



An Event Data Extraction Approach from SAP ERP for Process Mining

Alessandro Berti^{1,2(✉)}, Gyunam Park¹, Majid Rafiei¹,
and Wil M. P. van der Aalst^{1,2}

¹ Process and Data Science Group (PADS), RWTH Aachen University, Aachen, Germany

a.berti@pads.rwth-aachen.de

² Fraunhofer Gesellschaft, Institute for Applied Information Technology (FIT), Sankt Augustin, Germany

Abstract. The extraction, transformation, and loading of event logs from information systems is the first and the most expensive step in process mining. In particular, extracting event logs from popular ERP systems such as SAP poses major challenges, given the size and the structure of the data. Open-source support for ETL is scarce, while commercial process mining vendors maintain connectors to ERP systems supporting ETL of a limited number of business processes in an ad-hoc manner. In this paper, we propose an approach to facilitate event data extraction from SAP ERP systems. In the proposed approach, we store event data in the format of object-centric event logs that efficiently describe executions of business processes supported by ERP systems. To evaluate the feasibility of the proposed approach, we have developed a tool implementing it and conducted case studies with a real-life SAP ERP system.

Keywords: SAP · ETL · Process Mining · Object-Centric Event Logs

1 Introduction

Process mining is a branch of data science including techniques to discover process models from event data, so-called *process discovery*, check the compliance of data against the process models, so-called *conformance checking*, and enhance process models with constraints/information coming from the event logs, so-called *enhancement*. Such techniques have been adopted by various domains, including healthcare, manufacturing, and logistics. The first step of applying the techniques is to extract event logs from the target information systems, e.g., Enterprise Resource Planning (ERP) systems. This usually requires a connection to the database(s) supporting the information system. Afterward, the extracted event log undergoes pre-processing steps to resolve various data quality issues, including incomplete information, noise, etc. These steps are usually called ETL (Extraction, Transformation, and Load). The ETL phase is usually the most time-consuming part of a process mining project [12].

© The Author(s) 2022

J. Munoz-Gama and X. Lu (Eds.): ICPM 2021 Workshops, LNBP 433, pp. 255–267, 2022.

https://doi.org/10.1007/978-3-030-98581-3_19

ERP systems contain valuable data based on which process mining techniques provide insights regarding the underlying real-life business processes. In particular, the SAP ERP system has a significant share in the ERP market (22.5% in 2017, Gartner). Extracting data from an SAP ERP system is particularly challenging as it involves many different tables/objects. Due to its complexity, support to extracting event data from the SAP ERP system has only been limited to commercial vendors, e.g., Celonis and ProcessGold, which requires extensive interaction with domain experts. Moreover, the logs extracted by such extractors suffer from convergence/divergence problems [1]. This is due to the necessity to specify a *case notion*. A case notion is a criteria to group events that belongs to the same execution of a business process. In ERP systems, different case notions can be used for the same data. For example, in a procure-to-pay process, we could specify as case notion the order, the single item of the order, the delivery, the invoice, or the payment.

This paper proposes a novel approach to guide and ease the extraction of event logs from SAP ERP. The approach consists of two phases, i.e., 1) building *graph of relations* and 2) extracting object-centric event logs. We propose to use Object-Centric Event Logs (OCEL) as intermediate storage to collect the events extracted from different tables. OCEL does not require the specification of a case notion. Therefore, it provides flexible and comprehensive event data extraction. OCEL can be used with Object-Centric Process Mining (OCPM) techniques or flattened to traditional event logs by selecting a case notion out of objects. The proposed approach has been implemented as a prototypical extractor and evaluated using an SAP ERP system.

The rest of the paper is organized as follows. Section 2 presents some background knowledge. Section 3 presents the proposed approach. Section 4 presents a prototypical software implementing the ideas proposed in this paper. Section 5 evaluates the processes extracted by the prototypical software on top of an educational SAP instance. Section 6 presents the related work on extracting and analyzing event logs from SAP.

2 Background

This section presents some background knowledge on OCEL, convergence/divergence problems, and SAP systems.

