13th Conference on Learning Factories, CLF 2023

Digital Twins in Production: The integration of semi- and unstructured data

Michael Möhring[a]*

[a]*Reutlingen University, School of Informatics – HHZ, Alteburgstraße 150, 72762 Reutlingen, Germany*

Abstract

Digital twins deployed in production are important in practice and interesting for research. Currently, mostly structured data coming from e.g., sensors and timestamps of related stations, are integrated into Digital Twins. However, semi- and unstructured data are also important to display the current status of a digital twin (e.g., of a machinery or produced good). Process Mining and Text Mining in combination can be used to support the use of log file data to understand the current state of the process as well as highlight issues. Therefore, issue related reactions can be taken more quickly, targeted and cost oriented. Applying a design science research approach; here a prototype as an artefact based on derived requirements is developed. This prototype helps to understand and to clarify the possibilities of Process Mining and Text Mining based on log data for production related Digital Twins. Contributions for practice and research are described. Furthermore, limitations of the research and future opportunities are pointed out.

Peer Review statement: Peer-review under responsibility of the scientific committee of the 13th Conference on Learning Factories 2023.

Keywords: Digital Twin; Process Mining; Text Mining; Machine Learning; Production; Industry 4.0

1. Introduction

Digital Twins are an important key technology in general [1] and in particular for industrial production [2,3]. Digital Twins enable the possibility to show the current configuration of the physical entity and their current (live) state [4]. They consist of three important parts [5,6,7,8]: the real-world physical object or entity; the virtual model, and the data connecting both worlds. Using digital twins, enterprises can improve efficiency and costs [4, 9]. Furthermore, enterprises can thereby obtain a benefit from a better planning, monitoring, visualization in real-time and over the whole life cycle [10]. This can be used to show the real state and related data [4,10] of the machinery or goods within the production, in example.

In current configurations, mostly structured data like sensor data and timestamps per station are shown (e.g. [16]). However, when it comes to production related issues particularly semi- and unstructured data is needed to understand the circumstances and solve the malfunction. There are different possibilities to integrate non-structured data into a production related digital twin. In the following, an approach focusing on the combination of process mining and text mining is described; process mining finds the issue within the process and text mining analyzes and highlights the identified issues (e.g., from log files).

Therefore, this research paper aims to answer the question: *"How can a prototype analyzing log-file data with Process Mining and Text Mining be designed to enrich production related digital twins?"*

To answer this research questions, the paper is structured as follows. First an overview about the technological background is given. Based on a design science research approach according to Hevner et al. [11] a first prototype as an artefact is developed upon identified requirements and objects. Finally, a conclusion and discussion are given.

* Corresponding author. Tel.: +49 7121 271 4127
E-mail address: michael.moehring@reutlingen-university.de

2. Theoretical Background

The integration of different variants of data is a current challenge for implementing digital twins [10]. But, without the integration of also semi- and unstructured data like textual data (e.g., log files), images or sound into digital twins the intended big picture of the related physical object (e.g., machinery or production line) is not possible [18,17,10]. In general, companies mostly use within their digital twins structured data like temperature, pressure, time, etc. (e.g., [16] [17]). This data is commonly stored in different, heterogeneous systems and therefore, it must be centrally integrated into the Digital Twin [10].
Log files are an important source to show the live state of a machinery or production line [19,20]. If an issue occur, log files can be used to show and analyze the root cause. To analyze log files from a technical point of view process mining techniques can be helpful. According to van der Aalst [13,14] process mining can be consulted to discover, monitor and improve real processes based on real event data from event logs. For instance, log files from machines or production lines can be used as event logs through process mining. Based on the log files, information about the production process can be obtained (e.g., current state of the manufacturing of parts; issue while performing production step a,b,c). Here, process mining can help to reconstruct the different process steps through process *discovery* [15]. Furthermore, through the technique of process mining *conformance* [15], an existing process model can be checked in comparison with the real event log. Hence, the difference between the modelled production process and the real, executed production process can be comprehended and evaluated. Processes can be improved based on identified production bottlenecks, for example. Therefore, another process mining technique called *enhancement* [15] can be applied. To implement process mining techniques machine learning frameworks such as pm4py [22] from Fraunhofer and process mining tools e.g., from Celonis, ARIS or Signavio can be used.
Besides the discovery and reconstruction of the production process through process mining, it is important what happened within the log related machinery (e.g., is everything correct /or warnings /or errors?). The information from process mining only shows different steps of the machinery or related production process. But when it comes to an issue or a deeper analysis this is insufficient. It is not enough to know the step. Additionally, it is high important to know what was done exactly within the step(s). To gain these relevant insights text mining techniques [12] can be used to analyze the topics of a log file. First, the text must be pre-processed by converting it into different tokens [12]. Second, a case conversion, stop word removal etc. can be done upon individual configuration of the algorithm using text mining [12]. After, the implementation can be done e.g., through machine learning python libraries such as NLTK or data analytics tools like Rapid Miner. When it comes to automation, machine learning libraries should be chosen instead of graphical data analytics tools based on our experience. In the following the theoretical foundations are used to develop a prototype integrating not structured data into digital twins.

