

Course “Process Mining”

Disco

Content

Laboratory Session: Disco Software Tool	1
Step 1: Import Data	1
Step 2: Inspect Process	3
Step 3: Animate Process	4
Step 4: Inspect Statistics	5
Step 5: Inspect Cases	6
Step 6: Filter on Performance	8
Exercise 1 – Filter out running cases	10
Exercise 2 – Filter out running cases	10
Exercise 3 – Import XES files to Disco	10

Laboratory Session: Disco Software Tool

This laboratory session is intended to let students familiarize themselves with event logs and process concepts. An industrial tool, Disco by Fluxicon, will be used to analyze a real-life process. Steps 1 to 6 below illustrate how to conduct an analysis of an actual process based on event data (i.e. an event log). Exercises 1, 2, and 3 enable consolidating the acquired knowledge.

In the remainder, we will employ the terminology used in the Disco tool, where a “process map” is what we referred to in the lecture as a “[Dependency graph](#)”. Recall that an event log comprises traces (a.k.a. cases). Each trace is a sequence of events; each event of a particular trace t refers to the starting or completing a process activity within the case associated with t .

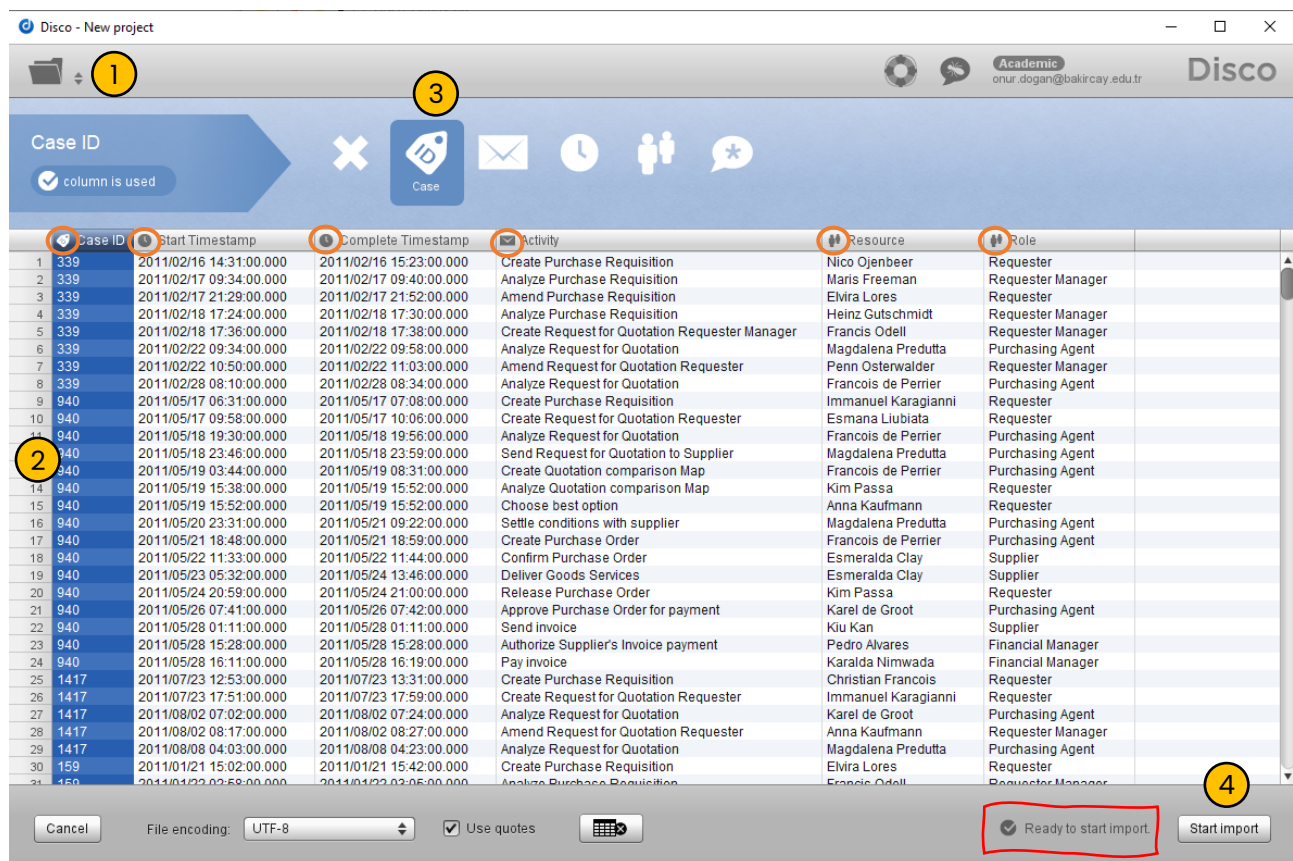
Step 1: Import Data

File [2-Disco-Dataset.csv](#) is an event log that records real executions of a process executed by a real organization to place orders to purchase goods.

This file is in the so-called [csv format](#), a delimited text file that uses a comma to separate values. A csv file stores tabular data (numbers and text) in plain text. Each line of the file is a data record.

Each record consists of one or more fields, separated by commas.¹ Often, the first line is the header, which assigns names to the different fields. It is the case in the file provided in this laboratory session. Each line identifies a separate event log for process mining: the case identifier column allows the grouping of events in cases.

You can upload the event-log file by clicking the 'Open' button in the top-left corner and selecting the csv file on the hard disk. Once you have chosen it, you will see a preview of the first 1000 rows of the data set in a tabular form, as shown in [Figure 1](#).



