# Process mining and data warehousing – a literature review

**Snježana Križanić**
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin
skrizanic@foi.hr

**Kornelije Rabuzin**
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin
krabuzin@foi.hr

***Abstract.*** *Process mining is a set of techniques which allow the automatic process analysis. Databases are places where all the data about process execution are stored. Data warehouses are databases of a special kind which collect and store data from different heterogeneous sources, including the data about processes. This article explores the existing literature about two related areas: process mining and data warehouses. The aim of this article is to research the methodologies used in previous researches, find out in which areas process mining combined with data warehousing is used, and what kind of tools researchers used to perform analysis.*

**Keywords.** process mining, process model, databases, data warehouses.

## 1 Introduction

In a time of changing work environment, possession of right information is important for business decision making. Information systems are used to support business processes in organizations, and business processes help to achieve the goals of the organizations. The flow and the success of business process execution need to be monitored and analysed with the aim of improvement.

Process mining is a set of techniques which allow the automatic process analysis. Process mining uses event logs which consist of events that are recorded during the process execution.

With the aim to produce useful information, huge amount of data needs to go through three stages:

1. transformation from a workflow log format into a format convenient for analysis,
2. loading into a data warehouse,
3. analysis (Koncilia, Pichler & Wrembel, 2015).

Data warehousing is an area which allows gathering business process information from different sources (González López de Murillas, Reijers & Van der Aalst, 2019). "A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of manager's decision making process" (Ghosh, Haider & Sen, 2015).

When we talk about process mining and data warehouses, there are two sides of a common model. The data side consists of next three elements: versions, objects and data models. The process side consists of: processes, instances and events (González López de Murillas, Reijers & Van der Aalst, 2019). Objects represent the appearance of an active actor in the process (Kassem & Turowski, 2018). A process model that is discovered from enhanced logs strives to become a generic model. That kind of generalization of the process model is usually popular to spaghetti-like models, or in situation when a business analyst wants to measure some key performance figures (Jareevongpiboon & Janecek, 2013). At the data level, e.g., different contexts by different content keywords, resources and people, involved in the events, can be distinguished (Štajner, Mladenić & Grobelnik, 2010).

It is expected that business analysts know about the features of how event data are stored, and that is an important challenge to overcome for every user (González López de Murillas, Reijers & Van der Aalst, 2019).

The motivation for writing this article stems from a curiosity to identify opportunities for possible contribution and improvement in future work in the domains of process mining and data warehousing.

The research aims are:

1. to identify the connection between process mining and data warehousing,
2. to identify the reasons for using data warehouses in the field of process mining,
3. determine the limitations of the data warehouse usage in process mining.

The purpose of this article is to detect a research idea for further work based on the results of this research.

Research questions that we are trying to answer in this article are:

1. Which methodologies are used in previous researches where the data warehousing terminology is used together with process mining?

2. What are the areas of research where the process mining is used as a technique for data warehousing?
3. Which tools are used or researched in previous researches that enable process mining in data warehouses?

This article is designed as follows: section 2 represents the literature review, where the basic ideas of process mining and data warehousing are presented. Section 3, Methodology, shows the methodology used in this article with the aim of gathering information about previous works. Then, there is a section 4 with results obtained from literature research. This article is also concluded with ideas about future research.

# 2 Literature review

Information systems, which support business processes, generate a huge amount of information about events which occurred during system execution. This information can help in analysis of business operations and performance over time (Belo et al., 2017). The process related data, which are arranged over multiple tables in the company's database, are assembled and transformed to the format that can be used in the process mining technique (Jareevongpiboon & Janecek, 2013). The role of process mining stays in a function how the data plays, and it is one of the most popular and latest technology for data scientist process analysis. In order to better understand business processes and improve them, there are some techniques that process mining provides. These techniques are: process discovery, conformance checking, compliance checking, performance analysis, process monitoring, prediction, operational support etc. (González López de Murillas, Reijers & Van der Aalst, 2019). Discovery is a step of searching through event logs according to which process model can be designed. Conformance can consist of an associated process model with an event log record. The aim of the conformance is to discover if there exist any inconsistencies, and commonalities between the modelled performance, and the empirical comportment (Ahmed, Faizan & Burney, 2019). The data contained in event logs are usually quite complex and often have ad hoc structures, what can make the process mining difficult (Belo et al., 2017). Sometimes, process mining results are used to enable resource prediction based on the users point in the process (Štajner, Mladenić & Grobelnik, 2010). For these reasons, the importance of a structure called data warehouse arises. Data warehouse could help in decision making and improving the business. The data warehouse is "a database that is organized according to other design principles than those that are used for good database design". Once a data warehouse is implemented, reports used for decision making can be built within seconds (Rabuzin & Škvorc, 2016).