3. Development of the Prototype

To answer the research question, a first prototype as an artifact according to Design Science research [11,23] is developed. In accordance with this method [11,23] the artifact is developed based on established methods (see section 2 and 3). The objective of this prototype is to analyze log file data with process mining and text mining to enrich production related digital twins. This is important for practice and research as well to understand the current production process and to realize a fast root cause analysis. A first evaluation of the prototype is done by testing the collected requirements with the developed prototype. Hence, this should be followed by further evaluations in different case studies and practical scenarios. After the development of the prototype is described in the following, the contribution is stated in section 4. Furthermore, the paper is part of the needed communication of the research and design as a search process [11,23]. The prototype should cover the following core requirements (using methods described in part 2 of the paper): (1.) analyze data from different production related log files with a pre-defined timestamp, (2.) reconstruct the current production related process through process discovery (process mining), (3.) show insights into the different identified process steps through text mining, and (4.) provide results to a central production related digital twin system.

The prototype is developed using the programming language python version 3. The different functional methods of the system can be accessed via a defined REST-API as microservices. For the implementation the fastapi framework is used. This allows for a scalable use in real production scenarios and a convenient integration into existing IT system landscapes.

To fulfill the requirement, (1.) the event log data can be loaded (e.g., csv or later as a .json file) into the system. The data is transformed to an analyzable form using the library “pandas”. The data structure should be defined in the following format: (a.) Timestamp, (b.)) ID, (c.) ProductionStep, and (d.) LogMessage (maybe additionally LogLevel). The timestamp is later used to reconstruct the concrete production related process and order of the

production steps based on the individual ID (e.g., ID/serial number of a produced good). The log message in the following case consists of the current logging state and its related description/message. It can be filtered by the loglevel, too [27, 26].

An example of the data structure is shown in Table 1.

Table 1. Example of the event log with log file message

(a.) Timestamp	(b.) ID	(c.) ProductionStep	(d.) LogMessage
26.05.2022 09:30	812	A	Error while loading data […]
26.05.2022 09:32	813	A	Everything correct
…	…	….	….

To reconstruct the process (requirement (2.)), the process mining discovery technique [15] is applied. Therefore, the collected event log data is analyzed with the process mining python framework pm4py [22]. As a result, a process model of the production process can be gained. An example is shown in Figure 1. Based on the example displayed in Figure 1, first insights into the process can be discovered. For instance, it shows that there are different process variants for the same process with a different order of production steps. Furthermore, loops at every production step might provide first insights into some possible issues during the production steps, too. For more insights about possible issues the implementation of the next requirement focusing on unstructured textual data is needed.

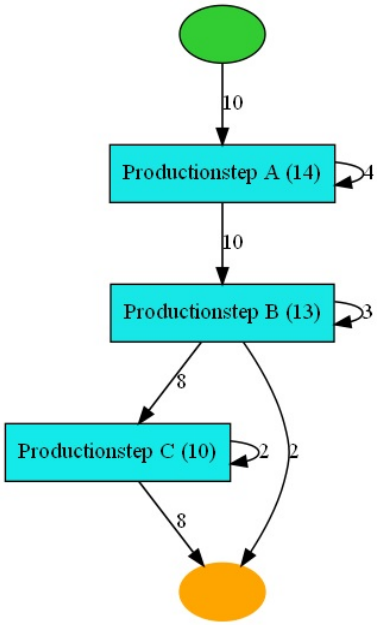

Figure 1: Example of a production process discovered with the usage of Process Mining discovery techniques via pm4py

Fulfilling requirement (3.), insights of the log messages of the different identified process steps through text mining need to be done. For implementation the log messages must be pre-processed using text mining techniques [12] like tokenization, stop word removal etc. Here, this was implemented via the NLTK framework [25]. Furthermore, a wordcloud technique [21] was used to visualize the occurrence of identified issues when a failure log occurs. An example of the analysis can be found Figure 2. LDA [24] or other analysis and pre-trained classified models can be also useful. Here it is necessary to note that they must be fine-tuned by human experts.

