Pre-print copy of the manuscript published by IEEE
(available at: https://ieeexplore.ieee.org/document/9842522)
identified by doi: 10.1109/COMPSAC54236.2022.00184

An Interactive Approach to Support Event Log Generation for Data Pipeline Discovery

Dario Benvenuti, Leonardo Falleroni, Andrea Marrella
Sapienza University of Rome, Italy
name.surname@uniroma1.it

Fernando Perales
JOT Internet Media, Madrid, Spain
fernando.perales@jot-im.com

***Abstract*—Process Mining is a discipline that sits between data mining and business process management. The starting point of process mining is an *event log*, which is analyzed to extract useful insights and recurrent patterns about how processes are executed within organizations. However, often its concrete application is hampered by the considerable preparation effort that needs to be conducted by human experts to collect the required data for building a suitable event log. Instead, event logs need to be extracted from different and heterogeneous data sources, often using customized extraction scripts whose implementation requires both technical and domain expertise. While this is recognized as a relevant issue in the process mining community, literature solutions tend to be ad-hoc for particular application contexts, or not enough structured to be easily applied in practice. In this paper, we tackle this issue by proposing an interactive and general-purpose approach to support organizations in generating simulated event logs that can be employed to discover the structure of the *data pipelines* executed within a business process. A data pipeline is a composite workflow for processing data that is enacted as part of process execution. To assess the practical applicability of the approach, we show the results of a preliminary evaluation performed in a digital marketing scenario in the range of the recently funded H2020 DataCloud project.**

I. Introduction

Process mining is a research discipline aimed to discover, monitor and improve business processes (BPs) by extracting knowledge from the data that are stored in information systems about how these systems are used to carry out BPs [1]. Differently from a-priori analysis, the focus is not on the assumed processes but on real processes in the way in which they are executed. Therefore, the starting point of process mining is an *event log*, which is analyzed to extract useful insights and recurrent patterns about how BPs are executed within organizations. Specifically, event logs consist of a sequence of events, each of which includes at least a timestamp, a case identifier and an activity reference.

Process mining has found applications to achieve relevant tasks in many real-world domains, including security auditing [2], healthcare [3] and emergency management [4]. In addition, since 2021, in the context of the recently funded H2020 DataCloud project,[1] process mining solutions are explored to tackle the challenge of semi-automatically discovering the structure of *data pipelines* from event logs. There is no unified specification of the concept of data pipeline; nonetheless, some common features that are inherently related to it can be identified: *(i)* data pipeline consists of chains of processing elements that manipulate and interact with datasets; *(ii)* the outcome of a processing element of a data pipeline will be the input of the next element; *(iii)* each processing element of a data pipeline interacts with "big" datasets, i.e., with at least one of the Vs dimensions that is verified to hold.

However, event logs are typically not readily available, and they need to be extracted from different and heterogeneous data sources, often using customized extraction scripts whose implementation requires both technical and domain expertise. For this reason, event log extraction is recognized as a relevant challenge in process mining in which human support does the heavy lifting [5]. Nonetheless, often organizations do not will to invest the time of human resources in the log extraction task, hampering the application of process mining [1]. Hence, to support the use of process mining techniques to the discovery of data pipelines, we designed a structured and general-purpose interactive approach that can be implemented and applied in general industrial scenarios with relatively small human effort. The aim is to produce a simulated event log that is able to approximate the behaviour of a pipeline execution and can be employed as input of process mining techniques to discover the structure of the data pipeline of interest. Since we rely only on the available domain knowledge, the proposed approach can be applied also in scenarios where no concrete data about pipelines' execution is recorded in any data source, without having to worry about heterogeneous technologies stacks, scarce human resources to invest or partial information about the workflow. Moreover, since the quality of the discovered pipeline depends on the available knowledge about its execution, our approach mitigates this issue by enabling a human-in-the-loop interaction during the pipeline discovery stage. Specifically, domain experts and BP analysts are asked to validate the structure of the discovered pipeline. If the structure is far from reality, or if the organization is not satisfied with the obtained insights, the approach can be repeated by filtering out infrequent discovered behaviours, refining the simulation parameters or constraining the domain knowledge towards a more precise pipeline discovery.

