Big Data Pipeline Discovery through Process Mining: Challenges and Research Directions[⋆]

Simone Agostinelli, Dario Benvenuti, Francesca De Luzi, and Andrea Marrella

Sapienza Universitá di Roma, Rome, Italy
firstname.lastname@uniroma1.it

Abstract. Big Data pipelines are essential for leveraging Dark Data, i.e., data collected but not used and turned into value. However, tapping their potential requires going beyond the current approaches and frameworks for managing their life-cycle. In this paper, we present the challenges associated to the achievement of the Pipeline Discovery task, which aims to learn the structure of a Big Data pipeline by extracting, processing and interpreting huge amounts of event data produced by several data sources. Then, we discuss how traditional Process Mining solutions can be potentially employed and customized to overcome such challenges, outlining a research agenda for future work in this area.

Keywords: Big Data Pipeline · Pipeline Discovery · Process Mining.

1 Introduction

With the recent developments of Internet of Things (IoT) and cloud-based technologies, massive amounts of data are generated by heterogeneous sources and stored through dedicated cloud solutions. Often organizations generate much more data than they are able to interpret, and current Cloud Computing technologies cannot fully meet the requirements of the Big Data processing applications and their data transfer overheads [4]. Many data are stored for compliance purposes only but not used and turned into value, thus becoming *Dark Data*, which are not only an untapped value, but also posing a risk for organizations [10]. Examples of Dark Data range from server log files, which can give clues related to the workflows enactment of an organization, to old files that may not seem relevant (e.g., drafts of internal projects) but are often interesting and valuable for external attackers who aim to exploit them for monetary gain.

Big Data pipelines (or simply *data pipelines*) are composite workflows for processing data with non-trivial properties, commonly referred to as the Vs of Big Data (e.g., volume, velocity, etc.) [17]. Tapping their potential is a key aspect to leverage and, at the same time, protect Dark Data [6]. In this direction, the DataCloud project[1] aims to realize novel methods and tools for effective management of the data pipeline life-cycle in the context of Cloud Computing.

[⋆]

[1] DataCloud is a Research and Innovation project funded by the European Commission under the Horizon 2020 program (Grant number 101016835). The project runs for three years, between 2021–2023. URL: https://datacloudproject.eu/

The main objective of the project is to develop a software ecosystem consisting of new languages, methods and tools for supporting data pipelines on heterogeneous resources. Six life-cycle phases will be covered: (1) pipeline discovery, (2) pipeline definition, (3) pipeline simulation, (4) resource provisioning, (5) pipeline deployment and (6) pipeline adaptation.

In this paper, we focus on the phase of *pipeline discovery*, whose target is to provide robust techniques to learn the structure of data pipelines by extracting, processing and interpreting huge amounts of event data produced by several data sources. To achieve this ambitious yet unexplored research goal, the idea is to employ (and potentially customize) existing Process Mining solutions to the discovery and analytics of data pipelines. In this paper, after presenting in Section 2 the background on data pipelines and the challenges to properly conduct the discovery task, in Section 3 we discuss how the application of existing process mining solutions can be exploited to tackle and overcome the identified challenges, towards the definition of novel approaches for pipeline discovery. Finally, in Section 4, we conclude the paper.

2 Background and Challenges on Pipeline Discovery

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 [15], a framework that reveals key layers and components to design data pipelines for manufacturing systems is presented. In [16], the authors derive a set of data and system requirements for implementing equipment maintenance applications in industrial environments, and propose an information system model that provides a scalable and resilient data pipeline for integrating, processing and analysing industrial data.

However, to date, there is no explicit research study that investigates the issue of pipeline discovery. Consequently, even if the concept of “Big data pipeline” can be traced back to 2012 (cf. [18]), the literature lacks a shared understanding of what a data pipeline is and how it can be defined. For instance, in [15] the authors refer to a data pipeline as the “path through which Big Data is transmitted, stored, processed and analyzed”. In [14], a data pipeline is defined as “a complex chain of interconnected activities from data generation through data reception, where the output of one activity becomes the input of the next one”. Similarly, in [11], a data pipeline is “a set of data processing elements connected in series, often executed in time-sliced fashion, where the output of one element is the input of the next one”. Then, in [19], data pipelines are described as a “mechanism to decompose complex analyses of large data sets into a series of simpler tasks, with independently tuned components for each task”.

The above definitions confirm that there is no unified specification of the concept of data pipeline; nonetheless, some common features that are inherently related to it can be identified:

- A data pipeline consists of chains of processing elements that manipulate and interact with data sets;
- The outcome of a processing element of a data pipeline will be the input of the next element in the pipeline;
- Each processing element of a data pipeline interacts with data sets considered as "big", i.e., with at least one of the Vs dimensions that is verified to hold.

