BPI Challenge 2019

Aislu Naran

Business Information Systems
Prof. Paolo Ceravolo
a.a. 2021- 2022

Case study description

Describes the process of processing purchase orders in a multinational coatings and paints company. The dataset includes information from 60 of the company's subsidiaries. The data focuses on four types of flows within purchase orders:
1) three-way matching with invoice after receipt of goods,
2) three-way matching with invoice before goods arrive,
3) two-way matching without goods receipt,
4) consignment items.

Each purchase order contains one or more line items, and each line item can have multiple POs and invoices that must match to be eligible.

The data set includes attributes such as concept name, purchasing document ID, item ID, item type, flags for GR based billing and three-way matching, source system, document category name, subsidiary, expense classification text, expense area text, subitem cost area text, vendor information, document type, and item category.

The data presented in BPI 2019 challenge

The data is collected from the Eindhoven university of technology 2019.
The files contains:

1) Over 1.5 million purchase order events.
2) Total 251,734 purchase items.
3) 42 activities within Purchase-to-Pay Process.
4) 76,349 purchase documents.
5) 627 users which include both human and batch users.

Various attributes are recorded for each purchase item. Most important for the analysis are:

1) Item Category
2) Case name
3) Timestamps

The data is split in datasets presented in the following table:

	User	org:resource	concept:name	Cumulative net worth (EUR)	time:timestamp	case:Spend area text	case:Company	case:Document Type	case:Sub spend area text	case:Purchasing Document	...
1590025	NONE	NONE	Vendor creates invoice	198.0	2017-12-16 22:59:00+00:00	Others	companyID_0003	Framework order	Utilities	4507075964	...
1590026	user_602	user_602	Create Purchase Order Item	198.0	2018-01-03 12:47:00+00:00	Others	companyID_0003	Framework order	Utilities	4507075964	...
1590027	user_603	user_603	Change Approval for Purchase Order	198.0	2018-01-03 13:56:00+00:00	Others	companyID_0003	Framework order	Utilities	4507075964	...
1590028	user_359	user_359	Record Invoice Receipt	198.0	2018-01-04 09:04:00+00:00	Others	companyID_0003	Framework order	Utilities	4507075964	...
1590029	user_604	user_604	Clear Invoice	198.0	2018-01-10 11:35:00+00:00	Others	companyID_0003	Framework order	Utilities	4507075964	...
1590030	user_602	user_602	Change Approval for Purchase Order	198.0	2018-04-06 08:38:00+00:00	Others	companyID_0003	Framework order	Utilities	4507075964	...
1590031	user_603	user_603	Change Approval for Purchase Order	198.0	2018-04-06 13:58:00+00:00	Others	companyID_0003	Framework order	Utilities	4507075964	...

Goals

Goals are an excellent indicator of the effectiveness of the analysis work. The main objective of the study is to analyze the data provided in order to obtain valuable information. The results of the analysis can be used by the organization to achieve its goals, improve the efficiency of processes, eliminate bottlenecks and detect anomalies in data.

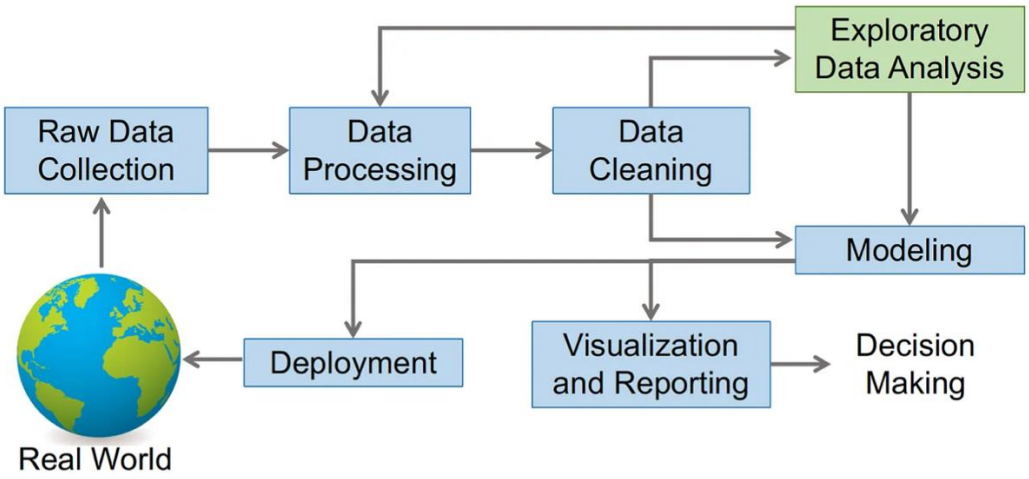


Figure 1 – Goal diagram

What questions should our analysis on the dataset answer?

1. Is there a set of process models that together properly describe the process in this data?

2. Which processes among them can be analyzed to find the difference between them?

Knowledge uplift trail

The Knowledge Uplift Trail is a sequence of analytical steps that will be carried out to gain new knowledge from the event log.

To begin with, we must prepare the data, or rather, use filtering, data cleaning. After preparing the data, we describe the analysis. In order to describe the analysis, you can use various visualizations, statistics. Performance analysis is used to answer questions related to throughput, cycle time, resource usage, and other performance metrics. And lastly, outlier detection. Used to detect deviations from the process.

In this process, I analyzed using Disco and PMTK. Used pm4py libraries. In this process, I analyzed the Item Category, which is one of the important elements. There are four categories of cases in our dataset:

1) 3-way matching, invoice after receiving the goods.
2) 3-way matching, invoice before receiving the goods.
3) 2-way matching (goods acceptance is not required)
4) Consignment

Illustrating the BPMN process of each categories of cases:

1) You need to compare the value of the goods receipt message with the value of the invoice receipt message and the value specified when the item was created.