The rest of the paper is organized as follows. In section II we look at the background knowledge in process mining and event log extraction. Then, in section III, we present the main steps of the approach. Next, in section IV, we discuss how

[1]DataCloud is a Research and Innovation project funded by the European Commission under the Horizon 2020 program (Grant number 101016835).

the questionnaires employed to collect the domain knowledge on pipelines' execution have been designed. In section V we analyze the results obtained from a preliminary evaluation conducted with an organization that manages data pipelines for massive digital marketing campaigns. Finally, in section VI, we provide a critical discussion on the approach, while in section VII we conclude by tracing future works.

II. Background

The literature on Big Data processing and analytics has often neglected the research on pipeline discovery, working with the assumption that the anatomy of data pipelines is already known at the outset, before running any Big Data processing feature. A couple of relevant approaches exists that aims at studying the structure of data pipelines. In [6], a framework that reveals key layers and components to design data pipelines for manufacturing systems is presented. In [7], the authors derive a set of data and system requirements for implementing equipment maintenance applications in industrial environments, and propose an information system model providing a scalable data pipeline for processing and analysing industrial data. However, to date, there is no explicit research study that investigates the issue of pipeline discovery.

Nonetheless, the discovery of data pipelines resembles the discovery of BPs [1], as both consist of flow of processing elements, with the main difference that data pipelines manipulate and interact with (big) data sets, while in BPs the concept of data is usually not considered as a first-class citizen. These considerations are pushing researchers to investigate how the use of process mining solutions may support the development of novel techniques to achieve the pipeline discovery task [8].

Process mining focuses on the real execution of BPs, as reflected by the footprint of reality logged and stored by the software systems in use within an organization [1]. The main type of process mining is called *Discovery*: it starts from an event log and automatically produces a BP model that explains the different behaviours observed in the log, without assuming any prior knowledge on the BP [9]. For process mining to be applicable, such information has to be structured in the form of explicit *event logs*. In fact, all process mining techniques assume that it is possible to record the sequencing of relevant events occurred within an enterprise, such that each event refers to an activity (i.e., a well-defined step in some process) and is related to a particular case (often called trace). Events may have additional information stored in event logs, such as the exact timestamp at which the event has been recorded, the resource (i.e., person or device) that generated the event, or data elements recorded with the event, cf. Figure 1. In 2016, the XES (eXtensible Event Stream) format emerged as the main reference format for the storage, interchange, and analysis of event logs [10].

Extracting data from heterogeneous resources to perform analytics is the core process behind data mining. We call data warehouse [11] a sink that collects data from many operational or external systems (its sources) to provide its end-users with access to integrated and manageable information. This data

case id	event id	properties				
		timestamp	activity	resource	cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	register request	Pete	50	...
	35654522	30-12-2010:15.06	examine casually	Mike	400	...
	35654524	30-12-2010:16.34	check ticket	Ellen	100	...
	35654525	06-01-2011:09.18	decide	Sara	200	...
	35654526	06-01-2011:12.18	reinitiate request	Sara	200	...
	35654527	06-01-2011:13.06	examine thoroughly	Sean	400	...
	35654530	08-01-2011:11.43	check ticket	Pete	100	...
	35654531	09-01-2011:09.55	decide	Sara	200	...
	35654533	15-01-2011:10.45	pay compensation	Ellen	200	...
4	35654641	06-01-2011:15.02	register request	Pete	50	...
	35654643	07-01-2011:12.06	check ticket	Mike	100	...
	35654644	08-01-2011:14.43	examine thoroughly	Sean	400	...
	35654645	09-01-2011:12.02	decide	Sara	200	...
	35654647	12-01-2011:15.44	reject request	Ellen	200	...