The screenshot shows the Disco software interface for importing a CSV file. The top bar displays 'Disco - New project' and a user profile. The main area shows a preview of the CSV data with columns: Case ID, Start Timestamp, Complete Timestamp, Activity, Resource, and Role. The 'Case ID' column is highlighted in blue. A red box highlights the 'Ready to start import' button at the bottom right.

Case ID	Start Timestamp	Complete Timestamp	Activity	Resource	Role
339	2011/02/16 14:31:00.000	2011/02/16 15:23:00.000	Create Purchase Requisition	Nico Ojenbeer	Requester
339	2011/02/17 09:34:00.000	2011/02/17 09:40:00.000	Analyze Purchase Requisition	Maris Freeman	Requester Manager
339	2011/02/17 21:29:00.000	2011/02/17 21:52:00.000	Amend Purchase Requisition	Elvira Lores	Requester
339	2011/02/18 17:24:00.000	2011/02/18 17:30:00.000	Analyze Purchase Requisition	Heinz Gutschmidt	Requester Manager
339	2011/02/18 17:36:00.000	2011/02/18 17:38:00.000	Create Request for Quotation Requester Manager	Francis Odell	Requester Manager
339	2011/02/22 09:34:00.000	2011/02/22 09:58:00.000	Analyze Request for Quotation	Magdalena Predutta	Purchasing Agent
339	2011/02/22 10:50:00.000	2011/02/22 11:03:00.000	Amend Request for Quotation Requester	Penn Osterwalder	Requester Manager
339	2011/02/28 08:10:00.000	2011/02/28 08:34:00.000	Analyze Request for Quotation	Francois de Perrier	Purchasing Agent
940	2011/05/17 06:31:00.000	2011/05/17 07:08:00.000	Create Purchase Requisition	Immanuel Karagianni	Requester
940	2011/05/17 09:58:00.000	2011/05/17 10:06:00.000	Create Request for Quotation Requester	Esmana Liubiata	Requester
940	2011/05/18 19:30:00.000	2011/05/18 19:56:00.000	Analyze Request for Quotation	Francois de Perrier	Purchasing Agent
940	2011/05/18 23:46:00.000	2011/05/18 23:59:00.000	Send Request for Quotation to Supplier	Magdalena Predutta	Purchasing Agent
940	2011/05/19 03:44:00.000	2011/05/19 08:31:00.000	Create Quotation comparison Map	Francois de Perrier	Purchasing Agent
940	2011/05/19 15:38:00.000	2011/05/19 15:52:00.000	Analyze Quotation comparison Map	Kim Passa	Requester
940	2011/05/19 15:52:00.000	2011/05/19 15:52:00.000	Choose best option	Anna Kaufmann	Requester
940	2011/05/20 23:31:00.000	2011/05/21 09:22:00.000	Settle conditions with supplier	Magdalena Predutta	Purchasing Agent
940	2011/05/21 18:48:00.000	2011/05/21 18:59:00.000	Create Purchase Order	Francois de Perrier	Purchasing Agent
940	2011/05/22 11:33:00.000	2011/05/22 11:44:00.000	Confirm Purchase Order	Esmeralda Clay	Supplier
940	2011/05/23 05:32:00.000	2011/05/24 13:46:00.000	Deliver Goods Services	Esmeralda Clay	Supplier
940	2011/05/24 20:59:00.000	2011/05/24 21:00:00.000	Release Purchase Order	Kim Passa	Requester
940	2011/05/26 07:41:00.000	2011/05/26 07:42:00.000	Approve Purchase Order for payment	Karel de Groot	Purchasing Agent
940	2011/05/28 01:11:00.000	2011/05/28 01:11:00.000	Send invoice	Kiu Kan	Supplier
940	2011/05/28 15:28:00.000	2011/05/28 15:28:00.000	Authorize Supplier's Invoice payment	Pedro Alvares	Financial Manager
940	2011/05/28 16:11:00.000	2011/05/28 16:19:00.000	Pay invoice	Karalda Nimwada	Financial Manager
1417	2011/07/23 12:53:00.000	2011/07/23 13:31:00.000	Create Purchase Requisition	Christian Francois	Requester
1417	2011/07/23 17:51:00.000	2011/07/23 17:59:00.000	Create Request for Quotation Requester	Immanuel Karagianni	Requester
1417	2011/08/02 07:02:00.000	2011/08/02 07:24:00.000	Analyze Request for Quotation	Karel de Groot	Purchasing Agent
1417	2011/08/02 08:17:00.000	2011/08/02 08:27:00.000	Amend Request for Quotation Requester	Anna Kaufmann	Requester Manager
1417	2011/08/08 04:03:00.000	2011/08/08 04:23:00.000	Analyze Request for Quotation	Magdalena Predutta	Purchasing Agent
159	2011/01/21 15:02:00.000	2011/01/21 15:42:00.000	Create Purchase Requisition	Elvira Lores	Requester
450	2011/01/22 02:58:00.000	2011/01/22 03:05:00.000	Analyze Purchase Requisition	Francis Odell	Requester Manager

Figure 1. Importing a csv file as an event log

You can now select each column (it will be highlighted in blue) and tell Disco how it should interpret this column: At the top, you find configuration options for the Case ID, the Activity name, Timestamps, Resource, and Other (which are additional attributes).