2.1 Object-Centric Event Logs

Traditional event logs in process mining have events associated with a single case/process execution. These event logs, extracted from information systems, suffer from convergence/divergence problems [1]. We have a *convergence* problem when the same event is duplicated among different instances. This happens, for example, in an order-to-cash process, when *item* is considered as the case notion, and an event of order creation can be associated with several items. We have a *divergence* problem when several instances of the same activity happen in a case

while not being causally related. This happens, for example, in an order-to-cash process, when *order* is considered as the case notion, and several instances of the same item-related activity are contained in the same order.

OCEL relax the assumption that an event is associated with a single case. Instead, in an OCEL an event can be related to several objects, where every object is associated with a type. This results in a more natural way to extract event data from a modern information system. For example, in ERP systems, the event of order creation can currently involve an order document and several items. This resolves the convergence problem (since we do not need to duplicate the events anymore) and the divergence problems (since activities related to items of an order are not associated with the case of the general order).

Recently, the OCEL standard¹ has been proposed as the mainstream format for storing object-centric event logs [5]. The format is supported by different implementations and libraries in various programming languages, e.g., Java (ProM framework) and Python. OCEL can be used to discover object-centric process models [2,3], which describe the lifecycle of different object types and their interactions. Moreover, conformance checking can be done on multiple object types [2].

2.2 SAP: Entities and Relationships

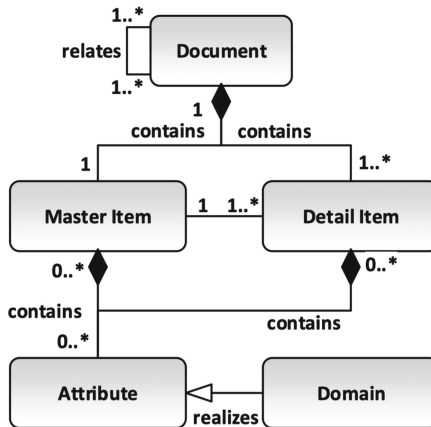


Fig. 1. Core entities of SAP ERP systems in UML 2.0 class diagram

In a broader sense, SAP ERP can be seen as a document management system. Therefore the concept of *document* is particularly important. Figure 1 introduces the document and its relevant entities and relationships among them, using UML 2.0 class diagram. First, a document represents a core business object, including

¹ <http://www.ocel-standard.org/>.

orders, deliveries, and payments. Each document contains a *master item* and *detail items*. For instance, a delivery document contains a delivery master item, corresponding to an order, and multiple delivery detail items, corresponding to materials in the order. A *master table* is a collection of the same type of master items, whereas a *detail table* is a collection of the same type of detail items. For instance, *EKKO* as a master table contains purchase order master items. *EKPO* as a detail table contains purchase orders detail items.

Both master and detail items contain a various number of *attribute values*, e.g., the total cost of a document or the cost of a single item. Each attribute belongs to a *domain* that encodes the type of information reported by the attribute, e.g., creation date and posting date of a document share the same domain because they are both dates.

3 Extracting Event Data from SAP ERP: Approach

Figure 2 describes an overview of our proposed approach to extract OCEL from SAP ERP systems. It consists of two phases: 1) building *Graph of Relations (GoR)* and 2) extracting OCEL. The former aims to construct a graph that describes all relevant tables of a business process. There are well-known business processes in SAP ERP, e.g., Purchase to Pay (P2P) and Order to Cash (O2C). For such business processes, target tables, where we extract event data regarding the process, are already known, e.g., *EKKO*, *RBKP*, *EKBE* for P2P and *VBAK*, *BKPF* for O2C. However, most business processes in an organization are mostly unknown and, thus, require the identification of relevant tables.

Based on the GoR, we extract OCEL by connecting them to the underlying database of SAP ERP systems. To this end, we first preprocess records of tables described in the GoR. Next, we define activity concepts relevant to the target business process using the relevant tables. Finally, based on the activity concept, we extract event data from the relevant tables.