Workflow Management Systems record all steps of a business process execution in a file named log file. "A workflow log records which task in which business process has been performed by which actor, and what was its processing time" (Koncilia, Pichler & Wrembel, 2015).

There exist several ways in which events can be selected and grouped into traces. According to the perspective to which the data are taken, the event logs should be extracted differently (González López de Murillas, Reijers & Van der Aalst, 2019). One of the most popular method of data unification in process mining is clustering. It is possible to cluster similar process models, and select a representative for each cluster "assuming that big differences are more relevant to the analyst than minor differences between the models" (Vogelgesang & Appelrath, 2017). Usually, process models which reflect similar behaviour are clustered, and then one representative process model per cluster is selected (Vogelgesang & Appelrath, 2015). "The model has a perfect fit if all suggestions in the record can be repeated by the model from start to end" (Ahmed, Faizan & Burney, 2019).

Often mentioned term in the field of process mining is multidimensional process mining. Limitations and challenges of multidimensional process mining are, e.g., comparison of cells, high effort for data integration, performance optimizations, interactivity, and handling of concept drifts (Vogelgesang et al., 2016). Multidimensional event log (MEL), on the other hand, is a specific data warehouse that stores all the available event data in a cube-like data structure, and maps the structure of event logs to a data cube while organizes events and cases on different levels (Vogelgesang & Appelrath, 2017). In this case, the purpose of a data warehouse is to physically integrate data from different sources and provide a trusted place for important pieces of information (Rabuzin, 2014).

Data warehouse is the basis for the powerful data analysis. Data warehouse uses cubes structure to contain the data that are imported. Cubes are the main objects in online analytic processing (OLAP) (Kassem, Turowski, 2018). OLAP is an important part of business intelligence that deals with a huge amount of data (Ghosh, Haider & Sen, 2015). Cubes in OLAP "divide the data into sub-sets that are defined by dimensions and the dimension is the descriptive attribute of a measure" (Kassem, Turowski, 2018). The OLAP queries (like *roll-up* and *drill-down*) are based on a set of basic operators like aggregation and selection (Vogelgesang & Appelrath, 2017). With the aim of maintaining a huge amount of data for analytical processing, data warehouse is the most effective solution (Ghosh, Haider & Sen, 2015). Using OLAP tools, Process Warehouse (data warehouse built for business process analysis) can implement information aggregating, analysis, comparing, mining new process model and it can improve the quality of an existing process model (Xia, Yao & Gao, 2013).

The event data are spread through multiple database tables, and it is necessary to join these tables with the aim to reconstruct the relationships between them. For that reason, the central tables (tables with facts, cases and events) are joined and the fact table and the event table need to be joined with the dimension tables in order to link the events with their respective dimension level values (Vogelgesang & Appelrath, 2017). The fundamental data warehouse structure that allows a user to conduct multidimensional data analysis is named snowflake schema (Sturm, 2012).

According to Vogelgesang & Appelrath (2017), events can have dimension attributes, which are stored in dimension tables similar to the case dimensions. Additionally, the event table is directly referenced to the dimension tables and dimension values can be different for events of the same case. Dimensional modelling technique is used to build different data marts by using one fact table, several dimension tables, and relationships between them. A fact table is used for storing numerical performance measurements of the business process, while dimension tables contain a description of the business process as its attributes describe it. The primary keys of the tables are used to construct the relationships (Sturm, 2012).

