An Interactive Digital Twin for Visual Querying and Process Mining

Spyros Loizou
Department of Computer Engineering and Informatics
Cyprus University of Technology
Limassol, Cyprus
e-mail: spyros.loizou@cut.ac.cy

Andreas S. Andreou
Department of Computer Engineering and Informatics
Cyprus University of Technology
Limassol, Cyprus
e-mail: andreas.andreou@cut.ac.cy

***Abstract* — Large volumes of structured, semi-structured and unstructured data are produced daily by industrial businesses which require analysis and processing with appropriate models and algorithms to obtain valuable knowledge. This paper introduces a framework for business technology that combines the notion of Digital Twins with Process Mining aiming at delivering a simple and efficient way to retrieve customized data and process it with the use of graphical techniques, the providing interactive visualization of process mining steps. More specifically, the proposed framework provides the ability to define different data sources and link these sources with a visual query generator which constructs, executes and depicts graphically the results of custom queries. The framework also includes sophisticated Artificial Intelligence / Machine Learning algorithms for data analysis, filtering and prediction. The framework is demonstrated through an interactive dashboard, which was implemented in Python to support a fully operational and visual process mining environment that facilitates decision making without the need of programming or data management skills.**

Keywords- Big Data; Digital Twin; Business Process Mining; Visual Querying; Visual Analytics.

I. INTRODUCTION

Nowadays, the new, most popular scientific trends worldwide are the Internet of things (IoT), big data, cloud computing and Artificial Intelligence (AI). All these technologies generate a large volume of data which may be structured, semi-structured and unstructured. Big data analysis models and algorithms are used to organize, analyze and mine raw data to obtain valuable knowledge. Data visualization represents data in some systematic form including attributes and variables for the unit of available information. Visualization of data allows users and businesses mash up data sources to create custom analytical views [1]. Data processing and visualization include data mining, data collection of various types, structured or unstructured, as well as their knowledge-based representation techniques to transform primary data to meaningful data.

A Digital Twin is a virtual representation of an object or system that spans its lifecycle, is updated from real-time data, and uses simulation, machine learning and reasoning to support decision-making. In addition, the digital twin can also make predictions about how an asset or process will evolve or behave in the future [2].

Process Mining is a technique relating the fields of data science and process management to support the analysis of operational processes based on event logs. The goal of process mining is to turn event data into insights and actions. Process mining techniques use event data to show what people, machines, and organizations are really doing. Process mining provides novel insights that can be used to identify the executional path taken by operational processes and address their performance and compliance problems.

The relevant open research challenge of this area is to investigate how Digital Twins may contribute to enhancing the applicability and efficiency of process models so as to prevent costly failures in physical objects or activities, and improve quality and productivity, by using advanced analytical, monitoring and predictive capabilities, test processes and services.

This paper proposes the integration of Process Mining with Digital Twins, which targets providing interactive visualization capabilities of data related process mining steps. In this context, a framework is described for defining the data sources, linking them to a visual query generator, and finally depicting graphically the results. The framework may also utilize sophisticated algorithms residing in its Artificial Intelligence / Machine Learning module, mostly for performing data analysis and prediction. This interactive approach essentially provides smart data processing (logs and events) and graphical visualization of the insights produced. The user environment is simple, self-explanatory and ergonomic, while its graphical form and usability allow non-expert users, that is, users without prior knowledge in the area of databases or process mining algorithms, to easily structure and run queries, execute steps for process mining modeling, and receive the result in a comprehensible, interactive form. These features make the proposed approach unique and quite appealing to personnel in industrial environments, such as production engineers and shift managers, who wish to be continuously informed about selected parts in a process cycle. A prototype tool has been developed as a proof of concept which provide the basic functionality described above. The tool is work in progress and enhanced.

The rest of the paper is structured as follows: Section 2 discusses related work and provides the technical background in the areas of Process Mining and Digital Twins. Section 3 presents the proposed framework and describes how Digital Twins are integrated with specific process mining phases. This is followed by a system

 ISBN: 978-1-61208-997-3

demonstration in Section 4. Finally, Section 5 concludes the paper and highlights future work directions.