Disco tries to guess the proper configuration for the input data. However, to make sure, it is better to go through each column and choose the correct structure at the top. For example, the first column is currently selected, and above, you can see that it is configured as the [Case ID](#). The two

¹ Note that, sometimes, the delimiter is not a comma but a semicolon (";") or other characters. Still, the format is called csv.

timestamp columns should be set as *Timestamp*, the Activity column as *Activity*, the Resource column as *Resource*, and the Role column as *Other*.

When you have configured all input data, you will see a message about the configurations left to the "Start Import" button at the bottom right of the window. If you forgot to define one of the necessary inputs (Case ID and Activity), Disco warns you with the message "No Case ID (or Activity) defined". Note that timestamp is not mandatory to import an event log. In this case, Disco assumes that the first occurred event is executed first. But defining timestamps allows us to explore time-based analysis such as bottleneck, waiting time, etc.

Step 2: Inspect Process

As soon as you click 'Start import', Disco will mine your data set and automatically display a process map showing how the process was really performed. See [Figure 2](#).

According to the process owners, the "*Amend Request for Quotation Requester*" activity is only supposed to be performed in exceptional situations because a change is being made to an existing request in this step. However, we can see now that this activity was performed 514 times in 608 cases, which is not an exception.

The activity "*Amend Purchase Requisition*" is conversely really an exception: it is performed 11 times out of 608 cases.

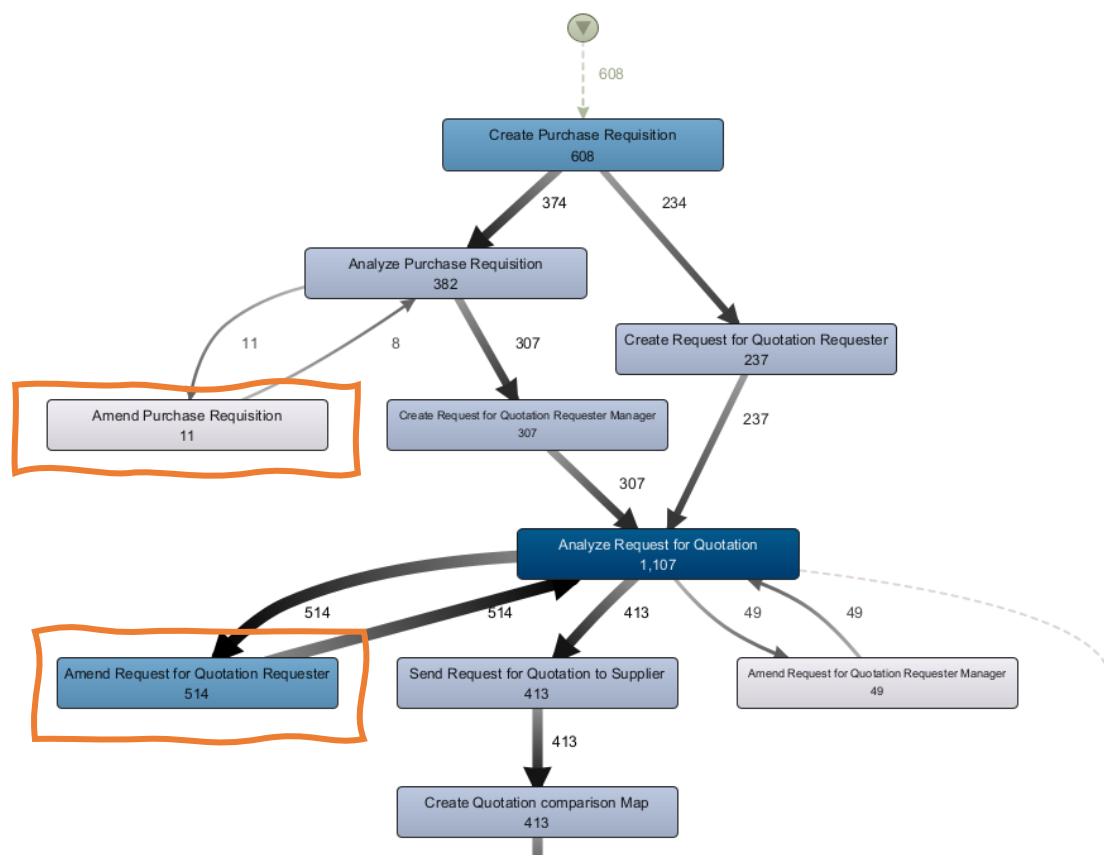


Figure 2. Visualization of a process map

If you really want to see the main process behavior (i.e., the most common activity paths of process execution), you can go to the slider Activities and take it down to 0%. See [Figure 3](#).

The question here: how do you see all possible behaviors?