Some of the most important reasons for using enterprise data warehouse systems (DWH) are quick response time, ad-hoc queries and reduction the load on existing productive systems (Kassem & Turowski, 2018). Extract-Transform-Load (ETL) system is important because of the fact that "it is the component responsible for populating the data warehouse, ensuring the right way how data is collected, processed and stored conveniently to support decision-making processes" (Belo et al., 2017). ETL allows organization of analytical data in a multi-dimensional format. Additionally, ETL is also used to migrate data from one database to another, to form data marts and data warehouses, and, finally, to convert databases from one format to another (Ghosh, Haider & Sen, 2015).

# 3 Methodology

In this section, the methodology used for the research is described.

The research was conducted by examining the existing literature in the fields of process mining and data warehousing. Two scientific databases were used: Scopus and Web of Science (WoS). The queries applied for document search are represented in table 1.

**Table 1** Strategies of research

| **Scopus strategy of research** | **Number of articles** |
|---|---|
| ALL ((“process mining” OR “process discovery”) AND “data warehouses”) AND PUBYEAR > 2009 | 229 |
| **WoS strategy of research** | **Number of articles** |
| ALL FIELDS: ((“process mining” OR “process discovery”) AND “data warehouse”) Timespan: 2010-2020. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC. | 11 |

As table 1 shows, almost same keywords were used in searching both databases. Timespan was from 2010 till 2020 year for both databases, as we wanted the newest articles. Search in Scopus resulted with 229 articles, while the search in WoS resulted with 11 articles. Results in Scopus were sorted by relevance. From the 229 results in Scopus, the first 40 articles were taken for “quick analysis”, where the titles, keywords and abstracts of the articles were analysed. From 40 examined articles, 12 best matching articles for this research were taken for deeper analysis. All the 11 articles, resulted in WoS, were taken for “quick analysis” where, once again, titles, keywords and abstracts of the resulted articles were examined. Some resulted articles in WoS overlapped with articles in Scopus. Precisely, 7 of 11 articles in WoS appeared also in Scopus. From the rest of 4 articles in WoS, 2 of them were good matching with the needs of this work, and were selected for deeper analysis. Totally, there were 14 articles selected for the analysis in this work: 12 from Scopus and 2 from WoS.

# 4 Results

In this section, the results of the examined literature are presented. As follows, there are two tables representing the results of the analysed literature. We've tried to determine:

1. the aims and the methodologies of the analysed literature,
2. tools used or researched in previous works and,
3. research area, where process mining and data warehousing were applied.

## 4.1 Methodologies used in previous works

Table 2 shows the aims and the methodologies of the analysed literature. Common in most literature is usage of OLAP tools to analyse data stemming and process variants. Multidimensional process mining and appliance of it is also often studied. Usually, to reach the data, relational databases and SQL are used. In process mining, one of the aims is to extract process execution data contained in logs in order to see the real state of process execution. This information can be used for further and deeper analysis. The most common methods for process mining are clustering and classification. Prediction of process models allows action discovery and process discovery based on real data.