3.1 Building Graphs of Relations

Figure 3 shows the conceptual model of three GoRs, each of which corresponds to a business process. A GoR is an undirected connected graph where the nodes are SAP tables containing the potentially interesting information and the edges show a relation among two tables based on a joint field/column. The node in the center of a GoR is a master table that is most relevant to the target process. The distance of each node from the master table shows the relevancy of the information contained in the corresponding table to the tables of interest and consequently to the corresponding type of process. Different colors in a GoR indicate different classes of tables. Each class has a unique way of defining activity concepts. As a result, different GoRs may be connected to each other. Below are the steps to construct GoRs:

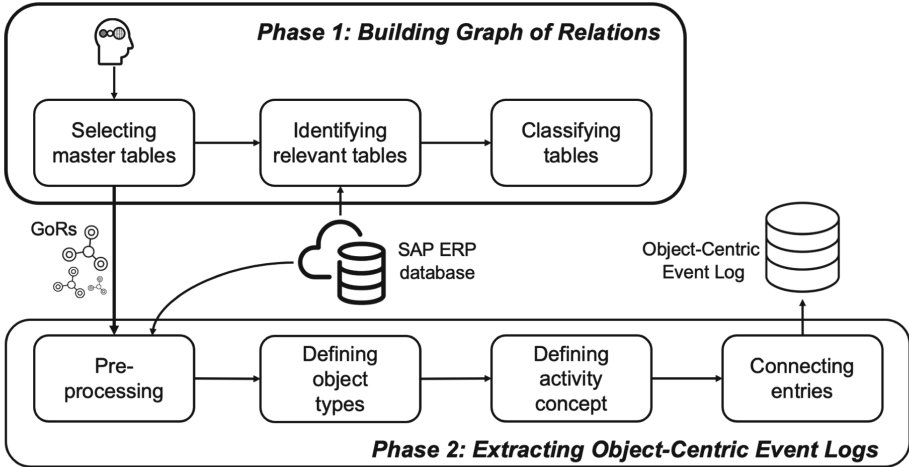


Fig. 2. Overview of extracting object-centric event logs from SAP ERP systems

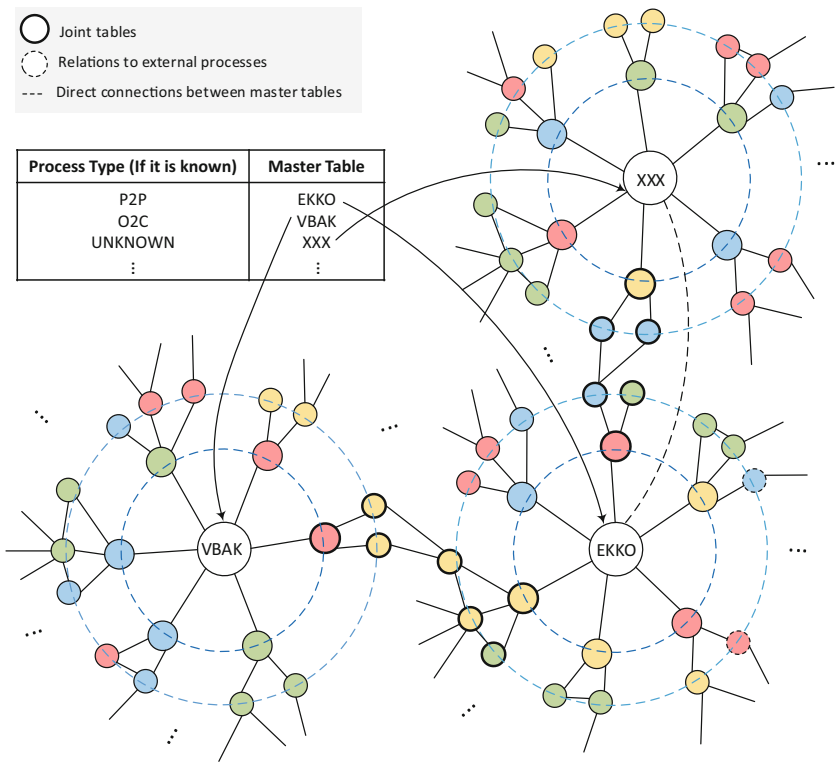


Fig. 3. Conceptual model of Graph of Relations (GoRs)

Selecting Master Tables. A GoR is built upon a master table relevant to a business process to analyze. In this work, we consider relevant master tables as users' input.