Fig. 1: *An example of an event log consisting of 4 traces.*

collection process must overcome several inherent problems: different sources structure data into different schemata; data quality issues, and the data warehouse needs to be updated frequently as new data comes from the sources. The software processes that facilitate the population of the data warehouse are commonly known as Extraction-Transformation-Loading (ETL) processes. They are responsible for: (i) the extraction of the appropriate data from the sources; (ii) their transportation to a special purpose area where they will be processed; (iii) the transformation of the source data to make it fit into the warehouse; (iv) the isolation and cleansing of problematic tuples, to guarantee that business rules and database constraints are respected; and (v) the loading of the cleansed, transformed data to the appropriate relation in the warehouse, along with the refreshment of its accompanying indexes and materialized views. [12] proposes a UML-based metamodel for data warehouses that covers both the back-end and the front-end of it. From 1999 up to now, several techniques have been proposed in the context of UML-based ETL, such as [13] and [14]. The close relation between ETL and UML highlights how much data warehouses are related to observing and analysing the structure and meaning of the data coming from heterogeneous sources. In process mining, the main interest is not in the structure of data, but rather in the activities that used and transformed it. Furthermore, we need to extract event data from data sources, which can differ quite substantially from the data extracted by classical data warehouses. That's why ETL processes cannot be applied to build data warehouses for BP discovery and there is the need to develop novel techniques for event data extraction [1]. As described in [5], many organizations use process mining techniques to visualize, analyse, and improve their BPs. While it has been shown to come with many benefits [15], its application is still often hindered by the considerable

preparation effort that needs to be conducted by humans [16]. One of the key challenges in this context is to obtain the input artifact for process mining techniques, i.e., the event log. Many information systems are not process-centric and, thus, do not record events or case identifiers explicitly. Depending on the specific IT landscape, the data that is required for building an event log must be collected and extracted from several data sources and then transformed into an appropriate format that can be processed by available process mining tools. The process of obtaining event logs is called event log extraction. It is a complex and time-intensive process, which requires human involvement at several stages [5]. To provide automated support for this challenge, many automated techniques have been developed. Recognizing the large variety of potential data sources and requirements in practice, available techniques differ considerably with respect to required inputs and their output. Among others, there are extraction techniques that build on ontologies [17], redo logs [18], and database objects [19]. All these techniques require human intervention or input at some stage. Unfortunately, the exact role of the human is not always clear. Since many of these techniques address rather specific problems of event log extraction, the required human involvement can often only be understood when these techniques are applied in practice. Given that the human involvement in event log extraction is both time and cost intensive, there is the need to develop a precise understanding of the respective human tasks. Furthermore, it can happen that the available raw data cannot be used at all as input to process mining techniques. This could happen due to several factors, like, for example, a lack of knowledge about the data from the system owner, of the resources needed to perform manual tasks, or due to privacy constraints.

III. Approach

As highlighted by [5], the human involvement during event log extraction is crucial, even if we do not know yet exactly when and how it should be applied. Though, the majority of the included manual activities are used to understand, enhance and transform data already available in event logs. In practice, it may happen to end up either in a situation in which we can not use at all that kind of data, or in which the organization does not dispose enough human resources to go through all the required activities. Those situations could be caused mainly by two factors: *(i)* not all organizations apply good modeling practices, documenting and logging their systems and *(ii)* event log extraction is not yet enough structured to allow an organization to know in advance the exact amount of resources that will be needed to produce the logs from data sources.

Hence, we decided to design an approach that could be applied even in the *worst case scenario* in which both factors *(i)* and *(ii)* hold, with the aim to produce a set of simulated event logs useful to feed a process mining tool to produce the structure of a data pipeline directly from the logs. The approach is based on four steps that can be repeated until the organization is not satisfied with the discovered pipeline.