II. Technical Background/Related Work

This section provides a short description of the technical background behind Digital Twins, visualization platforms for data processing and process mining, and outlines related studies. To the best of our knowledge, there are no studies in the literature reporting the integration of graphical environments for interactive, visual smart data processing with process mining based on event logs.

A. *Data Processing and Visualization*

The area of smart data processing comprises the ability to clearly define, interoperate, openly share, access, transform, link, syndicate, and manage data. Under this perspective, it becomes crucial to have various knowledge-based metadata representation techniques to structure datasets, annotate them, link them with associated processes and software services, and deliver or syndicate information to recipients. Smart Data Processing Systems are used to include various topics to fully utilize the aforementioned capabilities, such as data ingestion, data aggregation of an enormous variety of structured, unstructured and semi-structured datasets, knowledge-based meta-data representation techniques for the conversion of raw into smart data, data privacy and protection, automated deployment, run-time software performance monitoring and dynamic configuration.

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are the challenges tackled in different papers focusing on how to analyze massive amounts of information and make data-driven decisions. Visualization-based data discovery methods allow business users to mash up disparate data sources to create custom analytical views. Wang et al. [1] present new methods and advances of Big Data visualization through introducing conventional visualization methods and the extension of some of them for handling big data, discussing the challenges of big data visualization, and analyzing technology progress in big data visualization.

Steed et al. [3] describe and demonstrate a visual analytics system, called the Exploratory Data analysis ENvironment (EDEN), with specific application to the analysis of complex earth system simulation datasets. EDEN represents the type of interactive visual analysis tools that is necessary to transform data into insights, thereby improving critical comprehension of earth system processes.

B. *Digital Twins*

A Digital Twin (DT) is traditionally characterized by two-way interactions between the digital and the physical world. A DT offers error optimization to save money and time, reduces defects and manages the lifecycle of the Internet of things (IoT). A DT represents a powerful technology because it is able to stream, optimize and analyze data in the virtual and physical world. This paper utilizes the idea of DTs to graphically represent and interact with event data and process logs and applies this approach to industrial environments.

Several papers address the problem of monitoring real-time data and optimization of an industrial production line or product design. Vachalek et al. [4] describe a project which is a technological concept focusing on the continuous optimization of production processes, proactive maintenance, and continuous processing of process data. This project is essentially promoting the concept of Industry 4.0, while its basic goal is to support the existing production structures within the automotive industry and the most efficient use of resources by augmented production and planning strategies, such as DTs. Schluse et al. [5] introduce the concept of Experimental Digital Twins (EDTs) as a new structuring element for simulation-based systems engineering processes and their interdisciplinary and cross-domain simulation in Virtual Test Beds (VTBs). This enables comprehensive simulations on system level and allows for seamless connection between virtual and real worlds in hybrid scenarios. It also introduces new structures and processes to consistently use simulations in varying application scenarios through-out the life-cycle. Fuller et al. [6] present the challenges, applications, and enabling technologies for Artificial Intelligence, Internet of Things (IoT) and DTs. A review of publications relating to DTs is performed, producing a categorical review of recent papers which discusses a range of papers residing in different scientific areas and presents the current state of research, providing at the same time an assessment of the enabling technologies, challenges and open research issues for DTs.

C. *Process Mining*

Process mining is a technique designed to discover, monitor and improve real processes by extracting readily available knowledge from event logs stored in information systems. Process mining includes process discovery, that is, extracting process models from an event log file. Moreover, process mining techniques use different algorithms to extract and organize data and business flows, with the top 5 mining algorithms being Alpha Miner, Fuzzy miner, Heuristic miner, Inductive Miner and Genetic miner. Alpha Miner generates a Petri Net model in which all the transactions are visible, unique and correspond to the classified events.