Identifying Relevant Tables. Based on the given master table, we need to identify relevant tables to the master tables. Such tables become the candidates for constructing the GoR. Three different main approaches may be taken: *manual*, *automatic*, and *hybrid*.

- In the manual approach, the identification is conducted by domain experts who understand business processes and the technical details of SAP systems. In addition, the domain expert may provide a data schema to explain the entities and relationships among them.
- In the automatic approach, the identification is made automatically by exploiting existing information in the system. For instance, using the table *DD03VV*, one can extract the relationships between the tables.
- Finally, the hybrid approach exploits both manual and automated techniques. For instance, the data schema from domain experts can provide an initial set of relevant tables, which will be improved by including more relevant tables with the help of automatically generated relationships.

Classifying Tables. The last step is the classification of the identified tables into different classes. In the following, we describe five different classes.

- A *flow table* describes the status of objects that compose the target business process. It explains the creation, deletion, and update of such objects, e.g., VBFA explains the status of objects that are associated with the Order-to-Cash (O2C) process.
- A *transaction table* describes the execution of transactions (TCODE) in SAP systems.
- A *change table* describes the changes in objects of the target business process, e.g., CDHDR and CDPOS are primary change tables.
- A *record table* stores relevant attributes of objects of the target business process, e.g., the table EKKO contains the relevant attributes of purchase order documents.
- A *detail table* stores the relationships between different entities, e.g., the table EKPO stores the connection between purchase requisitions and purchase orders.

3.2 Extracting Object-Centric Event Logs

In this subsection, we explain how OCEL are extracted using GoRs. The extraction consists of four main steps; *pre-processing*, *defining activity concept*, *defining object types* and *connecting entries*.

Pre-processing. SAP tables contain a lot of data related to different companies or groups in the same company (multi-tenant system). Moreover, when invoicing/accounting tables are considered, documents are organized by their fiscal year. A pre-processing step must be performed to extract an event log of reasonable size, containing the desired behavior and a coherent set of information since document identifiers can be replicated across different organizations. To this end, the union of all the fields in the primary keys of the tables is considered, and for some of them, a filtering query is executed, e.g., on a specific company code or a specific fiscal year.

Defining Object Types. During the extraction, the entries of the master tables are transformed into events, having the columns as event attributes. Moreover, the values of all the columns except the dates and the numbers become objects of the object type given by the column's name.

Defining Activity Concept. To extract event data from GoRs, we take a divide-and-conquer approach. We first extract event data from each table and then combine them. The first step of extracting event data from each table is to define the activity concept. In the following, we explain how the activity is defined in each class of tables.

- Each row in flow tables contains a current document number, a previous document number, the type of the current document, and the type of the previous document. For instance, considering VBELN as the domain, VBELN, VBELV, VBTYP_N, and VBTYP_V in the VBFA table contain respectively the current document number, the previous document number and the current and previous document types. We define activities as the type of the current documents, i.e., the value in VBTYP_N.
- Each row in transaction tables contains a transaction code. We transform the transaction code into human-readable formats using the TSTCT table, e.g., VA02 is transformed to Change Order, which becomes the activity name.
- Each row in record tables describes the properties of an object. All the rows of the record tables are associated with the same activity, e.g., *Create document [...]* for all the rows in EKKO.
- For change tables, we suggest three approaches: (1) Transaction codes used for changes are transformed into activities, (2) Fields, updated after changes, are converted into activities, e.g., *Price Changed*, and (3) We consider both old and new fields' values and define activities, e.g., *Postpone Delivery*, by comparing old and new values of delivery dates.

Connecting Entries. In this step, the information of the detail tables is used to enrich events. For example, if an entry of the table RSEG, containing detailed information about invoices, associates an invoice identifier with an order identifier, every event associated with the invoice identifier is also associated with the order identifier in the subsequent step.