**Table 2** Aims and the methodologies of the analyzed literature

| Article name | Aim and a methodology of the analysed literature |
|---|---|
| "A Generic Data Warehouse Architecture for Analyzing Workflow Logs" | Usage of Sequence Warehouse (SeWA) architecture and OLAP tools to analyse data stemming from workflow logs, including process variants. Usage of PHP scripts, which parse and import a XES file into several tables in the PostgreSQL database and generate sequences of events. |
| "A Relational Data Warehouse for Multidimensional Process Mining" | The underlying relational concepts of PMCube, a data warehouse based approach for multidimensional process mining; generic query patterns which map OLAP queries to SQL to push the operations to the database management system. |
| "Connecting databases with process mining: a meta model and toolset" | A meta model in order to integrate process and data perspectives, which are related, is proposed. Automatically obtaining events from database systems. The meta model structure and the data, inserted into it, are stored in a SQLite file, while the process of extracting, transforming and querying data has been implemented in RapidProM workflow. |
| "A Process Mining Approach for Discovering ETL Black Points" | Using process mining to identify and characterize ETL black points (ETL bottleneck situations). Working example of an ETL system designed and implemented in one course at a university. |
| "An Integrated Approach to Deploy Data Warehouse in Business Intelligence Environment" | The proposed architecture that integrates business intelligence in On-Line Analytical Processing (OLAP) environment, that incorporates Data Warehouse, Data Mining, Data Mart and Virtual Data Warehouse. |
| "*Eventifier*: Extracting Process Execution Logs from Operational Databases" | Identification and extraction of process execution events from databases for reconstructing ready-to-use event logs for process discovery. |
| "Exploring contexts and actions in knowledge process" | An approach about how to automatically discover the contexts on which a business analyst is working in, and the numerous actions a business analyst executes across contexts. Except context discovery, an action discovery is also evaluated using predictive power of process models. Real data were used to resolve a clustering problem. |
| "Matching of Business data in a generic Business Process Warehousing" | Mapping the business data and integrating them in a data warehouse process model, with the aim to calculate different granularity level of the business process. A mechanism that allows keeping business data as a "separated table outside the multidimensional model and match it to the abstract multidimensional model based on executed events" is proposed. Workflow Mining is used to build a meta business process multidimensional model considering business objects and their attribute values. |
| "Multidimensional Process Mining with PMCube Explorer" | PMCube Explorer, a novel tool for multidimensional process mining, that allows an analysis of a process from various views, is presented. |
| "Problems and Challenges When Implementing a Best Practice Approach for Process Mining in a Tourist Information System" | A data warehouse component, that stores process logs in the tourism information system OHA, is implemented. The customer journey process through the tourism platform OHA is analysed. Process logs were gathered, modified, and converted into o format which was convenient for statistical calculations or process mining. |

| | |
|---|---|
| "Research and Design of Process Data Warehouse for Business Process Assessment" | A case study, where a process assessment-oriented warehouse for storage and management of process instance data is presented. A matrix is used to express the relationships among facts and dimensions as well as dimensions and dimensions. |
| "Supporting business process analysis via data warehousing" | A method for constructing data warehouse schemata from business process specifications is proposed. Snowflake Maker is a prototype in which the rules encompassed within the proposed method were developed. The aims were to facilitate the off-line analysis of the business process execution, and to identify potential improvements by querying the business process performance. Off-line analysis of business processes is done through OLAP using data warehousing. |
| "A data-mining based method for the Gait pattern analysis" | The method that captures procedures, creates an organized repository (data warehouse), standardizes the data, and develops the steps for process mining, is proposed. The proposed method consists of data classification automation. With the aim to find footprint signature patterns, a graphical analysis is used. |
| "Ontological approach to enhance results of business process mining and analysis" | A methodology that combines domain and company-specific ontologies with databases, in order to obtain multiple levels of abstraction for analysis and process mining technique. A real case study was performed using a prototype system and techniques developed in ProM. Approach used in the research combines domain ontologies, company ontologies and a company's database to form a knowledge source for process mining. |

## 4.2. Tools and research areas of previous works

Table 3 represents the tools and the research area of the analysed literature. As already noted by the methodologies of analysed literature, OLAP techniques are widely used by data warehousing and process mining. Relational database, SQL, and all its variations are also often used for data access and querying. In several works, the usage of RapidProM (RapidMiner), Disco and ProM is mentioned, so it can be supposed that these tools are some of the most popular for process mining. Further on, various tools for generating scripts and reports are also popular in the stage of business analysis (e.g., PHP). If there is a sign "-" (minus) in table 3, that means that no tool is mentioned in the work. Research areas, where the domains of process mining and data warehousing are usually applicable, are: health care, multidimensional process mining, and higher education. Reasons for wider application of these domains in these areas could be in good data availability, rapid fluctuation of event records, lower sensibility of data, and economic accessibility of data.