A Petri net is a tuple $N = (P, T, F, W, m_0)$, where,

$P=\{p_1, p_2, \ldots, p_m\}$ is a finite set of places,
$T =\{t_1, t_2, \ldots, t_k\}$ is a finite set of transitions,
places P and transitions T are disjoint ($P \cap T = \emptyset$),
$F \subseteq (P \times T) \cup (T \times P)$ is the flow relation (arcs),
$W : ((P \times T) \cup (T \times P)) \rightarrow \{1, 2, \ldots\}$ is the arc weight mapping (where $W(f) = 0$ for all $f \notin F$, and $W(f) > 0$ for all $f \in F$), and $m_0 : P \rightarrow \{1, 2, \ldots\}$ is the initial marking representing the initial distribution of tokens.

The Heuristics Miner algorithm deals with activities expressed as time-based intervals. Burattin et al. [7] introduced a new definition of the direct succession relation based on time-based intervals. A single event is represented as an activity. Accorsi et al. [8] promote the topic of process mining and define a set of guiding principles and list important challenges, to guide for software developers, scientists, consultants, business managers, and end-users. The goal is to increase the maturity of process mining as a new tool to improve the (re)design, control, and support of operational business processes. Bozkaya et al. [9] propose a methodology to perform process diagnostics, based on process mining. Given an event log of an information system within an organization, process diagnostics gives a broad overview of the organization's process(es) within a short period of time.

Recent studies have explored the use of Digital Twins for process mining: Park et al. [10] proposed a digital twin interface model as a representation of an organization that reflects the current state of business processes. Using this representation, process analysts are able to define constraints and actions that are continuously monitored and triggered to improve business processes. Pan et al. [11] presented a data-driven Digital Twin framework integrated with Building Information Modeling (BIM), IoT, and data mining for advanced project management, which can facilitate data communication and exploration to better understand, predict, and optimize the physical construction operations.

None of the studies on coupling Digital Twins with process mining thus far has been concentrated on defining, linking and analyzing data used for process modelling or enhancement through approaches that alleviate the need for expert knowledge. This paper addresses this challenge and provides the means for a totally different user experience based on visual querying and process mining data-driven tasks, which is characterized by simplicity, self-explainability, ease of use and graphical ergonomics.

III. METHODOLOGY

The basic idea of using visual analytics is to represent the information in a graphical and meaningful visual manner, allowing the user to interact with the information, gain insight, and make better decisions. The visual representation of the information reduces complexity in cognitive work needed to perform certain tasks and derives insight from massive, dynamic, and often conflicting data by providing timely, defensible, and understandable assessments [12].

The main target of this paper is to combine Digital Twins and process mining in a unified graphical and interactive dashboard to deliver visual representation of business data and logs, which allows for the execution of visual queries and the launch of intelligent (Artificial Intelligence and Machine Learnings) algorithms. This target is realized through a framework that describes how to transform data into a Digital Twin visual representation and then use this representation to construct and execute customized visual queries for selected process data describing the business flow. In this respect, the framework is rather generic to be able to apply in every business and data context. Figure 1 depicts the architectural structure of this approach, which involves two distinct parts: (i) the mechanism for importing data sources residing in a database (tables, fields, relationships and semantics), as well as process mining related data in the form of log files and event streams; (ii) the interactive dashboard, which offers multiple functions for constructing and executing queries in a simple, visual manner, without requiring knowledge on databases or Structured Query Language (SQL), enhanced by a set of AI/ML modules (e.g. Neural Networks, Evolutionary Algorithm and Fuzzy Logic) that provide data cleaning and/or prediction, and recommendations. The framework is generic in the sense that it does not pose strict development restrictions on the data importing mechanisms (e.g., format, files, data models, etc.) or to the mechanisms for performing visual processing of the data. Therefore, the proposed architecture essentially converts process mining into an interactive procedure that utilizes Digital Twins to visualize data and processes retrieved from logs and files stored in a database. This involves interacting with all information present in logs (activities, resources, cost, time performance, etc.) which may now be represented graphically, depicting all data connections and relations, as well as the hierarchical flow of data that is followed within the business.