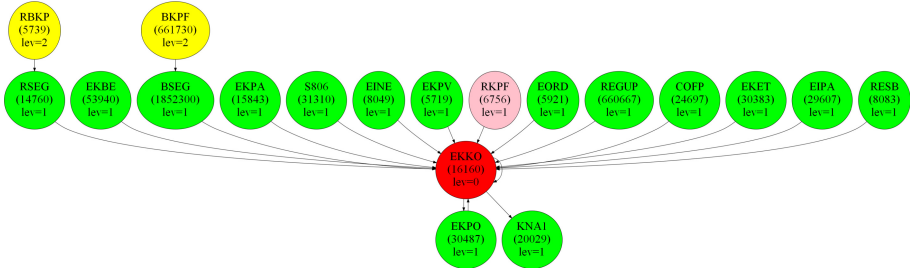


Fig. 4. A GoR built on our SAP IDES instance on the P2P process. Detail tables are colored by green, RKPF that is an additional record tables is colored by pink, and RBKP and BKPF that are additional transaction tables are colored by yellow. (Color figure online)

4 Extracting Event Data from SAP ERP: Tool

We implemented a tool in the Python3 language, available in the Github repository; <https://github.com/Javert899/sap-extractor>. The tool is available as a web application implemented using the Flask framework and can be launched with the command; *python main.py*. The web application can be accessed at the address; http://localhost:5000/new_extractor.html. First, the extractor asks the parameters of connection to the database supporting the SAP ERP instance. Then, it provides both a list of object classes contained in the database and a list of pre-configured sets of tables related to the mainstream processes. The next step is the construction of the GoR, which permits extending the set of tables. The following step is about pre-providing the values for the primary keys of the included tables, e.g., the client used during the connection and the fiscal year. After this step, the identification of the type of tables and the extraction occurs, which permits obtaining an OCEL, that can be flattened to a traditional event log or analyzed using object-centric techniques such as the ones provided in <https://github.com/Javert899/sap-extractor>.

5 Assessment

This section proposes an assessment of the proposed techniques on top of an SAP ERP IDES system. In particular, we will target the extraction of the well-known Purchase to Pay (P2P) system. A P2P process involves different steps including *approval of a purchase requisition, placement of a purchase order, invoicing from a supplier, and payment*. Therefore, it involves different tables in the SAP system.

5.1 Building a Graph of Relations

Selecting Master Tables. The first step in the tool is selecting a candidate table related to the process. In this case, we start from EKKO that is one of the

main tables in the P2P process and contains the master information. In building the GoR, represented in Fig. 4, several other tables that are connected to *EKKO* are found. Given the vast number of tables contained in SAP, we applied a simple filtering based on the number of entries in each table to show the main nodes in the GoR.

Identifying Relevant Tables. Figure 4 shows other tables containing event data meaningful to extract an event log for the P2P². The user needs to specify the tables to include along with the original set of tables. The GoR is therefore updated³. In our implementation, the master tables related to the detail tables are automatically included in the set⁴.

Classifying Tables. The tool needs to categorize the tables in the set between master tables and detail tables, as the master tables contain event data, while detail tables contain the connection between different entities:

- Some tables are recognized as transactions tables: *RBKP* (containing the transactions related to the invoices) and *BKPF* (containing the transaction related to the payments).
- Some tables are recognized as record tables: *EBAN* (in which a record is a purchase requisition), *EKKO* (in which a record is an order document), and *RKPF* (in which a record is a reservation).
- Some tables are recognised as detail tables: *EKPO*, *EKPA*, *EKET*, *EKBE*, *BSEG*, *RSEG*, *RESB*⁵.

5.2 Extracting Object-Centric Event Logs

In this section, we will explain the main steps of the log extraction process, including the definition of the object types and the activity concept for the extraction, and the connection between the entries given the information of the detail tables. Since we did not perform a pre-processing step, we will not assess the step here.

² Including *EKBE*, containing goods/invoice receipts, *BSEG*, containing detail table for payments, *RSEG*, containing detail data for invoices, *RKPF*, including inventory management data, *EKPO*, containing the detailed information about the purchase orders, *EKPA*, containing the partner roles in purchasing, and *EKET*, containing the scheduling agreement. We can see that the *EBAN* table, containing purchase requisition data, has not been included because of the filtering applied on the number of entries. However, it would be found by the method if the threshold is set to a lower value so that we will include it in the following steps.