With this knowledge at hand, we performed many rounds of interviews with the five business case partners involved in the DataCloud project (cf. also Section 4), which were useful not only to confirm the validity of the three above features that characterize a data pipeline, but also to identify four major challenges to be tackled towards the development of a robust pipeline discovery approach:

C1 **Event Data Extraction:** The challenge is to analyze and turn torrents of rough data stored in several data sources or exchanged within the underlying Cloud Computing infrastructures into valuable *event data* that reveal the events that concretely happened into day-to-day operations.

C2 **Event Log Generation:** Event data may contain interleaved information related to the enactment of different data pipelines, or of multiple instances of the same data pipeline. Moreover, the possibility exists that many events must be filtered out by the analysis, since they do not refer to processing elements that manipulate data (e.g., events that track the sending or receiving of notifications). Therefore, the generation of *event logs* from the set of event data is strongly required to (later) learn the structure of a data pipeline. Each entry of the generated event logs should possess at least the following characteristics: *(i)* a case identifier that maps each event to a case, *(ii)* a timestamp that records when the event happened, *(iii)* the processing element associated to the event, and *(iv)* the set of data processed or manipulated during the event enactment.

C3 **Pipeline Structure Learning:** This challenge is about the analysis and the interpretation of event logs to learn the pipelines' structure and to extract valuable insights related to their performance and compliance.

C4 **Dark Data Analysis:** This is the hardest part, because it requires to know what to look for and where to look within the event data, without deploying intrusive agents that manipulate the systems and networks of an organization. Identifying Dark Data through the analysis of data pipelines would enable to unlock their semantics and understand if some of them provide insights and, finally, a certain business value.

3 Pipeline Discovery through Process Mining

Even if the specification of a shared definition of a data pipeline is still a research challenge, from the previous section it is evident that many similarities exist between the concepts of "data pipeline" and "business process". With the main difference that any element of a data pipeline is thought to manipulate some (big) data set. Conversely, business processes include activities that do not necessarily interact with any kind of data. In fact, in the Business Process Management (BPM) field, data flow is usually not considered as a first-class citizen [8].

Nonetheless, the discovery of data pipelines resembles the discovery of business processes [3], as both require an event log as a starting point to enact the discovery task. For this reason, in the range of the DataCloud project, we investigate how the (customized) use of Process Mining solutions [1] may support the development of novel techniques to achieve the pipeline discovery task.

Process mining is a family of data analysis techniques that enable decision makers to discover process models from data (*process discovery*), compare expected and actual behaviours (*conformance checking*), and enrich models with information retrieved from data (*process enhancement*). Process mining focuses on the real execution of processes, as reflected by the footprint of reality logged (in the form of explicit event logs) by the software systems of an organization.

Within DataCloud, we are investigating and elaborating the following research solutions that are inspired from the process mining literature in order to tackle the challenges presented in Section 2:

S1 Human-in-the-Loop Methodology for Extracting Event Data. The literature on process mining provides a number of semi-automated methods to support organizations in extracting event data from data sources, such as PM^2 and L* [9][20]. However, often their application is hampered by the considerable preparation effort that needs to be conducted by human experts at different stages of the extraction procedure [7]. This issue is even more severe in presence of heterogeneous data sources that store huge amount of data, like in the case of data pipelines. Furthermore – to date – there is no deep understanding of how human experts should be involved in the process of event data extraction. Within DataCloud, we aim to tackle **C1** by enhancing the existing event data extraction methods through the identification and specification of the manual activities that the human experts need to perform in the context of event data extraction. This will include, for example, activities like the assessment of the quality of the available data sources, the detection of those data elements that relate to events, etc.

S2 Pre-processing, Clustering and Filtering techniques. To tackle **C2**, i.e., to reduce the overall dataset complexity by extrapolating only its relevant fragments for an effective event log generation, we will work on the realization of four techniques: *(i) Segmentation* pre-processes the event data to identify the events that belong to the same pipeline (i.e., case); *(ii) Aggregating* events reduces complexity and improves the structure of discovery results by merging multiple events into larger ones; *(iii) Clustering* partitions the event log to discover simpler models for each partition of a complex pipeline; *(iv) Filtering* removes potential outliers from the log. Concerning *(i)*, *(ii)* and *(iii)*, the literature on process mining provides many solutions that can be potentially customized and re-used in the context of data pipelines [12]. On the other hand, segmentation is a rather unexplored topic in process mining, since it is assumed that any event in a log is always associated to a known case. In practice, the majority of information systems do not record case identifiers explicitly. To mitigate this issue, we will leverage our previous works on segmentation performed in the Robotic Process Automation field [2] to semi-automatically detect the different pipeline cases from a log.

S3 **Pipeline Discovery algorithm.** To learn the structure of a data pipeline, we aim to leverage existing process discovery algorithms from process mining [3], enhancing them with other techniques coming from different areas, ranging from data mining to automated planning in Artificial Intelligence, like already experienced in [13]. Our solution is to realize a pipeline discovery algorithm that enables not only to efficiently build the sequence flow of the discovered pipelines, but also learning all data flows and event-based conditions that ruled their execution.