Furthermore, the framework couples AI and machine learning algorithms that are stored in its smart engine. These algorithms may be trained and executed in the background to enable smart data processing, including filtering and cleaning (e.g., removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted as it may hinder the process or provide inaccurate results), prediction and predictive analytics, classification, etc. Using visual task execution (querying and machine learning) in the background enables users that will use these steps to obtain the desired results without the need of programming skills or prior knowledge on intelligent algorithms. The smart engine is able to host as many intelligent algorithms as desired, while the process for adding them into the engine is simple and straightforward, following the principles of ML-Ops [13]. Currently, the smart engine is under development comprising several implementations of neural networks for prediction and recommendations.

A dedicated software tool was developed to demonstrate the proposed framework, which was built in Python, mainly using pm4py, pandas and Streamlit libraries. Pm4py was the main instrument for implementing well-known process mining tasks and algorithms. Streamlit, which is an open-source framework that creates applications for data science and machine learning in a short time, was utilized to support the development of an interactive, flexible and user-friendly dashboard. Finally, pandas, which is a dedicated library that offers data structures and operations for manipulating tables and time series, was responsible for managing all data from databases and events from log files. Lastly, Unity was the environment used to produce playful, aesthetically correct and user-friendly graphics for the presentation and interactive use of the dashboard. Unity [14] is a cross-platform game engine, which is primarily used to develop video games and simulations for computers, consoles and mobile devices.

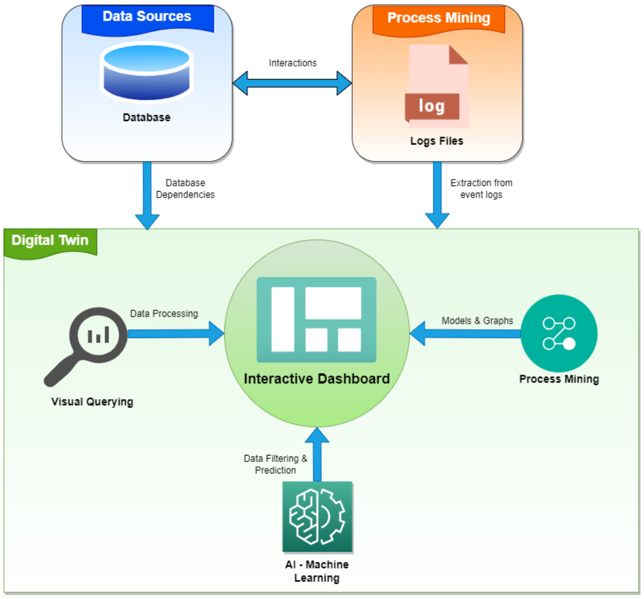

Figure 1. Architecture of the proposed framework for visual and interactive process mining.

IV. System Demonstration

The supporting software system is currently under development and its current form supports two process discovery algorithms, namely the Directly Follows Graph and the Heuristic Miner.

Figure 2 illustrates the Directly Follows Graph which is constructed using process mining logs that depict the average time for each process event. In this graph, each node represents an activity, and the arcs describe the relationships between various activities.

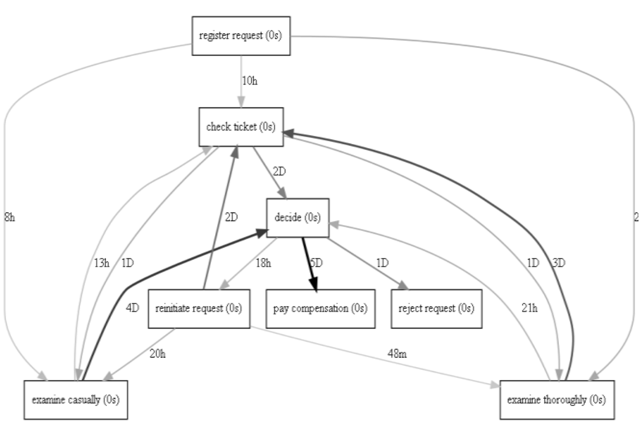

Figure 2. Directly Follows Graph based on average time.

Figure 3 presents the Heuristic Miner that is considered an improvement of the Alpha Miner algorithm and acts on the Directly Follows Graph. It provides a way to handle noise and to find common constructs. The output of the algorithm is a Heuristics Net, an object that contains both the activities and the relationships between them.