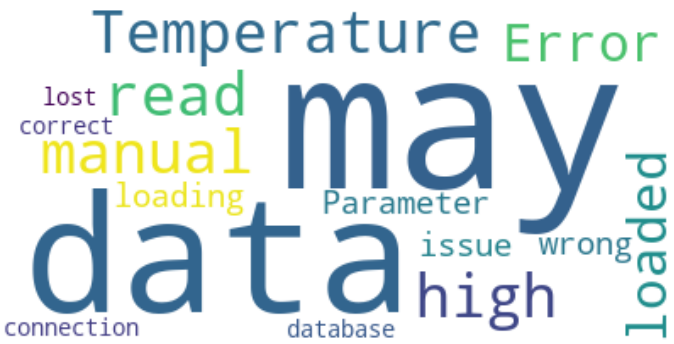

Figure 2: Example log message analysis of a failure

As a result, log files provide insights into the different identified process steps and related issues. And, the wordcloud vizualizes the most important topics of the issues (e.g, "data", "read" → database issues) for every process step. This can later taken into account for a root cause analysis and the production process's improvement based on the central production related digital twin view.

To implement and connect everything into the digital twin (requirement (4.)) the data can be transferred upon REST API request based on the individual ID (e.g., of the machinery or produced good). Therefore, semi- and unstructured data (like textual data/ log files) can be integrated into the digital twin like other existing sensor data such as temperature and/or pressure. The implementation can be done via the fastapi Python framework. Furthermore, more other unstructured data sources [18] like images, sound, etc. can be integrated via a REST-API into a digital twin. The data must be analyzed with related methods and results should be automatically transferred to the digital twin.

Figure 3 summarizes the steps necessary to cover requirement (4). As it shows, the production related digital twin is not only visualizing the current state of an e.g., machinery or good, based on the structured sensor data like temperature. It connects and integrates also to semi- and unstructured data [18,10] sources and enables a broader view and deeper investigations (e.g., for root cause analysis through process and text mining results). The digital twin's visualization can be realized in different ways like in a self-developed app, using a framework or standard data visualization solutions, or via dashboard tools like Tableau, QlikView, Grafana, Microsoft PowerBI, etc.

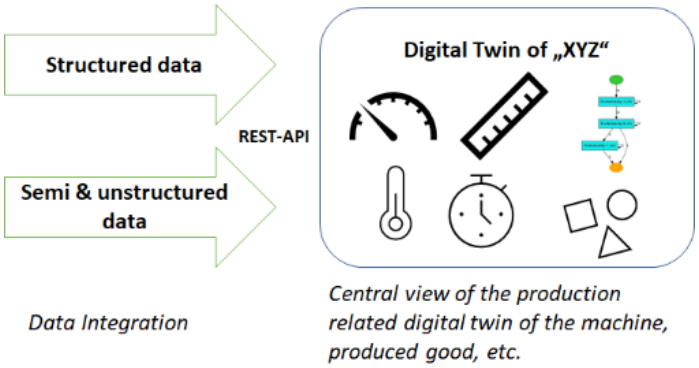

Figure 3: Overview / Summary Integration of different data types

These new integrated data sources and consecutively, the digital twin's improved informational content then, can be used at the shopfloor level to e.g., find bottle necks and possibilities for improvements. Furthermore, root cause analysis may speed up and enhance the competitiveness of the factory.

4. Conclusion & Discussion

The integration of different data sources with a different structure is a challenge in the current implementation of digital twins in practice [10]. This may hinder the benefits of using this important technology. Production state and failure prevention can be more accurate using such data.

The paper answered the research question how a prototype analyzing log-file data with process mining and text mining can be designed to enrich production related digital twins. The prototype was developed in python and

designed in a scalable way as microservices accessed and integrated via a REST API. The developed prototype integrates ideas from process mining [15] and text mining [12].

Different contributions can be derived. The scientific community can benefit from new insights how to improve production related digital twins based on semi- and unstructured data. The broader the database, the more production related insights can be gained. It extends the current work of the data integration challenges of digital twins and the usage of unstructured data in Industry 4.0 [10,18] by answering how to integrate log files. Furthermore, the prototype can be used to improve the current production processes as well as issue handling to react fast, targeted and cost oriented.

Due to the nature of a first prototype several limitations can be derived. The prototype was not evaluated in a productive industrial setup. Therefore, the prototype should be evaluated in an industry case using case study method. Furthermore, evaluations and variants due to different production sectors should be done.