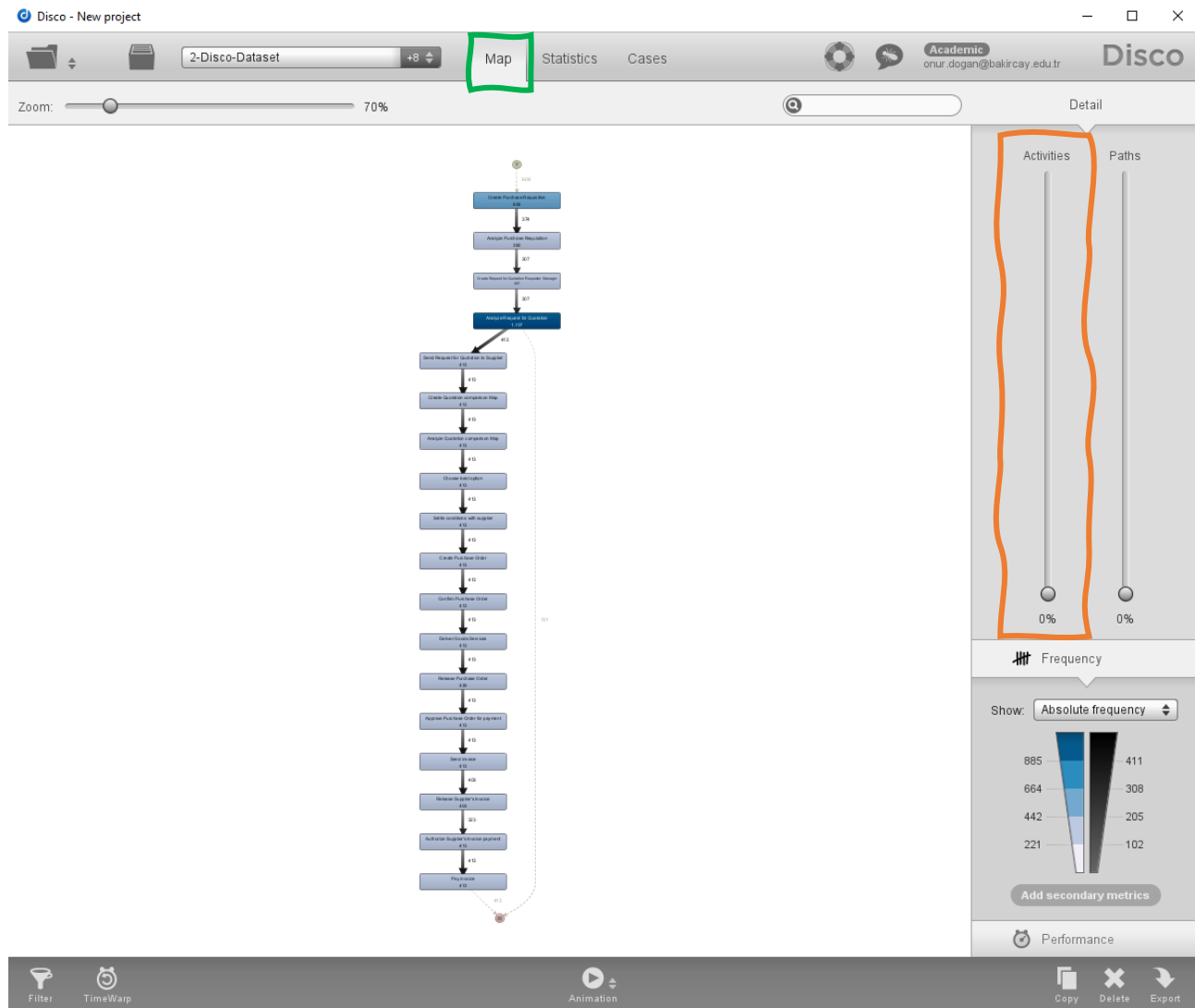


Figure 3. Process map

Step 3: Animate Process

The process executions in the event logs can be projected on the process maps and animated over time. The animation can be extremely helpful in communicating any process problems you have found.

Click on the “*Animation*” button in the middle at the bottom of the process map to get to the Animation view. Then press the “*Play*” button at the bottom left corner, thus obtaining something similar to what is depicted in Figure 4. Each circle represents a set of executed cases: larger circles indicate larger sets of cases (i.e., process instances) at a certain point of the process progression.

This way, we can make the discovered bottleneck really tangible for people and “bring it to life”. Animations can intuitively highlight them by looking at where circles (i.e., cases) are mostly at.

The question here: do you see any bottleneck?