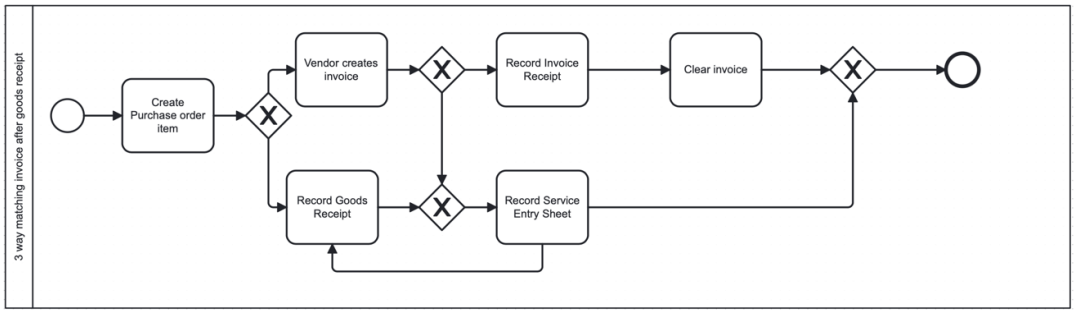


Figure 2 – BPMN of 3-way matching, invoice after receiving the goods

2) It is possible to enter invoices before the goods arrive, but they remain blocked until the goods are received. It is necessary to compare the goods receipt value with the invoice value and the value when the goods were created.

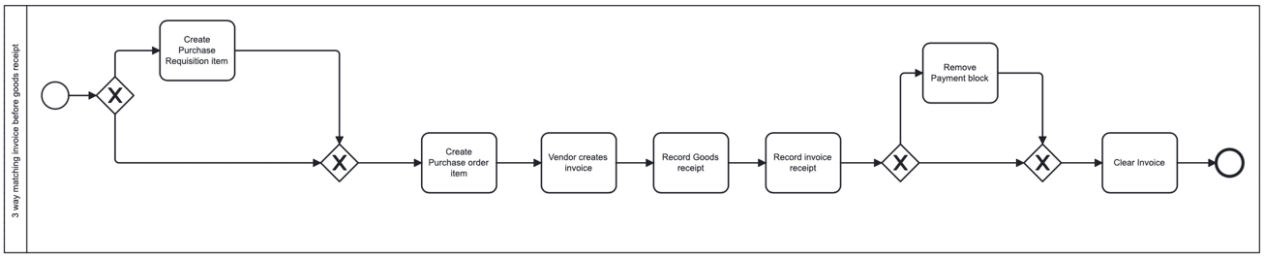


Figure 3 – BPMN of 3-way matching, invoice before receiving the goods

3) The value of the invoice must match the value specified when it was created, but a separate goods receipt message is not required.

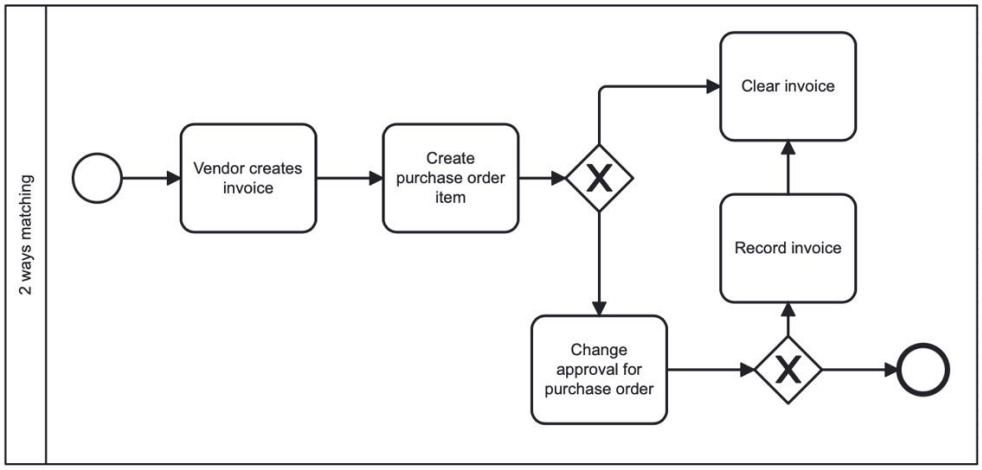


Figure 4 – BPMN of 2-way matching (goods acceptance is not required)

4) During purchase order processing, this process is performed entirely in a separate procedure.

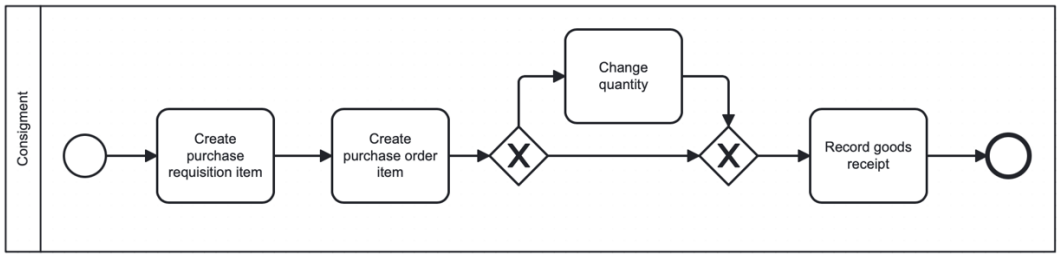


Figure 5 – BPMN of Consignment

Analysis of Item category

1) 2-way matching

I compared between 2-way matching and 3-way matching, invoice before receiving the goods. First of all, I cleared the data itself, removed duplicates, thereby reducing the data from 1595923 to 1463175, by 8.3%.
Having selected the data of 2-way matching and 3-way matching, invoice before receiving the goods, I saw that 2-way matching has much less data than 3-way matching, invoice before receiving the goods, or rather 5898 and 1231845. For comparison, I used this data to look at the difference between big data and small data.

