Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)
© 2022 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI220743

An Autoscaling Platform Supporting Graph Data Modelling Big Data Analytics

Athanasios KIOURTIS [1], Panagiotis KARAMOLEGKOS, Andreas KARABETIAN, Konstantinos VOULGARIS, Yannis POULAKIS, Argyro MAVROGIORGOU and Dimosthenis KYRIAZIS
Department of Digital Systems, University of Piraeus, Greece

Abstract. Big Data has proved to be vast and complex, without being efficiently manageable through traditional architectures, whereas data analysis is considered crucial for both technical and non-technical stakeholders. Current analytics platforms are siloed for specific domains, whereas the requirements to enhance their use and lower their technicalities are continuously increasing. This paper describes a domain-agnostic single access autoscaling Big Data analytics platform, namely Diastema, as a collection of efficient and scalable components, offering user-friendly analytics through graph data modelling, supporting technical and non-technical stakeholders. Diastema's applicability is evaluated in healthcare through a predicting classifier for a COVID19 dataset, considering real-world constraints.

Keywords. big data, cloud computing, user experience, graph modelling, analytics

1. Introduction

Big data analytics covers multiple domains, including healthcare where since the COVID19 pandemic, services enhancement is demanded. Big Data in healthcare has proved to be vast and complex, without being efficiently manageable through traditional architectures, data management tools and methods [1]. According to Statista, in 2020, the total data storage capacity worldwide was 985 exabytes, with the total amount of healthcare data being expected to increase to 2.314 exabytes in 2025 [2]. Also, based on another research [3], data analytics is among the most funded sectors of healthcare, creating a tremendous interest for healthcare advancements, requiring data analytics democratization, allowing anyone to perform analytics, without a commitment to further understand the low-level infrastructures required. This calls for data analytics platforms to lower their entry barrier and minimize technicalities, enhancing user experience.

In this paper, Diastema is being described as a single-entry point platform for providing a user-friendly solution to its technical and non-technical end-users to process big data, build analytical procedures, and visualize these results, through its wide variety of components. Diastema can horizontally scale any processes required by any end-to-end scenario, using a cloud-centric architecture [4], providing increased time-efficiency and high-end insights. Diastema can also scale in big data centers and computing clusters, considering all the available resources, towards energy-efficiency.

[1] Corresponding Author, Athanasios Kiourtis, 80, M. Karaoli & A. Dimitriou St., 18534 Piraeus, Greece; E-mail: kiourtis@unipi.gr.

For the remainder of this paper, Section 2 studies the related work, Section 3 depicts the Diastema flow, whereas Section 4 includes an evaluation, presenting our future goals.

2. Related Work

For Diastema to become a reality, in its core components several research works have been considered. *Process models* are exploiting diagrams to explain how processes interact and help in choosing the right actions for the desired procedures. The aim of [5] is the identification and modelling of processes for the healthcare sector, as well as to suggest ways to improve it. *Orchestration systems* in cloud platforms are responsible for creating and managing the computational and network bandwidth resources to the requesting services. The authors in [6], using Markov approximation and auction theory, proposed a fully distributed resource management scheme for data centers. Moreover, *Data as a Service* is often described as a collection of tools executed in a computing cloud towards data analytics. As datasets are increasing in complexity and size, it is crucial for these tools to scale efficiently without domain-specific technical micro-management [7]. Furthermore, *charts* and *graphs* can be exploited to interpret complex data sets, with data visualization being the star of COVID19 pandemic [8].

All these works are offering different capabilities, with most of them being siloed and not concurrently interlinked in real word cases. Thus, the end-user is not provided with a unified way to combine multiple data analysis, resource allocation, or data visualization tools, without requiring a tedious amount of technical know-how. Diastema aims to provide a scalable, intelligent, and user-friendly ecosystem of technologies that will utilize State-of-the-Art research, to provide accessible and manageable big data analytics tools to data scientists across all levels and sectors.

3. Methods

Diastema is divided into several components, each performing a specific task (Figure 1).