S4 **Conformance Checking technique for Dark Data analysis.** To tackle **C4**, we aim to customize existing conformance checking techniques [5] to replay the streams of event data filtered out during the event log generation phase (and stored into a dedicated Dark database) over the structure of the discovered data pipelines. The target is to understand if some of the discarded data can be potentially exploited to improve the quality or the business value of the identified data pipelines. Of course, the definition of specific threshold values to quantify if a dark data should be restored in an event log must be investigated and specified as well.

4 Concluding Remarks

The expected impact of the DataCloud project is to lower the technological entry barriers for the incorporation of Big Data pipelines in organizations' workflows and make them accessible to a wider set of stakeholders regardless of the hardware infrastructure. In this context, we discussed the key considerations around the concept of pipeline discovery and suggested a number of research challenges and potential ways to tackle them employing process mining solutions, to serve as a research agenda for the future.

All the proposed solutions for pipeline discovery will be validated through a strong selection of complementary business cases offered by four SMEs and a large company targeting higher mobile business revenues in smart marketing campaigns, reduced live streaming production costs of sport events, trustworthy eHealth patient data management, and reduced time to production and better analytics in Industry 4.0 manufacturing.

Acknowledgments. This work has been supported by the Horizon 2020 project DataCloud (Grant number 101016835).

References

1. van der Aalst, W.: Data Science in Action, pp. 3–23. Springer (2016)

2. Agostinelli, S., Marrella, A., Mecella, M.: Automated Segmentation of User Interface Logs. In: Robotic Process Automation, pp. 201–222. De Gruyter (2021)
3. Augusto, A., Conforti, R., Dumas, M., Rosa, M.L., Maggi, F.M., Marrella, A., Mecella, M., Soo, A.: Automated Discovery of Process Models from Event Logs: Review and Benchmark. IEEE Trans. on Know. and Data Eng. **31**(4) (2019)
4. Barika, M., Garg, S., Zomaya, A.Y., Wang, L., Moorsel, A.V., Ranjan, R.: Orchestrating big data analysis workflows in the cloud: Research challenges, survey, and future directions. ACM Comput. Surv. **52**(5) (Sep 2019)
5. Carmona, J., van Dongen, B., Solti, A., Weidlich, M.: Conformance checking. Springer (2018)
6. Chakrabarty, S., Joshi, R.S.: Dark Data: People to People Recovery. In: ICT Analysis and Applications, pp. 247–254. Springer (2020)
7. Diba, K., Batoulis, K., Weidlich, M., Weske, M.: Extraction, correlation, and abstraction of event data for process mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **10**(3), e1346 (2020)
8. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A., et al.: Fundamentals of Business Process Management, vol. 1. Springer (2013)
9. van Eck, M.L., Lu, X., Leemans, S.J.J., van der Aalst, W.M.P.: Pm^2: A process mining project methodology. In: Zdravkovic, J., Kirikova, M., Johannesson, P. (eds.) Advanced Information Systems Engineering. pp. 297–313. Springer International Publishing, Cham (2015)
10. Gimpel, G.: Bringing dark data into the light: Illuminating existing IoT data lost within your organization. Business Horizons **63**(4), 519–530 (2020)
11. Gressling, T.: Data Science in Chemistry: Artificial Intelligence, Big Data, Chemometrics and Quantum Computing with Jupyter. De Gruyter (2020)
12. Mannhardt, F., de Leoni, M., Reijers, H.A., van der Aalst, W.M., Toussaint, P.J.: Guided process discovery–a pattern-based approach. Information Systems **76** (2018)
13. Marrella, A., Lespérance, Y.: Synthesizing a library of process templates through partial-order planning algorithms. In: Enterprise, Business-Process and Information Systems Modeling, pp. 277–291. Springer (2013)
14. Munappy, A.R., Bosch, J., Olsson, H.H.: Data pipeline management in practice: Challenges and opportunities. In: International Conference on Product-Focused Software Process Improvement. pp. 168–184. Springer (2020)
15. Oleghe, O., Salonitis, K.: A framework for designing data pipelines for manufacturing systems. Procedia CIRP **93**, 724–729 (2020)
16. O'Donovan, P., Leahy, K., Bruton, K., O'Sullivan, D.T.: An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. Journal of Big Data **2**(1), 1–26 (2015)
17. Plale, B., Kouper, I.: The centrality of data: data lifecycle and data pipelines. In: Data analytics for intelligent transportation systems, pp. 91–111. Elsevier (2017)
18. Rabl, T., Jacobsen, H.A.: Big data generation. In: Rabl, T., Poess, M., Baru, C., Jacobsen, H.A. (eds.) Specifying Big Data Benchmarks. pp. 20–27. Springer (2014)
19. Raman, K., Swaminathan, A., Gehrke, J., Joachims, T.: Beyond myopic inference in big data pipelines. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 86–94 (2013)
20. Van Eck, M.L., Lu, X., Leemans, S.J., Van Der Aalst, W.M.: PM^2: a process mining project methodology. In: International Conference on Advanced Information Systems Engineering. pp. 297–313. Springer (2015)