Using Disco, I looked at the 2-way matching timeframe and chose the period when it was active, which was about 2 years from 2017-12-01 to 2019-01-01. I filtered the data by this date and noticed that the data decreased from 5898 to 4800, by 18.16%. Then I used several methods to compare key factors, or rather Alpha Miner, Heuristics, Inductive Miner.

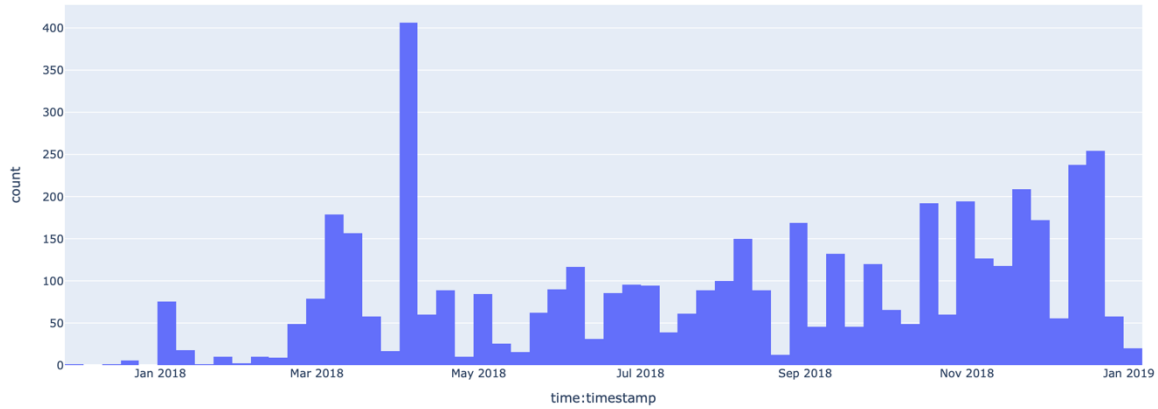
Alpha Miner helps to discover cause-and-effect relationships between events, and also captures temporal dependencies, which are suitable for analysis and workflow optimization. Heuristics provides a fast approximation and is very useful for data-driven decision making. Inductive Miner captures frequent sequences of events and dependencies. It handles noise and incomplete event logs, suitable for process detection and anomaly detection. A Directly-Follows Graph (DFG) is a visual representation that shows the order and frequency of actions or events in a process. DFG diagrams are commonly used in process analysis, a field that aims to gain insights into and analyze real processes based on event logs.

The table presented the attributes and descriptions for each attribute:

User	The user who performed the activity or event
Cumulative net worth (EUR)	The total value of net worth in euros accumulated over a certain period
Item Category	The category or classification of the item associated with the case or process
Item Type	The type or category of the item associated with the case or process
Company	The company or organization related to the case or process
Vendor	The vendor or supplier associated with the case or process
Purchasing Document	The document related to the purchasing process within the case

Activity Distributions over period

Most activities for the 2-way matching occurred on May 2018.



1.1. Alpha Miner

The Alpha miner algorithm is applied taking into account the frequency of events by the filtered year 2-way matching