**Table 3** Tools and the researched areas of the analyzed literature

| Article name | Tools used or researched in analysis | Research area |
|---|---|---|
| "A Generic Data Warehouse Architecture for Analyzing Workflow Logs" | OLAP techniques, DM, DDL for Sequential SQL (S-SQL), XES importer, Sequence Cube (data structure). | Workflow Management Systems (WfMS), traditional data warehouses. |
| "A Relational Data Warehouse for Multidimensional Process Mining" | OLAP queries, SQL, ROLAP (Relational OLAP). | Multidimensional process mining, health care. |
| "Connecting databases with process mining: a meta model and toolset" | RapidProM, SQLite. | Obtaining, transforming, organizing, and deriving data and process information from databases. Source environments: database redo logs, in-table version storage (Dutch financial organization), SAP-style change tables. |
| "A Process Mining Approach for Discovering ETL Black Points" | Disco. | Higher education. |

| | | |
|---|---|---|
| "An Integrated Approach to Deploy Data Warehouse in Business Intelligence Environment" | Global B. I. analyser. | Analytical processing in OLAP environment, business intelligence, knowledge & report generation. |
| "*Eventifier*: Extracting Process Execution Logs from Operational Databases" | *Eventifier* (an integrated platform that includes the components for *eventification*, correlation and process discovery). | Producing process execution events in a fundamentally different context (e.g., there is no access to the information system running the process). |
| "Exploring contexts and actions in knowledge process" | - | Context mining (collection of documents), Action mining. |
| "Matching of Business data in a generic Business Process Warehousing" | OLAP Engine, SAP BW (BEx Query Analyzer). | Generic solution evaluated by Business Process Warehousing. |
| "Multidimensional Process Mining with PMCube Explorer" | PMCube Explorer. | Health care domain. |
| "Problems and Challenges When Implementing a Best Practice Approach for Process Mining in a Tourist Information System" | Relational database, Java Web server, OHA platform. | Tourist Information System. |
| "Research and Design of Process Data Warehouse for Business Process Assessment" | XES. | Health care. |
| "Supporting business process analysis via data warehousing" | BPMN modelling language. | Business process management (BPM), Higher education. |
| "A data-mining based method for the Gait pattern analysis" | GaitRite system, RapidMiner. | Health care. |
| "Ontological approach to enhance results of business process mining and analysis" | Process Mining Framework (ProM). | Apparel domain. |

## 4.3. Fundamental aspects of literature analysis

During the research conducted in this article, there were some findings which were noticed and should be highlighted. The findings are consistent with the aims that this research sought to achieve. One of the reason for implementing a data warehouse is reducing the complexity of data contained in event logs which are used in process mining. Besides, the data can have variable structure which can make process mining difficult. The data warehouse is designed to store and prepare logs in order to apply further research techniques like process mining, with the aim of analysing the process and obtaining information for decision making. The data warehouse integrates information into data sources and that is a basis for data analysis. Other advantages of data warehouses are ad hoc queries and quick response time. According to the results from table 2 in this article, the importance of relational databases like PostgreSQL can be noticed, in the sense that those are still some of the most popular tools in data warehousing. The usage of relational databases for data storage facilitates modifying logs and data preparation. It is important to perform a data log analysis and determine the order between the tasks. Workflow log records which task in the business process was performed by which participant, the duration and the processing time of a task. The usage of process mining in data warehouses can increase the efficiency, correctness and performance of ETL systems. We can use multi-relational clustering algorithms, which can handle large data sets that contain information about differences in distributions between resources and events. What is often mentioned is the usage of semantics and semantic technologies in web services, which manage synonyms and different languages of such terms, and convert them into a normalized form to increase the quality of process mining performance. Standard data warehouses do not consider the sequential nature of data stemming from a workflow log. The place for improvement could be in

finding a suitable more advanced type of data warehouse that could consider the sequential nature of data contained in event logs. According to the results from table 3 in this article, most of the existing analytical tools are focused on process mining only or offer basic functionality analysis which is not enough for OLAP analysis. Limitations in data warehouse design may be the lack of guidelines for designing data warehouse schemes or the lack of appropriate analysis of a process perspective, which could be a place for improvement in the form of standardization of a design of a special type of data warehouse.