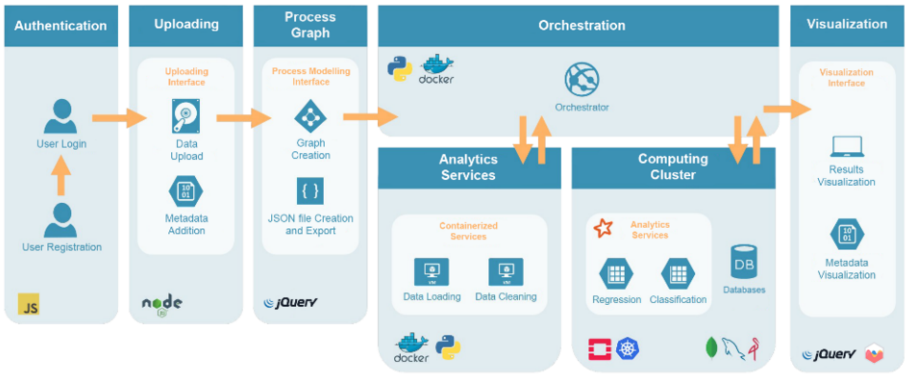

Figure 1. Diastema Architecture and Analysis Flow.

The *Authentication* component is handled through sign in and register forms, with credentials being stored in MongoDB [9] utilizing hashing functions. During the *Uploading* component, the user can provide the desired dataset to perform analytics,

where the dataset description can be enriched through metadata fields, and finally uploaded on the MinIO [10]. During the *Process Graph* component, a user-friendly interface, using LeaderLine library [10], provides the necessary tools to build a custom graph based on the dataset needs. This results in a flexible drag and drop node system, where the nodes can be connected to declare the process flow through a graph. This graph is provided to the *Orchestration* component, including a service for orchestrating all the procedures mentioned in it. This service, is used in a containerized environment [11] to be able to scale on computing clusters, thus making it possible manage as many calls as possible from the infrastructure on which it is deployed. The next component includes the *Analytics Services*, a collection of three (3) sub-components (Data Loading, Data Cleaning, Spark Engine). A typical use case involves a serial execution of these sub-components, starting with the data collection processed by the Data Loading, and its further enhancement through the Data Cleaning, before proceeding with the model's training through the Spark Engine [10]. For the scaling of the Diastema end-users, a *Computing Cluster* has been installed using Kubernetes [10], deposited into a private digital cloud to ensure data confidentiality. The Diastema cloud is built using OpenStack [10], making it able to scale on many physical machines and consequently in large data centers. It should be mentioned that the Diastema *Computing Cluster* supports storing systems like MongoDB and MinIO, as well as the Spark Engine for Machine Learning. The last step of the analysis procedure deals with the *Visualization* component, through which the gained insights are presented to the end-user. Through visualization functionalities [10], interactive charts are being illustrated, to describe the analyzed data.

4. Results, Discussion, and Concluding Remarks

For a preliminary evaluation, Diastema was set on a physical machine with 20 CPU cores, 64GB of RAM and 1TB of SSD storage. In the digital cloud, the computing cluster used three (3) virtual machine nodes with 3 vCPUs, 12GB of RAM, and 48GB of storage each. Experiments were performed on various datasets (Figure 2) provided by the European Centre for Disease Prevention and Control (ECDC) [11], regarding COVID19 and its infection rates, including the total number of incidents, and daily/weekly deaths.

	date	location	new_cases	new_deaths	total_cases	total_deaths	weekly_cases	weekly_deaths	biweekly_cases	biweekly_deaths
1	2021-02-23	Germany	5764.0	422.0	2405263.0	68785.0	52497.0	2956.0	103212.0	5779.0
2	2020-08-29	Canada	351.0	2.0	127620.0	9137.0	3130.0	40.0	5886.0	83.0
3	2020-03-04	Greece	2.0		9.0		8.0			

Figure 2. Snapshot of used dataset.