The interactive dashboard developed offers the ability to graphically analyze large volumes of information and facilitate data-driven decision making. A demonstration workflow of this graphical dashboard is provided as follows:

In step 1, the user is able to select the tables they wish to process from a pool of data available (Figure 4). This is performed by dragging and dropping in the central part of the screen called workspace or canvas, any table from the list appearing on the right.

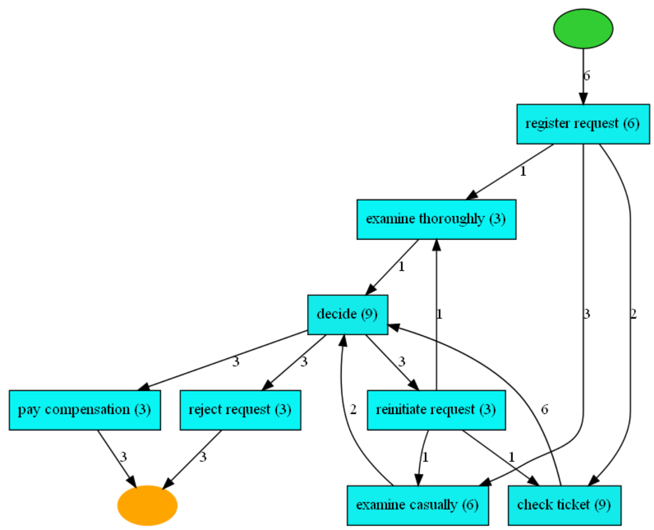

Figure 3. Heuristic Miner.

When tables are inserted into the workspace, any existing relationships are automatically depicted, showing the links between fields. The user may also select a field and reveal any links to other tables and the corresponding attributes. The dashboard supports the importing of disparate data sources in its pool of tables, allowing users to mash up different information to create custom analytical views.

Subsequently in Step 2, the user selects the fields they want to work with and defines the type of processing they wish to perform. To do so, they select an action from a toolbar at the bottom center of the screen (Figure 5), which dictates how a field is to be processed, (e.g., defines a threshold value or logical condition – Figure 6). In essence, the user constructs a query without the need of prior knowledge on SQL statements or semantics. They just define the action that needs to be executed and the system runs it and visualizes the results.

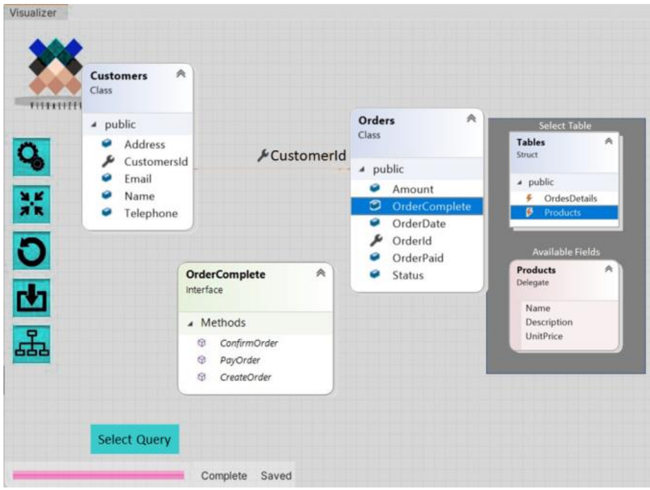

Figure 4. Selection of tables and attributes of interest.

The results are presented in various tabular and graphical forms (Figure 7). Each presentation form is interactive, as the user can point and click on a certain part of the visual information and be redirected to another part which relates new data from existing ones, such as flattening of data in process mining terms [8] or offers a different visual representation (e.g., from tabular form to a pie chart).