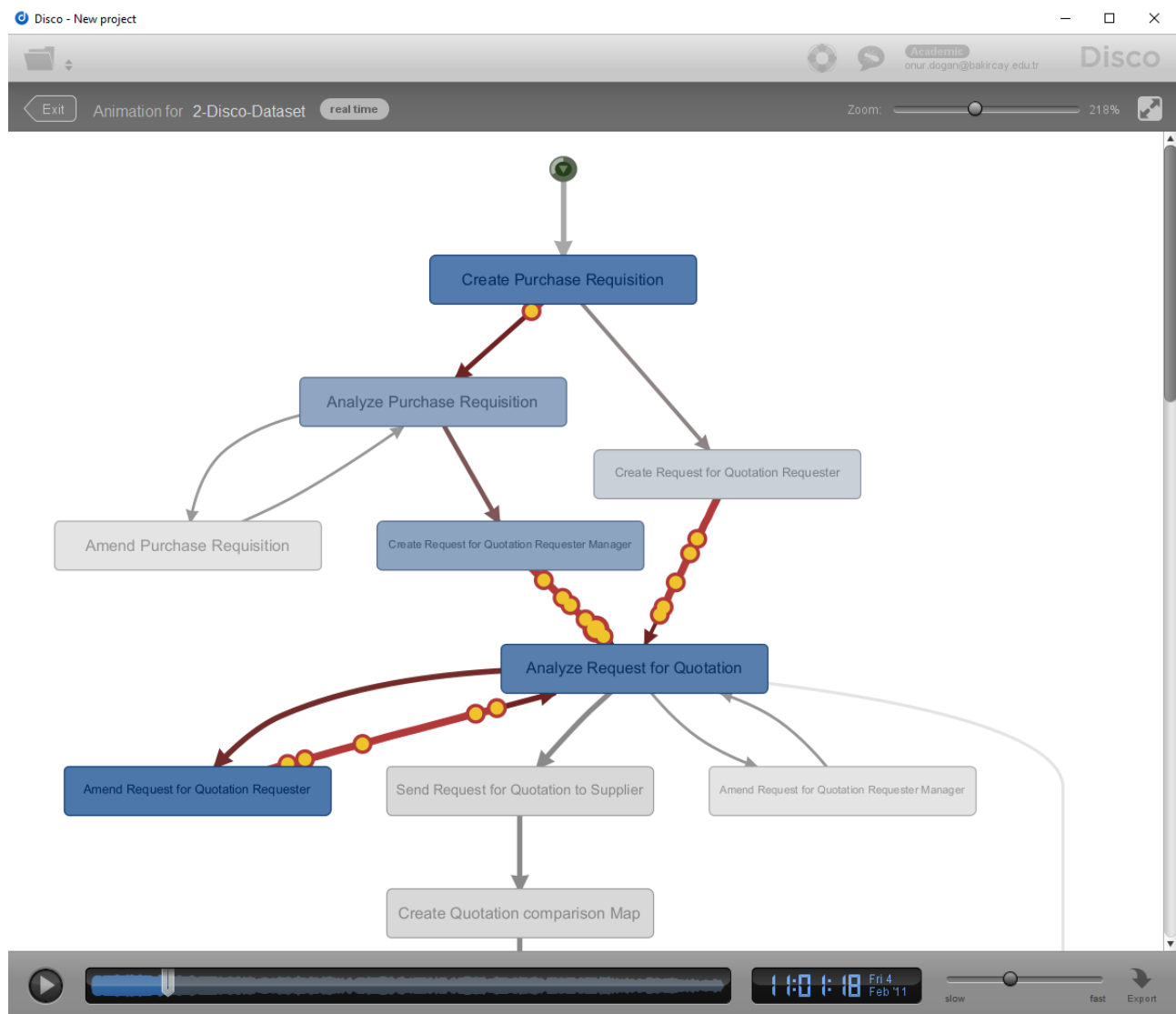


Figure 4. Process animation over time

Step 4: Inspect Statistics

With the process map, you have now obtained a bird’s eye view of the overall process. Next, let’s inspect some process statistics by changing to the ‘*Statistics*’ tab at the top.

- 1- You can find some overview statistics about your data set on the right. For example, we can see that there are 608 cases (purchase orders) and 9119 events (rows in the data set). You can also see the timeframe of the covered process: The data runs from January 2011 to October 2011, so there are about ten months of data.

- 2- Assume now to have received complaints about the throughput time for this process. So, to look at the performance, you can change from the "Events over time" to the "Case duration" statistics next to the chart (see [Figure 5](#)). The case duration shows the time from the very beginning to the very end of the case. When you move the mouse over the histogram, you can see that most cases are completed within up to 18 days in total. However, quite a few cases take much longer than 75 days to 110 days.

It does not seem like an exception, but it looks like we have a serious problem with the throughput time in this process. As the process owner, you want to know where we spend so much time in the process that we end up with such long case durations.