³ The set of tables to extract include: *EKKO*, *EKPO*, *EKPA*, *EKET*, *EKBE*, *BSEG*, *BKPF*, *RSEG*, *RBKP*, *RKPF*, *RESB*, *EBAN*.

⁴ This means that *BKPF*, the master table of *BSEG*, containing the master data about the payments, and *RBKP*, the master table of *RSEG*, containing the master data about the invoices, are included.

⁵ Because their primary key is contained in the primary key of *EKKO* (for *EKPO*, *EKPA*, *EKET*, *EKBE*), *BKPF*, *RBKP* and *RKPF* respectively.

Defining Object Types. Starting from the choices on the GoR and the identification of the type of tables, it is possible to extract different object types, including *BANFN-BANFN* (purchase requisition), *INFNR-INFNR* (purchasing record), *EBELN-EBELN* (purchase order), *BELNR-RE-BELRN* (the invoice number), *BELNR-BELNR_D* (the payment number), and *AWKEY-AWKEY* (a generic object type containing the ID of the object in SAP).

Defining Activity Concept. The activity concept is defined as follows:

- For the record tables, a unique activity is defined for all the events, that is *Create document (TABNAME)* (where TABNAME is the name of the corresponding record table, so it can be EBAN/EKKO/RKPF).
- For the transaction tables, the activity is given by the transaction code⁶. Mainstream transactions occurring are *Enter incoming invoice*, *Enter incoming payment*, *Enter outgoing payment*.

Connecting Entries. The detail tables are used to enrich the entries extracted from the master tables as follows:

- *BSEG* provides a connection from the payments to the purchase order items.
- *RSEG* connects the invoices to the purchase order items.
- *EKPO* provides a connection of the purchase order items to the corresponding purchase requisition.
- *EKPA* and *EKET* contain detailed information that does not provide meaningful links to other tables in the set. *EKBE* is a peculiar type of detail table, as it contains the information about goods/invoice receipts, so it could be seen as a master table. Still, it also links the purchase order items with the invoices through the goods/invoice receipts.

6 Related Work

This section presents the related work on data extraction from ERP systems for process mining purposes.

Data Extraction and Pre-processing from SAP ERP. In [6], an approach to extract traditional event logs from SAP ERP is proposed. The set of relevant business objects is identified, and the related tables and their relations are identified. A limitation is that the construction of the document flow is manual. In [7], the authors address the pre-processing challenges to extract event logs from SAP ERP by using tools such as EVS Model Builder. In [4], an ontology-driven approach for the extraction of event logs from relational databases is proposed, in which the user can express semantic queries which are then translated to relational queries. In [8], the effects of some decisions on the quality of the resulting event log are analyzed. In particular, the context of event log extraction from ERP system is considered.

⁶ Using the description of the transaction contained in the table TSTCT.

Artifact-Centric Models on ERP Systems. In [10], an approach to discover artifact-centric models from ERP systems is proposed. The approach is split into two main parts: 1) identifying a set of artifacts, extracting a traditional event log, and a model of its lifecycle; 2) discovering the interactions between artifacts. The set of tables to extract needs to be decided by the user and the specification of the activity concepts is not described in this work.

In [9], object-centric event logs (in the XOC format) are extracted from the Dollibar ERP system. These logs have been used to generate an object-centric behavioral constraints (OCBC) model. However, OCBC/XOC are not scalable.

OpenSLEX Meta-models. In [11], a meta-model is proposed to ease the extraction of process mining event logs from information systems supported by relational databases. The instances of the OpenSLEX meta-model can be built from different types of database logs (redo logs, SAP change tables). Hence, the meta-model is generic and not tailored to the peculiar features of an SAP ERP system. The main problem is that the extraction of an event log requires a case notion's specification, which leads to convergence/divergence problems.

Enterprise-Grade Connectors. Several commercial vendors of process mining solutions offer enterprise-grade connectors to SAP, that are able to ingest and process millions of events. Notable examples in the current landscape are Celonis⁷, Signavio⁸, LANA⁹, UIPath¹⁰.