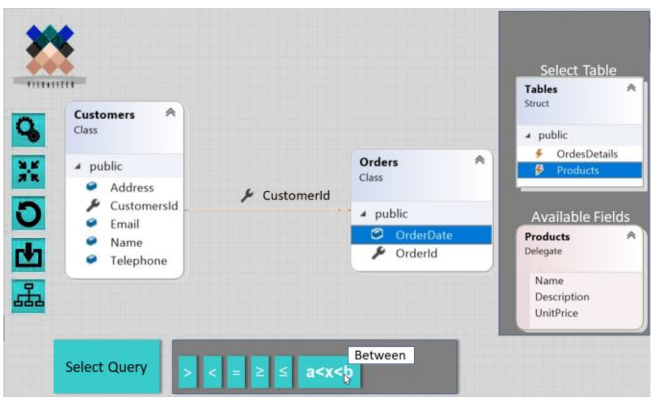

Figure 5. Selection of logical and arithmetic operators

This short presentation of the proposed framework demonstrates its usefulness and applicability. The user is able to utilize its functionality and depict graphically selected data values, connections and relations, as well as construct and execute queries in a simple, visual and interactive way. In addition, she may produce models presenting the hierarchical flow of data that is followed within the business, the sequence and dependencies of events within a business process, etc.

All this data manipulation and business process graphical information is produced requiring from users little or no prior experience and knowledge on either databases and SQL queries, or process mining and process models, thus making the framework suitable for practical use in real environments by workers that specialize in other tasks, such as manufacturing and production.

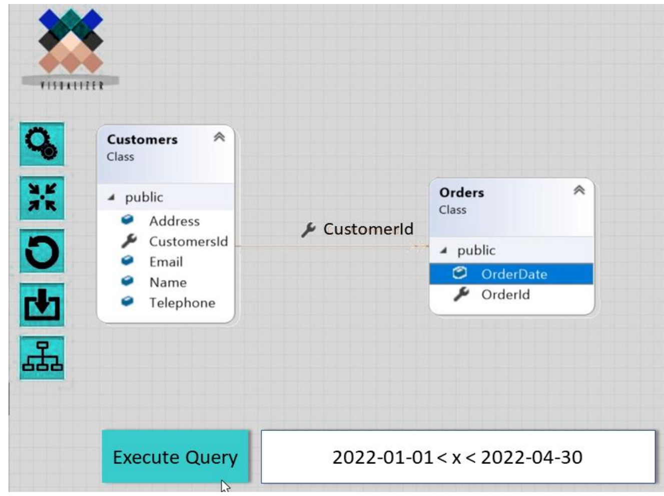

Figure 6. Creation and execution of a visual query.

CustomersId	Name	Email	Address	Telephone	OrderId	OrderDate
1451	John	john@example.com	Steet no 1	99111111	147854	2022-01-05
1001	Andrew	Andrew@example.com	Street no 10	99101010	254114	2022-02-10
1018	Chris	chris@example.com	Street no 22	99221122	189588	2022-02-22
1451	John	john@example.com	Street no 1	99111111	156040	2022-04-03
1451	John	john@example.com	Street no 1	99111111	156065	2022-04-10
1991	Helen	jelen@example.com	Street no 9	99121314	221478	2022-04-28

Figure 7. Results obtained from a visual query.

V. CONCLUSIONS

The area of Big Data Visualization and Analytics is in great need for efficient and simple ways to retrieve customized data mainly with the use of graphical techniques. This paper addressed the challenge of providing a framework which provides the ability to define data sources in the form of tables, attributes and relationships, and then process and analyze this data based on the concept of visual querying. More specifically, the framework builds upon the notions of Digital Twins to provide smart data processing and data analytics, and especially targets at bridging graphical methods with algorithms and activities from the area of process mining. Overall, this approach provides a fully operational and interactive environment in which meaningful business process data is depicted graphically to clients/organizations, and functional decision making, and predictive analytics are supported.

The combination of Digital Twins and Business Process Mining offers an interactive visual representation of business data and logs. Process data is first transformed into parts of a DT and then specific functionality is provided that gives the ability to the user to create and execute customized queries

 ISBN: 978-1-61208-997-3

of the data and the business flow and visualize results in a graphical and interactive form. Digital twins investigate in depth the relationships between the data mainly with a graphical technique. Algorithms are made available to users who will use standard steps to obtain the desired results without the need of programming skills.