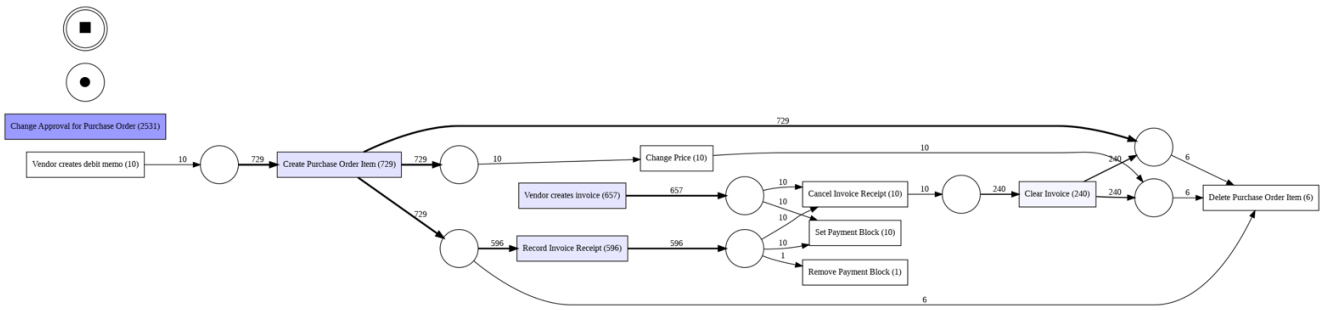


Figure 6 - Petri net for Alpha Miner

1.2. Heuristic Miner

Figure 7 - Process tree Heuristic Miner, dependency threshold: 0.8

1.3. Inductive Miner

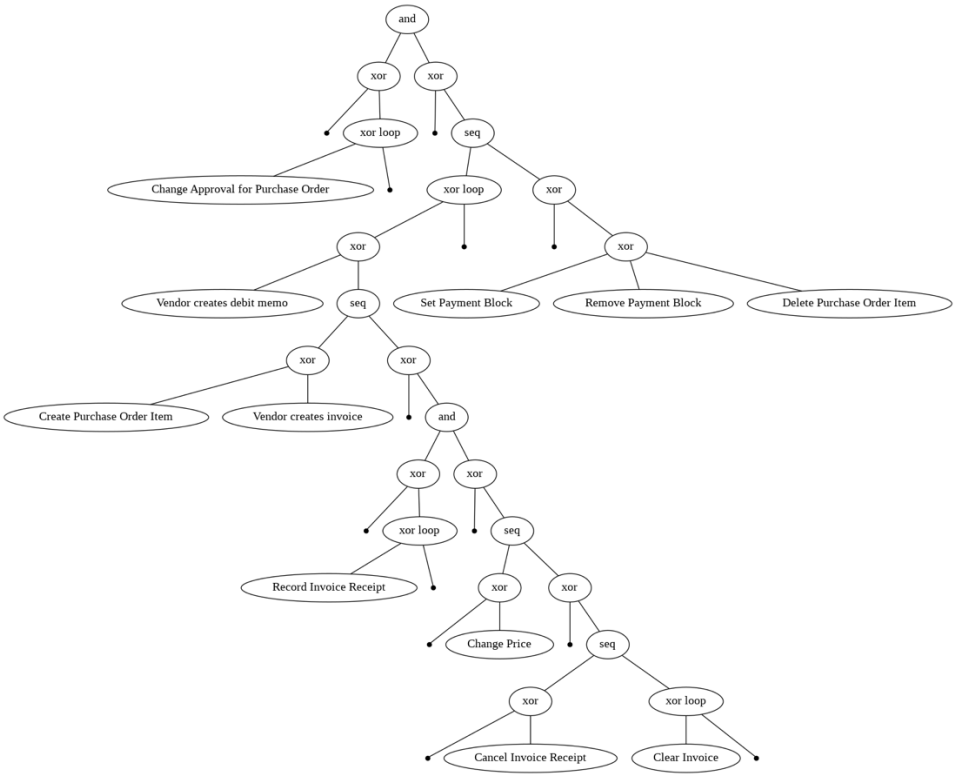


Figure 8 - Inductive Miner for 2-way matching

1.4. DFG diagram

The DFG diagram applied based on the event frequency of the filtered 2-way comparison of the year.

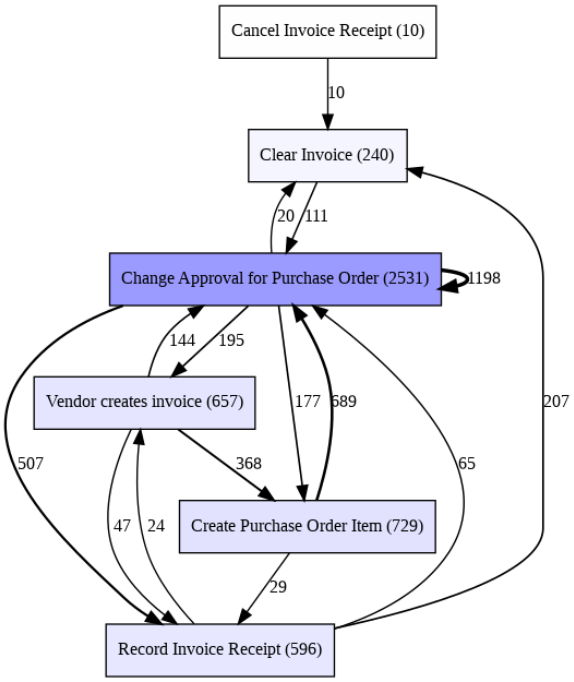


Figure 9 - Directly-Follows Graph diagram with maximum no edges 15

Conformance analysis is the process of evaluating whether a particular system, process, or product conforms to given requirements or rules. The purpose of the Conformance analysis is to determine the degree of compliance of the analyzed object with the established criteria.
Conformance analysis plays an important role in ensuring quality, reliability, safety and compliance in various areas.

Fitness is a suitability analysis in which a process or product conforms to specified requirements or regulations. It measures how well the analyzed object satisfies the established criteria and whether it meets the desired level of compliance.

Simplicity refers to the clarity and ease of understanding of the specified requirements or standards. This includes ensuring that the eligibility criteria are expressed in a concise and simple manner that facilitates their understanding and application by stakeholders.

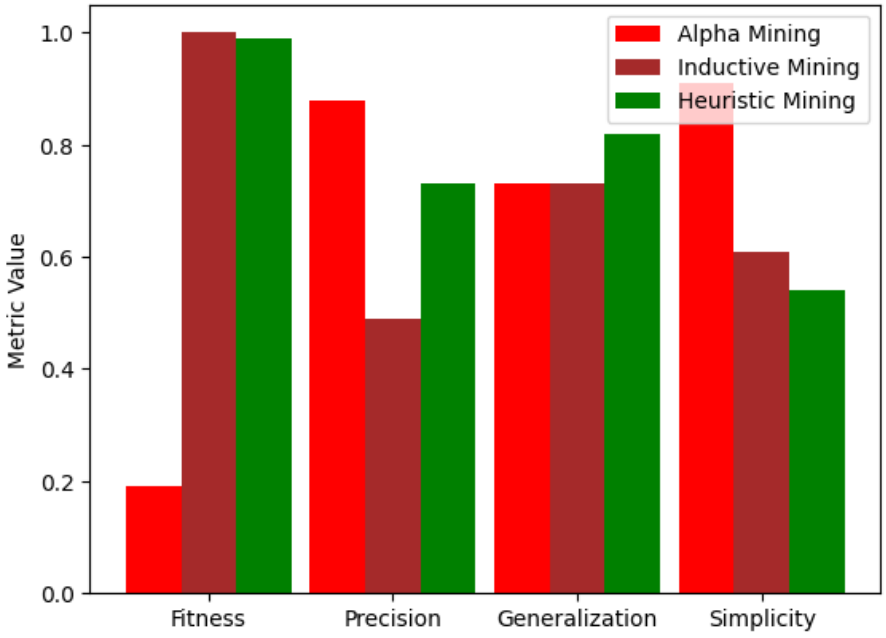
Precision is related to thoroughness and attention to detail in the evaluation process. It includes conducting an in-depth assessment of the subject to identify any inconsistencies, deviations with established criteria. Accuracy provides a

comprehensive and thorough check to provide an accurate representation of the level of compliance.