A. Administering Questionnaires

The first phase of the approach involves using two questionnaires (in sequence) to understand the characteristics and the structure of the data pipelines running within the organization workflows. Both questionnaires are meant to be administered to the business experts and BP analysts of the organization. The first questionnaire includes a set of questions to identify the amount of data pipelines to discover, their execution context within the organization and the Key Performance Indicators (KPIs) under which they are usually evaluated. The second questionnaire, which is dynamically generated depending on the answers collected in the first one, includes specific forms to collect data about the type of human/software resources involved in the pipeline execution, the (known) ordering relations between the pipeline steps and other quantitative information on the duration/costs of the single steps. We will discuss the details of both questionnaires in section IV.

B. Generation of the Simulation Environment

Starting from the collected data, the second phase consists of generating a simulation environment to investigate the evolution of each pipeline of interest over time. The environment requires the presence of a model that describes (even partially) the behaviors of the selected pipeline. Depending on the amount of knowledge collected on the structure of the pipeline, its model can be represented using prescriptive notations like BPMN (Business Process Modeling Notation [20]), which are useful when the pipeline's control-flow is well known at the outset, or declarative formalism such as Declare [21]. The latter enables to focus on the formalization of few relevant control-flow constraints that must hold true during the pipeline execution. Together with the "static" pipeline model, it is also required to define the dynamics of pipeline steps through specific simulation parameters (e.g., how long does the steps take and how many resources are available), whose values can be obtained from the questionnaires. It is worth to notice that the notation to define the model and its dynamic features depend on the simulation technology being employed.

C. Running the Simulation

Once the simulation environment is ready, the third phase of the approach consists in simulating each modeled pipeline, using one of the technologies available in the literature/market [22]. This results in a simulated event log containing many possible executions of the pipeline that are complaint with the simulation parameters. While the format of the event log can be different based on the simulation technology being used, a minimum set of parameters is guaranteed to be available for each event recorded in the log (specifically, a timestamp, an activity name and a case id per event). The simulated event log is then used to automatically generate a model of the pipeline through a BP discovery technique [9].

D. Validating the simulation results

Finally, the fourth and last phase of the approach consists of interviewing both the BP analysts and the business experts

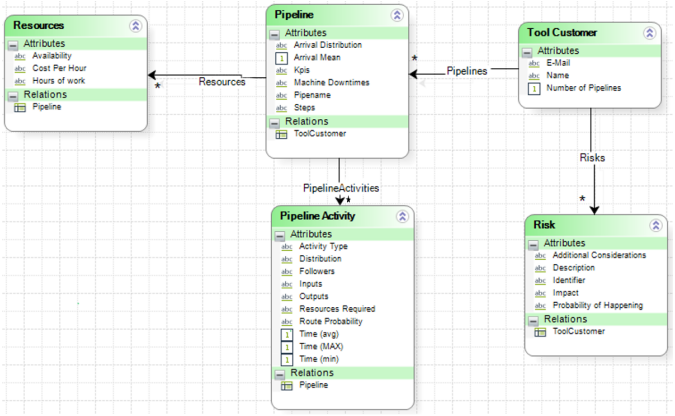

Fig. 2: *Schema of the data collected through the two questionnaires.*

of the organization to validate the results of the simulation. In practice, this phase: *(i)* highlights potential discrepancies between the discovered model of the pipelines and the real ones running behind the organization's workflow; *(ii)* leverages the KPI values obtained during the simulation process to make sure that it was in line with the reality, and *(iii)* can be used to discuss potential insights that emerge from the simulation.

IV. Questionnaires Design

The core component of the proposed approach is made up by the two questionnaires. As pointed up by [5], the literature acknowledges the importance of manual activities in event log extraction, but we miss rigorous guidelines on how and when to employ them. In response, we designed two questionnaires that must be: *(i)* general purpose, to be applicable even in the worst case scenario described in section III; *(ii)* well structured to be easily reproduced and answered; *(iii)* cheap with respect to the time and human resources that the organization needs to dispose; and *(iv)* useful even when the business experts knowledge on the pipeline is only partial. In the following, we discuss our design choices to achieve these goals.