References

[1] Accenture, Technology trends. https://www.accenture.com/gb-en/insights/technology/technology-trends-2021. (2021).
[2] Vachálek, J., Bartalský, L., Rovný, O., Šišmišová, D., Morháč, M., & Lokšík, M., The digital twin of an industrial production line within the industry 4.0 concept. 21st international conference on process control, (2017) 258-262.
[3] Uhlemann, Thomas H-J., Christian Lehmann, and Rolf Steinhilper, The digital twin: Realizing the cyber-physical production system for industry 4.0. Procedia Cirp 61 (2017) 335-340.
[4] Qi, Q., Tao, F., Hu, T., Anwer, N., Liu, A., Wei, Y., Wandg, L. Nee, AYC, Enabling technologies and tools for digital twin. J. Manuf. Syst. 58, (2021) 3–21.
[5] Qi, Q., Fei, T: Digital twin and big data towards smart manufacturing and industry 4.0, 360 degree comparison. IEEE Access 6, (2018) 3585–3593.
[6] Jones, D., Snider, C., Nassehi, A., Yon, J., Hicks, B, Characterising the digital twin. CIRP J. Manuf. Sci. Technol. 29, (2020) 36–52.
[7] Michael, W.: Grieves digital twin: manufacturing excellence through virtual factory replication-llc (2014).
[8] Hochhalter, J., Leser, W.P., Newman, J.A., Gupta, V.K., Yamakov, V., Cornell, S.R., Willard, S.A., Heber, G, Coupling damage-sensing particles to the digitial twin concept. No. NF1676L-18764 (2014).
[9] Rosen, R. von Wichert, G., Lo, G., Bettenhausen, KD, About the importance of autonomy and digital twins for the future of manufacturing. In: 15th IFAC Symposium on Information Control Problems in Manufacturing (2015)
[10] Möhring, M., Keller, B., Radowski, CF, Blessmann, S., Breimhorst, V., Müthing, K., Empirical Insights into the Challenges of Implementing Digital Twins. Human Centred Intelligent Systems, (2022) 229-239.
[11] Hevner, Alan R., Salvatore T. March, Jinsoo Park, and Sudha Ram, Design science in information systems research. MIS Quarterly (2004) 75-105.
[12] Tan, Pang-Ning, Hannah Blau, Steve Harp, and Robert Goldman, Textual data mining of service center call records." In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, (2000) 417-423.
[13] Aalst, W. van der, Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag (2011).
[14] Aalst, W. van der, Using process mining to bridge the gap between BI and BPM. IEEE Computer 44, (2011) 77-80.
[15] Van Der Aalst, Wil, Process mining: Overview and opportunities. ACM Transactions on Management Information Systems 3, no.2 (2012) 1-17.
[16] Mendi, A. F., A Digital Twin Case Study on Automotive Production Line. Sensors, 22(18) (2022).
[17] Harbart, T, Tapping the Power of Unstructured Data. MIT Sloan Management School (2021).
[18] Möhring, M., Keller, B., Schmidt, R., Schönitz, F., Mohr, F., Scheuerle, M., Analytics in Industry 4.0: Investigating the Challenges of Unstructured Data. Business Informatics Research, (2022) 113-125.
[19] Siemens, SINUMERIK, SINUMERIK Integrate, Shop Floor Integrate Handbook (2019).
[20] de Jesus Pacheco, D. A., Jung, C. F., & de Azambuja, M. C., Towards industry 4.0 in practice: a novel RFID-based intelligent system for monitoring and optimisation of production systems. Journal of Intelligent Manufacturing, (2021) 1-17.
[21] Heimerl, Florian, Steffen Lohmann, Simon Lange, and Thomas Ertl, Word cloud explorer: Text analytics based on word clouds. 47th Hawaii international conference on system sciences, (2014) 1833-1842.
[22] Berti, A., Van Zelst, S. J., & van der Aalst, W., Process mining for python (PM4Py): bridging the gap between process-and data science. arXiv preprint arXiv:1905.06169. (2019).
[23] Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S., A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), (2007) 45-77.
[24] Blei, D. M., Ng, A. Y., & Jordan, M. I., Latent dirichlet allocation. Journal of Machine Learning Research, 3, (2003) 993-1022.
[25] Bird, S., NLTK: the natural language toolkit. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions (2006).
[26] Ristic, I., Modsecurity handbook. Feisty Duck. (2010).
[27] Microsoft, Extension Logging, https://learn.microsoft.com/en-us/dotnet/api/microsoft.extensions.logging.loglevel?view=dotnet-plat-ext-7.0 (2022).