Generalization refers to the extent to which specified requirements, standards or rules can be applied universally or in different contexts. It includes an assessment of whether the conformance criteria are applicable to different systems, processes, products or activities in a particular area or industry, allowing for consistent assessment and comparison.

The table presented the all algorithms and their values:

Algorithm	**Fitness**	**Precision**	**Generalization**	**Simplicity**
Alpha Miner	0.19	0.88	0.73	0.91
Heuristic Miner	0.99	0.73	0.82	0.54
Inductive Miner	1.0	0.49	0.73	0.61



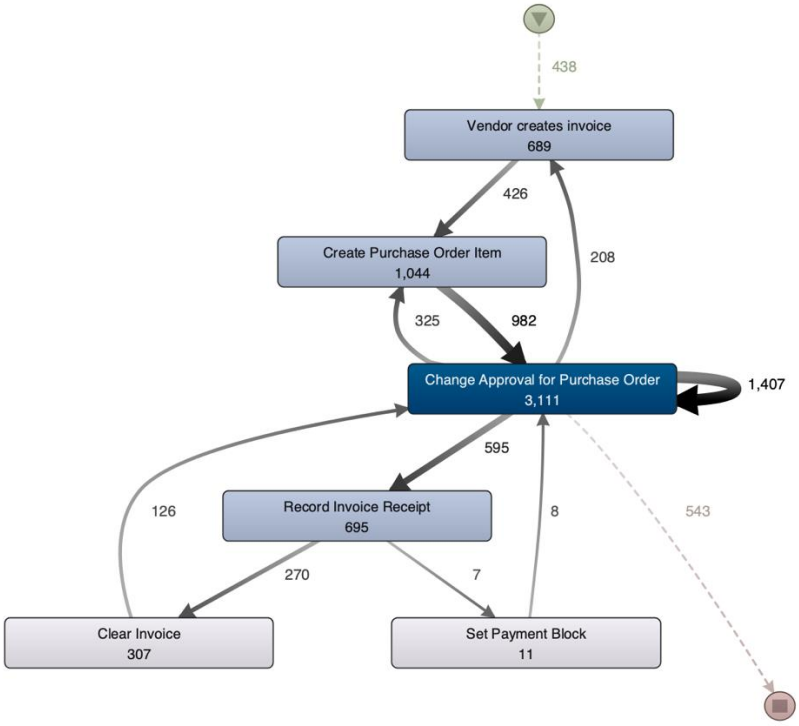


Figure 10 - 2 way matching in Disco

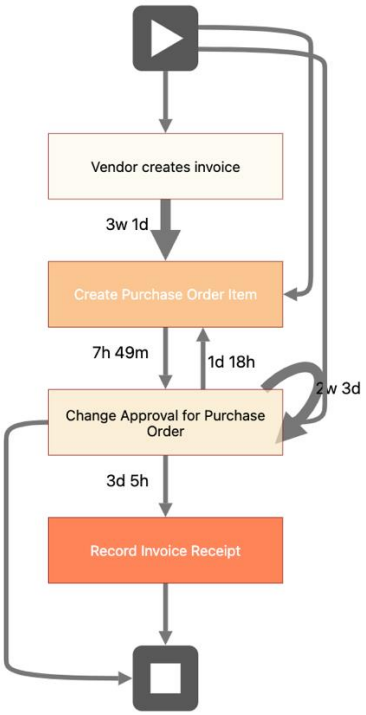


Figure 11 - 2 way matching in PMTK

2) 3-way matching, invoice before receiving the goods

Now we start looking at 3-way matching, invoice before receiving the goods. First of all, I used filtering by 3-way matching, invoice before receiving the goods, to see what time it was active in Disco. Used filtering by date from 2018-01-01 to 2019-02-01. I noticed that the data decreased by 0.01% from 1231845 to 1231631.

Here you can see that the data is much larger than in 2-way matching, by almost 99.5%. You can see that the graphs for 3-way matching, invoice before receiving the goods will come out very huge and incomprehensible, thereby showing the complexity of the actions.

1.1. Alpha Miner

The Alpha miner algorithm is applied taking into account the frequency of events by the filtered year 3-way matching, invoice before receiving the goods.