7 Conclusion

In this paper, we proposed a generic approach to extract event logs from SAP ERP, which exploits the relationships between tables in SAP to build Graphs of Relations (GoRs) and obtains Object-Centric Event Logs (OCEL) using GoRs. Figure 2 summarizes our approach. By storing extracted event data into OCEL, we permit the specification of multiple case notions, avoiding the convergence/divergence problems and simplifying the extraction process. An open-source tool implementing the approach and a case study on an educational SAP instance have been presented, showing the feasibility of identifying the relationships between different tables of the P2P process and extracting corresponding OCEL. As future work, we plan to deploy our approach on different instances of SAP systems running in real businesses to explore the connection between GoRs and underlying processes and to discover unknown processes. Moreover, we should further assess how good the extraction of a typical SAP process is in comparison to commercial-grade extractors.

⁷ <https://www.celonis.com/solutions/systems/sap>.

⁸ <https://www.signavio.com/products/process-intelligence/>.

⁹ <https://lanalabs.com/en/migration-to-sap-s-4-hana-with-lana/>.

¹⁰ <https://docs.uipath.com/process-mining/docs/introduction-to-sap-connector>.

References

1. Aalst, W.M.P.: Object-centric process mining: dealing with divergence and convergence in event data. In: Ölveczky, P.C., Salaün, G. (eds.) SEFM 2019. LNCS, vol. 11724, pp. 3–25. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30446-1_1
2. van der Aalst, W.M., Berti, A.: Discovering object-centric Petri nets. *Fundam. Inform.* **175**(1–4), 1–40 (2020)
3. Berti, A., van der Aalst, W.: Extracting multiple viewpoint models from relational databases. In: Ceravolo, P., van Keulen, M., Gómez-López, M.T. (eds.) SIMPDA 2018–2019. LNBIP, vol. 379, pp. 24–51. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-46633-6_2
4. Calvanese, D., Montali, M., Syamsiyah, A., van der Aalst, W.M.P.: Ontology-driven extraction of event logs from relational databases. In: Reichert, M., Reijers, H.A. (eds.) BPM 2015. LNBIP, vol. 256, pp. 140–153. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42887-1_12
5. Ghahfarokhi, A.F., Park, G., Berti, A., van der Aalst, W.M.P.: OCEL: a standard for object-centric event logs. In: Bellatreche, L., et al. (eds.) ADBIS 2021. CCIS, vol. 1450, pp. 169–175. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85082-1_16
6. van Giessel, M.: Process Mining in SAP R/3: A Method for Applying Process Mining to SAP R/3. Eindhoven University of Technology, Eindhoven, The Netherlands (2004)
7. Ingvaldsen, J.E., Gulla, J.A.: Preprocessing support for large scale process mining of SAP transactions. In: ter Hofstede, A., Benatallah, B., Paik, H.-Y. (eds.) BPM 2007. LNCS, vol. 4928, pp. 30–41. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78238-4_5
8. Jans, M., Soffer, P.: From relational database to event log: decisions with quality impact. In: Teniente, E., Weidlich, M. (eds.) BPM 2017. LNBIP, vol. 308, pp. 588–599. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74030-0_46
9. Li, G., de Murillas, E.G.L., de Carvalho, R.M., van der Aalst, W.M.P.: Extracting object-centric event logs to support process mining on databases. In: Mendling, J., Mouratidis, H. (eds.) CAiSE 2018. LNBIP, vol. 317, pp. 182–199. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92901-9_16
10. Lu, X., Nagelkerke, M., Van De Wiel, D., Fahland, D.: Discovering interacting artifacts from ERP systems. *IEEE Trans. Serv. Comput.* **8**(6), 861–873 (2015)
11. de Murillas, E.G.L., van der Aalst, W.M.P., Reijers, H.A.: Process mining on databases: unearthing historical data from redo logs. In: Motahari-Nezhad, H.R., Recker, J., Weidlich, M. (eds.) BPM 2015. LNCS, vol. 9253, pp. 367–385. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23063-4_25
12. van Eck, M.L., Lu, X., Leemans, S.J.J., van der Aalst, W.M.P.: PM²: a process mining project methodology. In: Zdravkovic, J., Kirikova, M., Johannesson, P. (eds.) CAiSE 2015. LNCS, vol. 9097, pp. 297–313. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19069-3_19

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

