users of the Information System after the execution of each analysis. In their ratings, users give a rating based on the experience they have according to the execution speed of their work. Using these ratings, the tool renews the parameters of its future forecasts. To make the Hetero-evaluation of the

Dimensioning Tool work, there needs to be a way from the User Interfaces of the Information System to gather the users' answers. Fig. 3 shows the component's architecture.

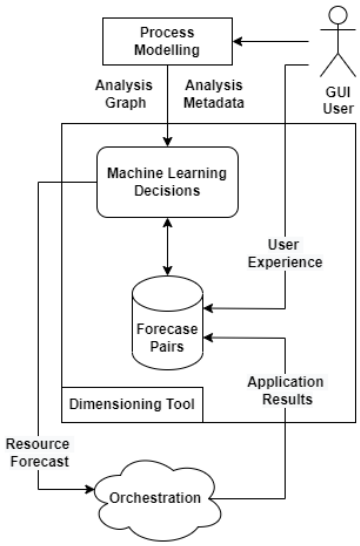

Fig. 3. Dimensioning Tool Architecture

IV. Discussion and Concluding Remarks

In this paper, a Dimensioning Workbench service was proposed, aiming to build a dynamic resource allocation service for cloud-computing environments. As it has been already specified, since the cloud computing industry is extremely growing in size and needs, such a tool is mandatory for cloud providers to trigger better resource management that will give benefits to both providers and their consumers. The Dimensioning Workbench was divided into two (2) components, that individually compose the final structure of the workbench, giving it the ability to improve over time to the needs of the current provider, due to their easily adjustable architecture. During the current research, we have proceeded to the specification and design of the Dimensioning Workbench, aiming towards its implementation with use cases considering the Healthcare and the Maritime domain, while it can generally cover all the Big Data industry. For that reason, and since feedback gathering and analysis is mandatory, it is within our plans to invite, through proper dissemination and communication measures (e.g., workshops, training sessions), other researchers who are interested in trying to implicate the Dimensioning Workbench on their own systems. These results can relate to various metrics for the Dimensioning workbench's evaluation, such as the speed of analysis of each graph, the accuracy of its decisions in cloud resource forecasts, or even the number of executions of different analytics until the point of accurate decisions on a given percentage. The importance given to end-users is also of high value, in terms of their evaluation of the system. Users who seem to be interested in cloud technologies with long term plans of use, but also users who have some kind of expertise on the subject, should have more weighted impact than the other types of users on their evaluations. Moreover, upon completion of a system based on the proposed architecture, cloud providers will have the ability to increase their profits, while decreasing their carbon footprint. This is because they will be able to share fewer resources with users who do not require as many as others, resulting into lower power costs, as well as minimizing heat levels' production. Finally, a cloud environment with these configurations will be able to maximize user experience via the personalization of each use case, by allocating the correct number of resources. For these goals, our future plans include the implementation of the Dimensioning Workbench, and its evaluation in multiple scenarios, considering the aforementioned different target groups and requirements, interlinking data interoperability [13] and security by design needs [14], also considering networking and quality of service challenges [15][16].

Acknowledgment

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: DIASTEMA - T2EDK-04612).

References

[1] F. A. M. Ibrahim, and E. E. Hemayed, "Trusted Cloud Computing Architectures for infrastructure as a service: Survey and systematic literature review", Computers & Security, vol. 82, pp. 196-226, 2019.

[2] S. Höhl, "Storage report 2021," Statista. [Online]. Available: https://www.statista.com/study/84962/storage-report.

[3] H. Naito, T. Yokogawa, N. Igawa, S. Amasaki, H. Aman, and K. Arimoto, "A Node-Style Visual Programming Environment for the nuXmv Model Checker," 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), 2020, pp. 71-75.

[4] A. A. Corodescu et al., "Locality-aware workflow orchestration for big data," in Proceedings of the 13th International Conference on Management of Digital EcoSystems, 2021, pp. 62-70.

[5] I. Galanis, D. Olsen, and I. Anagnostopoulos, "A multi-agent based system for run-time distributed resource management", in IEEE Interna. Symposium on Circuits & Systems (ISCAS), 2017, pp. 1-4.

[6] G. Kaur, A. Bala, and I. Chana, "An intelligent regressive ensemble approach for predicting resource usage in cloud computing," Journal of Parallel & Distributed Computing, vol. 123, pp. 1-12, 2019.

[7] M. Elhoseny, A. Abdelaziz, A. S. Salama, A. M. Riad, K. Muhammad, and A. K. Sangaiah, "A hybrid model of Internet of Things and cloud computing to manage big data in health services applications," Future Gener. Comput. Syst., vol. 86, 2018, pp. 1383-1394.

[8] S. Mirjalili, "Genetic Algorithm," in Studies in Computational Intelligence, Cham: Springer International Publishing, 2019, pp. 43-55.

[9] X. S. Yang, "Particle swarm optimization", in Nature-inspired optimization algorithms, Elsevier, 2014, pp. 99-110.

[10] P. K. Gupta, B. T. Maharaj, and R. Malekian, "A novel and secure IoT based cloud centric architecture to perform predictive analysis of users activities in sustainable health centres" in Multimedia Tools and Applications, vol. 76, no. 18, 2017, pp. 18489-18512.

[11] S. Amershi et al., "Software Engineering for Machine Learning: A Case Study," in IEEE/ACM 41st Intern. Conf. on Software Engineering: Software Engineering in Practice (ICSE-SEIP), 2019, pp. 291-300.

[12] M. Alloghani, et al., "A systematic review on supervised and unsupervised machine learning algorithms for data science," in Unsupervised & Semi-Supervised Learning, Springer, 2020, pp. 3-21.

[13] A. Kiourtis, A. Mavrogiorgou, and D. Kyriazis, "FHIR Ontology Mapper (FOM): aggregating structural and semantic similarities of ontologies towards their alignment to HL7 FHIR", in 20th Int. Conf. on e-Health Networking, Applications & Services, 2018, pp. 1-7.

[14] A. Kiourtis, A. Mavrogiorgou, and D. Kyriazis, "Towards a secure semantic knowledge of healthcare data through structural ontological transformations", in Joint Conference on Knowledge-Based Software Engineering (JCKBSE), 2018, pp. 178-188.

[15] E. Skondras, A. Michalas, A. Sgora, & D. Vergados, "QoS-aware scheduling in LTE-A networks with SDN control", in 7th International Conference on Information, Intelligence, Systems & Applications (IISA), 2016, pp. 1-6.

[16] E. Skondras, A, et al., "A Network Slicing Algorithm for 5G Vehicular Networks", in 12th International Conference on Information, Intelligence, Systems & Applications (IISA), 2021, pp. 1-7.