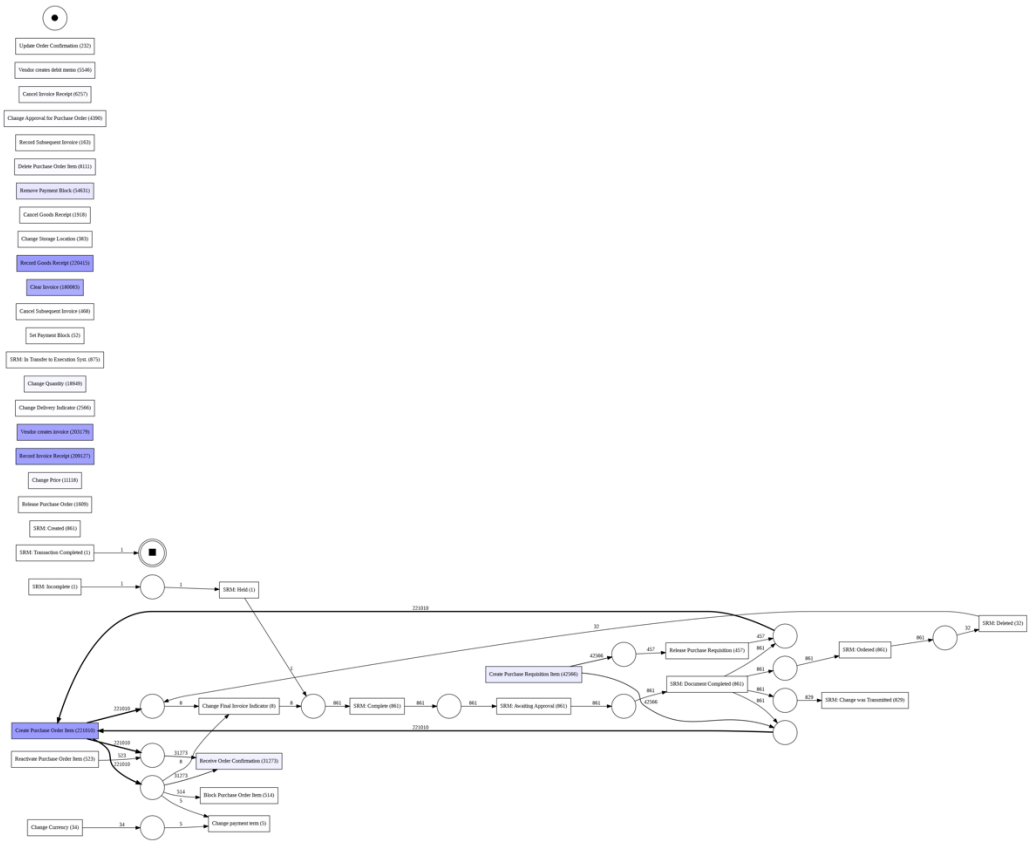


Figure 11 - Petri net for Alpha Miner

1.2. Heuristic Miner

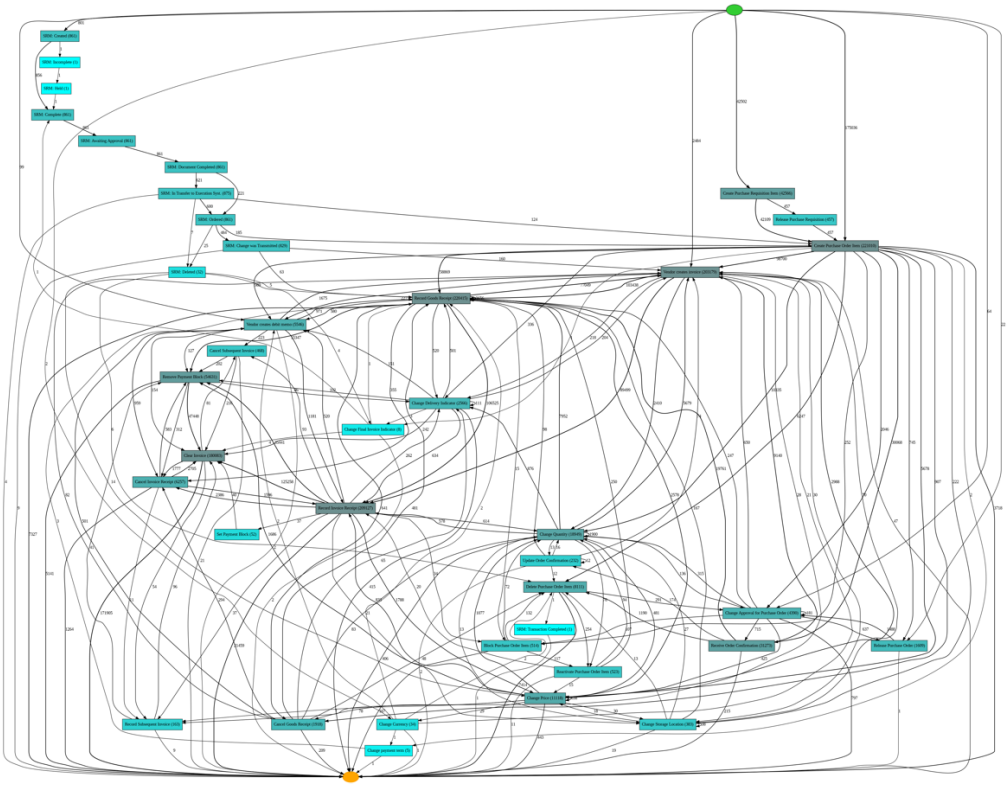


Figure 12 - Process tree Heuristic Miner, dependency threshold: 0.8

1.3. Inductive Miner

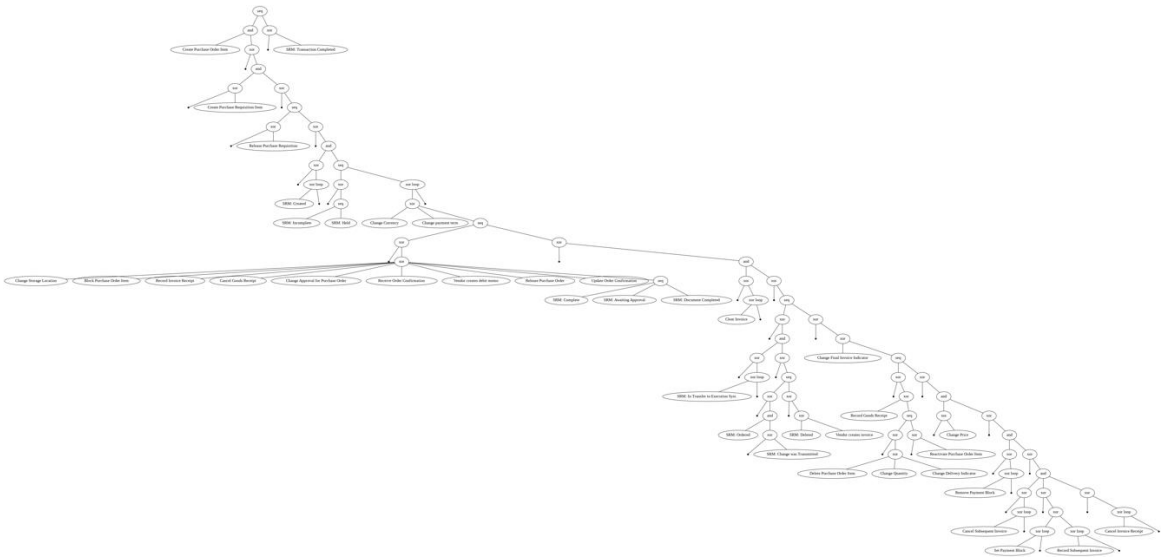


Figure 13 - Inductive Miner for 3-way matching, invoice before receiving the goods