(i). Rather than trying to assess the usefulness of the raw data available and to transform it into event logs, we focus the questionnaires on understanding the structure of the pipelines under investigation by asking to the business experts details on the pipeline steps, their average duration, the order in which they are performed and the branching probabilities (e.g., if step A is followed by step B or C, we need to know the probability of doing B, or C, after A). To build this high-level model, we rely on the knowledge owned by the business experts. As described in ***(iv)***, the approach is designed to give useful results even if such a knowledge is only partially available.

(ii). To make the questionnaires as structured as possible, the only open questions included are brief descriptions of the pipelines under investigation. Everything else is asked through closed question to make sure that both answering and reviewing the result is straightforward. Furthermore, each question comes with a brief explanation to make sure that the business expert(s) can properly answer.

(iii). We decided to split the questionnaire in two parts. The first one, which needs to be answered during an interview between the business expert(s) and the BP analyst(s) is used only to collect general information about the pipelines (e.g., the number of pipelines to investigate with a brief description, the KPIs that will be used to assess the simulation process) and about the risks that the pipeline execution can fail. Then, the second part is left to the business expert(s) to be answered offline, and will show only the questions that are known to be answered. In this way the coordinated work between BP analyst(s) and business expert(s) is reduced to the minimum.

(iv). Finally, if the information given from the organization is partial, we can simply run the other phases of the approach with approximate values, and exploit the result of the simulation to improve the next iteration. During the interview with the business expert(s), rather than asking again the information given in the previous iteration, we can look at the simulation results to infer the missing data.

The schema of the data collected through the questionnaires can be seen in Figure 2. It primarily covers: *(i) Resources*, with their availability and cost; *(ii) Pipeline Activities*, with their completion time, required resources, input, output, distribution probabilities; *(iii) Pipelines* as a whole, with their KPIs, down times and list of activities; *(iv) Risks*, with their impact level and probability of occurring.

V. Preliminary Evaluation

We performed a preliminary evaluation of the approach by applying it to one of the business cases involved in the H2020 DataCloud project (the one related to the implementation of data pipelines for massive digital marketing campaigns management). The goal was to validate: *(i)* its efficacy (e.g., if it can lead to good results or not) and *(ii)* its efficiency (e.g., how easy it is to answer the questionnaires or how good the results are compared to the disposed resources). A summary of the results is shown in Table I. We started by scheduling an online interview between the business expert and the BP analyst. Only two people were involved, and it took around 30 minutes. In this interview, the BP analyst explained to the business expert the context and the technical terminology needed to answer the questions. Discussing about the possible risks affecting the data pipelines, the two most relevant ones were identified and inserted in the table. Next, the business expert filled offline the second questionnaire. It took around 30 minutes and, during the procedure, questions and doubts about the usage of the tool were solved by consulting help

TABLE I: *Summary of the preliminary evaluation process.*

Attribute	Value
People involved	2
Time required for the organization	2 hours
Iterations required	2
Formal model	BPMN, Simio
Simulation technology	Simio
Easy to answer	Yes
Pipelines discovered	Validated after first iteration
KPIs during simulation	Realistic after second iteration