Future work will concentrate on the following: (i) Data management will be enhanced with the use of Data Lakes in which sorted and cleaned data will be hosted, thus ensuring the highest quality of data is fed into the AI/ML algorithms; (ii) Further expansion of the interactive dashboard with inclusion of more sophisticated visualization features supporting predictive analytics; and, (iii) Enhancement of the visualization component with interactive capabilities that will suggest business flow corrections to achieve better process results.

ACKNOWLEDGMENT

This paper is part of the outcomes of the CSA Twinning project DESTINI. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 945357.

REFERENCES

[1] L. Wang, G. Wang, and C. A. Alexander, "Big Data and Visualization: Methods, Challenges and Technology Progress," *Digit. Technol.*, vol. 1, no. 1, pp. 33–38, 2015, doi: 10.12691/dt-1-1-7.

[2] A. Rasheed, O. San, and T. Kvamsdal, "Digital twin: Values, challenges and enablers from a modeling perspective," *IEEE Access*, vol. 8, pp. 21980–22012, 2020, doi: 10.1109/ACCESS.2020.2970143.

[3] C. A. Steed *et al.*, "Big data visual analytics for exploratory earth system simulation analysis," *Comput. Geosci.*, vol. 61, pp. 71–82, 2013, doi: 10.1016/j.cageo.2013.07.025.

[4] J. Vachalek, L. Bartalsky, O. Rovny, D. Sismisova, M. Morhac, and M. Loksik, "The digital twin of an industrial production line within the industry 4.0 concept," *Proc. 2017 21st Int. Conf. Process Control. PC 2017*, pp. 258–262, 2017, doi: 10.1109/PC.2017.7976223.

[5] M. Schluse, M. Priggemeyer, L. Atorf, and J. Rossmann, "Experimentable Digital Twins-Streamlining Simulation-Based Systems Engineering for Industry 4.0," *IEEE Trans. Ind. Informatics*, vol. 14, no. 4, pp. 1722–1731, 2018, doi: 10.1109/TII.2018.2804917.

[6] A. Fuller, Z. Fan, C. Day, and C. Barlow, "Digital Twin: Enabling Technologies, Challenges and Open Research," *IEEE Access*, vol. 8, pp. 108952–108971, 2020, doi: 10.1109/ACCESS.2020.2998358.

[7] A. Burattin, *Process mining techniques in business environments: Theoretical aspects, algorithms, techniques and open challenges in process mining*, vol. 207. 2015.

[8] R. Accorsi, M. Ullrich, and W. M. P. Van Der Aalst, *Process mining*, vol. 35, no. 5. 2012.

[9] M. Bozkaya, J. Gabriels, and J. M. Van Der Werf, "Process diagnostics: A method based on process mining," *Proc. - Int. Conf. Information, Process. Knowl. Manag. eKNOW 2009*, no. 2, pp. 22–27, 2009, doi: 10.1109/eKNOW.2009.29.

[10] G. Park and W. M. P. Van Der Aalst, "Realizing A Digital Twin of An Organization Using Action-oriented Process Mining," *Proc. - 2021 3rd Int. Conf. Process Mining, ICPM 2021*, pp. 104–111, 2021, doi: 10.1109/ICPM53251.2021.9576846.

[11] Y. Pan and L. Zhang, "A BIM-data mining integrated digital twin framework for advanced project management," *Autom. Constr.*, vol. 124, no. July 2020, p. 103564, 2021, doi: 10.1016/j.autcon.2021.103564.

[12] C. Mehrotra, N. Chitransh, and A. Singh, "Scope and challenges of visual analytics: A survey," *Proceeding - IEEE Int. Conf. Comput. Commun. Autom. ICCCA 2017*, vol. 2017-Janua, no. 4404, pp. 1229–1234, 2017, doi: 10.1109/CCAA.2017.8229987.

[13] S. Makinen, H. Skogstrom, E. Laaksonen, and T. Mikkonen, "Who needs MLOps: What data scientists seek to accomplish and how can MLOps help?," *Proc. - 2021 IEEE/ACM 1st Work. AI Eng. - Softw. Eng. AI, WAIN 2021*, pp. 109–112, 2021, doi: 10.1109/WAIN52551.2021.00024.

[14] Unity, [online] https://unity.com/, [accessed Oct. 2022]