1.4. DFG diagram

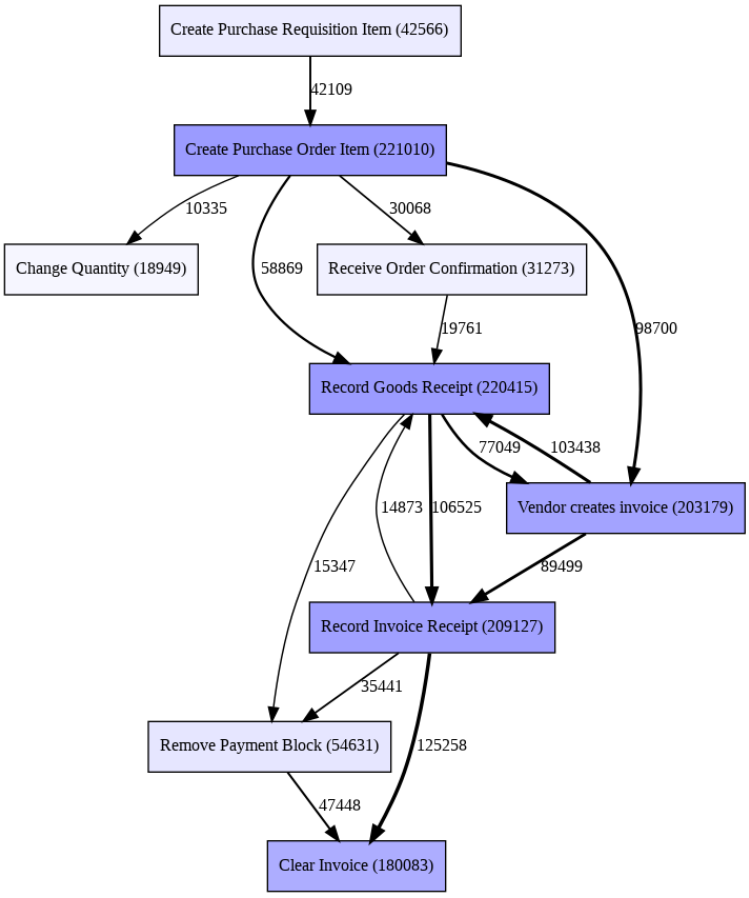
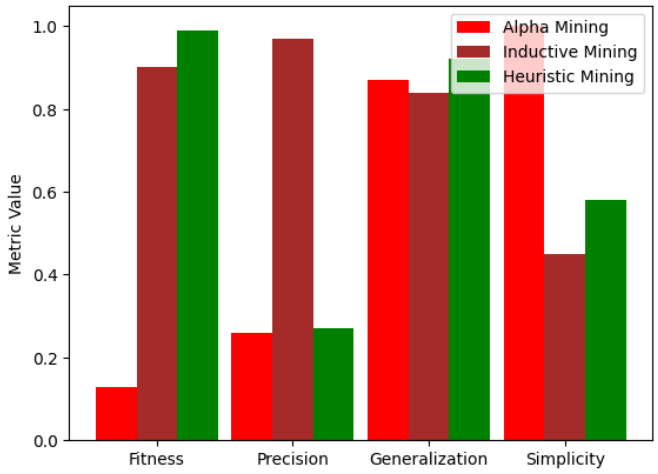


Figure 14 - Directly-Follows Graph diagram with maximum no edges 15

Conformance Analysis

The table presented the all algorithms and their values:

Algorithm	**Fitness**	**Precision**	**Generalization**	**Simplicity**
Alpha Miner	0.13	0.26	0.87	1.0
Heuristic Miner	0.90	0.97	0.84	0.45
Inductive Miner	0.99	0.27	0.92	0.58