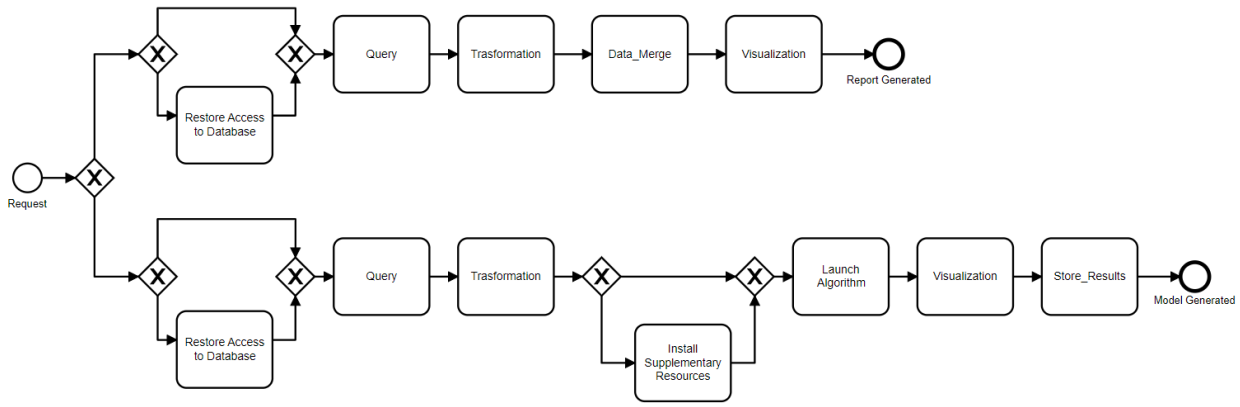

Fig. 3: *BPMN model of the pipelines under investigation.*

buttons and examples available. At the end of the procedure, personal considerations about the interview were asked to the domain expert, which found the questionnaires clear, purpose oriented and user-friendly.

The questionnaires were connected to a SQL server to store the results. Then the data were downloaded and processed through a python script. The BP analyst, by looking at the answers, started creating the simulation environment. Two pipelines were under investigation: *(i)* "Reporting Data", in which data are withdrawn and processed for performance monitoring purposes, and *(ii)* "Modeling Data", in which data are processed for the development of new modeling and predictive services for mobile marketing campaigns optimization purposes. These two pipelines shared three common resources: *(i)* SQL server; *(ii)* Cloud storage and *(iii)* PowerBI licenses. The Modeling Data pipeline required an additional resource: Jupiter Notebook licenses. The main sources of risks listed by the business expert are failures while connecting to the database or inadequate resources, like wrong data models. Then, leveraging the answers from the offline questionnaire, the BP analyst was able to create a BPMN model of the pipelines, shown in Figure 3, and a simulation environment on Simio [23], shown in Figure 4. Next, the BP analyst simulated the pipelines on Simio, generating event logs and performance measure that can be checked against the KPIs set by the business expert. Simio offers a good level of freedom in creating simulations, allowing the user to customize every single aspect of the model represented. Activities have been modeled with *servers* – generic elements used to model a point in time or space in which a unit of work is performed – and resources are modeled as entities belonging to the pipeline, with a measure of capacity – the number of units of that resource available at the same time – and of reliability – the amount of time that resource can work before needing maintenance. To model tokens, which represent the flow of the pipeline, we can use *entities*. Then, we need *sources* – the elements that generate the *entities* and introduce them into the pipeline – and *sinks* – the last point of the pipeline, where entities are destroyed. Risks have been modeled through the processes function of Simio, which allows the user to make certain events happen whenever some conditions are met. The simulations length was set to thirty days – a time frame in which it is possible to observe all different behavior of the BP in relation to the occurrences of risks – and the number of replications of the simulation was twenty, meaning that the result refers to the average of the results of all the simulations.

Finally, an online meeting between the business expert and the BP analyst was scheduled to validate the simulation results. It consisted of two phases: *(i)* the structure of the pipelines was shown to the business expert, who agreed on its fidelity, and *(ii)* the KPIs' value obtained in the simulation was shown to the business expert, who highlighted a discrepancy.

By looking at the details about the simulation environment, it was evident that some information about risks and resources were not clear enough at the outset, and that some partial information was given as approximations. Hence, the BP analyst asked the business expert to clarify the doubts, which were mainly related to internal pipeline criticalities, integration with other BP in the data pipeline (orchestration) and timing. This led to a new, more detailed, simulation environment that led to a new simulation. Finally, the KPIs newly obtained were validated with the business expert through emails, and new event logs were generated.