# 5 Conclusion and future work

In this paper, we examined the existing literature in order to find out which methodologies are used in previous works in domains of process mining and data warehousing. Besides, we thought it would be interesting to perceive the tools used for conducting the analysis in previous works, and to examine the research areas where the process mining is used as a technique together with data warehousing. After conducting the analysis, the importance of OLAP techniques in data warehousing was confirmed. Some of the most popular tools for process mining are RapidMiner, ProM and Disco. Research areas in which process mining and data warehousing are most commonly applied are, e.g., health care and higher education. Future work could be oriented on examination types of data warehouses and their popularity. The advantages and disadvantages of these data warehouses as so as the domains of their application could be explored. Based on selected data from data warehouse, various business process analysis can be performed, and the functionality, as so as the behavior of a process warehouse can be affirmed.

# References


Ahmed, R., Faizan, M. & Burney, A.I. (2019). Process Mining in Data Science: A Literature Review. MACS 2019 - *13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics*, Proceedings 9024806.

Belo, O. et al. (2017). A process mining approach for discovering ETL black points. Advances in *Intelligent Systems and Computing* 570, 426-435.

Ghosh, R., Haider, S. & Sen, S. (2015). An integrated approach to deploy data warehouse in business intelligence environment. *Proceedings of the 2015 3rd International Conference on Computer, Communication, Control and Information Technology*, C3IT 2015 7060115.

González López de Murillas, E., Reijers, H.A. & van der Aalst, W.M.P. (2019). Connecting databases with process mining: a meta model and toolset. *Software and Systems Modeling*, 18(2), 1209-1247.

Jareevongpiboon, W. & Janecek, P. (2013). Ontological approach to enhance results of business process mining and analysis. *Business Process Management Journal*, 19(3), 459-476.

Kassem, G. & Turowski, K. (2018). Matching of business data in a generic Business Process Warehousing. Proceedings - 2018 *International Conference on Computational Science and Computational Intelligence*, CSCI 2018 8947788, 284-289.

Koncilia, C., Pichler, H. & Wrembel, R. (2015). A generic data warehouse architecture for analyzing workflow logs. Lecture Notes in *Computer Science* (including subseries Lecture Notes in *Artificial Intelligence* and Lecture Notes in *Bioinformatics*) 9282, 106-119.

Lux, M. & Rinderle-Ma, S. (2017). Problems and challenges when implementing a best practice approach for process mining in a tourist information system. *CEUR Workshop Proceedings*, 1985, 1-12.

Rabuzin, K. (2014). Deductive Data Warehouses. *International Journal of Data Warehousing and Mining*, 10(1), 16-31.

Rabuzin, K. & Škvorc, D. (2016). Data Warehouses and Business Intelligence in Croatia: Do Managers Know How to Use Them? *International Journal of Business Analytics*, 3(2), 50-60.

Rodríguez, C. et al. (2012). Eventifier: Extracting process execution logs from operational databases. *CEUR Workshop Proceedings*, 936, 17-22.

Rudek, M. et al. (2015). A data-mining based method for the gait pattern analysis. *Mechanical Engineering*, 13(3), 205-215.

Sturm, A. (2012). Supporting business process analysis via data warehousing. *Journal of software: Evolution and Process*, 24(3), 303-319.

Štajner, T., Mladenić, D. & Grobelnik, M. (2010). Exploring contexts and actions in knowledge processes. *CEUR Workshop Proceedings* 626.

Vogelgesang, T. & Appelrath, H.J. (2015). Multidimensional process mining with PMCube explorer. *CEUR Workshop Proceedings*, 1418, 90-94.

Vogelgesang, T. et al. (2016). Multidimensional process mining: Questions, requirements, and limitations. *CEUR Workshop Proceedings* 1612, 169-176.

Vogelgesang, T. & Appelrath, H.-J. (2017). A relational data warehouse for multidimensional process mining. Lecture Notes in *Business Information Processing* 244, 155-184.

Xia, H., Yao, Q. & Gao, F. (2013). Research and design of process data warehouse for Business Process Assessment. Lecture Notes in *Electrical Engineering* 256 LNEE, 377-385.