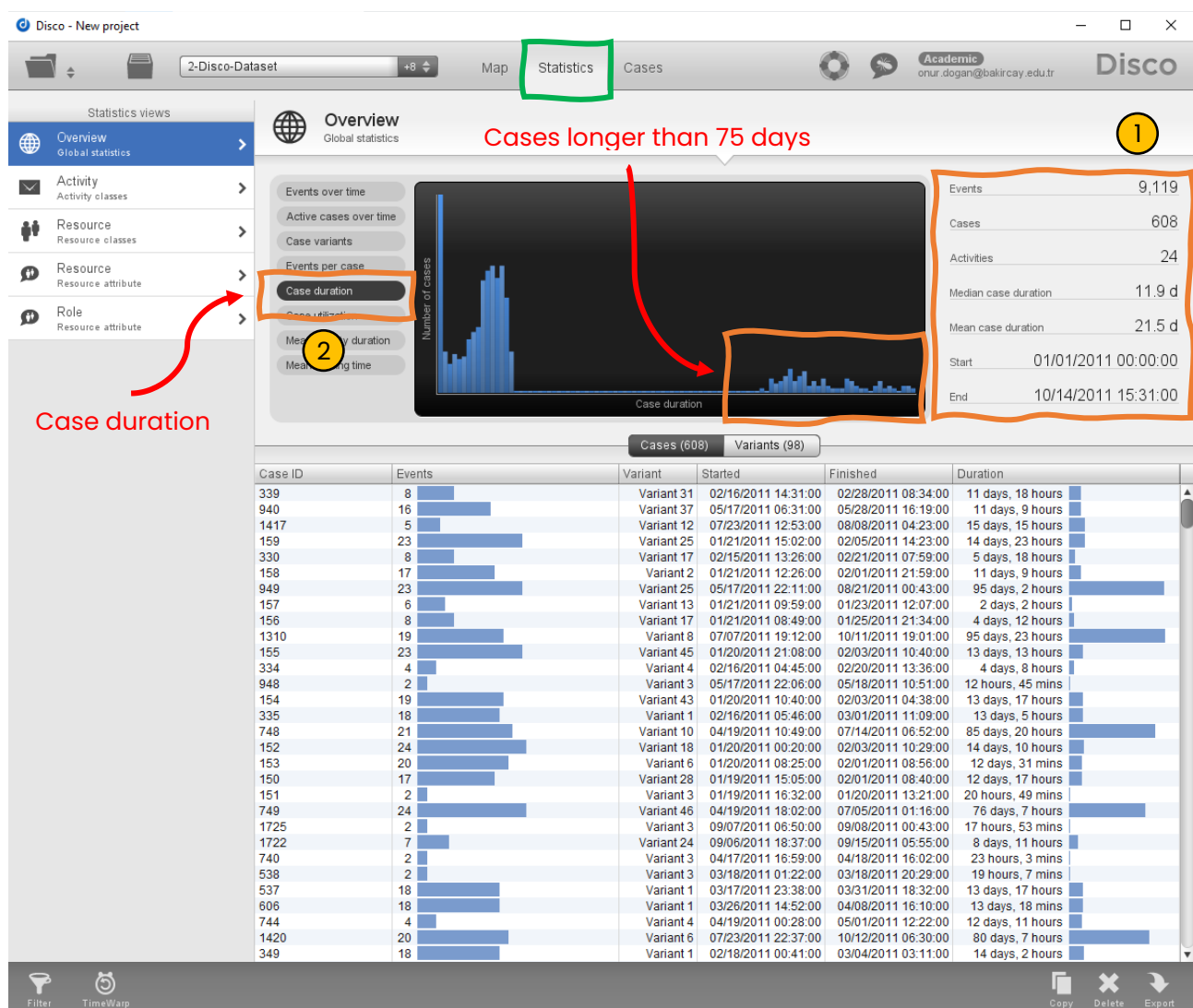


Figure 5. Process statistics

Step 5: Inspect Cases

Before focusing on the discovered performance problem, let's go one step deeper to inspect individual process instances by changing to the 'Cases' tab (see [Figure 6](#)).

You see in the right panel all the variants that include cases with the same activity order. (Switch from the 'Graph' view to the 'Table' view to get a more compact representation.) When you select a variant on the left panel, the second panel shows you a list of all cases belonging to that specific variant. For example, Figure 6 shows the history of *Case 151* in *Variant 3*. There are only two activities: i) "Create Purchase Requisition" and ii) "Analyze Purchase Requisition". Interestingly, we can find that in *Variant 3*, the process has been stopped after just two steps.

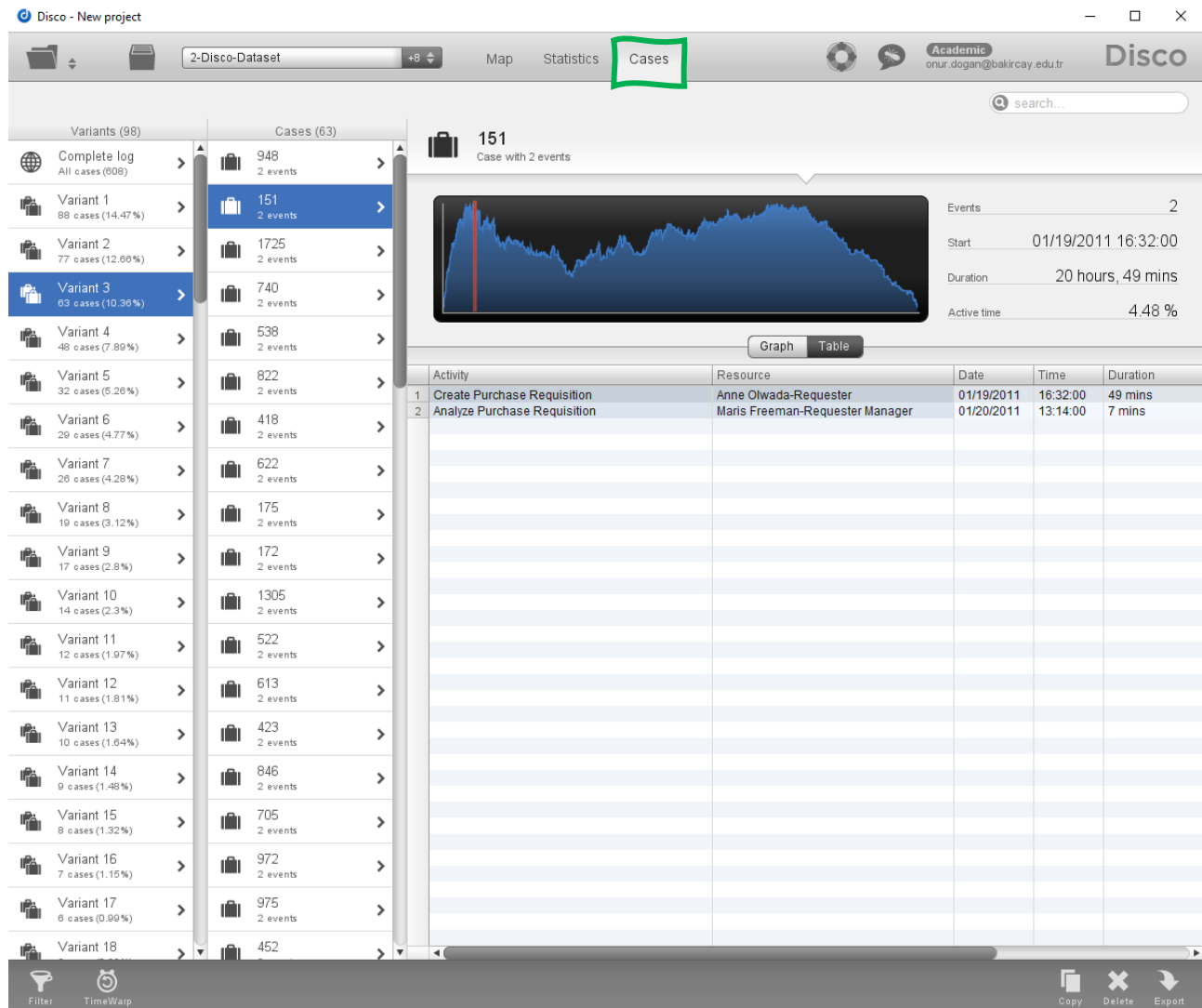


Figure 6. Process variants

Check some other cases under *Variant 3* to realize they all are in the same activity order, and check some other variants to see they all are of different activity orders.