VI. Discussion

The main issue addressed in this paper concerns the fact that event log extraction techniques can not always be used in practice. This mainly happens primarily due to organizations not disposing of enough raw data or human resources to apply manual techniques. In response, we proposed an interactive approach to the creation of event logs that consists of iterating four phases: *(i)* Administering Questionnaires; *(ii)* Generation

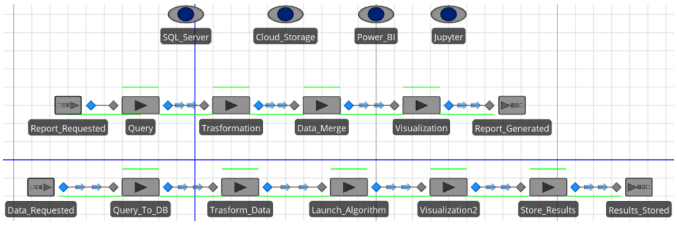

Fig. 4: *Simio environment for the pipeline simulation.*

of the Simulation Environment; *(iii)* Running the Simulation and *(iv)* Validating the simulation results. The goal of the approach is to generate a realistic simulation environment to synthesize event logs that can be passed as input to process mining techniques to realize automated pipeline discovery. Finally, we performed a preliminary evaluation by applying the approach to one real-world business case involved in the management of digital marketing campaigns. This allowed us to investigate its practical effectiveness.

On the one hand, having the opportunity of leveraging a real-world business case allowed us to easily apply the approach in a concrete scenario coinciding with the worst case described in section III. The organization kept logs with the finest possible granularity (e.g., mouse clicks), meaning that employing manual techniques to extract event logs would be too expensive (also because the organization only disposes of a single business expert, with not much time to invest in the pipeline). Furthermore, some information about the pipeline was partial or vague (e.g., the mean duration of steps). On the other hand, the organization already had good knowledge about its data pipelines and the technical terminology used in the questionnaires, and this has undoubtedly facilitated the application of the approach.

The main limitation behind the proposed approach is that the produced event logs are simulated. This has two main consequences: *(i)* it trivializes some process mining techniques, like process discovery (e.g., it would output the structure of the pipeline used to generate the logs) and *(ii)* it strictly ties the validity of output of process mining techniques to the accuracy of parameters identified in the simulation environment. Though, if we manage to get a realistic simulation environment, the ability to generate rich event logs (in terms of amount of simulated behaviours) allows to easily cover large time spans of execution without the need to wait to collect real logs. Since the approach is not tied to real-world event data, the organization can potentially test pipeline variants by simply changing the simulation environment and leveraging the simulated logs to apply process mining techniques.

VII. Concluding Remarks

Traditional process mining technologies are a great tool to understand what has happened during a BP execution, but the strict requirement of using event logs as input limits the opportunities to apply them in other contexts than BP management. In practice, event log extraction is costly and not structured enough to be applied to any context. In this paper, we have proposed an interactive approach that aims to mitigate those limitations by leveraging structured questionnaires and simulation techniques to generate event logs from domain knowledge. If the simulation environment is realistic enough, the approach would bring great benefits to process mining, allowing to obtain event logs also in worst case scenarios.

We envision three future works to improve the approach: *(i)* a validation with an organization having a not deep knowledge on its data pipelines; *(ii)* an analysis of the results that can be obtained by applying process mining techniques to simulated event logs, and *(iii)* the specification of a reference data model to define a standard way of describing the relevant properties of an event log for achieving data pipeline discovery. We will also investigate if existing AI solutions for BPM can support the semi-automated generation of simulated event logs [24].

Acknowledgments. This work has been supported by the H2020 project DataCloud and the Sapienza grant BPbots.