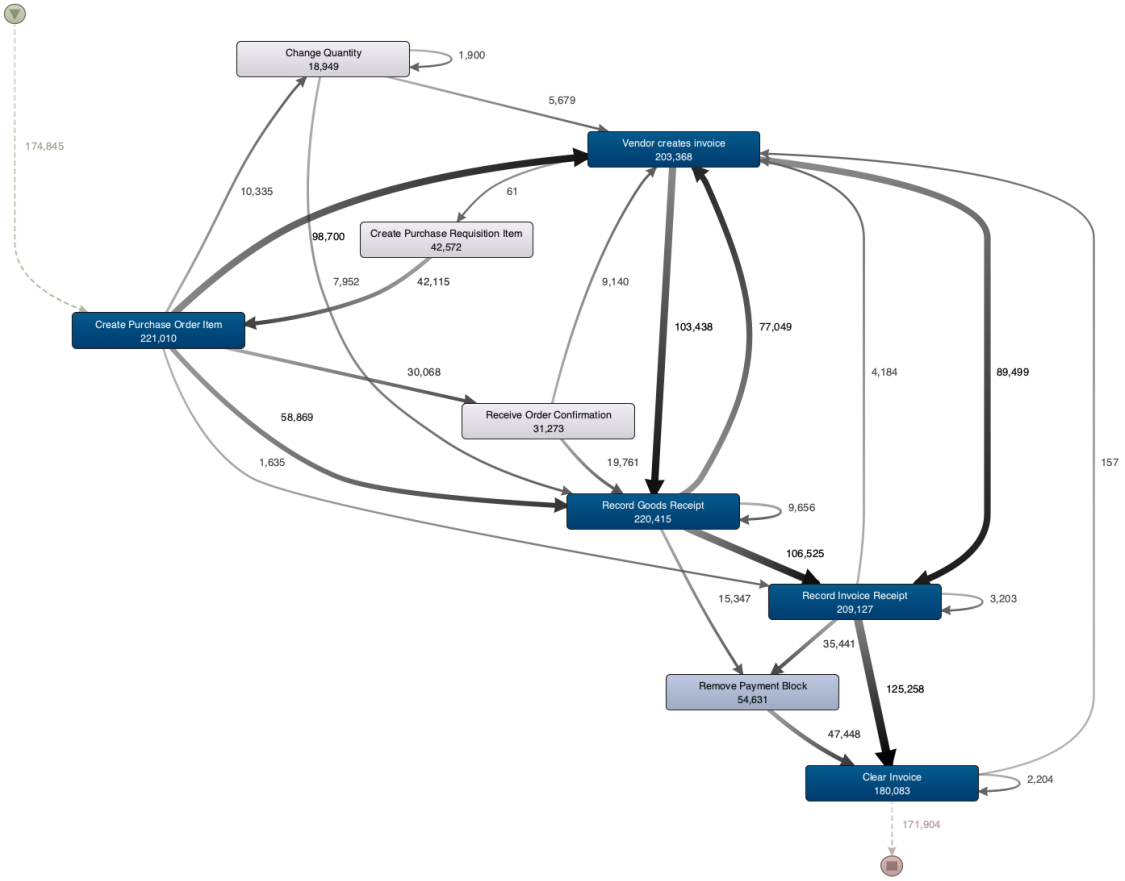


Figure 15 - 3-way matching, invoice before receiving the goods in Disco

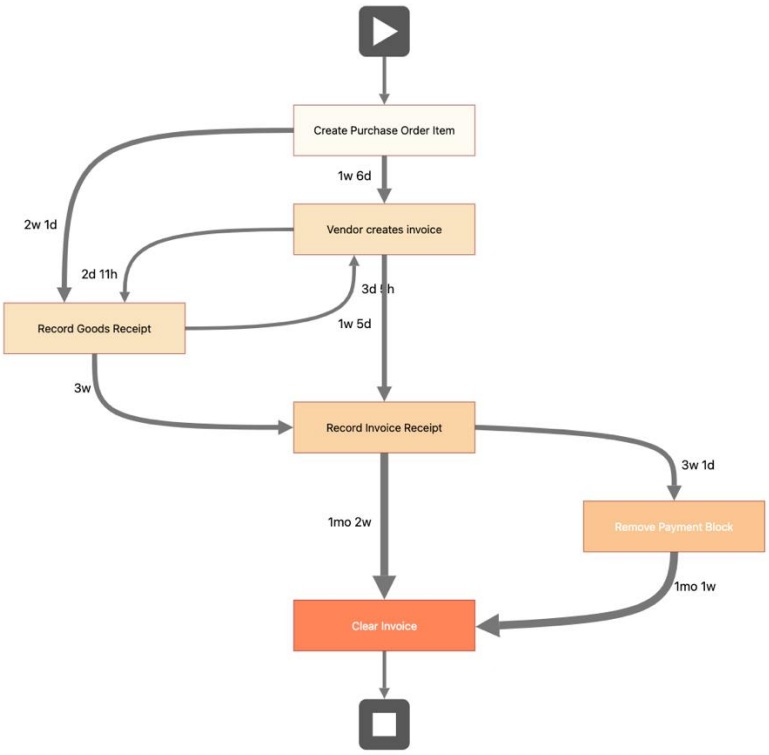


Figure 16 - 3-way matching, invoice before receiving the goods in PMTK

Conclusion

This business process was analyzed using various process analysis methods. Thanks to the analysis, it was possible to understand and answer questions.

In the product category, you could notice that we chose 2-way matching and 3-way matching, invoice before receiving the goods. Since 2-way matching was much smaller than 3-way matching, invoice before receiving the goods. This was necessary to compare between large data and between small ones.

In 2-way matching, you can see that Alpha Miner and Heuristic Miner, Inductive Miner worked well than in 3-way matching, invoice before receiving the goods.

In the context of Item category analysis, Alpha Miner is used to explore sequences of events and dependencies between them in a 2-way matching process. It can help identify major steps or milestones in a process, as well as identify which events occur most frequently and which events follow each other. Create purchase order item happens more in this case. In 3-way matching, invoice before receiving the goods, due to big data, many processes are not handled correctly in Alpha Miner and Heuristic Miner. Thanks to the Heuristic Miner, you can determine the most frequently occurring sequences of events. You can see that Create purchase order item and Change approval for purchase order are the most common sequences of events in 2-way matching. Thanks to the Inductive Miner in 2-way matching can help identify the most frequently occurring sequences of events, it automatically reveals the relationships between events to build a process model. Create approval for purchase order, set payment block, remove payment block, record invoice receipt, clear invoice are relationships between events, and in 3-way matching, invoice before receiving the goods, you can see that there is too much data, even using filtering was hard cut.

The process discovery step took place by creating data models. The correspondence analysis showed that the best models for 2-way matching were Heuristic Miner and Inductive Miner with a high dependency threshold, allowing to have a Fitness of 0.99. This high dependency allows us to observe the most successful trace.

The best model for 3-way matching, invoice before receiving the goods was Heuristic Miner and Inductive Miner with Precision 0.97 and Fitness 0.94. In this category of products, you can see that Alpha Miner is much less than in 2-way matching, almost by 86%.

References

1) https://data.4tu.nl/articles/_/12715853/1

2) https://pm4py.fit.fraunhofer.de/documentation

3) https://medium.com/@c3_62722/process-mining-with-python-tutorial-a-healthcare-application-part-2-4cf57053421f

4) https://medium.com/@c3_62722/process-mining-with-python-tutorial-a-healthcare-application-part-2-4cf57053421f