You can often understand the main scenarios by looking at the most frequent variants. *Variant 1* (and sometimes *Variant 2*) usually gives a clue about the ideal process flow. In our example, the purchasing process has 98 variants, and the most frequent variant (*Variant 1*) covers 88 cases (ca. 15% of the cases). *Variant 2* covers 77 cases (ca. 13% of the cases), and so on.

Remember that after the “*Analyze Purchase Requisition*” step, the process ended because the request had been rejected in *Case 151*. You can also find this early end point in the process map given by the arc highlighted in Figure 7. When you go back to the *Map view*, then you can see a dashed line leading from the activity “*Analyze Purchase Requisition*” to the end point. It is typically observed in many event logs: they may refer to cases that were still running when the event log was extracted from the organization’s information system. If we repeated the instruction now, we would see that those cases have likely been completed.

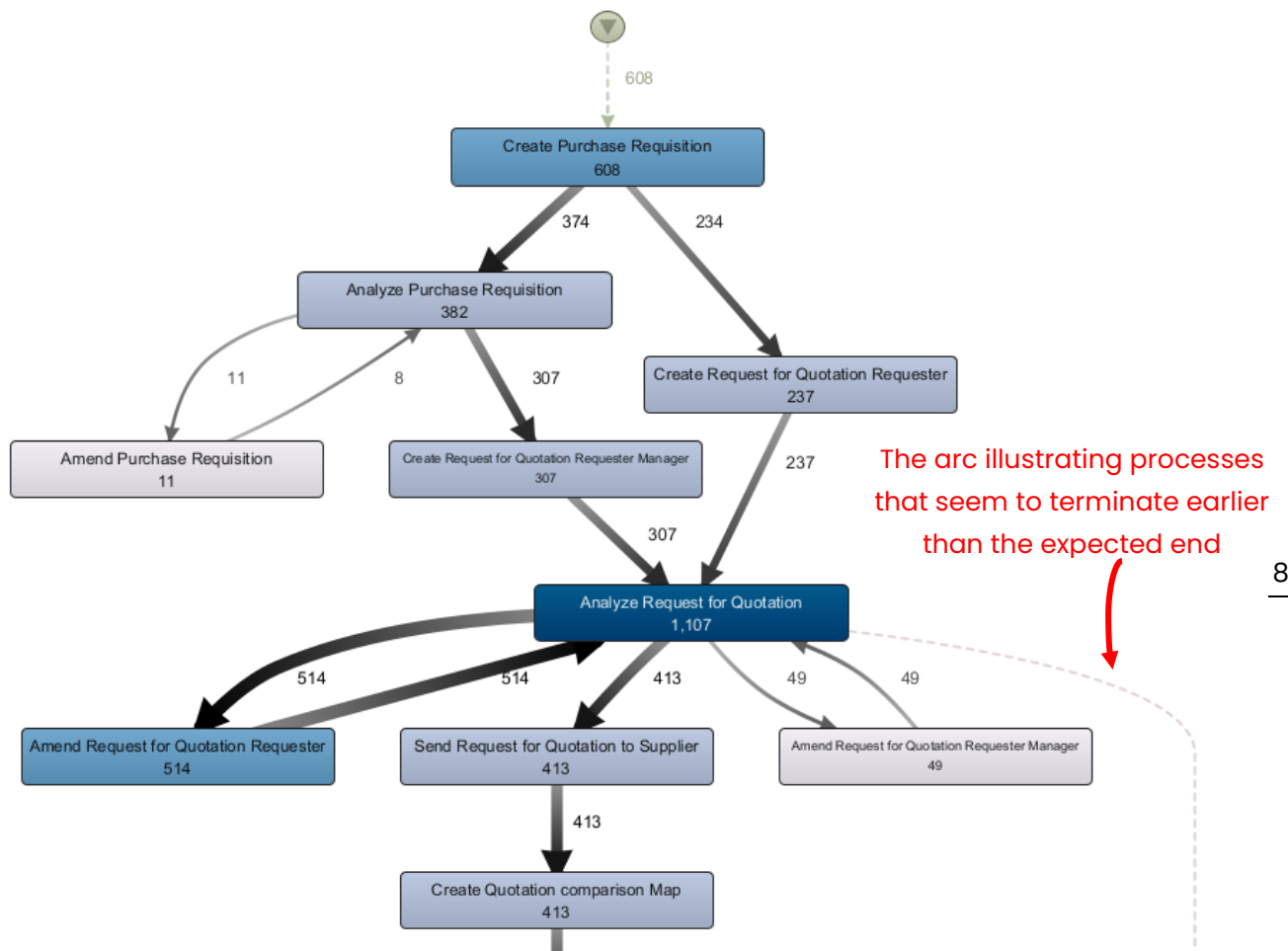


Figure 7. Early (unexpected) end

Step 6: Filter on Performance

You can use filters to focus on specific questions about your process. We will use the Performance filter to investigate why some of the cases are taking so long. You can add a performance filter by clicking on the filter symbol at the bottom left corner and then choosing the “Performance” from the list (see *Figure 8*).