References

[1] W. M. Van der Aalst, *Process mining: data science in action*, 2016.
[2] S. Coltellese, F. Maria Maggi, A. Marrella, L. Massarelli, and L. Querzoni, "Triage of iot attacks through process mining," in *27th Int. Conf. on Cooperative Information Systems (CoopIS'19)*, 2019.
[3] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare: A literature review," *Journal of biomedical informatics*, vol. 61, pp. 224–236, 2016.
[4] A. Marrella, M. Mecella, and A. Russo, "Collaboration on-the-field: Suggestions and beyond," in *8th Int. Conf. on Information Systems for Crisis Response and Management (ISCRAM)*, 2011.
[5] V. Stein Dani, H. Leopold, J. M. E. van der Werf, X. Lu, I. Beerepoot, J. J. Koorn, and H. A. Reijers, "Towards understanding the role of the human in event log extraction," in *BPM'21 Workshops*. Springer, 2021.
[6] O. Oleghe and K. Salonitis, "A framework for designing data pipelines for manufacturing systems," *Procedia CIRP*, vol. 93, pp. 724–729, 2020.
[7] P. O'Donovan, K. Leahy, K. Bruton, and D. T. O'Sullivan, "An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities," *Journal of Big Data*, vol. 2, no. 1, pp. 1–26, 2015.
[8] S. Agostinelli, D. Benvenuti, F. D. Luzi, and A. Marrella, "Big Data Pipeline Discovery through Process Mining: Challenges and Research Directions," in *1st Italian Forum on Business Process Management (ITBPM'21)*, ser. CEUR Workshop Proceedings, vol. 2952, 2021.
[9] A. Augusto, R. Conforti, M. Dumas, M. L. Rosa, F. M. Maggi, A. Marrella, M. Mecella, and A. Soo, "Automated discovery of process models from event logs: Review and benchmark," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 4, pp. 686–705, 2019.
[10] C. Günther and E. Verbeek, "IEEE standard for extensible event stream for achieving interoperability in event logs and event streams," 2016.
[11] P. Vassiliadis, "A survey of extract–transform–load technology," *Int. Journal of Data Warehousing and Mining*, vol. 5, no. 3, 2009.
[12] T. Stöhr, R. Müller, and E. Rahm, "An integrative and uniform model for metadata management in data warehousing environments," in *Int. Workshop on Design and Management of Data Warehouses*, 1999.
[13] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos, "Conceptual modeling for etl processes," in *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*, 2002, pp. 14–21.
[14] J. Trujillo and S. Luján-Mora, "A uml based approach for modeling etl processes in data warehouses," in *International Conference on Conceptual Modeling*. Springer, 2003, pp. 307–320.
[15] T. Grisold, J. Mendling, M. Otto, and J. vom Brocke, "Adoption, use and management of process mining in practice," *BPM Journal*, 2020.
[16] K. Diba, K. Batoulis, M. Weidlich, and M. Weske, "Extraction, correlation, and abstraction of event data for process mining," *Wiley Int. Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, 2020.
[17] D. Calvanese, M. Montali, A. Syamsiyah, and W. M. van der Aalst, "Ontology-driven extraction of event logs from relational databases," in *Int. Conf. on Business Process Management*. Springer, 2016.
[18] E. G. L. d. Murillas, W. M. van der Aalst, and H. A. Reijers, "Process mining on databases: Unearthing historical data from redo logs," in *Int. Conf. on Business Process Management*. Springer, 2016.
[19] E. H. Nooijen, B. F. v. Dongen, and D. Fahland, "Automatic discovery of data-centric and artifact-centric processes," in *BPM*, 2012.
[20] S. A. White, "Introduction to BPMN," *IBM Cooperation*, vol. 2, 2004.
[21] M. Pesic, H. Schonenberg, and W. M. Van der Aalst, "Declare: Full support for loosely-structured processes," in *11th IEEE Int. Ent. Distributed object computing Conf. (EDOC 2007)*. IEEE, 2007.
[22] K. Rosenthal, B. Ternes, and S. Strecker, "Business Process Simulation: A Systematic Literature Review," in *ECIS*, 2018, p. 199.
[23] "Simio homepage." https://www.simio.com, 2022.
[24] A. Marrella, "What Automated Planning Can Do for Business Process Management," in *Business Process Management Workshops*, 2018.