Each experiment started with the authentication of a non-technical end-user. Then, the COVID19 datasets were uploaded to be further analyzed. Afterwards, the *Process Graph* component was used to describe the desired analysis by creating a graph of the connected processes (Figure 3a). This contained processes to normalize and clean the given datasets, and then execute a Random Forest classification algorithm [12] and visualize the analysis results. After the graph was successfully executed by the *Orchestration* component, the analytics results were visualized (Figure 3b). Through the derived results (Figure 3b), it can be identified that from the provided graph model, following efficient resource management and scaling, Diastema offered valuable insights about COVID19 epidemiological data. Shortly, across the four (4) locations (classes) that the test dataset contained, the classifier correctly predicted the upcoming classes of

each newly incoming COVID19 datum with high accuracy, facilitating cases predicting data mismanagement or errors. Hence, Diastema's experience can help technical and non-technical researchers to efficiently access a user-friendly platform for on-demand disease analysis, for better decision making via high-end detection and decision schemes.

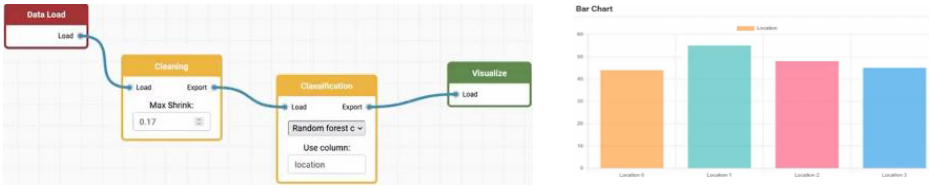

Figure 3. (a) Process Graph model, (b) Visualization.

Future work includes the exploration of the possibility to incorporate additional techniques into Diastema (e.g., streaming data) and conduct experiments considering Internet of Medical Things [13]. Additions to the *Process Graph* component could be provided, such as setting time related constraints in the analysis workflow, to be more configurable. Another future step includes the provision of additional information to the machine learning pipelines, making Diastema more efficient to repeating tasks, and its deployment on multiple physical machines, to take advantage of the system's scalability.

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] Raja R, et al. A Systematic Review of Healthcare Big Data. Scientific Programming. 2020; p. 1-15.
[2] Healthcare Data Storage Constraints Globally 2020 Forecast, https://www.statista.com/statistics/1038042/global-healthcare-data-storage-limitations/
[3] Digital Health, https://www.statista.com/statistics/736163/top-funded-health-it-technologies-worldwide/
[4] Mazumdar S, Seybold D, Kritikos K, Verginadis Y. A Survey on Data Storage and Placement Methodologies for Cloud-Big Data Ecosystem. Journal of Big Data. 2019;6(1):1-37.
[5] Marques IT, et al. Process Modelling (BPM) in Healthcare - Breast Cancer Screening. In: International Conference on Human-Computer Interaction. 2020; p. 98-109.
[6] McCormick B, Halabian H, Fung CJ. Distributed Orchestration in Cloud Data Centers. In: IFIP/IEEE Symposium on Integrated Network and Service Management (IM). 2019; p. 346-352.
[7] Wullianallur R, Raghupathi V. Big data analytics in healthcare: promise and potential. Health information science and systems. 3024;2(1): 1-10.
[8] Leung CK, et al. Big Data Visualization and Visual Analytics of COVID-19 Data. In: 24th International Conference Information Visualisation. 2020; p. 415-420.
[9] Mavrogiorgos K, Kiourtis A, Mavrogiorgou A, Kyriazis D. A Comparative Study of MongoDB, ArangoDB and CouchDB for Big Data Storage. In: 5th ICCBDC. 2021; p. 8-14.
[10] Petrova M, et al. Big data tools in processing information from open sources. In: 1st SAIC. 2018; p. 1-5.
[11] Rad BB, Bhatti HJ, Ahmadi M. An introduction to docker and analysis of its performance. International Journal of Computer Science and Network Security (IJCSNS). 2017;17(3):228.
[12] EU Centre for Disease Prevention and Control, https://www.ecdc.europa.eu/en/covid-19/data.
[13] Mavrogiorgou A, Kiourtis A, Kyriazis D. A plug 'n'play approach for dynamic data acquisition from heterogeneous IoT medical devices of unknown nature. In: Evolving Systems. 2020;11(2):269-289.