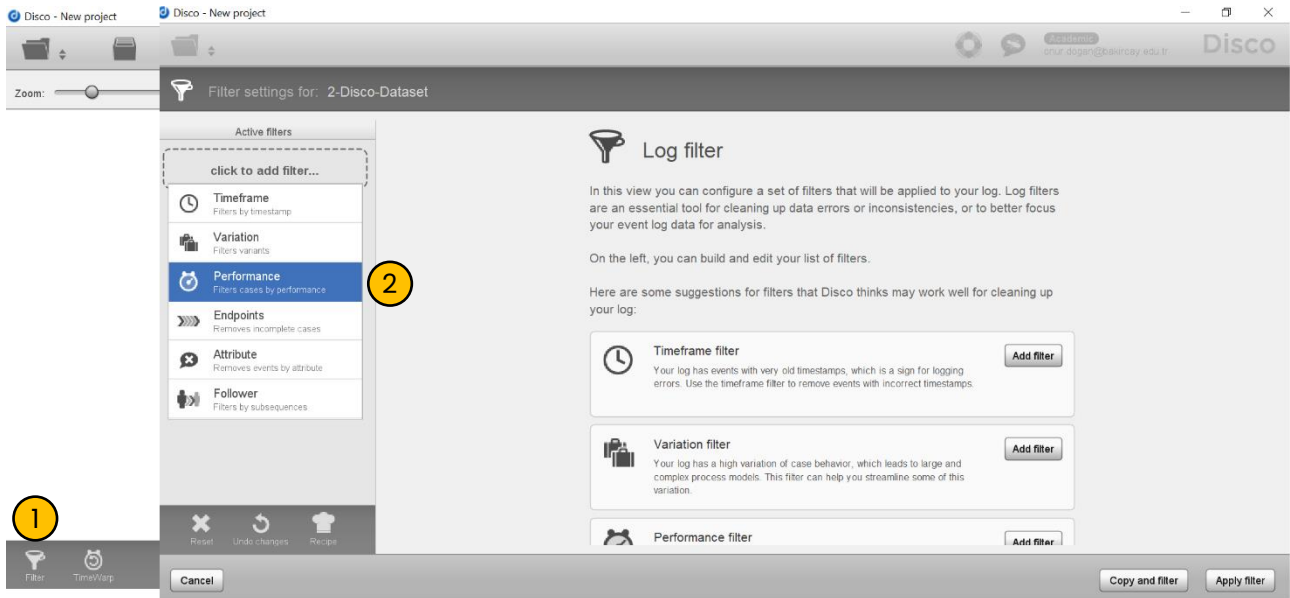


Figure 8. Performance filter

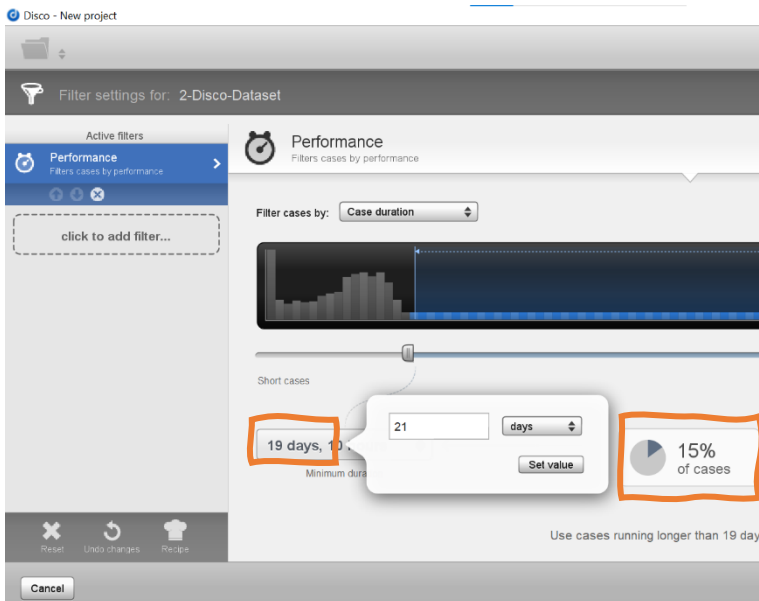


Figure 9. Adjust a certain case duration

Then move the left end of the slider to the right around the 21 days or adjust the exact duration that you want to set by clicking "*Minimum duration*" from the drop-down menu (see Figure 9). The blue area now covers all cases you want to focus on: The cases that take longer than 21 days. You can see that about 15% of all cases in the data set fall outside of the service level target for this process. It will produce a new process map with only the long case duration behaviors. Play around and see whether you observed differences.

Exercise 1 – Filter out running cases

Repeat the analysis mentioned above after removing the incomplete cases. Namely, the cases are still running. It requires selecting an endpoints filter, which can be achieved by selecting the "Endpoint" filter and choosing the activity(ies) with which cases are expected to complete. An endpoint filter will then exclude all the traces referring to cases that end with an activity different from those specified in the endpoint filter.

Exercise 2 – Filter out running cases

Repeat the abovementioned analysis, using the resource as the activity name when importing from the csv file. Think about the reason on what you see in this case.

Exercise 3 – Import XES files to Disco

Perform the analysis of a second process that refers to applications for loans. Use the file *2-Disco-financial.xes*. Note that, differently from a csv file, this file is already in a log format, namely in the XES format. The *eXtensible Event Stream (XES)* format is the IEEE standard to store event logs.